Passing the Test: Improving Learning Gains by Balancing Spacing and Testing Effects

Hedderik van Rijn (D.H.van.Rijn@rug.nl)[a,b], Leendert van Maanen (leendert@ai.rug.nl)[b], Marnix van Woudenberg[b]

[a] Experimental Psychology
University of Groningen

[b] Artificial Intelligence
University of Groningen

Abstract

Where the spacing effect promotes longer intervals between facts that need to be memorized, the testing effect argues for intervals that are short enough to recall the facts. As the ease by which facts are memorized differs greatly between students, an individual assessment of how well certain facts are represented in memory is required to successfully balance spacing and testing effects. We present a model that adapts itself to the abilities of the student, and show in a real-world experiment that this model outperforms other approaches to spacing.

Keywords: spacing-effect; testing-effect; subsymbolic model tracing; cognitive model.

Introduction

The last couple of years have seen a renewed interest in applying insights from fundamental memory research in real-world settings. One of the most visible lines of work are studies to the application of the spacing effect. The spacing effect, first described by Ebbinghaus (1913/1885) at the end of the 19th century, is the positive effect on factual recall that is observed when study trials are temporally separated. Thus, the probability of recall of facts learned in a spaced sequential order (e.g., abcabcabc or abc-*break*-abc-*break*-abc) is higher than the probability of recall of facts that are learned massed (e.g., aaabbbccc). The consequence of this finding is that the presentation sequence of a to-be-memorized list of facts partly determines how well these facts will be recalled on a later test: items on a list that presents the items with wider spacing will be recalled better than items on a list that presents the items as many times as the first list, but massed instead of spaced.

This observation was central to much applied research in the 1960s and early 1970s. Using the possibilities provided by digital computers, scientists tried to construct optimal learning schedules. Although some of this work has stood the test of time from an applied or commercial point of view (e.g., the Pimsleur and Leitner methods are still available commercially), the methods used by these early systems are relatively simple and the learning gains often did not outweigh the extra investment associated with using these systems. This led to a decline in applied research on the spacing effect, although over the decades, more fundamental research on this effect has thrived (for reviews see Dempster, 1988, Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). Only recently has attention again shifted to using algorithms to determine the optimal schedule for learning (Wozniak & Gorzalanczyk, 1994, Pavlik, 2007, Pavlik & Anderson, 2008).

Another finding that has a potentially large effect on how an optimal sequence has to be constructed, is the testing effect. This effect can be described as: "If students are tested on material and successfully recall or recognize it, they will remember it better in the future than if they had not been tested" [but merely studied the same material] (Roediger & Karpicke, 2006, p.249, see also Carrier & Pashler, 1992). As it is generally assumed that memory decays over time, increasing the interval between successive presentations makes it more likely that an item cannot be recalled. Therefore, spacing beyond a certain interval will be associated with lower learning gains because of failing the testing effect (c.f., the inverse u-shape often observed when the performance on a test is plotted as a function of the interval between two presentations, Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008).

When it comes to computing an optimal presentation sequence for fact learning, spacing and testing have different interests. For the spacing effect, increased spacing is theoretically preferred. But for the testing effect, small to no spacing would theoretically provide the best results. One of the aims of the study reported here is to reconcile these seemingly conflicting requirements.

An interesting observation in almost all work on the spacing effect is that the "optimal schedule" is defined as the schedule that reaches the best performance (often defined as the highest probability of recall) over a longer timeframe. Although this is of course what the goal of all learning *should be*, the goal of learning in a real-world situation is often more pragmatic: passing the next day's test by studying for a limited, often more or less fixed amount of time. So, although the results of more than a century of spacing results can be used for the real-world situation of having to learn numerous vocabulary word pairs for a foreign language test that is scheduled a couple of weeks or months in advance (c.f., Wozniak & Gorzalanczyk, 1994), these results do not necessarily tell us anything about the pragmatic goal of learning: What method should a student use to learn a set of 20 vocabulary word pairs for a potential test tomorrow, knowing that, because of soccer practice, favorite TV-shows and other homework, all he or she has is 15 minutes to spare?

Note that this real-life situation differs quite a bit from typical experimental setups: First, to prevent effects of prior knowledge, the learning materials in experimental contexts are often selected in such a way that none of the participants has any relevant prior knowledge (by either learning sequences of nonwords, e.g., Ebbinghaus, 1913/1885, very

obscure facts, e.g., Cepeda et al, 2008, or word-pairs from languages previously unstudied, e.g., Pavlik & Anderson, 2008). In contrast, when learning for a vocabulary test, most students bring additional knowledge to the learning session from earlier experiences with that language. Second, in most studies the list of word pairs presented to the participants is much longer than the 10 to 30 words that typically have to be learned in a single real-life learning session. Third, the retention interval (defined as the time between the final test on the learned materials and the last study of the materials) in most studies is less than a day (221 out of 254 studies reviewed in Cepeda et al, 2006, used an interval less than a day). Fourth, where many experimental studies aim for finding a general law that describes the effects of different types of spacing on performance in general, the goal of an individual student is not striving for the best performance of a larger group, but for the optimal results on his or her test. As the speed and ease by which vocabulary is learned differs greatly between individuals (e.g., Baddeley, 2003), settings that are optimal for the group as a whole might not be the optimal settings for an individual. These differences are less substantial with respect to the spacing effect than with respect to the testing effect. That is, irrespective of the individual expertise in vocabulary learning, the spacing effect predicts that increased spacing provides better scores. However, with respect to the testing effect, individual differences greatly determine the probability of recall of a particular item. Since successful recall is associated with better learning gains, it is important to account for individual differences in such a way that facts are presented before they cannot be recalled anymore.

To test whether the general findings associated with spacing and testing effects hold when these issues are taken into account, we ran an experiment that closely mimics everyday learning contexts. In this experiment, pre-university level students were asked to memorize Dutch translations of French words in a computer-supported learning session of 15 minutes. During learning, the schedule of presentations of the Dutch-French word pairs was computed according to one of four algorithms.

Algorithm 1 was based on a flashcard strategy: the study items were clustered in sets of 5 which were presented individually until all items in the set had been responded to correctly once. After all sets had been presented, the sequence was started anew until time ran out. **Algorithm 2** is an implementation of the spacing method proposed by Pavlik and Anderson (2005), which will be discussed below. **Algorithm 3 and 4** are adaptations of the original Pavlik and Anderson algorithm in that the model that is used to determine the optimal sequence is dynamically adapted on the basis of the observed performance of the student while taking the testing effect in account. Before turning to these algorithms, we will first discuss Pavlik and Anderson's spacing model and how this model can be applied to provide an optimal learning sequence.

Pavlik & Anderson's Spacing Model

The spacing model proposed by Pavlik and Anderson (2005, referred to as the PA model) is based on the work of Anderson and Schooler (1991). Anderson and Schooler demonstrated that the "availability of human memories for specific items shows reliable relationships to frequency, recency, and pattern of prior exposures to the item" (Anderson & Schooler, 1991, p.396). Eventually, the following formula was proposed to express the availability (or activation) A of a certain item i at a certain time (t) as a function of prior encounters:

$$A_i(t) = \sum_{j=1}^{n} (t - t_j)^{-d_j}$$

According to this equation, which has become central to all memory related models created in the ACT-R cognitive architecture (Anderson, 2007), all previous encounters ($t_1..t_n$) of the item i contribute to its current activation. However, the older an encounter (t_j represents the time of encounter j), the smaller the contribution of that encounter to the total activation. The speed of this decline is expressed by $-d_j$, the decay parameter. Although initially $-d_j$ was assumed to be variable for different encounters j (Anderson and Schooler, 1991, provided an equation to account for some spacing effects but downplayed its importance by noting that "its exact form is a bit arbitrary", p.407), it quickly became a parameter that was treated as a constant (d=.5) as different values for different encounters did not add much explanatory power for most tasks to which this equation was applied. However, in contrast to the original work of Anderson and Schooler, in none of these later tasks was spacing a factor of importance. To account for a broader range of spacing phenomena, the PA model reintroduced individual decay values for individual items.

Pavlik and Anderson proposed to relate the decay values for the individual encounters to the activation of that particular item at the time of the encounter (c.f., Rescorla-Wagner's, 1972, model of learning). As recently presented items have a high activation, the second encounter of an item presented twice in quick succession will be associated with a high decay value. Therefore, the long-term influence of this item will be small as its activation will decay quickly. On the other hand, an encounter of an item of which the last presentation was longer ago (and therefore has a lower activation) will receive a lower decay value, resulting in more long term impact on the activation of that item. The proposed equation calculates the decay, d, for encounter j of item i by calculating the activation of that item (A_i) at the time of encounter j.

$$d_{ji} = ce^{A_i(t_j)} + \alpha$$

In this equation, alpha represents the decay intercept. This intercept is the minimum decay for an encounter that will also be used as decay value for the first encounter. The decay scale parameter c determines the relative contribution of the activation dependent component. Pavlik and Anderson (2003, 2005, 2008, Pavlik, 2007) have shown in a series of studies that these equations account for a wide range of spacing-related learning phenomena.

In the PA model the activation of a fact determines both the probability of recall of that fact and the latency associated with recalling that fact. For the probability of recall, the activation of the fact is compared to the retrieval threshold while taking into account the noise that is associated with declarative memory. If the activation of a fact is higher than the retrieval threshold, that fact can be recalled. However, if the fact is below the retrieval threshold, it is unavailable for further processing. Apart from the probability of recall, the activation also determines the latency of a retrieval at each point in time (t) according to the following formula:

$$L_i(t) = Fe^{-A_i(t)} + \text{fixed time}$$

In this equation, F is a scaling factor and "fixed time" refers to the time cost of all non-fact-retrieval processes required in giving the answer.

Applying the Spacing Model

In Pavlik and Anderson (2008), the spacing model is used to actively determine the optimal sequence for learning a list of Japanese-English word-pairs. In this paper, Pavlik and Anderson do not explicitly discuss the testing effect (although it is partly accounted for), but instead focus on presenting a sequence of items that have the highest activation gain per second of practice. Thus, the positive effects of increased spacing intervals on the probability of recall are balanced against the negative effects that increased intervals have on accuracy of immediate recalls. This results in a series of complex formulae to determine the learning gains of test-trials and study-trials.

An alternative and simpler approach is to determine the optimal sequence on the basis of the activation of the word-pairs in relation to the retrieval threshold. That is, if we assume on the basis of the combination of spacing and testing-effects that the time between two encounters is optimal *just before* the activation of the fact drops below the retrieval threshold, an optimal sequence can be determined on the basis of the activation of all facts.

Algorithm 2: Default PA

On the basis of the approach discussed above, the default PA model (i.e., pre-2008) can be used to determine the optimal spacing sequence: as soon as a fact is about to fall below the retrieval threshold, it has to be presented again. If no previously presented fact is close to the threshold, a new fact can be introduced. More precisely, as it could be that a fact drops below the retrieval threshold while another fact is being tested, the algorithm computes the activation of all facts 15 seconds ahead to determine whether to introduce a new fact or present a previous one. If all facts have been introduced, the fact with the lowest activation is selected for presentation. The performance of this algorithm is highly dependent on the accuracy of the internal activation representations, which are in turn dependent on the choice of parameter values. Although the PA model has been tested extensively, no fixed set of parameter settings have emerged yet. The values for the decay scale (c) range (Pavlik & Anderson, 2005, 2008) from 0.143 to 0.495, and for the decay intercept (alpha) from 0.058 to 0.300. The threshold parameter is typically set at -0.704. As these parameters have been fit to experiments with longer study session than used in the current experiment, we explored the effects of different settings on the resulting sequences. As a threshold that is too low results in extended spacing (e.g., in our explorations, sometimes all word-pairs were presented before the first word-pair was repeated), we decided to raise the retrieval threshold to -0.500. Following similar reasoning, the decay intercept and the decay scale were set at .25. With respect to the latency equation, we decided against separate estimations for F and the "fixed time". In Pavlik and Anderson (2008), F is set at a value larger than 1 (1.29) indicating an enhanced effect of A_i on the latency. At the same time, using a "fixed time" diminishes the effect of A_i on the latency. Therefore, we set F to 1, and the "fixed time" to 0.

Using the default PA algorithm, we can create an optimal schedule. However, this schedule will be similar for all participants: if the first word-pair is repeated after 5 trials because it will drop below the retrieval threshold within 15 seconds, this holds for all participants. Obviously, this does not match real performance profiles: some participants will have a higher overall performance level than other participants, but it might also be that some words are recalled better by some participants, but a different set of words is recalled better by other participants. However, each time an item is presented the learner provides us with additional behavioral data, which we can use to dynamically adapt the model to the individual learner. This approach can be described as subsymbolic model tracing.

Subsymbolic Model Tracing

In the traditional model tracing account (Anderson, Boyle, Corbett, & Lewis, 1990), the behavior of a student is matched against all knowledge available in a tutoring system. For example, if a student has shown accurate performance in a number of subtraction problems in which carrying is required, the knowledge in the tutoring system that represents carrying is marked as mastered. Thus, the tutoring system keeps a representation of all knowledge the student has mastered by updating the internal representation each time new behavioral information becomes available. The behavior that the learner displays can similarly be used to update the subsymbolic activation of facts (Jastrzembski, Gluck, & Gunzelmann, 2006).

Given that each time a student has to answer a test trial both accuracy and latency information becomes available, we can, in principle, use this information to determine what the current activation of the retrieved chunk is. If we know the latency and therefore the activation at the time of encounter j, and we also know the latency/activation at the time of encounter j-1, we can calculate what the decay for encounter j-1 should have been. By this rationale, we can minimize the difference between the predicted activation and the observed latency and use the behavioral data of the student to update our model that represents the state of the student.

However, given the general assumption that the retrieval process is inherently noisy, using this direct relation might be problematic when the response is fast. That is, when t_j - t_{j-1} is relatively long and the latency for t_j is short because of

a temporal boost in activation due to noise, the calculated decay for t_{j-1} will be very low (or even negative). As a very low decay results in facts that are predicted to be highly active over a very long period of time, this temporal noise-boost will ruin the scheduling of the fact. Therefore, we have chosen not to use the outcome of the algorithm described here directly, but instead change the d_{j-1} with a fixed, small amount in the direction indicated by the mismatch between predicted activation and observed latency (c.f., hill-climbing optimization algorithms).

Algorithm 3: Threshold-based Adaptation

Given the issues related to the noisy observations, using the more fine-grained subsymbolic model tracing method described above might result in overfitting. To minimize the chances of overfitting, a coarser algorithm might prove beneficial. Therefore, Algorithm 3 adapts the PA model by only modifying the decay parameter for a certain encounter when at test the word-pair cannot be correctly recalled (c.f., Pavlik & Anderson, 2008). As the system always presents word-pairs of which the estimated activation is above the retrieval threshold, a failure to recall indicates that the estimated activation was too high. Thus, the decay for that particular item should be higher, which is reflected in increasing the alpha parameter with 0.01.

Algorithm 4: Latency-based Adaptation

The threshold-based adaption algorithm focusses on maximizing the testing-effect. Each time a fact cannot be recalled, its decay is increased, ensuring that it will be presented with shortened spacing in subsequent trials. Although this will result in better testing effects because of shorter spacing for facts that could not be recalled, this algorithm does not adapt itself to the inverse situation when facts are better learned than expected. That is, where a failure to retrieve is a marker of lower than expected activation, a faster response than expected is a marker of a higher than expected activation. This idea is captured in the latency-based adaptation of Algorithm 4 which extends the threshold-based adaptation algorithm by comparing the expected latency with the observed latency. To prevent overfitting, the decay intercept is only changed if the difference between expected and observed latency is more than 0.5 seconds. Instead of a constant modifier, the decay intercept is changed according to:

$$\Delta\alpha = \max(0.01, \frac{\text{observed} - \text{expected}}{1000})$$

where observed and expected are the latencies expressed in seconds.

Experiment

Four classes of approximately 15-year old pre-university level pupils were asked to memorize Dutch translations of 20 French words. Each word pair was presented first in a study trial in which both the French and the Dutch word were presented. During a test trial, only the French word was presented, and the participant had to type the Dutch translation. After the initial presentation, the next trial was scheduled on the basis of one of the four algorithms discussed above.

Procedure Study trials were presented for 5 seconds. After each initial study trial, a test trial of the same word-pair was presented. During a test trial, only the French word was presented and students had 15 seconds to reply by typing in the correct Dutch translation. After pressing Enter, students were presented with a 2-second feedback screen stating "Correct", "Incorrect" or "Almost correct" (which was given if the Levenshtein-distance to the correct answer was smaller than 3). If the participant did not respond in time, or an incorrect answer was given, the study trial was presented to refresh the participant's memory. The four algorithms determined which word pair to present next. The learning session lasted 15 minutes, irrespective of the number of trials or words presented. After the learning session on Day 1, all words were tested by means of a traditional paper-and-pencil test on Day 2. The post-test was graded on a scale from 0 to 10. Each incorrect response deducted .5 point from the maximum score of 10. Participants were naive with respect to the experimental manipulation and did not know that they would be tested on Day 2.

On Day 1, participants were tested in groups in a classroom equipped for computer-supported education. Each participant operated his or her own computer. The paper-and-pencil test on Day 2 was conducted during normal class-hours.

Materials A list of 20 words was compiled for each class separately. All words were selected from a textbook chapter that would not be discussed until one week after the experiment.

Participants Ninety-one pre-university-education level students (all students of four 3rd year HAVO/VWO classes) of approximately 15 years of age participated, of which 85 took part in both tests. Participants were semi-randomly distributed over conditions to ensure that in each class an equal number of participants used each algorithm. All participants were instructed that their results would be stored anonymously and that the results would not be communicated to their school or teachers on an individual level.

Results

Of the 85 students who took part in both sessions, six students were removed from further analyses because they did not respond in more than 5% of all trials and gave a number of answers that did not fit the instructions (e.g., "I'm bored", or names of rock bands). Four participants were removed because their performance in terms of correct responses during the learning session deviated more than 2 standard deviations from the average of their group. One participant was removed because of scores on the final test that deviated more than 2 standard deviations from the average score for his or her group. This leaves 74 participants, 18 in the flashcard condition, and 19 participants in each of the three spacing conditions.

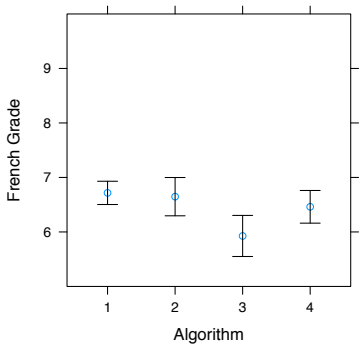

Figure 1: Average grades on French per algorithm condition

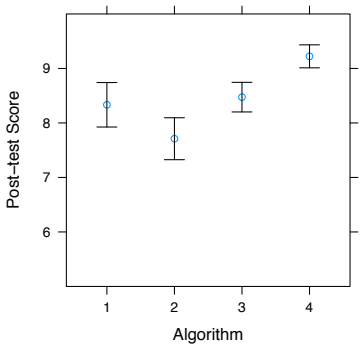

Figure 2: Raw scores on post-test per algorithm condition

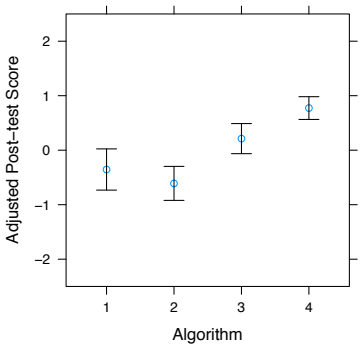

Figure 3: Post-test scores adjusted for covariates mentioned in text

All errorbars depict standard errors.

Covariates As we tested participants in a domain in which they have a significant amount of prior knowledge, it is important to control for potential differences in prior knowledge between groups. Hereto, we analyzed the students' school grades for French (graded on a theoretically linear scale from 0 to 10, with a 6 representing the grade required to pass that class), see Figure 1 ($F(3,73)= 1.27$, $p=0.29$). Although this effect is far from significant, the value of the F-statistic is larger than we hoped for. Therefore, we decided to include the grades for French as covariate in all subsequent analyses.

Given that we limited the amount of time to learn 20 word-pairs and the algorithm determined when a new word pair was introduced, not all participants might have seen all 20 word pairs in the non-flashcard conditions (algorithm 2 to 4). This did indeed turn out to be the case in all three conditions. The average number of word pairs presented to the participants was 19.5, 19.6 and 19.8 for the default PA, the threshold-based and the latency based conditions respectively. Although these differences (when compared to the 20 words seen by the students in the flashcard condition) fail to reach significance (ANOVA $F(3,69)=2.6$, $p=0.057$, post-hoc pairwise t-tests with pooled standard deviations: flashcard vs default PA algorithm, $p=0.08$, all other comparisons $p > .1$), this does give the flashcard-based condition an advantage when comparing scores on the post-test, as some participants in the other conditions will not have seen all word pairs. Therefore, we also included the number of words seen by the student as covariate in subsequent analyses.

To account for possible effects associated with the session in which the study was run or peculiarities of a particular class, another factorial covariate was included representing group.

Post-test Scores Figure 2 shows the raw scores on the post-test, and Figure 3 shows the scores on the post-test adjusted by the covariates French grade, group and number of words seen.

Analyzing the data presented in Figure 3 shows that the algorithm has a significant effect on the post-test scores ($F(3,70)=4.19$, $p=0.009$). Testing the individual effects by conducting pairwise comparisons using t-tests with pooled standard deviation and Benjamini and Hochberg's (1995) p-value adjustment method showed that students in the latency-adaptation group, Algorithm 4, score significantly higher than students in the flashcard ($p=0.032$) or in the PA model ($p=0.010$) group. None of the other comparisons reached significance ($p>0.100$).

Adaptions The observed differences between the more static PA model (Algorithm 2) and the latency adaptation condition (Algorithm 4) suggests that the adaptations resulted in different decay patterns for different participants. Figure 4 shows the average estimation of the alpha parameter associated with the last encounter per word-pair. As can be seen, different participants required different alphas, with, for example, participant 1 and 4 requiring relative low decay values and participant 13 requiring a very high decay value. If these three participants would have been set at the average alpha (0.259), the estimated activation for participant 13 would be too high, resulting in many retrieval failures - and violating the testing-effect constraints. At the same time, participants 1 and 4 would have had a too low estimated activation, resulting in a sequence with too low spacing, violating the spacing-effect constraints.

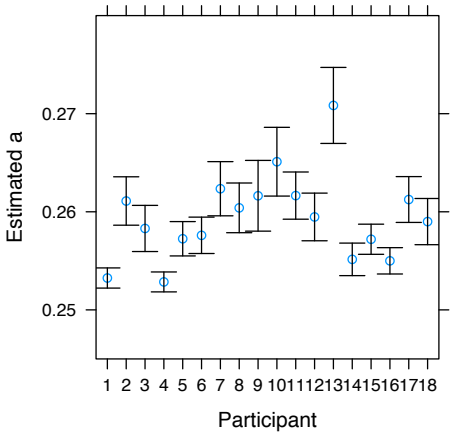

Figure 4: Effects of adaption per participant

Discussion

The current study set out to answer two questions. The first was to test whether the results obtained in the scientific literature on the spacing effect would also hold in the more real-life case of learning a small set of vocabulary items in a small period of time. The second question was to examine whether the learning gain would improve when the algorithms that construct the learning sequence take individual differences into account. With respect to the first goal, the significant difference between the

flashcard and the latency adaption conditions illustrates that a learning sequence that is based on an algorithm that takes spacing and testing-effects into account outperforms a more traditional flashcard sequence. However, *only* the condition that optimizes the sequence on the basis of individual latency differences significantly outperforms the flashcard condition, answering the second question.

It is striking to see that the default (pre-2008) PA spacing condition scores - in absolute terms - worse than the flashcard condition. This result for the default PA spacing condition might be caused by the parameter settings chosen for this study: alternative parameter settings might improve the PA model. However, it is difficult to come up with the parameter settings required. The first candidate for change would be the retrieval threshold, as in most PA studies the threshold is set at -0.704 instead of -0.5. However, decreasing the threshold would increase the spacing between two presentations of the same item. This will probably have a negative effect on the data as the sole difference between the default PA algorithm (2) and the threshold adaptation algorithm (3) is *decreased* spacing and algorithm 3 fares considerably better than the PA algorithm. With respect to changes in the parameters involved in calculating d_{ji}, it is most likely that these changes would benefit the other algorithms as well. Thus, although changes in the parameter settings might diminish the gap between the different spacing algorithms, it is hard to imagine how the default PA model would outperform the alternative algorithms proposed here.

With respect to Algorithm 3 and 4, although the differences in performance are not significant, the performance profiles favor the latency-based Algorithm 4 over the accuracy-based Algorithm 3. This suggests that Pavlik and Anderson's 2008 implementation might be further refined by incorporating the information that can be deduced from the latencies (c.f., Pavlik, Presson, & Koedinger, 2007).

Finally, it is interesting to note that Pavlik and Anderson (2008, p.102) discuss a very similar approach they call "performance tracking" and mention that this method will "add considerable power". Nevertheless, they conclude that this approach will make scheduling much more complex.

In this study we have shown that performance tracking is possible, but also that adapting the sequence to the characteristics of individual learners improves learning gains considerably, even if the learning session takes only 15 minutes.

Acknowledgments

Parts of this work has been conducted in the context of Van Woudenberg's (2007) Master thesis. The authors would like to thank the students and teachers of het Belcampo College, het Gormarus College and het Werkman College for their participation and Philip Pavlik and two anonymous reviewers for helpful remarks regarding the argument and clarity of the paper.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Anderson, J. R., Boyle, C. F., Corbett, A., & Lewis, M. W. (1990). Cognitive modelling and intelligent tutoring. *Artificial Intelligence, 42*, 7–49.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2 (6), 396–408.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4 (10), 829–839.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 289–300.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, 20, 633–633.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132 (3), 354.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psyc.Science*, 19, 1095-1102.

Dempster, F. N. (1988). The spacing effect. *American Psychologist*, 43, 627–634.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. Translated by Henry A. Ruger and Clara E. Bussenius (1913).

Jastrzembski, T. S., Gluck, K. A., & Gunzelmann, G. (2006). Knowledge tracing and prediction of future trainee performance. In *Proceedings of the 2006 interservice/ industry training, simulation, and education conference* (p. 1498-1508). National Training Systems Association.

Pashler, H., Rohrer, D., Cepeda, N., & Carpenter, S. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, 14 (2).

Pavlik Jr, P. I. (2007). Understanding and applying the dynamics of test practice and study practice. *Instructional Science*. *35*, 407-441

Pavlik Jr, P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In Proceedings of the 5th international conference of cognitive modeling (pp. 177–182).

Pavlik Jr, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29 (4).

Pavlik Jr, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied, 14 (2), 101–117.

Pavlik Jr., P. I., Presson, N., & Koedinger, K. R. (2007). Optimizing knowledge component learning using a dynamic structural model of practice. In R. Lewis & T. Polk (Eds.), Proceedings of the eighth international conference of cognitive modeling. Ann Arbor: University of Michigan.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (p. 64-99). Appleton-Century-Crofts: Appleton-Century-Crofts.

Roediger, H. L., & Karpicke, J. D. (2006). Taking memory tests improves long-term retention. *Psychological Science*, 17 (3), 249–255.

Van Woudenberg, M. (2008). *Optimal word pair learning in the short term: Using an activation based spacing model*. Unpublished master's thesis, University of Groningen.

Wozniak, P. A., & Gorzelanczyk, E. J. (1994). Optimization of repetition spacing in the practice of learning. *Acta Neurobiologicae Experimentalis*, 54 (1), 59–62.