Using Optimally Selected Drill Practice to Train Basic Facts

Philip I. Pavlik Jr.[1], Thomas Bolster[1], Sue-mei Wu[2], Kenneth R. Koedinger[1], and Brian MacWhinney[3]

[1] Human Computer Interaction Institute
[2] Deparment of Modern Languages
[3] Department of Psychology, Carnegie Mellon Univeristy, Pittsurgh, PA 15213
{ppavlik,tib,suemei}@andrew.cmu.edu, {keodinger,macw}@cmu.edu

Abstract. How to best sequence instruction in a collection of basic facts is a problem often faced by intelligent tutoring systems. To solve this problem, the following work details two tests of a system to provide drill practice (test trials with feedback) for foreign language vocabulary learning using a practice schedule determined to be optimal according to a cognitive model. In the first test, students chose between an optimized version and a version that merely cycled the vocabulary items. Examination of the time on task data revealed a preference for practice based on the decisions of the cognitive model. In the second test, the system was used to train the component parts of Chinese characters and measure the transfer of knowledge to subsequent learning of Chinese characters. Chinese character learning was improved for students with the relevant optimized training.

Keywords: computer assisted instruction, practice scheduling, prerequisites.

1 Introduction

Because many domains rely on basic facts, this paper addresses a general method for how deficits in basic facts can be addressed through efficient scheduling of practice. To illustrate, consider the case of vocabulary in foreign language learning. The importance of vocabulary knowledge to success in foreign language learning is emphasized by experts in foreign language instruction [1]. However, students are not always motivated to spend time in the repetitive exercises necessary to produce fast and accurate recall of vocabulary items in a language they are learning. Indeed, while taking advantage of the spacing effect (the advantage to long-term learning when practice is distributed in time) is recommended by many authorities [e.g. 2], the high numbers of errors produced during learning that uses spaced repetition might further reduce motivation for extended practice thus making spaced practice a difficult method to apply [3].

To address this dilemma, we have been developing a system that delivers practice of facts scheduled according to the predictions of a cognitive model. This optimal practice scheduling algorithm provides more efficient practice in the laboratory [4] and this paper reports two classroom experiments with the system in a college level Chinese I language class.

B. Woolf et al. (Eds.): ITS 2008, LNCS 5091, pp. 593–602, 2008.

2 Practice Model and Optimization Algorithm

We will begin by describing the ACT-R (Adaptive Control of Thought – Rational) model variant that we use to make scheduling decisions [5]. While ACT-R is best known for its production rule system that models the flow of procedural execution, the model in this paper uses the ACT-R memory equations to predict performance based on a history of learning events.

2.1 ACT-R Variant Practice Model

The model characterizes the strength of an item in memory (vocabulary pairs in the experiments in this paper) by a quantity referred to as "activation". Activation is a continuous real valued quantity specified by Equation 1.

$$m_n(t_{1..n}) = \beta_s + \ln\left(\sum_{i=1}^{n} b_i t_i^{-d_i}\right) \tag{1}$$

In equation 1, n is the number of prior practices for an item for which activation is being computed. The t_i values are the ages (in seconds) for each prior practice of the item by a learner. The summation of these t_i values captures the benefit of frequency of practice while the power function decay parameter (d) models how more recent practice contribute more to the summation. The b value multiplied by each t_i captures the influence of the result of each practice in later predictions. In the case where the past practice was a successful recall, b is typically high, whereas in the case of a failure, b is typically low. The β_s value is initially set at 0 for each student and is estimated during learning to dynamically improve the overall fit of the model for individual differences. These incremental adjustments occur every 50 trials. The d values in the superscript are computed according to Equation 2.

$$d_i = ce^{m_{i-1}} + a \tag{2}$$

In Equation 2 a and c are fitted parameters and m_{i-1} is equal to the activation at the time practice i originally occurred. For example, if we are computing d_7 (decay value for the 7th practice drill), we need to know the activation at the time this drill occurred m_6. Keep in mind that this is recursive since to compute m_6 we need to know ds 1 thru 6. (Since m_0 = -infinity, $d_1 = a$, according to Equation 2.) Equation 2 captures the spacing effect. It represents practice being forgotten more quickly when an item is easier due to narrow spacing of prior practices.

For activation, it has also proven necessary to scale the times between sessions as if it passes more slowly than time within practice sessions [5]. So, if a practice occurred during a previous session its t_i is modified by subtracting a fraction of the intersession from the actual t_i value. For example, if t_6 occurred 100000s ago and 99000s of this period was spent outside the practice sessions, the modifier (0.00046 for the experiments here) is multiplied by the inter-session time and the result (45.5s) is added to the within session practice duration (1000s). For this example t_6 is computed to be 1045.5s according to this procedure. Theoretically this mechanism captures the idea that memory interference (from other items in the set of items being

learned) may be much less intense when practice is not occurring as compared to when it is. Further, at least in the case of classroom experiments where students can be expected to practice the items outside the tutor for classroom work, this parameter also captures some of this learning that occurs without the tutor between sessions. This picking up of classroom learning in this parameter is unintentional and a more principled solution to issue will be sought in future versions.

Equations 3 and 4 are the functions that model recall probability and recall latency as a function of activation. In Equation 3, s captures the variance of the logistic transformation of activation into probability while τ captures the threshold of recall. If s = 0 it implies that activations above threshold result in perfect recall, while activation below threshold means recall always fails. As s increases the transition from 0% to 100% recall becomes increasingly graded. In Equation 4, F scales the latency, which is an exponential function of activation. Equation 4 captures the variable time necessary to perform a correct recall. Fixed costs of each practice are often summed to this variable cost function to capture perceptual motor costs of responding. Like the β_s parameter, F is incrementally adjusted every 40 trials.

$$p(m) = \frac{1}{1 + e^{\frac{\tau - m}{s}}} \tag{3}$$

$$l(m) = Fe^{-m} \tag{4}$$

2.2 Optimized Practice Scheduling

These ACT-R equations model the effect of practice history (including practice spacing, frequency, recency, and success) on both success and latency of later performances. This model allows fine-grained trial by trial predictions of which item of a learning set is optimal to practice next. These predictions are made by computing the long-term efficiency of practice for each item in the set, and then selecting items for practice when they are at a time when their efficiency is maximal. Efficiency is a value computed directly from the model and is equivalent to the long term learning gain divided by the expected time cost of practice for a particular item. Long term learning gains are shown by increase in the "activation" value, which is the strength of an item in memory. Expected time cost depends on activation's effect on latency of recall and on the probability of failure (which results in feedback time). Equation 5 shows the efficiency score equation used for the experiments in this report. The variable r is the retention interval desired for the optimization and is scaled like the t values for the reduced effect of between session forgetting. We set the raw r equal to 30 days, which, scaled by the 0.00046 between session adjustment fixed r at 1191s.

$$eff_m = \frac{p_m b_{suc} r^{-d_m} + (1 - p_m) b_{fail} r^{-d_m}}{p_m (Fe^{-m} + fixedcosts) + (1 - p_m) fixedfailcosts} \tag{5}$$

Figure 1 graphs this function at the parameter values used and shows the inverted u-shaped relationship between efficiency and activation (memory strength). The initial

increase in efficiency, as activation increases, is due to the reduction in failed drill opportunities with higher memory strength. Failed drills consume more time because they require the presentation of feedback for the item, and also because failure itself is typically slower than recall.

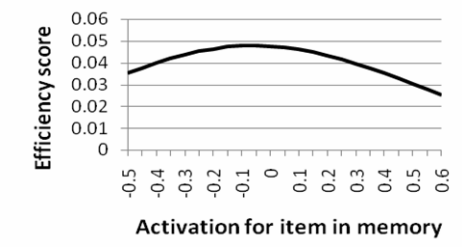

Fig. 1. Graph of the efficiency function for optimized practice

We can also see in Figure 1 that at some point increasing activation no longer provides benefits, but rather leads to less efficient learning. This effect is due to the model of spaced practice, which assumes that the learning from each drill would be more permanent if the drill occurs when activation is less. Each repetition in widely spaced practice reduces activation, as there is more time to forget between repetitions relative to more massed practice. Therefore, spaced practice causes more long-term gain for each trial despite resulting in lower correctness during learning.

Figure 1 illustrates how the interaction of the speed advantage effect for narrow spacing and the long-term learning spacing advantage translate to predict an optimal activation point at which a drill is optimal. The optimal scheduling in the following experiments used an algorithm created using this model to schedule items for practice at this point of maximally efficient learning. To do this, before each trial, the change in the efficiency score as a function of time (the first derivative of efficiency) is computed for every item. Items are selected for immediate practice when the change in the efficiency score approaches 0. If no items have approached 0 either because the derivatives are all positive (in which case more will be learned by waiting for spacing of practice to increase) or because no items have been introduced, the algorithm introduces a new item into the set. After all items have been introduced and the change in efficiency score for all items is positive the algorithm selects the item with the smallest change in efficiency score.

One interesting consequence of Equation 5 is that given a value for r it specifies a specific activation that corresponds to a specific percent correct performance level that should result in maximal efficiency. For Figure 1 this corresponds to -0.04 activation and a percent correct level of 92.6%. However, while the algorithm predicts this specific level at which practice should be optimal, the spacing of practice for each item increases as the learner accumulates practice. Spacing increases because of the increasing stability of the activation value as practice accumulates. This increasing stability is caused by the power function model of forgetting, which indicates that learning from older practices decays more slowly than recent learning. Figure 1 was computed for the following parameter values: s = .261, τ = -0.7, F = 2.322, b_{suc} = 2.497, *fixed success cost* = 0.63s, *fixed failure cost* = 9s, a = .17, c = 0.21 and

r = 1191s. These parameters were estimated by fitting the model for past classroom experiments not reported here. b_{fail} was estimated to be 1.526 for the first practice of any item with later practices $b = 0$.

3 Experiment 1

The first between-subjects experiment described here compares two computerized practice delivery systems in a Chinese I university level course. The control system drills the course vocabulary by cycling through each unit items in random order. The experimental system uses an identical drill procedure, but cycles through items according to the predictions of the model based algorithm designed to maximize learning. In laboratory studies, this optimized model results in large gains in performance (effect size ≈ 1), but it has proven hard to show similar performance advantages in the classroom. This difficulty seems to be mainly due to the relatively small amount of time for which students are assigned to use the tutor.

As a partial resolution to this problem, the following study compares two conditions using time on task as the dependent variable rather than final performance. By doing this we can avoid the problem of having limited classroom time allocated to either condition by simply looking at the total time students spend in each condition. An advantage for time on task is then taken to be evidence for improved motivation and compliance in that condition.

The two practice systems were made available to students by the professor, and students were asked to complete 15 minutes per unit for approximately 1-2% of their semester grade. The webpage that students used to access the two conditions gave simple instructions that were identical in each system. Further, the webpage randomized the order the two tutors appeared on the page so that one condition would not be selected more frequently because of its position on the page. However, students were not blind to condition since the optimized condition was described as, “Optimized Version -- In this version of the practice software, a model of learning is used to choose which flashcard to give for each trial. The model is set to provide approximately optimal spacing and repetition. You can choose to go through either the full set of flashcards for the class, or any particular unit. Your progress is saved, and you can return to where you left off at any time”, while the flashcard condition was described as, “Flashcard Version -- In this version of the practice software, flashcards are delivered in random order, but drop out of the deck when you get them right. You can choose to go through either the full set of flashcards for the class, or any particular unit. Your progress is saved, and you can return to where you left off at any time”.

Students were free to switch back and forth between systems by saving their data and returning to the webpage to continue with another version. This meant that students were not locked into their choice, but rather could change their preference at any time. Other than the practice scheduling the only difference between the versions was that the optimized version listed immediate and one month recall predictions based on the model while the control version kept track of the remaining pairs to be answered correctly to finish one repetition of the selected learning set.

3.1 Participants, Stimuli and Procedures

We only analyzed data from students that had produced complete data on the pre-quiz and post-quiz. According to this criterion there were 90 participants each of which distributed their practice between the two practice methods.

The practice items in each condition were identical for each between subject condition, however the item set varied from 540 pairs in 8 of the 9 class sections (the sections that included regular classroom meetings) to 555 pairs in 1 section of the 9 (the online only section with limited face to face meetings). Since the online section had a correspondingly small sample size it was aggregated with the classroom sections. There were three vocabulary pairs of stimuli and responses for each semantic item: Chinese sound file → English response, Chinese sound file → pinyin response, Hanzi character → pinyin response. (Hanzi are the Chinese characters, while pinyin is the English character orthography for the Chinese pronunciation.) This indicates there were 180 (540 / 3 pairings) classroom semantic items and 185 online semantic items. Every pairing was modeled by an activation value in the optimized condition.

There were 7 units of practice in the one online sections and 10 units of practice in the eight classroom sections. Order of introduction of items was randomized within unit for each version. Additionally, the optimized condition was set so that no related pairs (where the underlying semantic item was the same) were presented with a spacing of less than 2 intervening items. Further, while the flashcard condition randomized all pairings independently to determine the introduction order, the optimized condition randomized units by the groupings of 3 related pairs for each item. Having items introduced in groups was not an explicit model-based decision, but respects the spirit of the model since the model implies that related items should be spaced narrowly at introduction. In contrast, the standard spacing effect suggests practice should be maximally spaced as items were in the flashcard condition.

Practice was distributed according the algorithm in each condition. In the flashcard control condition practice for a particular unit simply involved randomizing the order of the vocabulary items and presenting them one by one. If an item was responded to correctly it was "discarded" and if an item was responded to incorrectly it was put at the end of the order. This procedure continued for each unit until all items were answered correctly for the unit (at which time the order was rerandomized and practice began again) or until the student quit using the tutor.

The drill procedure used for each trial was identical in both conditions. Each drill presented a stimulus on the left side of the screen and allowed the student 20 seconds to respond. If the response was correct there was a 0.5s presentation of a "check mark" to indicate correctness. If the response was incorrect there was a 3s presentation of the correct stimuli and response. If the response was incorrect but the student provided an answer to another item the system gave a 6s study opportunity of both the pair tested and the pair which the student provided a response for.

Pre- and post-quizzes tested the ability to translate 54 items randomly selected from the full item set for each student. Items did not repeat from pre-quiz to post-quiz.

3.2 Results

The main effect of interest was time spent practicing by students in each condition. Using the raw data there was no significant effect (M = 1.26 hours for optimized practice and

M = 1.12 for flashcard practice). However, these values overestimate the preference for the flashcard version since some students merely allowed the flashcard tutor to run without practicing and this happened more frequently for the flashcard condition. To alleviate this, we choose to only consider practice in a condition if probability correct was greater than 0.1 overall. This conservative analysis filtered out students that allowed the tutor to run without practicing. Using this filtered data there was a preference for the optimized version (M's equal 1.24 and 0.86 for optimized and flashcard conditions, t = 2.37, p = .020, Cohen's d effect size = .25).

While the post-quiz results could not be used to establish whether practice was more effective in one condition, it was possible to determine the correlation of practice in each version with gain in post-quiz scores. To do this we computed the correlation of 8 measures of practice in the tutor with the improvement in performance from pre-quiz to post-quiz. These measures included for each condition: total time (raw value), total time filtered to remove p(success) < 0.1 values, total count of correct responses, and probability correct during practice. Only 2 correlations were significant. First, the count of correct responses in the optimized condition correlated with pre-test to post-test gain, r = 0.421 (p = 0.000036). Second, the raw time spent in the flashcard condition was negative correlated with r = -0.366 (p = 0.00121) pre-test to post-test gain. This negative correlation was driven by the few subjects that used the flashcard condition but did not attempt to respond to the drills as discussed above.

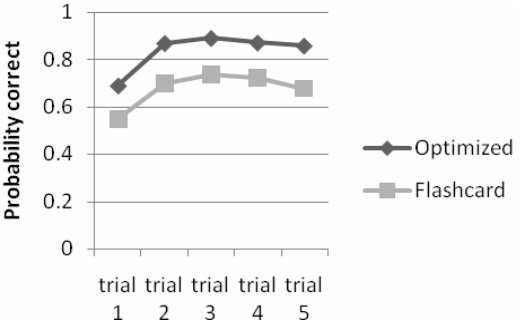

Fig. 2. Average learning curves across the first 5 practices for items in either condition

Figure 2 (created from the raw data) helps us understand why the preference occurred. The figure illustrates the average correctness for items in each condition as a function of the repetition. As the figure illustrates, students found practice in the optimized condition to be easier due to the narrower scheduling used by the optimization condition. In contrast, the lower performance for the flashcard condition showed it was more difficult, which we also take to be an effect of scheduling. Curiously, we also see an advantage for the first drill of practice when the algorithm was simply introducing the item. This benefit is different than the optimized scheduling benefit and was probably due to the procedure of randomizing the related pairings in groups of three and introducing them relatively close together (minimum of 2 intervening trials) in the schedule. However, this grouping would not have significantly affected

later trials because item schedules were dependent on performance after the first trial in the optimized condition.

4 Experiment 2

This between-subjects experiment applied the optimization algorithm to teach learning components rather than to directly train the items that would be tested. To do this experiment we relied on the structure of Chinese characters. Each Chinese character contains one radical item that forms part of the character (sometimes characters have more than one radical, but one radical is always primary). Figure 3 provides an example. As we can note the "see" radical forms part of the verb for "to think". This experiment tested the idea that learning these radical components would improve learning of the characters. While this particular benefit has been found before [6], and theory of part-whole transfer makes it seem likely it can be reproduced [7], we thought it would be interesting to explore this paradigm as a prototype for one type application of the optimal training of basic facts. So, while no single part of this experiment is novel, it remains an important test because it shows real world applicability by bringing a part whole training paradigm into the classroom using a design that measures long term effects on future learning rates.

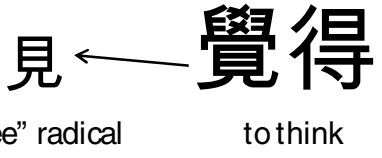

Fig. 3. Example of a Hanzi character and a constituent radical it contains

This experiment was run concurrently with Experiment 1; however, this study independently randomly assigned subjects into either an experimental "radical" training condition or a control "Hanzi" training condition. Both conditions used optimally scheduled practice. The hypothesis was that the "radical' components learned in the experimental condition would produce better learning of Hanzi characters. In contrast, practicing in the Hanzi condition should not improve learning since assessment used a different set of Hanzi than was used in practice. (The Hanzi control condition was intended to rule out the possibility that experience with the software by itself might cause any effects found.) The students were asked to complete one hour practice in the condition they were placed; however, some students practice more or less than that amount.

4.1 Participants, Stimuli and Procedures

We only analyzed data from students that had produced complete data on the pre-quiz and post-quiz. According to this criterion there were 94 participants, 46 of which were randomized into the radical condition and 48 of which were randomized into the Hanzi condition.

The radical set used for training was provided by the Chinese professor and included 154 radical pairs that were identified as components of the Hanzi characters students had to learn for Chinese I. Since some radicals appeared in multiple characters, we choose to introduce the radicals in the order from most frequent to least frequent. For the Hanzi control condition the practice set corresponded to the Hanzi characters for the first three units of the course. In the classroom version of the experiment, the Hanzi set contained 90 practice items, while the online only version contained 106. In both radical and Hanzi practice conditions, the practice items were both radical/Hanzi → pinyin trials and radical/Hanzi → English trials. Thus, for example, there were 77 radicals trained, each of which appeared in 2 pairings.

For the pre and post-quizzes, we tested randomly selected Hanzi items from the last 3 units of the course. Since the post-quizzes occurred at mid semester, this meant that the items were unfamiliar and unlearned for the majority of students. Both pre-quizzes and post-quizzes had the same structure, with 27 items tested each with 2 drill trials. Items did not repeat from pre-quiz to post-quiz. The goal of the quizzes was to produce a learning rate score for each quiz that was a measure of the average correctness on the second drill of a character minus the average correctness of a first drill. For example, on a pre-quiz, a student might get only 1 of 27 items correct for the first drills on the pre-quiz and then get 10 of 27 items correct for the second drills. This would indicate a 33% learning rate for the pre-quiz for this student.

4.2 Results

We were interested in comparing learning rates from the pre-quiz and post-quiz to see if there was an advantage for the post-quiz learning rate as compared to the pre-quiz learning rate. First we ran an ANOVA that compared the gain in learning rate from pre-quiz to post-quiz using the pre-quiz learning rate result as the covariate. This result showed the significant advantage for radical training ($F(1, 91) = 5.62$, $p = 0.020$, $d = 0.49$). The mean gain in learning rate was 12.8% in the radical condition and 6.6% in the Hanzi practice condition. Raw pre-quiz learning rates were 28.8% for radical training and 27.2 for Hanzi training. Raw post-quiz learning rates were 41.6% for radical training and 33.8% for Hanzi training.

Additionally, we were interested in whether this more accurate learning also translated to faster performance on the post-quiz. To look for this we compared the reduction in time for the post-quiz compared to the pre-quiz using the pre-quiz duration as a covariate. This result showed a significant benefit for the radical condition ($F(1, 91) = 4.04$, $p = 0.048$, $d = 0.42$). The mean time saved on the post-quiz was 46.1s for the radical condition and 17.6s for the Hanzi condition. These values are considerable since the average completion time on the pre-quiz was only 394 seconds.

Finally, we wanted to make sure that the result was not driven merely by better time on task in the radical condition. The difference for total practice time was not significant, nor was it significant when pre-quiz duration was used as a covariate ($M = 3771$s for the optimized condition and $M=3428$s for the flashcard condition).

5 Discussion

To address the importance of these basic facts, this paper described a theoretically based, algorithmic method of scheduling performance for such basic facts. This

method takes advantage of well know properties of memory such as the benefits of recency, frequency and spacing to optimize the efficiency of such fact learning. We tested this method of practice in 2 experiments. In Experiment 1 we were able to show that students tend to use the optimized practice more often when given the opportunity to choose an alternative more conventional schedule of practice. This result suggests that students either found the system more effective or more enjoyable. The higher level of correct performance for the optimization condition shown in Figure 2 may be the one reason why people prefer the optimized practice. Experiment 2 focused on efficacy, showing that learning using this method may automatically transfer to new contexts in a naturalistic classroom setting.

Acknowledgments. This research was supported by NIMH R01 MH 68234 and a grant from Ronald Zdrojkowski for educational research.

References

1. Nation, I.S.P.: Learning vocabulary in another language. Cambridge University Press, Cambridge (2001)
2. Dempster, F.N.: The spacing effect: A case study in the failure to apply the results of psychological research. American Psychologist 43, 627–634 (1988)
3. Rea, C.P., Modigliani, V.: The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. Human Learning: Journal of Practical Research & Applications 4, 11–18 (1985)
4. Pavlik Jr., P.I., Anderson, J.R.: Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied (accepted)
5. Pavlik Jr., P.I., Anderson, J.R.: Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. Cognitive Science 29, 559–586 (2005)
6. Taft, M., Chung, K.: Using radicals in teaching Chinese characters to second language learners. Psychologia 42, 243–251 (1999)
7. Singley, M.K., Anderson, J.R.: The transfer of cognitive skill. Harvard University Press, Cambridge (1989)