Stability of Individual Parameters in a Model of Optimal Fact Learning

Florian Sense (f.sense@rug.nl)
Department of Experimental Psychology and Department of Psychometrics and Statistics,
Groningen, The Netherlands

Friederike Behrens (f.behrens@student.rug.nl)
Research School of Behavioral and Social Sciences,
Groningen, The Netherlands

Rob R. Meijer (r.r.meijer@rug.nl)
Department of Psychometrics and Statistics,
Groningen, The Netherlands

Hedderik van Rijn (d.h.van.rijn@rug.nl)
Department of Experimental Psychology and Department of Psychometrics and Statistics,
Groningen, The Netherlands

Abstract

We are using an algorithm based on a computational model of human memory to optimize the scheduling and repetition of individual items within a learning session. The model estimates the rate of forgetting for each participant to determine the order in which items should be repeated and to decide when previous items have been learned well enough to introduce a novel item. To improve the model further, we conducted an experiment to test how stable the parameter estimates are over time and across different materials. We have found that estimated rates of forgetting are stable over time *within* one type of material but not across different types of material. This finding has important implications for how information about a learner should be preserved between study sessions.

Keywords: spacing effect; testing effect; cognitive model; learning; parameter stability.

Introduction

Fact learning is a big part of learning a new skill. In many school curricula, students are evaluated based on how well they learned a certain array of facts. With the advance of computers into classrooms and workplaces, tutoring systems have been developed to help learners master the required material. Over a hundred years of memory research have singled out two robust effects that developers of such systems can use to enhance that goal: the spacing effect and the testing effect. By making optimal use of both of them *and* adjusting the system to the individual learner, such systems can make learning a lot more efficient. As of now, however, most optimizing systems treat each learning session in isolation; user-specific characteristics are estimated during a learning session to optimize each learning session but are not preserved *between* learning sessions. In this study, we investigated to which extend user-specific parameters relevant to such a tutoring system are stable over time and across different materials to gauge to which extent they can be preserved between learning sessions.

The tutoring system used here works by balancing the benefits of the spacing and the testing effect. The spacing effect describes the finding that performance on tests of recall is improved when study time is distributed over multiple sessions with time in-between rather than massed study (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Dempster, 1988). The optimal spacing schedule ultimately depends on how much time is available and when the material is tested (Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). However, it has been shown convincingly that long-term retention can be increased by spacing items within a single learning session (Lindsey, Shroyer, Pashler, & Mozer, 2014; van Rijn, van Maanen, & van Woudenberg, 2009) as well as spacing individual learning sessions (Cepeda et al., 2006).

The testing effect, on the other hand, describes the finding that active memory retrieval during practice is more beneficial for long-term retention than passive study (Karpicke & Roediger, 2008; Roediger & Butler, 2011). That is, being forced to retrieve the answer from memory leads to better learning than simple re-studying (i.e. looking at) the cue-answer pair (Carrier & Pashler, 1992). This effect has been studied extensively in the laboratory (Cepeda et al., 2008) but also holds in more realistic classroom settings (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008; van Rijn et al., 2009).

Given our knowledge of the spacing and testing effects and the quasi-lawful behavior of memory, it seems possible to devise a learning schedule that would make optimal use of each effect's benefits. This would require balancing two seemingly opposing goals: (1) maximizing time between repetitions of an item to get the biggest spacing effect, and (2) minimizing time between repetitions of an item to make sure it can still be retrieved from memory to take advantage of the testing effect. Such computer adaptive practice models have been developed and have been shown to outperform flashcard control conditions (Nijboer, 2011; van Rijn et al., 2009).

As a starting point for the development of such models, Anderson and Schooler (1991) showed that data on memory performance (i.e. practice and retention) across time courses ranging from seconds to years can be fit nicely by power functions. Interestingly, this corresponds closely with environmental relationships (Anderson & Milson, 1989). That is, the likelihood that people still remember a non-sense syllable they learned today at a certain point in the future (i.e. the original Ebbinghaus data) can be described with the same power functions that can be used to describe the likelihood of receiving an e-mail (Anderson & Schooler, 1991). This leads Anderson and Schooler (1991) to conclude that the human "memory system is adapted to the structure of the environment" (p. 400).

Based on this assumption, it is argued that the practice and retention of facts can be approximated using the same equations that can be used to describe the behavioral effects in the data. Pavlik and Anderson (2003, 2005) developed a model that formalizes this process and show how it can be used to compute the optimal schedule of practice, taking into account the effects of practice, retention, and spacing (Pavlik & Anderson, 2008). Their model assumes that there is some stable effect based on each individual's *rate of forgetting* and additional effects based on *item difficulty*. In this original work, it is assumed that these effects are stable over time and, for the rate of forgetting, across knowledge domains/materials. That is, someone's rate of forgetting is assumed to be a property of their memory and therefore stable, regardless of whether they study vocabulary, topographical information, or word definitions.

The success of such models (Nijboer, 2011; Pavlik & Anderson, 2008; Van Rijn et al., 2009) is very promising but the stability of participants' rate of forgetting across time and knowledge domains has never been demonstrated empirically. The goal of the present study is to investigate to which extent participants' rate of forgetting varies over the course of three weeks as well as across four different types of material.

Methods

The Model

The model used in this experiment is based on ACT-R's declarative memory equations (Anderson, 2007). In the ACT-R framework, each item that is learned is assigned an *activation* value. Activation is highest at the moment an item is encountered and then decays as a function of time. The activation of an item at any point in time can be computed using the following equation:

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

According to this equation, the activation of item *i* at time point *t* depends on all previous time points at which item *i* has been encountered. After each previous encounter *j* the activation associated with that encounter decays with d_j, which translates to a smaller contribution to the current activation if encounter *j* has occurred long before time point *t*. The rate with which the activation decays after each encounter is calculated as follows:

$$d_{ji} = ce^{A_i(t_j)} + \alpha_i \qquad \text{Eq. 2}$$

In this equation, *c* is the decay scale parameter that determines the relative contribution of the activation component. Alpha represents the decay intercept, which represents a minimum decay value (and will be used as the decay value for the first encounter). This equation has been developed by Pavlik and Anderson (2008) to deal with the spacing effect. In the ACT-R framework, an activation value can be directly converted to an estimated response time by scaling the activation and adding a *fixed time* that accounts for non-memory related processes. The following equation is used to convert the activation of item *i* at time point *t* to an estimated reaction time:

$$RT_i(t) = Fe^{-A_i(t)} + \text{fixed time} \qquad \text{Eq. 3}$$

Pavlik and Anderson (2003, 2005, 2008) have shown that the three equations outlined here can be used to fit a wide range of data from learning-related experiments and can account for additional benefits gained through the spacing effect. The system has not only been used to describe collected data but also to devise a system that predicts, in real-time, the order in which items should be repeated to yield optimal retention. More recently, Van Rijn and colleagues (2009) and Nijboer (2011) have developed the system further and showed that a scheduling algorithm that compares observed with predicted reaction times (derived from an item's estimated activation) leads to even better learning than the Pavlik and Anderson (2008) model. The same algorithm is used in this study and a graphical representation of the procedure is depicted in Figure 1.

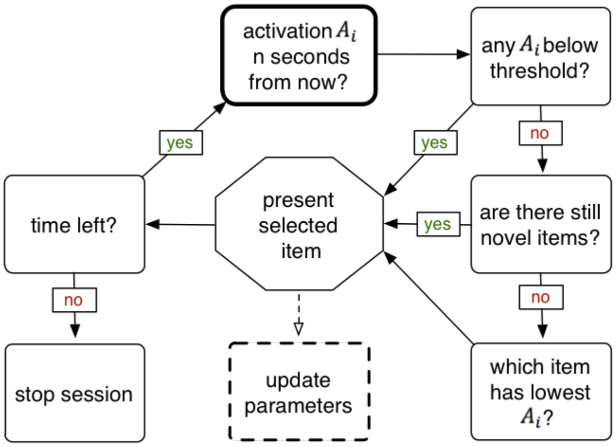
activation A_i n seconds from now?
any A_i below threshold?
yes
yes
no
time left?
present selected item
yes
are there still novel items?
no
no
stop session
update parameters
which item has lowest A_i?

Figure 1: A graphical representation of how the model determines the order in which to repeat old and present new items.

The algorithm selects the order in which items are presented to the learner dynamically and adjusts the order of repetitions based on the learner's behavior. This is done as follows: The model simulates the activation of all items that have already been encountered *n* seconds from now using the equations described above. If *n* seconds from now, the activation of any item is below the retrieval threshold, that item will be presented next (because this indicates that the item is about to be). If no item is below the threshold, a new item is presented as long as novel items are still available. Otherwise, the item with the lowest activation *n* seconds from now is presented. At the time the selected item is presented, the model uses the estimated activation of the item in the learner's memory to compute the estimated reaction time (see Eq. 3). The item's alpha parameter is updated by comparing the estimated reaction time with the observed reaction time. If the estimated reaction time was too slow, this indicates that the estimated activation was too low. That, in turn, indicates that the decay value for the previous encounter was estimated to be too high. To compensate for this discrepancy, the alpha parameter for the given item is adjusted in a step-wise procedure to improve the model's estimate on the following trial (see Nijboer (2011) for details). After the parameter has been updated, the model checks whether the learning session should continue and then either stops or starts the next repetition. The mechanism is depicted graphically in Figure 1.

Procedure

Each person participated in the study for three sessions on three days, each session spaced one week apart. Within each session, there were two blocks. Each block was made up of a 20-minute study session, a five-minute distraction task, and a test of the studied material that took about five more minutes. At the beginning of the first session, each participant also completed a short questionnaire regarding demographic information (age, gender, nationality, and language skills). The five-minute distraction was a simple variation of the puzzle game Tetris which participants played until they were automatically re-directed to the test that concluded each block.

During a study block, novel items were presented on *study trials* and subsequent repetitions were presented on *test trials*. On a *study trial*, participants saw both the cue and the correct response and had to type in the correct response to proceed. On a *test trial*, participants only saw the cue and had to type in the correct response. Feedback was provided in both trial types and lasted 0.6 and 4 seconds for correct and incorrect answers, respectively. The feedback always resembled a *study trial* and displayed both the cue and the correct response. Jang, Wixted, Pecher, Zeelenberg, & Huber (2012) have shown that for non-retrievable items, an additional *study trial* is very effective because participants do not benefit from the testing effect (but unsuccessful retrieval attempts can still enhance learning (see Kornell, Hays, & Bjork, 2009). Furthermore, they showed that four-second *study trials* yield the highest benefit. During the test at the end of each block, participants were provided with a list of all possible items and could provide their responses in any order they preferred.

Material

For each block, a list of 25 items was compiled. The lists of items were identical for all participants but during each study block, the model randomized the order in which items were presented based on their participant numbers. There were four types of material that were studied by each participant:

Vocabulary. There were 75 Swahili-English word pairs that were taken from Van den Broek, Segers, Takashima, & Verhoeven (2014). Swahili-English word pairs are common stimuli in vocabulary learning (e.g. Carpenter, Pashler, Wixted, & Vul, 2008; Pyc & Rawson, 2010; Van den Broek et al., 2014) because most university students do not have any prior knowledge.

Flags. A list of 25 items was compiled from Wikipedia's list of sovereign states. The authors strived to pick the flags of countries that were not likely to be known by the participants, using their own familiarity with the countries' flags and a pilot study as a benchmark.

City Locations. A list of 25 items was compiled by searching for smallish cities on Google Maps, making sure the cities are more or less evenly spaced across the continental United States of America. Cities were picked so their names are unique, not too difficult to spell, and do not contain information about their geographical location.

Bio-Psychology Facts. A list of 25 bio-psychology facts was compiled from the Glossary in Kalat (2012). The facts were chosen so that the answer would always be a single word and that there is some variations in how difficult the words are to spell.

Participants

Of the 76 first-year psychology students from the participant pool of the University of Groningen that signed up for this study, 71 completed all three sessions and 70 fulfilled the minimum requirement of having seen at least 10 unique items in each study block. It was assumed that if the participant was not able to perform well enough to be presented with at least 10 unique items within a 20-minute study block, there would likely be a problem that was beyond them being a poor learner so their data was dismissed. Participants were also removed when they answered less than 25% of the items they had seen during the study block correctly on the subsequent test, which applied to 3 participants. Of the remaining 67 participants, 50 were female and the median age was 20 ($SD_{age} = 1.73$; $range_{age} = [17; 26]$). No one indicated familiarity with Swahili, 35.8% were Dutch, and 52.2% were German. All participants indicated to be fluent in English and gave informed consent.

Results

Figure 2 summarizes the performance on the final test that concluded each block. The black bars in the plot are traditional box plots and the white dots highlight the group's median. The colored areas are scaled density plots that depict the distribution of each group's values. The data suggest that performance was very high overall. Especially in the three vocabulary sessions, performance was very high for most participants with a little more variation in the other blocks. The city location block (CI) seems to have been the most difficult followed by the bio-psychology fact block (BIO). The overall excellent performance suggests that participants actively engaged with the material during the study session, which makes us confident that the alpha parameters that were obtained during the study sessions contain meaningful information about the participants' engagement with and acquisition of the studied material.

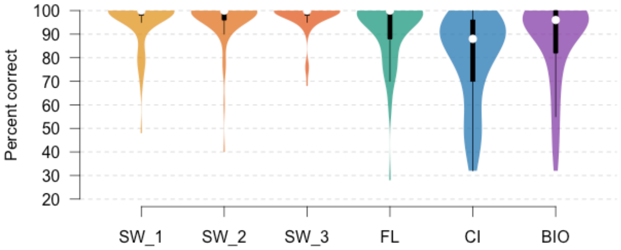

Figure 2: Performance on the final test across the six blocks. SW_1 and FL were tested in the first session, SW_2 and CI in the second session, and SW_3 and BIO in the third session. The sessions were spaced one week apart.

As described in the sub-section The Model, each item that each participant studied is assigned one alpha value. Each item starts with an alpha value of 0.3 but then the alpha value is adjusted on each repetition of trial, depending on how the participant responds to the item and how well that response matches up with the model's prediction. To address the research question of whether the estimated rate of forgetting is stable (1) over time and (2) over materials, we looked at the variation in alpha values across time and materials. For this analysis, the alpha values of items that have been presented at least three times have been included. That is, each participant contributed multiple alpha values and the exact number depended on how many items that participant encountered at least three times within each block.

For the analysis, the alpha values were log-transformed to satisfy assumptions of homoscedasticity and normality. The aim of the analysis was to check whether alpha scores differed across time and materials and whether both factors influenced each other in their effect on the alpha values. To test this, we used linear mixed-effects model regression with dummy coding. The mixed-effects model allows accounting for the interdependency between observations due to by-subject and by-item variation. Three variables were included in the model to test our research question: The first variable coded the *session* (that is, the day) on which the blocks were completed. This allowed us to check whether there is any significant variation over time across all blocks. The second variable was coded 0 for blocks in which participants studied Swahili words and 1 for those in which non-Swahili material was studied. This allowed us to directly compare the differences between multiple blocks of learning Swahili to non-Swahili blocks. The third dummy was coded -0.5 for the flags block (FL), 0.5 for the city location block (CI), and 0 for all other blocks. This allows us to compare the individual blocks (that is, types of material) in more detail. The results of the analysis are shown in Table 1.

Table 1: Results of the linear mixed-effects regression.

	beta	SE	df	| t |	p
intercept	-1.394	0.040	112	34.38	<0.001
session	-0.036	0.028	73	1.30	0.198
SW vs. ¬SW	0.182	0.011	9506	16.50	<0.001
FL vs. CI	0.364	0.013	9457	27.87	<0.001
session * SW v. ¬SW	-0.051	0.028	82	1.81	0.074

The alpha scores do not significantly differ between sessions (t(73)=1.3, p=0.198). However, the contrasts between the Swahili and non-Swahili blocks and between the flags and city location blocks significantly influence the alpha values (t(9506)=16.5, p<0.001; t(9457)=27.87, p<0.001, respectively). More specifically, participants had smaller alpha values in the Swahili blocks compared to the flag and city location blocks (a decrease of 0.049) indicating a faster forgetting rate for the latter two blocks. This effect was stronger for the city location block, which is suggested by the positive coefficient of the flags vs. city locations contrast. Specifically, the forgetting rate increases by 0.109 in the city compared to the flags block. The interaction between the sessions and the contrast Swahili vs. non-Swahili is not significant (t(82)=1.81, p=0.074). In other words, the increase of the forgetting rate in performing the flag or city task compared to the Swahili task is independent of when (that is, in which session) one performs the task.

While the regression analysis examined overall effects of difference in alpha values, it might also be informative to take a closer look at the development of estimated alpha values throughout the course of a study session. Figure 3 shows how the alpha values for each item change as a function of time. The items are color-coded (the legend is shown at the top of the graph) and it can be seen that each item has an alpha value of 0.3 when it is first introduced. On each subsequent repetition, the alpha value is adjusted and the magnitude of the change depends on the discrepancy between the estimated and the observed reaction time. The data shown in Figure 3 come from a very good participant so that many items end up with an alpha value lower than the default they started with. It can be seen, however, that there are substantial differences in alpha values within this participant, indicating that some items were more difficult to learn than others. The peak and frequent rehearsal of the 25th item is particularly obvious. This pattern was likely caused by a series of incorrect responses, which led the model to believe that the item was not learned yet, in response to which the alpha was corrected upwards step-by-step. The higher alpha than resulted in a more frequent rehearsal (see Figure 1). The plot also makes clear that the model does a good job of interleaving items that were learned early in the session with those learned later on.

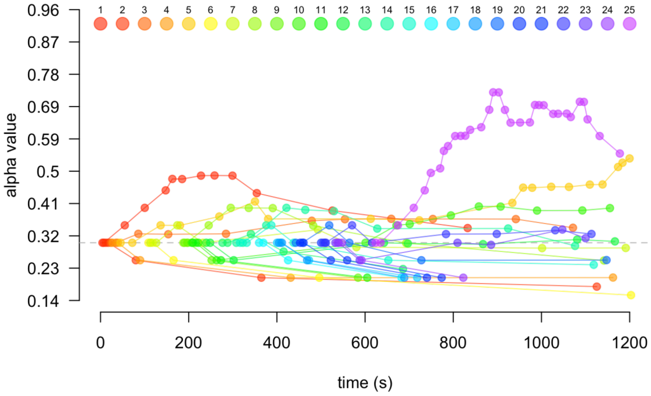

Figure 3: Development of the alpha values for each item for one participant in one block as a function of time.

Discussion

In this study, we investigated the stability of individual rate of forgetting parameters in a model of optimal fact learning. The emphasis is on scrutinizing the stability of the parameter values across time and across different materials. Knowing more about the circumstances under which a learner's estimated rate of forgetting is stable in time and across materials will enable us to further develop the model by carrying over what we learned about the participant in one learning session to the next.

The results of the analysis demonstrate that the estimated rates of forgetting do not differ significantly over time. There is a difference in estimated rates of forgetting, however, when different types of material are studied. Given the non-significant interaction between time point of study and type of material, differences between materials seem to be independent of time.

When looking at the data of the performance on the final test depicted in Figure 2, one can see that there was a clear ceiling effect. The effect is especially pronounced in the three Swahili blocks and the block in which participants learned flags. This might be considered to be an issue because it would facilitate the stability of results within those Swahili-learning blocks. It should be noted, however, that by using the parameter values that were estimated throughout the learning session instead of the *results* of the learning session (test performance), one gets a much more fine-grained view on the differences between conditions. There is much more variation in estimated rates of forgetting than the corresponding results on the test suggest. This conclusion is further supported by the fact that there was no significant difference across the three sessions (see Table 1) even though the comparison *did* include the blocks for which final performance was not at ceiling. In addition to that, using the estimated rates of forgetting for each item from each block can also serve as a diagnostic tool to get a better idea of the inner mechanics of the model and detect ways in which the model might not perform optimally and why. By plotting the development of the parameters over time for a single participant in one of the six blocks (see Figure 3) can indicate problems that would not be apparent from measures taken at the end of a learning session. Therefore, we think an analysis based on the estimated parameter values is much more interesting and insightful than one based on the performance on the final test.

As discussed in the Introduction, the model does not currently preserve estimated parameter values across multiple study sessions. That is, when a learner uses the model to study a number of Swahili-English word-pairs and then returns to the system the day after and starts another study session, the model will revert back to the default parameter values at the beginning of the second session. This seems both wasteful and inefficient. One would think that by observing the learner's behavior in the first session and comparing it to the model's estimates (which are based on the current parameter values), we have learned something about that particular learner. And updating the internal parameters of the model dynamically to capture this learning-about-the-learner is an essential part of the model. Therefore, it would be a logical next step to determine a way in which we could preserve what we have learned about the learner in the first session. That way, we can give the model a head start at the beginning of the second session instead of forcing the model to start from scratch.

The data reported here show that there is substantial stability of parameter values over time, especially if the same type of material is studied: Swahili vocabulary. We reckon it is reasonable to assume that these findings generalize to languages other than Swahili. It would be interesting, however, to test whether a transfer from Swahili to, for example, French is better than the transfer from Swahili to bio-psychology. A challenge for the future will be to determine the optimal transfer of parameter values between sessions that do not deal with the same type of material. In this study, we made an effort to devise material that is very different from each other (word-pairs (Swahili), visual information (flags), topographical information (city locations), and factual knowledge (bio-psychology)) but this leaves open the question of how similar material has to be to still allow smooth transfer of suitable parameter values.

Conclusion

The data presented here suggest that participants' rate of forgetting varies between materials but is relatively stable within a domain over time. This indicates that rate of forgetting is not purely a feature of a learner's memory system but also influenced by the type of material studied. If the same material is studied, though, the data suggest that the rate of forgetting is stable over time. Therefore, we should be able to improve the model further by carrying over what we learned about a learner from one session to the next, given that the sessions deal with the same type of material.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

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

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408. doi:10.1111/j.1467-9280.1991.tb00174.x

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514–527. doi:10.1080/09541440701326097

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438–448. doi:10.3758/MC.36.2.438

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–42.

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), 354–80. doi:10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095–102. doi:10.1111/j.1467-9280.2008.02209.x

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627–634.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*(5), 795–805. doi:10.1037//0021-9010.84.5.795

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences*, *20*(4), 155–156. doi:10.5214/ans.0972.7531.200408

Jang, Y., Wixted, J. T., Pecher, D., Zeelenberg, R., & Huber, D. E. (2012). Decomposing the interaction between retention interval and study/test practice: the role of retrievability. *Quarterly Journal of Experimental Psychology*, *65*(5), 962–75. doi:10.1080/17470218.2011.638079

Jastrzembski, T., Gluck, K., & Gunzelmann, G. (2006). An ACT-R predictive model of performance: Applications in education and training. In *Proceedings of the Society for Mathematical Psychology Annual Meeting*. Vancouver, Canada.

Kalat, J. W. (2012). *Biological psychology* (11th ed.). Wadsworth, Cengage Learning.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–8. doi:10.1126/science.1152408

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989–98. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19586265

Lindsey, R. V, Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–47.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494–513.

Nijboer, M. (2011). *Optimal fact learning: Applying presentation scheduling to realistic conditions*. University of Groningen.

Pavlik, P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In *Proceedings of the 5th International Conference on Cognitive Modeling* (pp. 177–182). Bamberg.

Pavlik, 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), 559–86. doi:10.1207/s15516709cog0000_14

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–17. doi:10.1037/1076-898X.14.2.101

Pavlik, P. I., Bolster, T., Wu, S., Koedinger, K. R., & MacWhinney, B. (2008). Using optimally selected drill practice to train basic facts. In B. Woolf, E. Aimer, & R. Nkambou (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 593–602). Montreal, Canada.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335.

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–7.

Van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*(7), 803–12.

Van Rijn, H., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 110–115).