PASTEL: Evidence-based learning engineering method to create intelligent online textbook at scale

Noboru Matsuda and Machi Shimmei

Center for Educational Informatics
Department of Computer Science
North Carolina State University
Noboru.Matsuda@ncsu.edu

Abstract. An extension of online courseware with the macro-adaptive scaffolding, called CyberBook, is proposed. The macro-adaptive scaffolding includes (1) a dynamic control for the amount of formative assessments, (2) a just-in-time navigation to a direct instruction for formative assessment items on which a student failed to answer correctly, and (3) embedded cognitive tutors to provide individual practice on solving problems. The paper also proposes two learning-engineering methods to effectively create a CyberBook: a web-browser based authoring tool for cognitive tutors and a text-mining application for automated skill-model discovery and annotation. A classroom evaluation study to measure the effectiveness of the CyberBook was conducted for two subjects: middle school science (Newton's Law) and high school math (Coordinate Geometry). The results show that students who used the fully functional Science CyberBook outperformed those who used a version of CyberBook where the macro-adaptive scaffolding was turned off. However, the same effect was not observed for Math Cyberbook. On both subjects, students on the CyberBook with the macro-adaptive scaffolding answered on fewer number of formative assessments due to the dynamic control. Further data analysis revealed that those who asked for more hints on the formative assessments achieved higher scores on the post-test than students who asked for fewer hints. The effect of hint usage was more prominent for students with the low-prior competency.

Keywords: Online Courseware, Macro-adaptive Scaffolding, Learning Engineering.

1 Introduction

One of the challenges in current online textbooks is a lack of individual support that apparently hinders students' learning. For example, competency-based scaffolding is desired for students who need tailored scaffolding based on their competency. The lack of an embedded student model results in the excessive training—i.e., all students being exposed to a fixed amount of assessments regardless of their competencies, which severly impacts students' learning [1, 2]. Studies say that the excessive training decreases students' motivation and causes an early course termination [3-6].

A technology innovation to drive individualized scaffolding on a large-scale online textbook that can be plugged-in to existing online-course platforms is therefore critically needed. Without such technology, the online textbook will not fully show its potential to impact a large body of students' learning.

We hypothesize that the lessons learned from long-standing research on intelligent tutoring systems, in particular the *skill-model based pedagogy*, will apply to the large-scale online textbook. The skill-model based pedagogy requires a skill-model (aka a knowledge-component model) that consists of skills each representing a piece of knowledge that students ought to learn. Given that each individual instructional content will be tagged with a skill, the system will compute a proficiency of each skill for each individual student to decide an appropriate pedagogical action. Cognitive tutors, for example, deploy the model-tracing and knowledge-tacing techniques upon a given skill model to drive micro (flagged feedback and just-in-time hint) and macro (problem selection) adaptive instructions [7, 8].

A major technical challenge in this line of research concerns the scalability of existing techniques for creating a skill model. A transformative technique to fully automatically create a skill model and annotate instructional materials on actual online courseware is desired.

A primary goal of the current paper is to introduce a platform-agnostic suite of learnking-engineering methods (called PASTEL; **P**ragmatic methods to develop **A**daptive and **S**calable **T**echnologies for next generation **E**-**L**earning) that allows courseware developers to efficiently create a particular type of online courseware called CyberBook. The CyberBook is an intelligent textbook that provides students with macro-adaptive pedagogy driven by an embedded skill model and cognitive tutors. As a proof of concept, we have made two instances of CyberBook, one for middle school science and another for high school math, and tested their effectiveness with actual students.

2 Solutions: CyberBook and PASTEL

2.1 CyberBook

CyberBook is a structured sequence of instructional activities organized in multiple chapters, sections and units. CyberBook contains three types of instructional elements: (1) direct instructions that convey subject matters (e.g., skills and concepts) usually with written paragraphs and videos, (2) formative assessments typically in the form of multiple-choice or fill-in-the-blank questions, and (3) cognitive tutors that provides mastery practice on solving a particular type of problem. These cognitive tutors are equipped with hint messages that are also considered as instructional elements.

On CyberBook, the three types of learning activities may be placed on multiple pages that compose a "unit." Multiple units become a "section," and a collection of sections becomes a "chapter." Each page has a navigation for page forward/backward, but students may freely visit any page in any order through a table of contents that is also available from any page.

A current version of CyberBook provides students with two types of the macro-

adaptive scaffolding: (a) a dynamic control for the amount of formative assessments and cognitive tutors, and (b) a just-in-time navigation to a direct instruction for a formative assessment that a student failed to answer correctly.

The first type of adaptive scaffolding, *the dynamic control for the amount of assessments and practice*, is to determine which formative assessment items and cognitive tutors should be given to individual students based on their competency. This dynamic control may reduce the number of unnecessary assessments, i.e., the *excessive training*. To judge if an assessment item (either a cognitive tutor or a formative assessment) is beneficial to a particular student, the system applies the knowledge-tracing technique [9] to compute the probability of an individual students answering the next formative assessment item correctly. Based on the student model, those assessment items and cognitive tutors that the student would highly likely (> 0.85) answer/perform correctly are automatically hidden from the students' view.

The second type of adaptive scaffolding, *the just-in-time navigation*, is to provide students with a link (called a dynamic link) to a corresponding direct instruction for (and only for) formative assessments and cognitive tutors that they failed to answer correctly.

To provide this macro-adaptive scaffolding, all the written instructional elements, i.e., text paragraphs, assessment items, and cognitive tutors are tagged with skills. This skill tagging is automatically done by the SMART method introduced in the next section.

The concept of CyberBook is platform independent hence it can be implemented on any online learning platform. Currently, as a proof of concept, we have prototyped CyberBook on Open edX (edx.org) and Open Learning Initiative (Carnegie Mellon University).

2.2 PASTEL

PASTEL is a collection of learning-engineering methods to efficiently build online courseware with embedded the skill model and cognitive tutors. In this paper, we describe two PASTEL methods that are used for the current study: a text-mining application for an automated skill-model discovery (SMART) and a web-browser based cognitive tutor authoring tool (WATSON).

SMART: Skill Model mining with Automated detection of Resemblance among Texts

SMART is a method for automatic discovery of a skill model from a given set of instruction texts. The unit of analysis is a "text," which is either a written paragraph, question sentences for a single assessment item, or hint messages for a single cognitive tutor.

SMART first applies the k-means text clustering technique [10] to divide *assessment items* into clusters with similar semantic meanings. Prior to clustering and keyword extraction, all "texts" are distilled by removing the punctuations and stopwords, which are a set of words that have little grammatical values (e.g., articles, conjunctions, and prepositions, etc.). Distilled "texts" are split into words (aka tokens). Each tokenized "text" is converted into a Term Frequency (TF) vector showing weighted frequency of the total tokens in all given texts (called the *token space*). For example, the *i*-th element

of a TF for a "text" corresponding to a written paragraph, shows the frequency of the *i*-th token in the token space appearing in the "text" (or zero if the "text" does not contain that token).

Our naïve assumption is that if a set of "texts" are all about a same latent skill (e.g., paragraphs explaining a concept X and assessment items asking about the concept X), the latent skill can be identified from the set of "texts." We then hypothesize that assessment items in a particular cluster are assumed to assess the same particular skill. Each cluster of assessments is therefore given a label, that becomes a skill name, by applying keyword extraction technique, TextRank [11]. Finally, each written paragraph and a hint message (of a cognitive tutor) is paired with the closest cluster, i.e., a skill, using the cosine similarity [12]. As a result, *the instructional elements on CyberBook are fully automatically tagged with skills, and a three-way skill mapping among written paragraph, assessment items, and cognitive tutors (through their hint messages) is formed*.

WATSON: Web-based **A**uthoring **T**echnique for adaptive tutoring **S**ystem on **On**line courseware

WATSON is a web-browser based authoring tool to create cognitive tutors by demonstration. Fig. 1 shows an example screenshot of WATSON. Cognitive tutors allow students to practice solving problems while providing the double-looped micro-adaptive scaffolding—scaffolding between problems (aka, the outer-loop) and within a problem (aka, the inner-loop) [7]. The tutor continuously provides students with problems, while the students solve a given problem step by step, until they show mastery in solving the problems. The outer-loop scaffolding uses domain pedagogy to pose a problem to be

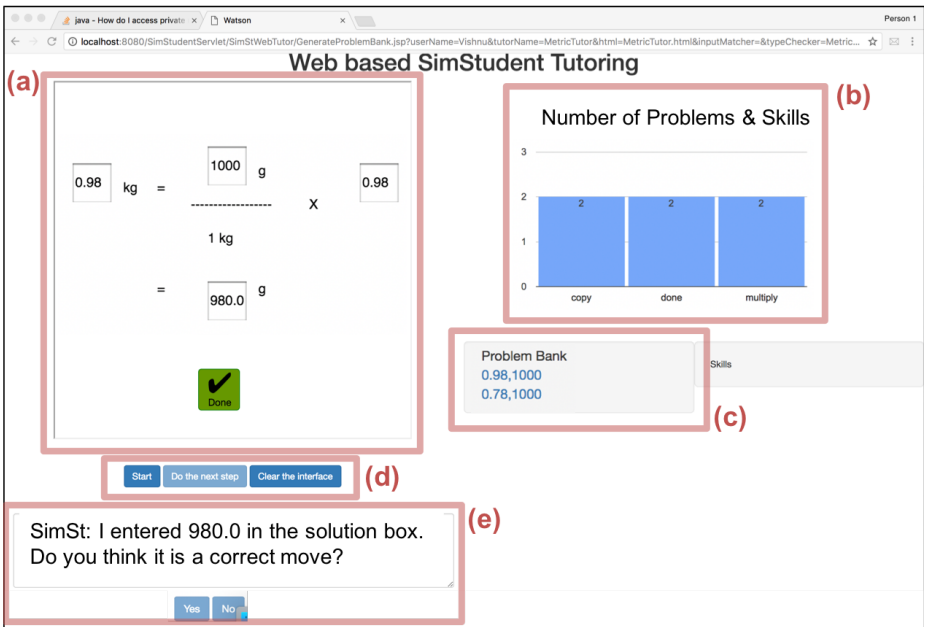

Fig. 1: An example screenshot of WATSON.

solved next that maximizes students' likelihood of achieving the mastery. The inner-loop scaffolding uses domain knowledge that consists of the immediate flagged feedback on the correctness of steps performed, and the just-in-time, on-demand hint on how to perform a next step.

The outer-loop is driven by the knowledge-tracing technique [9] whereas the inner-loop is driven by the model-tracing technique [13]. Both techniques are task independent and rely only on a given *domain expert model* that is written as a set of production rules each of which represents a piece of skill that students ought to learn. This implies that *creating a cognitive tutor is reduced to creating a domain expert model, a set of problems used for tutoring, and a tutoring interface*.

WATSON is built on a third party Cognitive Tutor Authoring Tool (CTAT) [14]. To build a cognitive tutor using WATSON, an author first uses CTAT to create a tutoring interface diretly on a web-browser. CTAT outputs an HTML5 code for the tutoring interface. WATSON renders the HTML5-based tutoring interface on a web-browser with additional graphical user interface for the author to interactively create a domain expert model (Fig. 1-a). To create the domain expert model, the author interactively tutors a machine learning agent, called SimStudent [15], through the tutoring interface. The author poses a problem for SimStudent and asks SimStudent to solve the problem (Fig. 1-d). SimStudent may attempt to solve the problem by performing one step at a time (which in this example, corresponds to entering a value in a text box shown on the tutoring interface). When SimStudent performs a step, it asks the author to provide feedback on the correctness (Fig. 1-e). The author responds by clicking the [yes/no] button. When SimStudent gets stuck on performing a step, it asks the author to demonstrate the next step. The author then performs that step on the tutoring interface.

Through the interctive tutoring, SimStudent produces a set of production rules each of which corresponds to a single step reified on the tutoring interface. Each production rule therefore represents a particular skill that is sufficient to perform a particular step. The author provides a name of each production rule while tutoring. Those production rules will be used as a domain expert model for the cognitive tutor with the exact HTML5 tutoring interface used to tutor SimStudent.

While authoring, a list of skills and problems tutored are displayed (Fig. 1-b and c respectively). When the author clicks on a skill name on the graph (Fig. 1-b), a current production rule written in the Jess language [16] is shown on a separate browser tab. When the author mouse-over a name listed in the problem bank (Fig. 1-c), names of the skills used to solve the corresponding problem are shown in a pop-up dialogue.

3 Evaluation Study

To measure the effectiveness of CyberBook, we conducted an evaluation study at our partner schools using the instances of CyberBook for middle school science and high school math. The Science CyberBook had 11 sections (40 units) with 17 videos and 83 formative assessments. There were no cognitive tutors embedded for the Science CyberBook due to a constraint on the timing of the study and the development cycle. All of the adaptive scaffolding functionalities mentioned earlier were available. The

Math CyberBook, on the other hand, had 23 sections (26 units) with 179 formative assessments and 14 cognitive tutors. No video was used for the Math CyberBook, partly because the in-service math teacher who led the curriculum design believed that videos would not be necessariy if the curriculum has robast and rich instructions and graphics.

3.1 Method

The school study was a stratified randomized controlled trial with two treatment conditions—fully functional CyberBook (the *Adaptive* condition, hereafter) vs. a version of CyberBook without the macro-adaptive scaffolding (the *Non-Adaptive* condition). For Science, two public middle schools in Texas, USA participated with 131 and 34 students in 6 and 2 science classes respectively. For Math, 143 and 25 students in 5 and 2 math classes from two public high schools in Texas, USA were participated. The study was conducted in their usual science and math class periods as a part of their business-as-usual classroom activities.

The school study sessions involved 5 days, one classroom period per day. On Day 1, all students took a pre-test. For Science, the test lasted for 20 minutes with 18 multiple-choice questions. For Math, the test lasted for 30 minutes with 21 multiple-choice questions. For both subjects, the test was printed on paper, but students were asked to enter their answers through an online form.

After taking the pre-test, students were randomly assigned to one of two conditions using the stratified randomization based on the pre-test score—i.e., the difference in the mean pre-test scores between two conditions was aimed to be the minimum; for Science, $M_{Adaptive} = 0.68 \pm 0.23$ vs. $M_{Non\text{-}Adaptive} = 0.70 \pm 0.25$, $t(140) = 0.76$, $p = 0.45$; for Math $M_{Adaptive} = 0.36 \pm 0.19$ vs. $M_{Non\text{-}Adaptive} = 0.36 \pm 0.19$, $t(132) = -0.47$, $p = 0.64$.

On Day 2 through Day 4, students used their assigned version of CyberBook. During this phase, students worked on CyberBook at their own pace while they were encouraged to ask questions to teachers, if necessary. Equally, teachers were encouraged to interact with their students in the same way as they usually do in their classrooms.

On Day 5, students took the post-test that was isomorphic to the pre-test, i.e., the same number and types of problems that can be solved by applying the same knowledge. For post-test items, the difference is in their cover stories and quantities used.

Two researchers attended each of the classroom sessions to take field observation notes and help students overcome any technical issues. Those researchers did not provide students with any instructional scaffolding (but only encouraged students to ask their teachers for an assistance when needed).

3.2 Results

There were no particular exclusion criteria for participants during the study—all students were welcomed to participate in any part of the study. In the following analysis, however, we include only those students who took both pre- and post-tests, and attended all three days of intervention. Table 1 shows the number of participants who

Table 1: The count of students who participated the study and who meet the inclusion criteria (take both pre- and post-tests, and attend all three days of intervention).

	Class Enrollment	Pre-test	Post-test	Sufficient Intervention	Students Included
Science	165	155 (77/78)	154 (79/75)	157 (80/77)	144 (73/71)
Math	168	148 (76/72)	153 (76/77)	159 (78/81)	134 (67/67)
Total	333	303 (153/150)	307 (155/152)	316 (158/158)	278 (140/138)

Note. Parentheses show a breakdown into conditions (Adaptive/Non-Adaptive)

Table 2: Mean test scores.

	Science		Math	
	Pre-test	**Post-test**	**Pre-test**	**Post-test**
Adaptive	0.68(0.22)	0.77(0.16)	0.35(0.18)	0.45(0.19)
Non-Adaptive	0.71(0.25)	0.74(0.19)	0.37(0.19)	0.46(0.23)

took pre and post-tests respectively, and those who attended all three days of intervention (i.e., Day 2 through Day 4). The table also show the number of participants who meet the inclusion criteria.

Test Scores: Table 2 shows mean test scores comparing two conditions both for Science and Math. To see if there was an effect of the macro-adaptive scaffolding on students' learning, a repeated-measures ANOVA was conducted for each subject independently, with post-test score as the dependent variable and test-time (pre vs. post) and condition (Adaptive vs. Non-Adaptive) as fixed factors.

For Science, there was an interaction between condition and test-time; $F(1,142) = 5.61$, $p < 0.05$. A post-hock analysis revealed that *only the adaptive condition shows an increase in the test score from pre- to post-tests*; for Adaptive condition: *paired-t*$(72) = -5.18$, $p < 0.001$, $d = 0.46$; for Non-Adaptive condition: *paired-t*$(70) = -1.52$, $p = 0.13$, $d = 0.13$. *In the science classes, students who used a version of CyberBook with the macro-adaptive scaffolding outperformed students without adaptive scaffolding on the post-test*.

For Math, there was a main effect of test-time ($F(1,132) = 61.57$, $p < 0.001$), but condition was not a main effect ($F(1,132) = 0.23$, $p = 0.63$). *In the math classes, students' scores on the test increased from pre- to post-tests equally regardless of whether the macro-adaptive scaffolding was available*.

Behavior Analysis: To understand why only the Science CyberBook showed the effect of the macro-adaptive scaffolding, we analyzed the process data showing detailed interactions between students and the system while they were working on the CyberBook.

The process data contain the clickstream data (including the information about the assessment items such as problem ID and the skills associated with each problem) and the correctness of the students' answers.

We first hypothesized that there was a condition difference in the way students watched the video vs. answering formative assessments on the Science CyberBook (there was no video on Math CyberBook). Not surprisingly, there was a notable condition difference in the number of formative assessments students answered in Science CyberBook; $M_{Adaptive}$ = 62.2±16.93 vs. $M_{Non\text{-}Adaptive}$ = 74.7±12.69, $t(133) = -5.04$, $p < 0.001$, $d = 0.84$. *The dynamic control for the amount of problems effectively reduced the number of formative assessments for the Adaptive students*. There was, however, no statistically reliable relationship between the number of formative assessments answered and the post-test score when the pre-test was entered as the primary factor to a regression model; $F(1,141) = 3.08$, $p = 0.08$. There was no condition difference in the number of videos watched either; $M_{Adaptive}$ = 25.6±26.23 vs. $M_{Non\text{-}Adaptive}$ = 25.0±28.92, $t(128) = 0.145$, $p = 0.89$. The doer/non-doer effect that predicts that *learning by doing* (i.e., working on formative assessments) better facilitates *learning than by watching videos* [17] did not present in the current study on the Science CyberBook. As a side note, for Math, there was no condition difference in the number of formative assessments students answered; $M_{Adaptive}$ = 55.5±12.79, $M_{Non\text{-}Adaptive}$ = 51.0±15.65, $p = 0.07$, $d = 0.32$.

Second, we hypothesized the dynamic link, which was only available for Adaptive students, effectively facilitated learning on Science CyberBook. This hypothesis was not supported. To our surprise, the average number of dynamic-link clicked was quite low; $M = 0.4 \pm 1.8$. It turned out that, in the instances of CyberBook used in the current study, most of the linked contents are placed on the same page as the assessment item, at relatively in a close distance. The field observation notes collected during classroom sessions mentioned that students noticed that they were simply able to scroll up the page to review a related content instead of clicking on the dynamic link.

Third, we explored students' engagement in learning by doing—i.e., how seriously students worked on formative assessments. In particular, we investigated whether Adaptive students worked on multiple choice questions more seriously than Non-Adaptive students. On Science CyberBook, about 2/3 of the formative assessments are multiple-choice questions. The degree of engagement on the multiple-choice question might have had a significant impact on students' learning [18]. We operationalized the "seriousness" as the number of choice items submitted before making a correct answer. Since CyberBook provides the immediate feedback on an answer submission, when students were not engaged in learning, they might merely try choice items one-by-one until they see affirmative feedback. We hypothesized that the ratio of choice items (RCI) submitted before submitting a correct answer on multiple-choice questions is lower among Adaptive than Non-Adaptive students. This hypothesis was not supported. There was no condition difference in the average RCI per student; for Science, $M_{Adaptive}$ = 0.48±0.06 vs. $M_{Non\text{-}Adaptive}$ = 0.47±0.07; $t(140) = 0.78$, $p = 0.44$. The same trend was observed for Math; $M_{Adaptive}$ = 0.52±0.08 vs. $M_{Non\text{-}adaptive}$ = 0.50±0.07; $t(129) = 0.78$, $p = 0.20$.

Fourth, we investigated the difference in the hint usage between Adaptive and Non-

Adaptive students. In particular, we hypothesized that Adaptive students used hints more frequently when they failed to answer a formative assessment item correctly. We operationalize the hint usage on failed assessment items (per student) as the ratio of assessment items on which a student failed to answer correctly *and* asked for a hint to the total number of assessment items that the student failed to answer correctly—denoted as Hint on Failure Ratio (HFR). This hypothesis was supported only for Science. When aggregated across all students within each condition, there was a condition difference on HFR for Science; $M_{Adaptive}$ = 0.32±0.23 vs. $M_{Non\text{-}Adaptive}$ = 0.23±0.19; $t(138) = 2.40$, $p < 0.05$, $d = 0.40$. For Math, the condition difference was weaker; $M_{Adaptive}$ = 0.31±0.26; $M_{Non\text{-}Adaptive}$ = 0.24±0.24; $t(130) = 1.71$; $p = 0.09$; $d = 0.30$. However, a regression analysis did not confirm a correlation between HFR and post-test score when pre-test was entered to the model as the primary factor; for Science, pre-test was a significant predictor, $F(1,141) = 177.9$, $p < 0.0001$; HFR was not, $F(1,141) = 1.28$, $p = 0.26$. The same trend was observed for Math; pre-test $F(1,130) = 167.4$, $p < 0.0001$; HFR $F(1,130) = 3.17$, $p = 0.08$.

A further analysis revealed that the correlation between HFR and post-test score was negative; for Science $r(142) = -0.34$, $p < 0.001$; for Math $r(131) = -3.99$, $p < 0.001$. Though this finding was controversial at the beginning, we hypothesized that (1) students with a low prior competency (measured as pre-test score) needed more hints on failed assessments—i.e., HFR and pre-test score were negatively correlated, and (2) pre-test and post-test scores were highly positively correlated as is almost always the case in school evaluation studies.

If this hypothesis is true, then we should see more evident condition difference of HFR among low prior students than high prior students. This hypothesis was supported as shown in Table 3. It was only for Science that condition (Adaptive vs. Non-Adaptive) was the main effect for HFR for low-prior students. Although, understanding the exact reason why Adaptive Low-Prior students used more hint than Adaptive Non-Adaptive students requires further investigation. We suspect the fact that the dynamic link (available only for Adaptive students) was located physically close to the hint button might have had a positive influence on students' hint usage. Interestingly, Table 3 shows the similar HFR values for low prior students among Science and Math students. Yet, the lack of statistical significance for Math is arguably due to a larger variance.

Table 3: The average ratio of asking hint on formative assessment items on which students failed to answer correctly (HFR). Numbers in parentheses show standard deviations. The condition difference is only statistically significant for low prior students on the Science CyberBook.

	Science		**Math**	
	Adaptive	**Non-Adaptive**	**Adaptive**	**Non-Adaptive**
Low Prior	**0.41(0.24)[a]**	**0.27(0.20)[a]**	0.39(0.28)[b]	0.27(0.28)[b]
High Prior	0.22(0.19)	0.19(0.18)	0.23(0.21)	0.20(0.17)

a: $t(69) = 2.57$, $p < 0.05$; b: $t(65) = 1.69$, $p = 0.10$.

4 Discussions

It is not entirely clear why the doer/non-doer effect was not confirmed in the current study hence needs more investigation. A previous study [17] reported that learning by doing (i.e., answering formative assessments and receive feedback/hints) is six times more effective than watching videos and reading texts. One potential hypothesis for the doer effect not shown in the current study is that almost all students in our study might have worked on the sufficient number of formative assessments hence they were all rather equally doers (hence no correlation between the number of assessments and test score observed). The students' competency might be another factor. In the current study, the number of assessments is determined based on the student's competency—those who have lower competency received more assessments. Therefore, it might not be surprising to see a negative correlation between the number of formative assessments answered and learning outcome (pre-test is strongly correlated with post-test after all).

The dynamic link was not apparently functioning as expected in the current study, anecdotally because students noticed that related contents were just a scroll away. Unfortunately, there was no logging made for this type of behavior (i.e., scrolling through pages and reviewing related contents). Therefore, it is technically challenging to comprehensively evaluate the effect of the dynamic link in the current data. The improvement of the logging function to track the precise usage of dynamic link is one of the subjects for future system improvement.

The effect of the proposed macro-adaptive scaffolding was not replicated between the Science and Math courseware. The current study is somewhat confounded due to the difference in the availability of videos (available only science) and cognitive tutors (only math). Further investigation is needed to conduct a thorough study on when and how the macro-adaptive scaffolding facilitates students' learning.

5 Conclusion

We found that the online courseware with the macro-adaptive scaffolding including a dynamic controlling of the mount of formative assessments and cognitive tutors amplified students learning on a middle school science course, but the effect was not replicated on a high school geometry course. The major differences between these two instances of courseware include that: (1) only the science courseware contained 17 videos, and (2) only the math courseware contained 14 cognitive tutors. The current data suggest that it was the use of hint for formative assessment items on which students failed to answer correctly that correlated with learning outcome, and this effect was present only among those students who had a low prior competency, measured as the pre-test score. This effect was only observed for the science course. Understanding why the same was not the case for the math course requires a further analysis.

Creating effective online courseware at scale is one of the most demanded challenges in the current cyberlearning era. The current paper demonstrated the fidelity of implementation for two learning-engineering methods to build practical online courseware with the macro-adaptive scaffolding. More studies are needed to understand what exactly is needed to develop practical learning-engineering methods with a firm impact on students' learning with a diverse population.

References

1. Koedinger, K.R., E.A. McLaughlin, and J.C. Stamper, *Automated student model improvement*, in *Proceedings of the 5th International Conference on Educational Data Mining*, K. Yacef, et al., Editors. 2012. p. 17-24.
2. Martin, B., et al., *Evaluating and improving adaptive educational systems with learning curves.* User Modeling and User-Adapted Interaction, 2011. **21**(3): p. 249-283.
3. Gates, S.J., et al., eds. *Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engi- neering, and Mathematics*. 2012, PCAST STEM Undergraduate Working Group: Office of the President, DC.
4. Goodman, I.F., *Final Report of the Women's Experiences in College Engineering (WECE) Project*. 2002, Cambridge, MA: Goodman Research Group.
5. Seymour, E. and N.M. Hewitt, *Talking About Leaving: Why Undergraduates Leave the Sciences*. 1997, Boulder, CO: Westview Press.
6. Watkins, J. and E. Mazur, *Retaining students in science, technology, engineering, and mathematics (STEM) majors.* J Coll Sci Teach, 2013. **42**(5): p. 36-41.
7. VanLehn, K., *The Behavior of Tutoring Systems.* International Journal of Artificial Intelligence in Education, 2006. **16**.
8. Ritter, S., et al., *Cognitive tutor: Applied research in mathematics education.* Psychonomic Bulletin & Review, 2007. **14**(2): p. 249-255.
9. Corbett, A.T. and J.R. Anderson, *Knowledge tracing: Modeling the acquisition of procedural knowledge.* User Modeling and User Adapted Interaction, 1995. **4**(4): p. 253-278.
10. Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A K-Means Clustering Algorithm.* Journal of the Royal Statistical Society, 1979. **C-28**(1): p. 100–108.
11. Mihalcea, R. and P. Tarau, *Textrank: Bringing order into texts*, in *Proceedings of EMNLP*, D. Lin and D. Wu, Editors. 2004: Barcelona, Spain. p. 404–411.
12. Salton, G. and M.J. McGill, *Introduction to modern information retrieval*. 1983, Auckland: McGraw-Hill.
13. Anderson, J.R. and R. Pelletier, *A development system for model-tracing tutors.* Proc. of the International Conference on the Learning Sciences, 1991: p. 1-8.
14. Aleven, V., et al., *The Cognitive Tutor Authoring Tools (CTAT): Preliminary evaluation of efficiency gains*, in *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, M. Ikeda, K.D. Ashley, and T.W. Chan, Editors. 2006, Springer Verlag: Berlin. p. 61-70.
15. Matsuda, N., W.W. Cohen, and K.R. Koedinger, *Teaching the Teacher: Tutoring SimStudent leads to more Effective Cognitive Tutor Authoring.* International Journal of Artificial Intelligence in Education, 2015. **25**: p. 1-34.
16. Friedman-Hill, E., *Jess in Action: Java Rule-based Systems*. 2003, Greenwich, CT: Manning.
17. Koedinger, K.R., et al., *Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC*, in *Proceedings of the Second ACM Conference on Learning@Scale*. 2015, ACM. p. 111-120.
18. Marsh, E.J., et al., *The memorial consequences of multiple-choice testing.* Psychonomic bulletin & review, 2007. **14**(2): p. 194-199.