How Spacing and Variable Retrieval Practice Affect the Learning of Statistics Concepts

Jaclyn K. Maass, Philip I. Pavlik, Jr., and Henry Hua

Institute for Intelligent Systems and Department of Psychology,
University of Memphis, Memphis, TN, USA
{jkmaass, ppavlik, hhua}@memphis.edu

Abstract. This research investigated key factors in learning conceptual material about statistics, and tested the effect of variability during retrieval practice. The goal was to build a model of learning for schedule-based interventions. Participants (n = 230) completed multiple reading and test trials with fill in the blank sentences about basic statistics concepts. The experiment was a 2 (trial type: read or drill) x 3 (learning trial spacing: wide medium, or narrow) x 2 (fill-in term during learning: variable or constant) x 2 (fill-in term during posttest: variable or constant) within-subjects design. The model of the results captures the data with recent and long-term components to explain posttest transfer and the testing and spacing effects. These results, and data on the conceptual confusions amongst statistical terms, are discussed with respect to implications for future intelligent learning systems.

Keywords: Fill in the blank · learner modeling · testing effect · spacing effect

1 Introduction

Over the history of spacing effect research, many experiments involved memorizing a set of items that were completely unfamiliar to the learners [e.g. 1, 2]. This procedure stems from an effort to avoid confounding the participant's familiarity with the topic with the effects of spaced practice, but the use of completely novel stimuli does not aptly simulate student learning in a classroom setting. Recently however, there has been a growth in the number of studies using more educationally relevant, ecologically valid, and complex material [e.g., 3]. Oftentimes, the spacing effect goes hand in hand with the testing effect (i.e., retrieval practice). Like the spacing effect, the testing effect is a fairly consistent phenomenon well replicated with different formats and domains, such as vocabulary [e.g., 4], paired-associates [e.g., 5], procedural knowledge [e.g., 6], and text materials [e.g., 7]. However, much of this type of research uses the same items for retrieval practice and posttest. The problem is that this provides a measure of memory for verbatim responses and does not offer a measure of deeper, integrated learning or transfer.

While many current intelligent tutoring systems focus on procedural knowledge [e.g., 8], certain domains require the prerequisite of a strong conceptual knowledge

base. Therefore, we have chosen to focus our work on learning semantic knowledge. We aimed to use model-based discovery [9] to describe learning in this task in a way that will aid in the creation of an tutoring system which implements simple reactive artificial intelligence to optimally schedule the repetition of retrieval practice for conceptual knowledge. This paper focuses on three key features of such a model in this domain: spacing of practice, retrieval practice, and variable practice. The current experiment tested those factors using a free entry cloze (fill-in-the-blank) task with a collection of 18 sentences about statistics. This is the second experiment in a series of cloze item research aimed at building an intelligent tutoring system (ITS) focused on didactic, verbal and/or conceptual information. Unlike the previous study, the current experiment used related sentences to measure spacing effects in an educationally relevant domain, rather than with decontextualized trivia facts [10]. Participants completed 162 drill and reading trials with sentences about statistics concepts (e.g., definition of a sample, characteristics of a normal distribution, etc.). For this paper we developed a learner model that predicts performance given the prior accuracy, spacing, repetition, testing, and concept difficulty.

In addition to the above goals, we had four experimental hypotheses: 1) wider spacing will lead to worse performance during learning but better performance at posttest; 2) spacing effects will be larger for sentences in which participants fill in the same term during the learning trials; 3) testing trials will have more impact on learning than reading trials; and 4) testing with variation in the retrieval term will lead to more generalizable learning. Further, the free entry nature of the cloze task provided rich insight into students' confusions, which has direct implications for creating customized hint and/or error messages in the next stage of this research project.

2 Methods

2.1 Design

A brief overview of terminology specific to this experiment may be helpful. Each trial involved the presentation of a sentence either to be read by the participant (reading trials) or with a missing word to be filled in by the participant (drill trials). Retrieval term refers to which one of four key terms in each sentence can be left blank for a drill trial. Variability in retrieval term refers to having to fill in different retrieval terms across drill trials for a sentence. The two main portions of the experiment are referred to as the learning portion (consisting of six trials for each sentence), and the posttest portion (consisting of three drill trials for each sentence).

This study used a within-subject design with the following factors: 2 (trial type: read or drill) x 3 (learning trial sentence spacing: wide medium, or narrow) x 2 (learning trials retrieval term: variable or constant) x 2 (posttest trials retrieval term: variable or constant). While the design was fully factorial for the other factors, the assignment to reading or drill trial for each sentence during the learning portion was selected randomly, with a 50% chance of reading on Trials 1 and 4, and a 25% chance of reading for Trials 2, 3, 5, and 6. The posttest portion (Trials 7-9) contained only drill tri-

als. If the retrieval term was constant during the learning or posttest portions, the participant would only see one of the four key terms left blank in the sentence, and this retrieval term would remain constant, e.g. always the 3rd of 4, for all trials. If the retrieval term was variable, any one of the four key terms could be left blank for the participant to fill in for each trial. For the learning portion, half of the sentences were randomly selected to have a constant retrieval term and half were selected to have variable retrieval terms. Retrieval term variability was manipulated for the learning portion and for the posttest portion, independently, meaning each of the nine sentences that had a constant retrieval term during learning was individually assigned as having variable or constant retrieval terms during posttest. Therefore, on average, 4.5 of the constant retrieval term sentences during learning were tested as variable at posttest and the other 4.5 continued as constant at posttest). Note that when the posttest retrieval term varied, it was randomly selected from only the other three fill in locations, excluding the ”constant” location for that learner. Therefore, the constant learning to variable posttest had no repetition of responses (i.e., all transfer posttest drills). Conversely, the constant to constant condition only ever filled in one of the key terms (i.e., no transfer posttest drills). The other two conditions, variable to variable and variable to constant, would have had some posttest items they had seen before and some they had not (i.e., some transfer posttest drills). Of course by “transfer” we are not referring to completely novel sentences; rather, the retrieval term for that sentence would not have been previously retrieved during the learning drill trials. This type of “within-sentence” posttest transfer measure has been used by others [11].

Therefore, participants received some sentences in each of the four drill trial conditions: variable retrieval terms during learning and variable retrieval terms during posttest (variable-variable), variable retrieval terms during learning and a constant retrieval term for posttest (variable-constant), constant retrieval term for all learning trials and variable retrieval terms for posttest trials (constant-variable), and a constant retrieval term for all learning trials and constant retrieval term for posttest trials (constant-constant). Each sentence was also randomly placed in one of the three spacing conditions during practice, resulting in 12 possible retrieval practice conditions.

2.2 Participants

The experiment was delivered through the Amazon Mechanical Turk (MTurk) service, an online data collection platform. A total of 231 people participated, but one subject was excluded (n = 230) because they produced no correct responses. The requirements for participation were for the person to be at least 18 years of age, a native English speaker from the United States or Canada, and to be a reliable MTurk “worker.” This last qualification was to ensure quality results. It requires participants to have previously completed at least 50 tasks (referred to as “Hits”) on the website, with at least 95% of those tasks approved (i.e., the person had done adequately enough to not be refused payment). Although this sample is not restricted to formal students, studies have reported that MTurk participants appear to produce qualitatively and quantitatively similar results to university and other online participants [12]. MTurk users were paid $4 for approximately 45 minutes of participation.

2.3 Materials

The 18 sentences used for the experiment were developed from basic statistics content authored by the three authors of this study. The sentences were not designed to be explicitly related, but naturally resulted in the reuse of several key terms; for example, several sentences mention the concept of *standard deviation* with regards to various other concepts. For each sentence we chose four crucial words (or two-word phrases) to create four versions of each sentence in which each version has one word or phrase removed (i.e., left blank). For example, in the sentence, "Although samples are variable, they are intended to represent the population from which they come," the four selected key terms were *samples*, *variable*, *represent*, and *population*. Only one key term was left blank for any given trial (i.e., the participants never saw a sentence with two blanks, except to indicate the same word used twice).

2.4 Procedure

The experiment was delivered online, using the Fact and Concept Training (FaCT) system, which is designed to handle complex designs of this sort [13]. The sentences were delivered in the center of the screen for study trials and for drill trials. For drill trials, participants were told whether their answer was correct or incorrect; after an incorrect response, the correct response was displayed on the screen for a review period. Participants were given eight seconds to read each sentence both for reading trials and for the review period after incorrect drill trials. Reading trials never indicated which terms could be left blank during drill trials.

For the drill trials, participants had 12 seconds to begin to type their response for the missing word, otherwise the trial timed-out, was counted as an incorrect response, and the system continued to the review. However, as long as the participants were trying to answer (i.e., if they had started to type in the answer box), the system allowed the participants as long as they wanted to finish typing. For the review that occurred after an incorrect drill trial, the computer program showed the whole sentence again, but with no missing words (i.e., without indicating which word or phrase had just been blank) for eight seconds, identical to reading trials.

The experiment consisted of the informed consent, instructions, learning portion (consisting of the 18 sentences, presented six times each in a combination of drill and reading trials, each at different spacing intervals), a 5-minute distractor task (an N-back task, data for which was not available at the time of writing due to a parsing error), and a posttest (three drill trials for each sentence). The learning portion lasted an average of approximately 30 minutes, the distractor task (i.e., retention interval) was approx. five minutes, and the posttest lasted an average of approx. ten minutes.

3 Results and Discussion

In order to eventually build an "intelligent" program that can adapt and respond to an individual student's progress, we must first understand the underlying features affect-

ing memory. To begin this process, we ran two repeated-measure ANOVAs to determine the changes in performance between the learning portion and posttest portion, based on 1) spacing of practice trials, and 2) variability during practice. The levels of the dependent measure, referred to as the trial variable, were proportion correct during the last learning trial (Trial 6) and proportion correct on the first posttest trial (Trial 7). The results of the first ANOVA indicated a presence of the spacing effect (i.e., a significant interaction between spacing and trial), $F(2, 458) = 29.28$, $p < .001$. Namely, during the last learning trial, performance was best for those sentences with narrow spacing, followed by medium spacing, with worst performance for the wide spacing (all spacing pairwise comparisons had *p*s < .001). However, these differences between spacing conditions disappeared at the first posttest trial (all *p*s >.2).

Similarly, the second repeated measures ANOVA, comparing change in performance from learning to posttest between sentences with variable retrieval terms during practice and sentences with a constant retrieval term during practice, showed a significant retrieval variability by trial interaction, $F(1, 230) = 138.11$, $p < .001$. During learning, participants scored higher on sentences that had a constant retrieval term than on those sentences with variable retrieval terms, pairwise comparison $p < .001$. However, this difference was no longer significant at posttest, pairwise $p = .44$. This shows a much steeper forgetting rate over the 5-minute retention interval for those with a constant retrieval term during practice. Our final preliminary analysis confirmed the presence of the testing effect through a simple logistic regression model (not shown) using only two parameters: count of prior reading trials for the sentence and count of prior drill trials for the sentence. This revealed drill trials to be about three times as effective as readings ($z = 7.531$, $p < .001$). In order to build a tutoring program, however, we will need more than these preliminary analyses can offer. Such a program will require a predictive model of student performance which would act as the main source of "intelligence" in optimally scheduling retrieval practice.

The process of developing a predictive model of student performance required an iterative, theory driven method known as model based discovery [9]. Specifically, we started by entering in factors we presumed, based on previous memory research thus far discussed, would be most influential in learning. Through a process of elimination, focused on parsimony and correspondence with prior theory, we were able to narrow down which factors had the greatest influence on learning. For example, we expected to see fast forgetting in addition to durable learning, as a function of spacing of practice, based on prior work with other verbal material [e.g., paired associates; 14]. Other features were implemented based on prior learner modeling. For instance, a distinction between the effects of incorrect and correct responses was based on Performance Factors Analysis [15], and a general track of all prior performance, which is similar to work with Bayesian knowledge tracing [16]. The final model is a logistic regression with the following parameters or factors: prior performance, term difficulty, recency, pure massing effect, spacing for drill trials, spacing for incorrect drill trials, and spacing for key terms. Table 1 contains the coefficients and *z* scores obtained for each factor, which is explained in more detail below. The *z* scores are a particularly useful measure to compare the coefficients and understand their relative importance in the model. For example, the most predictive features of the model, described in detail

below, are recency (z = 51.50) and term difficulty (z = -42.71). The model fit with an R^2 of .399 and a mean absolute probability deviation of .298 for the predictions of each of the 28,912 drill trial observations across the 230 learners. We also cross-validated the model using ten runs of 10-fold cross-validation. We observed an R^2 of .399 in the training and an R^2 of .357 in testing, which means we retained 89% of the fit when generalizing. It should be noted that the model was found using the entire dataset (rather than a portion); thus, the cross-validation might suffer some inflation. However, we did not further tune the model to optimize the cross-validation results. Additionally, we minimized the number of parameters relative to the number of conditions, to help protect against overfitting.

Table 1. Coefficients, standard errors, and z scores for each model parameter (all ps < .001).

Factor	Coefficient	*SE*	*Z* score
Intercept	-2.53	0.04	-61.95
Prior performance	0.92	0.02	40.56
Recency	8.78	0.17	51.50
Term difficulty	-1.48	0.03	-42.71
Pure massing	-0.15	0.03	-4.78
Spacing for drills	0.37	0.01	28.71
Spacing for incorrect trials	-0.35	0.01	-26.09
Spacing of terms	0.05	0.01	8.47

Prior performance is measured as the proportion[1] of previously correct trials which is transformed into a logit, ranging from, -∞ to +∞. This variable is seeded with a value equal to the overall grand mean to prevent the prior probability from ever actually taking an infinite value. The transformation to a logit serves to increase the predictive influence as prior learning nears 0 or 1. The high z value for this parameter indicates that transformed previous accuracy is one of the biggest predictors of future success. The recency coefficient in the model scales the influence of the most recent trial as a function of how many seconds ago that trial occurred according to the function 1/ sqrt (recency in seconds). This function was computed for both the most recent trial of any type (read or drill) and the most recent same response drill trial. These values were summed before being log transformed. This method essentially double counts the amount of learning for the specific recent drill trial compared to a recent reading trial. Otherwise, this method is equivalent to the base level activation computation in the ACT-R computational modeling system [17]. The term difficulty parameter is the average number of incorrect responses for that key term across all participants. This is averaged across the sample, with the expectation that the values for these terms will generalize to another sample in the future. Again, we use the logit of this proportion to transform the scale in a way that increases the influence of propor-

[1] To eliminate the possibility of calculating a logarithm of 0 (which would be undefined), all logarithmic transformations were calculated from the observed value plus 1.

tions near zero. The difficulty of the terms (i.e., their error rates) is a key component in this model and will be discussed in more detail below.

The pure massing effect parameter accounts for the effect of having prior back to back drills of the same sentence with no spacing at all. This effect is computed by taking the log of the number of prior back-to-back same-response drills, multiplied by the log of the total number of back-to-back repetitions of the sentence (i.e., both drill and reading trials). The negative coefficient for this parameter captures the negative effect of fully massed practice and the increasingly negative effect as more back-to-back drills are added. While this coefficient had the least significance, it was necessary to model the crossover interactions seen in Figure 1.

One feature of this model is that the spacing is not categorical (e.g., narrow, medium, wide), but rather is a function of the mean average time between prior repetitions of the sentence. This means that the model can be used to make predictions for spacing as a function of time rather than as a nominal effect, which would be less useful for inferring pedagogical decisions from the model. The last three parameters account for different spacing effect interactions, and are calculated by taking the log of some count of prior practices times the log of mean average spacing. The first of these parameters uses the log of the count of all same-response drill trials (i.e., previous drill trials for the same sentence, with the same retrieval term), which was found to predict a strong positive gain that increases with wider spacing. The second uses the count of *incorrect* same-response drill trials; its negative coefficient indicates that incorrect drills reduce the prediction of future successes for the sentence, indicating that incorrect retrieval attempts do not add much benefit to learning. This parameter, along with the first prior performance parameter, allows the model to adapt to student responses on both a student and item level. This capability will be key to using the model in the adaptive learning system we plan for future work. Finally, the last spacing coefficient scales the added effect of the prior repetitions of a *specific retrieval term*. Since some key terms were repeated in multiple sentences, this parameter demonstrates that specific retrieval term practice transfers between different sentences. It also provides preliminary evidence, combined with the semantic errors discussed below, that participants treated this as a meaningful task rather than as rote memorization.

In addition to modeling learning, our experiment also aimed to investigate the effect of variability during retrieval practice. Specifically, we hypothesized that the varying of retrieval terms in a sentence would have an effect on posttest performance for key terms they had not retrieved during practice. Figure 1 (left) illustrates this result by showing the change in performance from Trial 6 (the last learning trial) to performance on Trial 7 (the first posttest trial). The decrease in performance is likely attributed to the 5-minute distractor task between the learning and posttest portions. Figure 1 (right) illustrates the model's fit to Trials 6 and 7. Although only Trials 6 and 7 are graphed, the model fit just as well for the other trials; a full graph could not be included due to space restrictions.

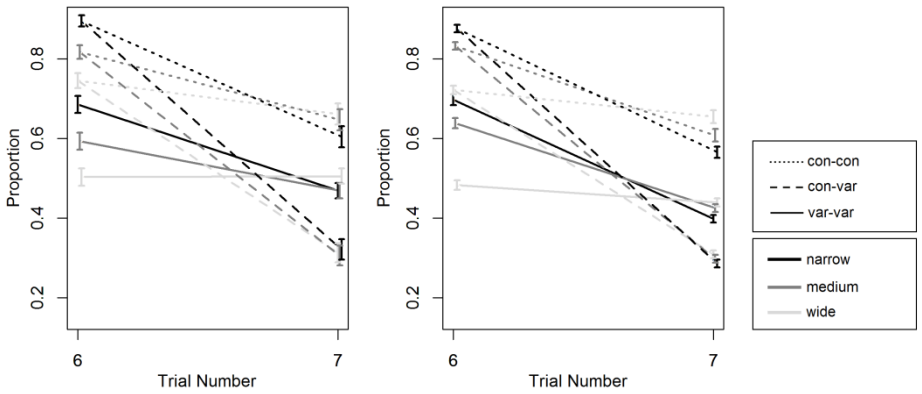

Fig. 1. Fig. 2. Actual performance by participants (left) and predicted performance by model (right) for the last learning trial (Trial 6) and first posttest trial (Trial 7). Line color indicates spacing level. Line style indicates retrieval term condition: constant during learning and constant during posttest (con-con), constant during learning and variable during posttest (con-var), and variable during learning and variable during posttest (var-var).

In Figure 1 (left), we see that participants performed best during learning when they were repeatedly tested on the same terms (con-con); in this condition, all posttest retrieval terms were the same as they had been tested on during learning, which explains why performance is the best for these sentences during both learning and posttest. Perhaps more interesting is the performance of the other groups on this figure. First we can see that when constant retrieval term practice switches to variable retrieval terms at posttest (which included no overlap between retrieval terms during learning and posttest; con-var), posttest performance is much worse than when retrieval terms were variable during learning and variable during posttest (var-var). In other words, variable practice resulted in better transfer to the varied retrieval terms at posttest. Note that the var-var lines in the figure include the var-con condition trials because the var-con was effectively equivalent for Trials 6 and 7. We also see a reduction in the spacing effect when there is less constancy, similar to other results showing reductions in spacing with variable learning [18].

The last component of this experiment we will discuss is the confusions witnessed in participants' incorrect responses. These confusions may have important implications for maximizing the potential of cloze practice. Figure 2 provides a partial graph of the most common confusions participants made between terms related to variance. The proportion of times a specific confusion was made is denoted on the links between terms. For example, participants fairly often (incorrectly) used the term *variance* in place of *standard error* (7.7% of the time), *density* (10.6% of the time), and *standard deviation* (6.9% of the time). This information about commonly confused terms will be indispensable for creating specific hint and error messages in an ITS.

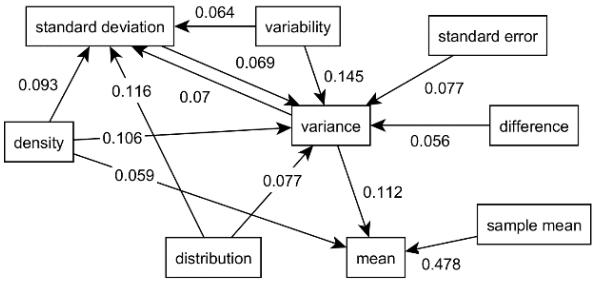

Fig. 3. The proportion of confusions participants made between terms. The figure is read as [correct answer] → [student response].

4 Conclusion

Our model captures many facets of learning conceptual material with retrieval practice and may be used within an adaptive learning system to optimally schedule retrieval practice. The model is able to accurately capture the effects of interest with relative parsimony through the integration of short term practice effects and long-term spacing effects to explain specific and general learning, while accounting for individual differences in performance. The results of this experiment replicate and extend results from an earlier study with trivia facts, showing similar short-term learning and spacing effects for variable posttest fill-ins [10]. The results also support the current experiment's hypotheses. Specifically, spacing effects were larger when the retrieval term was constant during learning; drill (i.e., test) trials had significantly more impact on learning than reading trials; and variation in retrieval terms during learning led to better performance on posttest trials with variable retrieval terms. Our hypothesis that wider spacing would lead to worse performance during learning but better performance at posttest, was also supported, although the difference at posttest was not significant. We attribute this to the relatively brief retention interval (5 minutes), since the spacing effect is usually more pronounced over time [17]. Lastly, the slower forgetting rate for sentences with variable retrieval terms during practice (rather than constant retrieval terms) may suggest more durable learning when practice is variable. However, this is a tentative conclusion. Future experiments will attempt to replicate these results with multiple retention intervals in an effort to get closer to our goal of developing an ITS to effectively schedule practice to maximize conceptual learning.

5 References

1. Carpenter, S.K., Pashler, H., Wixted, J.T., Vul, E.: The Effects of Tests on Learning and Forgetting. Memory & Cognition: 36**,** 438-448 (2008)
2. Ebbinghaus, H.: Memory: A Contribution to Experimental Psychology. Teachers College, Columbia University, New York (1913/1885)

3. Smith, M.A., Karpicke, J.D.: Retrieval Practice with Short-Answer, Multiple-Choice, and Hybrid Tests. Memory, 1-19 (2013)
4. Karpicke, J.D., Roediger, H.L., III: The Critical Importance of Retrieval for Learning. Science: 319, 966-968 (2008)
5. Underwood, B.J., Ekstrand, B.R.: Effect of Distributed Practice on Paired-Associate Learning. Journal of Experimental Psychology: 73, 1-21 (1967)
6. Rohrer, D.: The Effects of Spacing and Mixing Practice Problems. Journal for Research in Mathematics Education: 40, 4-17 (2009)
7. Roediger III, H.L., Karpicke, J.D.: Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. Psychological Science: 17, 249-255 (2006)
8. Leelawong, K., Biswas, G.: Designing Learning by Teaching Agents: The Betty's Brain System. International Journal of Artificial Intelligence in Education: 18, 181-208 (2008)
9. Baker, R.S.J.d., Yacef, K.: The State of Educational Data Mining in 2009: A Review and Future Visions. Journal of Educational Data Mining: 1, 3-17 (2009)
10. Pavlik Jr., P.I., Geno, A.: Deconstructing Cloze Practice. Manuscript submitted for publication (2015)
11. McDaniel, M.A., Anderson, J.L., Derbish, M.H., Morrisette, N.: Testing the Testing Effect in the Classroom. European Journal of Cognitive Psychology: 19, 494-513 (2007)
12. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running Experiments on Amazon Mechanical Turk. Judgment and Decision making: 5, 411-419 (2010)
13. Pavlik Jr., P.I., Presson, N., Dozzi, G., Wu, S.-m., MacWhinney, B., Koedinger, K.R.: The Fact (Fact and Concept Training) System: A New Tool Linking Cognitive Science with Educators. In: McNamara, D., Trafton, G. (eds.): Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society, 1379-1384. Lawrence Erlbaum, Mahwah, NJ (2007)
14. Pavlik Jr., P.I.: The Microeconomics of Learning: Optimizing Paired-Associate Memory. Dissertation Abstracts International: Section B: The Sciences and Engineering: 66, 5704 (2005)
15. Pavlik Jr., P.I., Cen, H., Koedinger, K.R.: Performance Factors Analysis -- a New Alternative to Knowledge Tracing. In: Dimitrova, V., Mizoguchi, R., Boulay, B.d., Graesser, A. (eds.): Proceedings of the 14th International Conference on Artificial Intelligence in Education, 531-538. Brighton, England (2009)
16. Corbett, A.T., Anderson, J.R.: Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. User Modeling and User-Adapted Interaction: 4, 253-278 (1995)
17. 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)
18. Appleton-Knapp, Sara L., Bjork, Robert A., Wickens, Thomas D.: Examining the Spacing Effect in Advertising: Encoding Variability, Retrieval Processes, and Their Interaction. Journal of Consumer Research: 32, 266-276 (2005)