Optimizing the Amount of Practice in an On-Line Learning Platform

Kim M. Kelly
Worcester Polytechnic Institute
Worcester, MA USA
kkelly@wpi.edu

Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, MA USA
nth@wpi.edu

L@S 2016, April 25–26, 2016, Edinburgh, UK.
ACM 978-1-4503-3726-7/16/04.
DOI: http://dx.doi.org/10.1145/2876034.2893393

Abstract
Intelligent tutoring systems are known for providing customized learning opportunities for thousands of users. One feature of many systems is differentiating the amount of practice users receive. To do this, some systems rely on a threshold of consecutive correct responses. For instance, Khan Academy used to use ten correct in a row and now uses five correct in a row as the mastery threshold. The present research uses a series of randomized control trials, conducted in an online learning platform (eg., ASSISTments.org), to explore the effects of different thresholds of consecutive correct responses on learning. Results indicate that despite spending significantly more time practicing there is no significant difference on learning between two, three, four, or five consecutive correct responses. This suggests that systems, and MOOCS, can employ the simple rule of two or three consecutive correct responses when determining the amount of practice provided to users.

Introduction
Online platforms (Khan Academy, MOOCs, Webwork.org, etc.) are widely used in education to help improve student learning. One feature of such systems is their ability to personalize the learning experience for students by customizing the amount of practice provided to students. However, determining the optimal amount of practice, while essential, is not a trivial task. Clearly, determining the correct amount of practice is

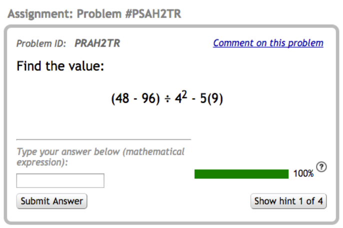

Figure 1. A sample order of operations problem students completed as part of Study 1.

Condition (Students who met the threshold)	Mean Post-Test Score	Mean # of questions	Mean minutes
2-CCR (n=69)	33% (48)	5.7 (3.9)	14.30 (10.90)
3-CCR (n=79)	34% (48)	7.5 (3.9)	17.36 (11.75)
4-CCR (n=66)	39% (49)	9.1 (5.1)	16.55 (12.48)
5-CCR (n=44)	36% (49)	13.1 (7.8)	23.45 (15.72)

Table 1. Study 1 mean post-test scores, number of questions completed by students, and number of minutes it took for students to complete the assignment by condition. Standard deviations are included in parenthesis.

critical because under-practice might not provide enough opportunities for a student to learn a skill, while over-practice might cause students and teachers to disengage with the system. To determine the correct amount of practice, systems attempt to identify the point in time when students have "learned" the skill, otherwise referred to as reaching mastery. Defining mastery may vary between systems.

Systems use a variety of methods to predict this latent variable, mastery. For example, some systems, such as Cognitive Tutor [4,6] rely on knowledge tracing. Knowledge tracing uses student performance to compute the probability that the student will answer the next problem correctly, as well as the probability that the student is in the known state [1]. More complicated methods, such as deep knowledge tracing [5] are currently being developed as well. However, a simpler method also exits. Systems, such as Khan Academy [2] and ASSISTments [3] use a predetermined number of consecutive correct responses (N-CCR) to predict mastery. This method is based on the assumption that if a student can answer a set threshold of consecutive questions correctly than the student has mastered the skill and additional practice is not needed.

N-CCR has the potential benefit of being simple to understand and interpret. Teachers, parents and students can easily understand what is required to complete an assignment. Additionally, a consecutive correct threshold is easy to implement in that it doesn't require complicated programming like knowledge tracing. This makes it particularly useful for simple systems and even MOOCs that are looking to customize learning experiences for thousands of users across the world.

So what is the optimal threshold? The present study uses the random assignment feature in ASSISTments to investigate whether additional attempts, due to a higher mastery threshold, will lead to improved learning as measured by performance on a transfer task.

Study 1- Method

Students from four different middle school classes were randomly assigned to one of four conditions: 2-CCR (n=103), 3-CCR (n=109), 4-CCR (104), or 5-CCR (n=96). The number indicates the threshold of consecutive correct responses that were required to complete the assignment. For example, a student in the 3-CCR condition would be required to answer three consecutive questions correctly without hints or support before reaching mastery. The topic for this assignment was order of operations (see Figure 1). Questions were all morphologically similar to each other. Once students reached the threshold for consecutive correct responses, they were immediately given a post-test that consisted of one transfer item. The transfer item required students to apply the current topic in a novel situation.

Results

In analyzing the data, it was noted that completion rates seemed to vary by condition. A chi-square test of independence was performed to examine the relation between students who finished and those who did not by the extreme conditions (2-CCR and 5-CCR). As expected, the relation was significant X^2 (2, N=199) = 9.06, $p<0.01$. This means that there was a disproportional amount of students who completed the assignment by condition. Reaching the 5-CCR threshold required students to complete more problems than the 2-CCR condition. Therefore more students in the 5-CCR condition quit the assignment before finishing than in the 2-CCR condition. This introduces a selection effect.

Presumably, students who persevered and completed the assignment with a higher threshold were higher knowledge students. This means that differences in post-test scores might be caused by student differences rather than the experimental manipulation.

Despite the probable selection effect, when examining post-test scores (see Table 1), an ANOVA revealed that there was no significant difference in performance on the post-test by condition ($F(3,258) = 0.21$, $p = 0.89$). Yet students in the 5-CCR condition completed on average twice as many questions and spent significantly more time than students in in the 3-CCR condition. To account for the differential completion rate by condition all students who did not complete the assignment were automatically assigned a post-test score of zero. As expected, post-test scores dropped. An ANOVA again revealed, no significant difference in post-test performance by condition ($F(3, 412) = 0.86$, $p = 0.46$).

A review of the 95% confidence interval shows that mean post-test performance for 2-CCR, 3-CCR, and 4-CCR conditions overlap. Remember, to account for the differential completion rate, students who failed to reach the threshold were assigned a post-test score of zero. As a result, the mean post-test score for the 5-CCR condition is greatly impacted by the increased number of students who did not complete the assignment.

Study 2- Method

In an attempt to generalize the findings of Study 1, an additional study was conducted simultaneously with a different mathematical topic. Again, students were randomly assigned to one of four conditions: 2-CCR (n=64), 3-CCR (51), 4-CCR (59), or 5-CCR (42). The topic for this assignment was the distributive property. Once students met the N-CCR threshold, they were immediately given a post-test that consisted of two transfer questions. The transfer questions required students to apply the skill in a more challenging problem.

Results

As expected, there was a differential completion rate by condition. However unlike Study 1, the difference between the two extreme conditions (2-CCR and 5-CCR) was not significant (X^2 (2, N=106) = 0.30, $p = 0.59$).

When considering only the students who met the threshold requirement for their condition, an ANOVA revealed that there is no significant difference in performance on the post-test by condition ($F(3, 162) = 1.35$, $p = 0.26$). See Table 2 for mean post-test scores.

Despite the lack of improved learning, students in the 5-CCR condition completed significantly more questions than those in the 2-CCR condition ($F(3, 163) = 25.24$, $p<0.0001$). An analysis of the 95% confidence intervals for number of questions completed reveals that 3-CCR does not appear different than 4-CCR in the amount of practice students receive.

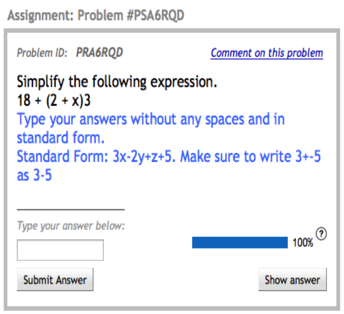

Figure 2. A sample distributive property problem students completed as part of Study 2.

Condition	Mean Post-Test Score	Mean Number of Questions
2-CCR (n=53)	46% (38)	2.9 (1.2)
3-CCR (n=38)	55% (39)	5.7 (3.8)
4-CCR (n=42)	38% (37)	6.6 (3.4)
5-CCR (n=33)	49% (39)	8.8 (4.1)

Table 2: For students in Study 2 who reached the N-CCR threshold and completed the assignment, the mean post-test scores by condition and mean number of questions completed by students are shown. Standard deviations are provided in parenthesis.

Conclusion

Determining the correct amount of practice students require to learn a skill is essential for the success of on-line learning environments. The present studies investigate different thresholds (including two, three, four, and five) of consecutive correct responses and find that there is no significant impact on learning as measured by performance on transfer tasks. However, as expected, higher thresholds force students to complete significantly more practice. This becomes a problem when it affects homework completion rates. Specifically, higher thresholds deter students from completing assignments, which may impact their future engagement with system and potentially their self-esteem.

This is a significant contribution to the field because as systems attempt to educate learners en masse, they must be able to easily, yet accurately customize practice opportunities. Future research should attempt to look at retention rates as an alternative measure of learning. Specifically, is there a difference in performance by condition on a post-test that is administered a week after students reach the N-CCR threshold? This might indicate that there is in fact a benefit to higher thresholds of N-CCR.

Acknowledgements

We thank multiple current NSF grants (ACI-1440753, DRL-1252297, DRL-1109483, DRL-1316736, DGE-1535428 & DRL-1031398) , the US Dept. of Ed (IES R305A120125 & R305C100024 and GAANN), and the ONR.

References

1. Corbett, A., Anderson, J. (1995). Knowledge Tracing: Modeling Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction,* 4, 253-278.
2. Faus, M. (2014). Improving Khan Academy's student knowledge model for better predictions. *MattFaus.com* [web log]. Retrieved October, 2014, from http://mattfaus.com/2014/05/improving-khan-academys-student-knowledge-model-for-better-predictions/
3. Feng, M., Heffernan, N. T., Koedinger, K. R.: Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. *User Modeling and User-Adapted Interaction*, 19, 243-266 (2009)
4. Koedinger, K.R., Corbett, A.T. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In Sawyer, R.K. (ed.). The Cambridge handbook of the learning sciences. Cambridge University Press, New York.
5. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. & Sohl-Dickstein, J. Deep Knowledge Tracing. Retrieved from http://web.stanford.edu/~cpiech/bio/papers/deepKnowledgeTracing.pdf.
6. Timeline of Cognitive Tutor History. (2003-2013). Cognitive Tutoring Authoring Tools. Retrieved from http://ctat.pact.cs.cmu.edu/index.php?id=timeline