TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 69–82
Copyright © 2017 Cognitive Science Society, Inc. All rights reserved.
ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12244

This article is part of the topic "Best of Papers from the 2016 International Conference on Cognitive Modeling," David Reitter and Frank E. Ritter (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-1/issuetoc.

An Account of Interference in Associative Memory: Learning the Fan Effect

Robert Thomson,[a,b] Anthony M. Harrison,[b] J. Gregory Trafton,[b] Laura M. Hiatt[b]

[a]*Army Cyber Institute, United States Military Academy, West Point, NY*
[b]*Naval Research Laboratory, Washington, DC*

Received 30 September 2016; received in revised form 6 November 2016; accepted 15 November 2016

Abstract

Associative learning is an essential feature of human cognition, accounting for the influence of priming and interference effects on memory recall. Here, we extend our account of associative learning that learns asymmetric item-to-item associations over time via experience (Thomson, Pyke, Trafton, & Hiatt, 2015) by including link maturation to balance associations between longer-term stability while still accounting for short-term variability. This account, combined with an existing account of activation strengthening and decay, predicts both human response times and error rates for the fan effect (Anderson, 1974; Anderson & Reder, 1999) for both target and foil stimuli.

Keywords: Fan effect; Associative learning; Memory; Priming; Cognitive modeling

1. Introduction

Associative learning is an essential component of human cognition, thought to be part of many mental phenomena such as classical conditioning (e.g., Rescorla & Wagner, 1972) and memory recall (e.g., Kahana, 1996; Klein, Addis, & Kahana, 2005). Despite

Correspondence should be sent to Robert Thomson, Army Cyber Institute, United States Military Academy, West Point, NY 10996. E-mail: robert.thomson@usma.edu

its ubiquity, it is difficult to model directly due to its entangled ties to other aspects of cognition (e.g., memory decay).

Perhaps associative learning's most studied effect is that of *priming* (and its converse *interference*; e.g., Stanovich & West, 1983; Hiatt & Trafton, 2013). Priming occurs when the retrieval of one memory facilitates the retrieval of another. Conversely, interference occurs when a memory primes multiple other memories instead of just the ones that are useful or relevant to the current situation. Those other memories are said to *interfere* with the useful one (MacLeod, 1991). When there is high interference, recognition accuracies are relatively lower and recognition response times relatively longer when compared to situations where there is low interference, ostensibly due to having lower overall activation in memory. Assuming that the degree of interference is positively correlated with the number of competing associations, then having more competing associations (i.e., a higher *fan*) will lead to relatively higher error rates and latencies than memories having relatively fewer competing associations. This effect is most popularly known as the *fan effect* (Anderson, 1974).

In this paper, we will extend our account of associative memory embodied in a cognitive architecture (Thomson, Bennati, & Lebiere, 2014) to explain interference in the fan effect experiment. This theory of associative memory has already successfully predicted the complicated results of a multi-trial free and serial recall task, including asymmetric contiguity effects that strengthen over time (Klein et al., 2005; Thomson et al., 2015). We include link maturation to balance associations between longer-term stability while still accounting for shorter-term variability. We then use the theory as part of a cognitive model that performs the fan effect experiment, using the same stimuli and presentation times as the human participants.

By doing this, we propose the first theory of associative memory to explain how associations are learned and updated throughout the fan effect experiment that capture the results on both target and foil trials. Previous models considered only associations at the end of the experiment (Anderson, 1974; Anderson & Reder, 1999; Rutledge-Taylor & West, 2008; Schneider & Anderson, 2012), but our model extends their description of associative memory by describing the process of how these end-state association strengths are learned.

2. Associative learning in memory recall

Our account of associative learning is situated in the cognitive architecture ACT-R/E (adaptive character of thought-rational/embodied; Trafton et al., 2013), a version of the ACT-R cognitive architecture (Anderson et al., 2004) written in Java with added functionality to be embodied in robotics platforms. ACT-R is an integrated theory of human cognition in which a "production system operates on a declarative memory" (Anderson, Bothell, Lebiere, & Matessa, 1998). In ACT-R, recall and latency depend on three main components: activation strengthening, activation noise, and associative activation. These three values are summed together to represent an item's total activation. When a recall is requested, the item with the highest total activation is retrieved, subject to a retrieval threshold; if no item's activation is above the threshold, the retrieval is said to *fail* and no item is recalled. The latency of the recall is also inversely correlated to the recalled item's activation.

2.1. Activation strengthening

Adaptive character of thought-rational's well-established theory of activation strengthening (also called *base-level* activation) has been shown to be a very good predictor of human memory recall (Anderson, 2007; Anderson et al., 1998). Intuitively, activation strengthening depends on how frequently and recently a memory has been relevant in the past, and it is calculated as follows:

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right), \tag{1}$$

where n is the number of times an element i has been accessed in the past, t_j is the time that has passed since the jth access, and d is the learning parameter, specifying an element's rate of decay. Importantly, this equation predicts that items that have occurred recently, or have been rehearsed more, are more likely to be recalled than those that have not.

2.2. Associative activation

In our account, associative strengths are learned, strengthened, and weakened over time as new elements are learned or prior elements re-experienced. These associations are learned between relevant working memory items within temporal proximity to one another, leading from earlier to later items (Thomson & Lebiere, 2013a, 2013b; Thomson, Bennati, & Lebiere, 2014). The strength of the learned association (or how strongly an existing association is increased) is influenced by the amount of time that passes between when the items were each in working memory. If one item is immediately followed by another in working memory, they will become very strongly associated; on the other hand, if an item has been out of working memory for a while before another is added, they will be only weakly associated. Additionally, associations are *asymmetric*; an association can be stronger from an item i to an item j, for example, than the association from item j to item i (or, there could be no association from item j to item i at all).

To balance the rate of associative learning between long-term stability and short-term variability, we propose a link maturation parameter. Associative link maturation slows the rate of strengthening and weakening based on the number of times the link has been used. This supports long-term stability of well-experienced associative links while allowing for rapid short-term learning of new associative elements. In neural networks, maturation is equivalent to the process of settling to reach a stable equilibrium (Eliasmith, 2005;

Wills, Cacucci, Burgess, & O'Keefe, 2005). Maturation is set using a logistic function:

$$M = 1 - \frac{1}{1 + e^{-(\ln(timesInContext) \times MaturationRate)}} \tag{2}$$

The maturation rate controls the steepness of the curve, or, in other words, controls how quickly links will stabilize.

To compute associative strength from an item j to an item i, the learning mechanism computes an increment I_{ji}:

$$I_{ji} = lr \times w \times M, \tag{3}$$

where lr is a learning rate parameter, w is the weight of the increment determined by the strength of the items in working memory (scales from 0 to 1), and M is maturation. This increment is used to update link strength as follows:

$$S_{ji} = S_{jiPrior} \times \left(1 - I_{ji}\right) + (I_{ji} \times S), \tag{4}$$

where S_{ji} is the strength of the link from j to i, $S_{jiPrior}$ is the prior strength of S_{ji}, I_{ji} is the learning increment from above, and S is a parameter controlling the maximum possible associative strength.

When a new link is learned or existing link updated that shares a source j with other existing links, then each of those other links are discounted proportionally to the *weight* that the original link is updated (e.g., S_{ji} is updated so S_{jk} is discounted):

$$S_{jk} = S_{jkPrior} \times \left(1 - I_{jk}\right), \tag{5}$$

where I_{jk} is computed using the *weight* from the link from j to $i,$ but using the maturation M from the link from j to k. Equation 5 normalizes the amount link j to k is discounted based on the degree to which it has settled. This allows for newer links to rapidly change while providing for long-term stability for more mature links.

This discounting function attenuates link strengths consistent with interference accounts of memory. As more concepts compete in memory, the amount of associative strength from each concept is reduced. In a balanced environment, this discounting will approximate the statistical likelihood $P(i/j)$, which is the odds of perceiving or retrieving i immediately prior to j.

Armed with an understanding of our modeling framework, we now turn to the fan effect experiment itself.

3. The fan effect experiment

To understand the fan effect, we consider the Anderson and Reder (1999) classical fan experiment. They capture the fan effect in a recognition task where participants begin by learning 48 pairs of persons and places. Persons and places could appear in multiple pairs, and each pair was shown for 5,000 ms. Then, during testing, participants respond yes (*target*) or no (*foil*) to whether presented statements were previously studied: The *person* is in the *place* (e.g., "the *hippie* is in the *park*"). In the testing phase, participants were provided a monetary reward based on their total score. The score was computed by providing one point for each correct response, plus an additional point for each 100 ms of response times faster than 1,500 ms. This induced a speed-accuracy trade-off into the experiment.

The experiment proceeded according to three phases: a study phase, drop-out training, and then a testing phase. In the study phase, each stimulus pair was presented once on the screen for 5,000 ms. In the drop-out training phase, participants were presented with a series of questions from each person and place concept of the type "Who is in the *location*?" and "Where is the *person*?" For each question, participants were instructed to respond with all persons associated with the queried location (or vice versa). If they were successful in recalling the requested persons or locations, the current question was dropped from the series and they moved onto the next question. If not, they were presented with the stimulus pairs from the study phase matching the question (i.e., the set of stimulus pairs for the given place or location). The question was then repeated and participants continued this cycle of question and study until they could successfully recall all the persons in the given location (or vice versa), and thus have that question dropped from the series. Once all questions in the series were correctly answered once, participants went through the series of questions a second time according to the same procedure. After successfully completing the second series of questions, participants moved onto the testing phase. Finally, in the testing phase, participants would respond *yes* or *no* to queries "the *person* was in the *location*" with participants receiving feedback on their response.

The experiment manipulated the test stimuli in two different ways. The first was to manipulate the fan of the persons and places. In this experiment, fan is the number of persons associated with a place, and vice versa. Fan is controlled by varying the number of persons in each place, or the number of places with each person (e.g., "the ***hippie*** is in the *bank*" or "the *soldier* is in the ***park***"). Here, the fan of one concept (person/place) was fixed at 2, while the fan of the other concept (place/person, respectively) was varied to be either 2 (low-fan) or 4 (high-fan).

The second manipulation was to control the composition of the set of test stimuli shown to participants by manipulating different target and foil conditions. There were four target conditions: facilitation, interference, suppression, and control. In the facilitation condition, each target (e.g., "the *biker* is the *tower*") was "facilitated" by being repeated five times each in the stimulus set. In the interference condition, each target was repeated only once in the stimulus set and was considered interference because the target's person or place overlapped with a target from the facilitation condition (i.e., "the *biker* is in the *factory*," or "the *doctor* is in the *tower*").

The other two conditions were the suppression and control conditions. In these conditions, each target appeared once in the stimulus set and consisted of concepts that were seen in the interference (but not facilitation) condition, such as *factory* and *doctor* in the above examples. Examples of suppression targets included "the *writer* is in the *factory*," and "the *doctor* is in the *bank*." Due to a particularity in the original study, there is limited difference between suppression and control stimuli, because the controls were designed such that they would functionally suppress stimuli from the suppression condition (e.g., "the *monk* is in the *bank*"). They are different insofar as the suppression stimuli were effectively two steps removed from the facilitation condition, while the control trials were effectively three steps removed.

Foils were classified according to three conditions: high-frequency foils, which used person/place concepts from the facilitation condition but with novel pairings, and were repeated four times each in the stimulus set; low-frequency foils, which had novel pairings of person/place concepts from the interference, suppression, or control conditions and appeared once each in the stimulus set; and mixed foils, which created novel pairings, using one high-frequency concept from the facilitation condition and one low-frequency concept from the interference, suppression, or control conditions and were repeated only once in the stimulus set. In total, there were 48 target sentences and 54 foil sentences in the stimulus set.

The test stimulus set was presented three times in successive blocks, and all stimuli were presented in each block. Feedback was provided for 1,000 ms after participants' responses, with an additional 1,000 ms inter-trial interval.

The results of the Anderson and Reder (1999) study were consistent with interference effects: There were longer latencies and more errors in the high-fan (i.e., fan of 4) conditions relative to the low-fan (i.e., fan of 2) conditions for both targets and foils, with both high-frequency (i.e., facilitation) targets and foils having relatively higher accuracy and quicker latencies than their corresponding low-frequency counterparts. They also predicted lower relative accuracy in the interference condition relative to the suppression and control conditions, and no difference between suppression and control.

3.1. Prior models of the fan effect

There have been several mathematical models of fan effects (Anderson, 1974; Anderson & Reder, 1999; Rutledge-Taylor & West, 2008). Most prominent is Anderson and Reder's (1999) model whose equations were grounded in the ACT-R cognitive architecture (Anderson & Lebiere, 1998). This model can be broken down into three related equations.

$$S_{ji} = S + \ln(1/fan_j) \tag{6a}$$

$$S_{ji} = S + \ln(P(i/j)) \tag{6b}$$

$$A_i = B_i + \sum_j W_j S_{ji} \tag{7}$$

$$T = I + Fe^{-A_i} \tag{8}$$

Equation 6a describes the spread of activation (S_{ji}) from element j to i as a function of associative strength intercept S attenuated by the fan of j, which is the number of concepts to which j is associated. In Eq. 6b, $P(i/j)$ is the frequency-adjusted form of $1/fan_j$, which is a simplification assuming that there were equal frequencies of i and j. In Anderson and Reder (1999), frequencies were not equal, and were instead set ahead of time according to the objective probabilities of perceiving the stimuli in the training phase. Equation 7 relates an activation function A_i to the base-level activation from Eq. 1, and the sum of spreading activation from Eq. 7 multiplied by an attentional weigh W_j. Prior efforts set B_i to 0 on the assumption that the drop-out testing would balance out base-level activation between stimuli. Finally, Eq. 8 computes retrieval time T based on an intercept I, time scale offset F, and the activation function A_i from Eq. 3. The estimates for each parameter were as follows: I was 1,197 ms, F was 773 ms, S was 2.5 ms, and W was .33 (reflecting an even weighting of "*person*" "*in*" "*place*"). Using these parameters, Anderson and Reder report a strong correlation with response times, r = .956. This model did not attempt to fit error patterns, although the dynamics of ACT-R (that response time is solely related to activation strength) imply that their model could theoretically fit the end-state human error pattern data with the appropriate retrieval threshold.

Our model will address the assumption that the recency and frequency of all stimuli will be equated across conditions by the end of the experiment. For instance, Anderson and Reder's (1999) use of objective statistics for i and j for each target condition in Eq. 6b was based on stimulus pair presentation in the study phase, with the assumption made that the number of presentations per condition during drop-out training and testing would be roughly equivalent. This may not be necessarily the case, as the high-fan conditions would have lower activation (thus be more prone to error). This means that they would not all be recalled with equal probability and would need to be rehearsed more often. Their assertion also does not necessarily account for the recall of incorrect pairs (i.e., pairs that were not studied) due to intrusion and confusion of recently studied concepts. Now, considering that Anderson and Reder did not attempt to model error rates, our concerns are not a criticism of their model but instead are an attempt to extend their theory and reconcile their assumption, using an updated account of associative memory.

Our approach is grounded within the ACT-R/E cognitive architecture along with the constraints it places on cognition (Trafton et al., 2013), using a production system simulating the time course of perception, encoding, retrieval, and response. Once base-level activation is included as a factor, then the time course of stimulus presentation and training becomes important in determining overall accuracy and response time. This added fidelity (and complexity) may test assumptions made in prior modeling efforts, and also may provide new insights or hypotheses about how participants learn the task. To that end, our model performs the experiment analogously to participants and learns associations over time. This supports our theory of associative memory explaining how associations are learned and adapt.

4. Learning the fan effect

The model starts with the assumption that all concepts used in the experiment have been encoded in the past with equal frequency[1] and that the model is equipped with the procedural knowledge necessary to perform the experiment. It has no underlying knowledge of the concepts of person and place, and thus has no knowledge of targets or foils. As we have said, the model is presented with the same experimental paradigm as the human participants.

The model uses the same procedural knowledge at the start of each phase to perceive the person and place concepts: When it attends to two concepts on the screen together, then they are encoded as a single concept-pair, with each constituent person and place priming the concept-pair. This concept-pair is represented in memory as a single chunk of information containing three features: *person*, *place*, and *was-target*. The Was-target is a binary *true/false* decision, where *true* indicates that the stimulus was a *target* and *false* indicates that the stimulus was a *foil*.

The external environment is a simulated computer screen. When attending to stimuli, the model randomly attends to one symbol first and then the other, encoding the whole as a single concept-pair. As task instructions do not specify an encoding strategy, we decided that—due to the short size of the sentence—the entire sentence (i.e., concept-pair) should be encoded holistically. The model categorizes these symbols according to two features: *person-symbols* and *place-symbols*. These representations are functionally identical and are distinguished only for lexical purposes to simply categorize incoming words. Since stimuli are of the form "the *person* is in the *place*" we present person-symbols on the left of the display and place-symbols on the right of the display.

When a concept-pair is learned or updated, then the associative strengths between the person/place concepts and the pair are strengthened while the strengths between the concepts and their other related concept-pairs are weakened. Since concepts are related to more concept-pairs on high-fan trials than on low-fan trials, high-fan concepts tend to have lower associative strengths to their related concept-pairs than low-fan concepts. This lower associative strength predicts that concept-pairs involving high-fan concepts will have slower response latencies and increased error rates. Also, since high-frequency stimuli have been seen more often, their strengths will be stronger than low-frequency stimuli; however, maturation controls the degree to which these stimuli increase in strength (e.g., a link seen 4 times as often is not 4 times as strong).

In the study phase, the model automatically encodes all concept-pairs as targets. After encoding the stimuli and generating a concept-pair representation, the model then repeats this encoding until the stimuli are no longer presented on the display, averaging 2–3 rehearsals over the 5,000 ms presentation time.

In the drop-out training phase, the model goes through the same two-round drop-out testing as the participants. For each stimulus in the set, the system displays either a *person* or *place* reflecting the question "Where is the *person*?" and "Who is at the *location*?" respectively. The model then attempts to retrieve all *places* where that *person* is (or all *persons* in that *place*). If the model correctly perceives all required *places* or *persons*, then the model moves onto the next stimulus in the set; otherwise it studies those stimuli again (by re-running those trials in the study phase) and returns to the drop-out training to be asked the same query again. Once the model has successfully retrieved all elements in both run-throughs of the drop-out training phase, the test phase begins.

The test phase is where all response times and error rates were recorded. Similar to the study phase, the model encodes the concept-pair for *person* and *place* from the display as an analogue to perceiving: Was *person* in the *place*? The model then attempts to retrieve any concept-pair containing said *person* and/or *place*. We consider the model's decision to be the value of the was-target slot in the retrieved chunk. A response is then generated according to the following criteria: (a) if the model is unable to retrieve any concept-pair due to all pairs being below threshold, then it responds *foil*; (b) if the model correctly retrieves the matching person and place, then it examines the *was-target* decision and responds with the respective decision: target for true and foil for false; or (c) if the model retrieves a mismatching concept-pair containing one correct *person* or *place* but not both, then it assumes that the response is a *target*. After the model responds it receives feedback from the display, which it uses to encode the correct concept-pair. This includes encoding foils, which allows the model to be capable of correctly retrieving that an item was a foil seen in an earlier testing phase. This is a unique behavior of our model and reflects the fact that participants cannot "ignore" decisions they have made and must encode feedback that they have seen. Finally, if the model was incorrect or had retrieved a mismatched element, then for the feedback period it rehearses the correct response. This process is repeated across all trials through all three test phases.

In the testing phase, response times are recorded from the stimulus onset time until the model has responded with the appropriate decision (target or foil). It is important to note that as a fully implemented production system model, the complete time to respond includes two relatively fixed durations: approximately 600 ms to encode the stimuli from the display, and approximately 350 ms to prime the motor command and press the response key. This is in a similar range to the structural offset I of 1,197 ms that Anderson and Reder (1999) used in Eq. 9. This means that fan effects in latencies occur mainly in the approximately 200–800 ms timeframe where the concept-pairs are retrieved. It is the retrieval of the concept-pair that determines fan effects in both latencies and accuracy.

One final difference between the present model and prior efforts is that our model incorporates base-level activation B_i (see Eq. 8) but replaces the S_{ji} from Eq. 7 with our learned S_{ji} from Eq. 5. As previously mentioned, base level activation reflects the recency and frequency of use of elements, and it is not a given that base level would be equivalent for items across the different target and foil conditions, especially for the high-frequency versus low-frequency elements where frequency is necessarily varied.

4.1. Results

The present model was run for 200 iterations with the parameters described in Table 1. As is apparent from Figs. 1 and 2, the model captured fan effects in both accuracy and latency, respectively, reflected by slower response times and higher error rates for high-fan concepts compared to low-fan concepts. Similar to human performance, we predicted relatively higher accuracy for high-frequency conditions (facilitation for targets and high-frequency for foils) compared to the rest of the conditions; lower accuracy in the interference condition relative to the control and suppression conditions, and no difference

Table 1
Parameters used in fan effect experiment

Parameter	Value
Base-level learning (B_i)	0.4
Learning rate (*lr*)	1.1
Maturation rate (*M*)	0.5
Maximum associative strength (*S*)	7.25
Mismatch penalty	4

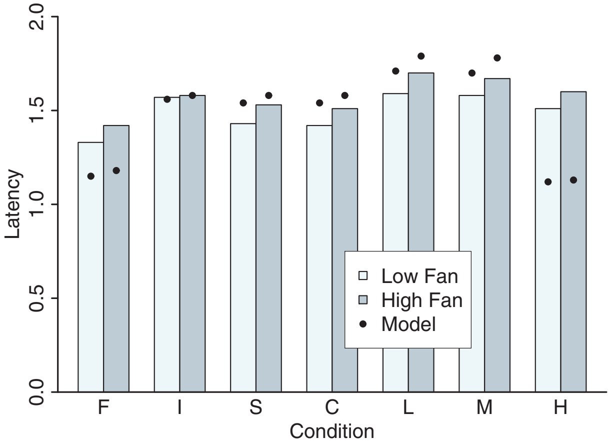

Fig. 1. Latencies across target and foil conditions. Target conditions are facilitation, interference, suppression, and control. Foil conditions are low, mixed, and high.

between suppression and control target conditions. One difference is that we predict a smaller average fan (.03 s instead of .09 s) than did Anderson and Reder (1999).

Overall, our model fits closely to track human performance, deviating more in the high-frequency conditions. Our model substantially predicted human performance in accuracy (R^2 = .75) and latency (R^2 = .38). Our fits are depressed by the relatively higher accuracy and quicker response times in the high-frequency condition. Excluding this condition (which still predicted fan effects), our latency fit significantly improves (R^2 = .75). It is important to note that our fits are lower than those of Anderson and Reder (1999); however, our model simulates nearly 2 hours of task performance with learning throughout and captures both accuracy and response times.

Interestingly, the source of errors is different between targets and foils. Target errors are due to failures in retrieving any concept-pairs, whereas foil errors are due to confusing foils with previously seen similar stimuli.

4.2. Discussion and conclusions

The present model describes the emergence of fan effects in both accuracy and latency, using a theory of associative memory including an account of interference by discounting

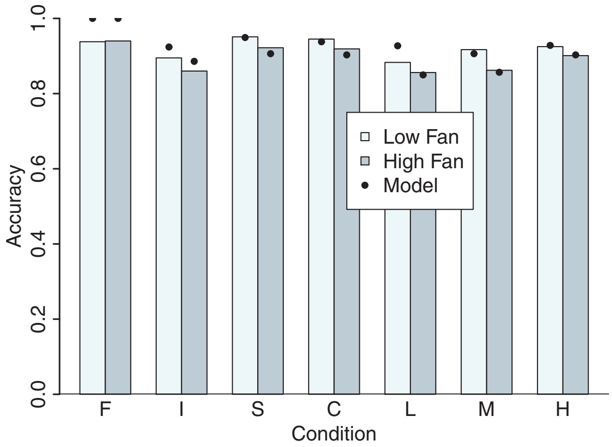

Fig. 2. Error rates across target and foil conditions. Target conditions are facilitation, interference, suppression, and control. Foil conditions are low, mixed, and high.

link strengths. As more stimuli (persons or places) are presented together (or within a short temporal window), they interfere with each other's associative strength, reducing overall activation. This has the effect of lowering overall accuracy and increasing response times. While prior explanations (Anderson & Reder, 1999; Schneider & Anderson, 2012) have presented good fits to human latency based solely on associative weights, the present model learns both base-level activation (reflecting recency and frequency of use) and associative weights throughout the entire experiment (including the testing phase) and predicts the presence of fan effects in both latency and accuracy.

Many mathematical models (e.g., Alpaydin, 2014; Anderson & Reder, 1999; Sohn, Anderson, Reder, & Goode, 2004) assume a fixed testing phase (where the model does not continue to learn. An advantage of modeling the fan effect experiment while learning throughout the experiment is that we are able to assess some of the assumptions made in prior modeling efforts. Most interesting is that the assumption that base-level activation would be similar between high-fan and low-fan stimuli was valid in aggregate, but it was highly variable between conditions and throughout the different phases of the experiment (see Table 2). This variability was due, in part, to the differences in stimulus presentation frequency seen in the study phase (e.g., facilitation vs. interference, suppression, and control). Perhaps counterintuitively, high-fan targets generally have higher base-level activation than their low-fan equivalents due to the extra rehearsals they require to reach criterion in the drop-out phase and their relatively increased error rates (due to lower associative activation) in the testing phase.

A novel decision that we made was to encode foils. This was done to reflect the fact that participants must encode feedback to which they have explicitly attended (Muzzio, Kentros, & Kandel, 2009). Furthermore, it maintains consistency across conditions: Stimuli perceived similarly on the display are also encoded similarly. Encoding foils causes retroactive interference (i.e., reduced associative strength) for existing targets, as reflected in Table 2.

Table 2
Average base-level activation (B_i) and average associative strength per link (S_i) per condition across drop-out training and testing. Activations are used in Eq. 7 to determine accuracy and in Eq. 8 to determine response time

	Drop		Test 1		Test 2		Test 3	
	B_i	S_i	B_i	S_i	B_i	S_i	B_i	S_i
F2	0.44	1.33	0.54	0.97	0.73	0.91	0.86	0.82
F4	0.91	0.69	0.87	0.55	0.83	0.48	0.97	0.44
I2	0.40	1.34	0.25	0.99	0.59	0.94	0.59	0.86
I4	0.70	0.70	0.11	0.58	0.40	0.60	0.60	0.47
S2	0.32	10.35	0.11	1.03	0.67	1.00	0.34	0.95
S4	0.41	0.72	0.26	0.61	0.53	0.60	0.40	0.49
C2	0.47	1.40	0.08	1.17	0.25	1.05	0.35	0.98
C4	0.44	0.71	0.37	0.59	0.60	0.52	0.60	0.52
L2	N/A	N/A	0.13	1.06	0.57	1.00	0.40	0.92
L4	N/A	N/A	−0.65	0.57	0.29	0.60	0.47	0.48
M2	N/A	N/A	−0.57	1.08	0.42	1.01	0.44	0.92
M4	N/A	N/A	−0.15	0.61	0.38	0.52	0.25	0.51
H2	N/A	N/A	0.63	3.21	0.75	3.22	0.78	3.20
H4	N/A	N/A	0.47	2.65	0.55	2.69	0.55	2.71

In our model, associative activation was comparable between the high-frequency *facilitation* condition and the low-frequency *interference*, *suppression*, and *control* conditions. Instead, the difference in model performance was based on difference on the base-level activation of recalled concepts. This difference in activation between conditions stands in contrast to the prediction of the Anderson and Reder (1999) model, who assume that base-level activations would be equivalent and that associative activation would vary between the facilitation and other conditions due to differences in the frequency of perceiving facilitation stimuli in the study phase. While both approaches attribute changes between conditions to the frequency of perceiving stimuli, the difference is that our model's maturation mechanism stabilizes targets' link strength over the course of the experiment while the well-justified base-level activation (Anderson, 2007; Anderson et al., 2004) varies according to the recency and frequency of concept perception and recall.

There are some aspects of our model that will be addressed in future work. We did not attempt to model a strategy of speed-accuracy trade-offs reflecting the time-pressure-based reward system of the original experiment. ACT-R does not have a mechanism to distinguish recognition from recall: Recall is an all-or-nothing event. It was not obviously possible to have a meta-awareness of stimulus familiarity buildup throughout the retrieval process (i.e., a *feeling of knowing*; Reder & Ritter, 1992), something which could be leveraged to induce speed-accuracy tradeoffs. This speed-accuracy trade-off may result in less difference in response times between the high- and low-frequency conditions, improving our fit to human data.

It is fair to argue that our model is substantially more complex than prior efforts, but we argue that this complexity is necessary to understand how fan effects arise from learning. Our model predicts that differences in response time and accuracy between the high-frequency *facilitation* condition and the low-frequency conditions are primarily due to the frequency of use of the high-frequency concepts rather than their associative activation to other concepts. By having our model perform the study equivalently to human participants and by having participants learn associative weights throughout the experiment, we present a model that supports our theory of associative memory and explain how associations are learned and adapt over time.

Acknowledgments

This work was supported by the Office of the Secretary of Defense/Assistant Secretary of Defense for Research and Engineering (LH) and the Office of Naval Research (LH). The views and conclusions contained in this paper do not represent the official policies of the U.S. Navy, the Department of Defense, or the United States Military Academy.

Note

1. We did not bias the beginning recency and frequency of the concepts based on prior word frequency (e.g., Kučera & Francis, 1967). We also assumed all concepts were unassociated at the beginning of the experiment, although it is possible to determine such relationships ahead of time (e.g., latent semantic analysis; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). While these are measures, we are interested in incorporating in future research, unless there were systematic biases in the frequency of, or relationships between, stimuli across conditions, we expect the impact to our results to be minor.

References

Alpaydin, E. (2014). *Introduction to machine learning*. Cambridge, MA: MIT Press.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*, 451–474.

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

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review*, *111*, 1036–1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341–380.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*, 186–197.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–408.

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation*, *17*(6), 1276–1314.

Hiatt, L. M., & Trafton, J. G. (2013). The role of familiarity, priming and perception in similarity judgments. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 573–578). Austin, TX: Cognitive Science Society.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*, 103–109.

Klein, K. A., Addis, K., & Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition*, *33*(5), 833–839.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Muzzio, I. A., Kentros, C., & Kandel, E. (2009). What is remembered? Role of attention on the encoding and retrieval of hippocampal representations. *The Journal of Physiology*, *587*(12), 2837–2854.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with questions terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *18*, 435–451.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Rutledge-Taylor, M. F., & West, R. L. (2008). Modeling the fan effect using dynamically structured holographic memory. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 385–390). Austin, TX: Cognitive Science Society.

Schneider, D. W., & Anderson, J. R. (2012). Modeling fan effects on the time course of associative recognition. *Cognitive Psychology*, *64*, 127–160.

Sohn, M. H., Anderson, J. R., Reder, L. M., & Goode, A. (2004). Differential fan effect and attentional focus. *Psychonomic Bulletin & Review*, *11*(4), 729–734.

Stanovich, K. E., & West, R. F. (1983). On priming by a sentence context. *Journal of Experimental Psychology*, *112*(1), 1–36.

Thomson, R., Bennati, S., & Lebiere, C. (2014). Extending the influence of contextual information in ACT-R using buffer decay. In P. Bello, M. Guarini, M. McShane, & B. Scasselatti (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1592–1597). Austin, TX: Cognitive Science Society.

Thomson, R., & Lebiere, C. (2013a). Constraining Bayesian inference with cognitive architectures: An updated associative learning mechanism in ACT-R. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the Conference of the Cognitive Science Society* (pp. 3539–3544). Austin, TX: Cognitive Science Society.

Thomson, R., & Lebiere, C. (2013b). A balanced Hebbian Algorithm for associative learning in ACT-R. In R. West, & T. Stewart (Eds.), *Proceedings of the international conference on cognitive modeling*. Ottawa, ON.

Thomson, R., Pyke, A., Trafton, J. G., & Hiatt, L. M. (2015). An account of associative learning in memory recall. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2386–2391). Austin, TX: Cognitive Science Society.

Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello II, F., Khemlani, S. S., & Schultz, A. C. (2013). ACT-R/E: An embodied cognitive architecture for human-robot interaction. *Journal of Human–Robot Interaction*, *2*(1), 30–55.

Wills, C. L., Cacucci, F., Burgess, N., & O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, *308*, 873–876.