Modeling a Three Term Fan Effect

Matthew F. Rutledge-Taylor (mrtaylo2@connect.carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive
Ottawa, Ontario, K1S 5B6, Canada

Aryn A. Pyke (apyke@ccs.carleton.ca)
Department of Psychology, Carleton University, 1125 Colonel By Drive
Ottawa, Ontario, K1S 5B6, Canada

Robert L. West (robert_west@carleton.ca)
Institute of Cognitive Science, Department of Psychology, Carleton University, 1125 Colonel By Drive
Ottawa, Ontario, K1S 5B6, Canada

Hana Lang (hlang@connect.carleton.ca)
Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive
Ottawa, Ontario, K1S 5B6, Canada

Abstract

A fan effect experiment where participants perform recall and recognition tasks on a study set of sentences with three content words was conducted. The aggregate results confirm a fan effect (Anderson, 1974). A model of the recall and recognition tasks was created using Dynamically Structured Holographic memory (DSHM). A comparison to the human data is presented. A discussion of the current resonance based mechanisms in DSHM for generating recognition accuracy and reaction time data is presented. This is contrasted with a previously employed retrieval based mechanism.

Keywords: cognitive modeling; the fan effect; holographic reduced representation.

Introduction

The purpose of this paper is to report the results of a fan effect style experiment and to demonstrate that these results can be captured by Dynamically Structured Holographic memory (DSHM). The experiment conducted was similar to the classic fan effect paradigm (Anderson, 1974).

In Anderson's original experiment, participants studied a set of sentences that contained two content words: a person and a place (e.g., "the hippie is in the park"). Each content word appeared in one, two, or three different sentences. The number of sentences in which a word appears is the *fan* of that word. Each sentence is assigned a fan, which is the sum of the fans of the content words in the sentence. For example, if 'hippie' appeared in three sentences while 'park' appeared in one sentence, 'hippie' would have a fan of three, 'park' would have a fan of one, and the sentence 'the hippie is in the park' would have a fan of four. The results of a recognition task performed on the sentences (and an equal number of foils) demonstrated that the time required to affirm or reject a sentence as a member of the study set was correlated with the fan of the sentence.

The present work extends prior research on the fan effect, and models thereof. We explore the generality of the fan effect by examining memory performance for sentences with three content terms rather than just two (e.g., Anderson, 1974). Additionally, our sentences had a wider range of fans than have typically been studied (or modeled).

The Three Term Fan Experiment

Method

Twenty seven participants (12 males and 15 females: mean age 20.0 years, SD = 2.2) were recruited from introductory psychology courses a Carleton University to take part in the experiment. Participants received course credit for their time. Participants took part in the experiment one at a time. The experiment was divided into three main phases: A study phase, a recall phase and a recognition phase.

In the study phase each participant was assigned one of three unique sets of study sentences and was instructed to memorize the sentences in the list. Once the participant indicated that he or she was prepared to proceed, the recall portion of the experiment began.

The study set consisted of sixteen sentences of the form, "The *color thing* is in the *place*". The color term was one of ten colors; the thing was one of ten house-hold items; and the place was one of ten locations in/around a typical home. Very typical item/locations combinations, such as 'comb'/'bathroom', were omitted when generating the study set sentences. Eight terms from each category appeared in one study sentence each, while two terms from each category appeared in four sentences each. No two terms appeared together in more than one sentence. For example, if "The orange comb is in the garage" was a member of the study set, no other sentence in the study set described an orange comb, a different colored comb in the garage, or any other orange object in the garage. However, these combinations could occur in foil sentences.

The fan of a sentence is the sum of the fans of the terms in the sentence. Thus, the four possible sentence fans were:

3, 6 9, and 12. The fan effect predicts that judgments for sentences with higher fans should take longer (i.e., have higher reaction times) than for sentences with lower fans. Additionally, the truth of sentences with a higher fan should be recognized with less accuracy than sentences with a lower fan.

Recall Task Method

Each participant engaged in three iterations of the recall task. Each iteration began with the participant trading the study sentences list with the experimenter for a new list of sentences identical to the study set, but with one term from each sentence replaced with a blank, and the order of the sentences randomized. The participant's task was to correctly fill-in each of the blanks with the missing word. The participant was given as much time as he or she needed to do so. The experimenter then recorded the number of correct responses and for each error, provided the correct missing word to the participant. The participant was then given the opportunity to review the study set again. The three iterations were balanced such that each term from each sentence in the study set was replaced with a blank exactly once. After the third iteration the recognition phase began.

Recognition Task Method

The recognition task was conducted on a computer using the Experiment Builder software package from SR Research. Sentences were presented one at a time, centered on a 17" CRT monitor (in black font on a white background). Participants judged whether each presented sentence was a member of the study set, or not. To respond, participants hit either the z-key or the /-key, respectively. Accuracy and reaction time were recorded for each trial. After each trial, the screen blanked for 1 second, and then the word "READY" appeared for 1 second to prepare the participant for the next trial.

The participant was presented with 96 test sentences, which consisted of three exposures to each of the study set sentences, and 48 foil sentences which were not from the study set. Participants were told that they should consider sentences from the study set to be *true*, while all others should be considered *false*. Each false sentence was generated by replacing one of the three terms from a true sentence with another term from the same category (e.g., color, thing, or place) and with the same fan. For example, for a true sentence like "The blue hat is in the garage", one false counterpart might be "The green hat is in the garage". Each true sentence was used to generate three different false sentences. Thus, for each exposure of a true sentence there was a corresponding false test sentence with the identical fan.

Results

The data from one participant was excluded from the analysis below. This participant's recognition reaction time was significantly longer than all the other participants by a large margin ($P < .001$). The results below reflect the data collected from the remaining 26 participants.

Human Recall Performance

Performance in the recall task improved, on average, with each of three iterations. Table 1 presents the mean number of correct responses (out of 16), the standard deviation, and the accuracy measured as a percentage for each of the three iterations of the recall phase.

Table 1: Recall accuracy

	Iteration		
	1	2	3
Correct (/16)	10.9	13.4	14.6
SD	3.7	3.1	1.8
Percentage	68.1	83.8	91.4

This result is important because an intended purpose of the recall task was to confirm that the participants had memorized the study set before entering the recognition phase. By the end of the third iteration the participants were correctly completing the sentences 91.4 percent of the time.

Human Recognition Performance

Overall, participants' accuracy and reaction time results were consistent with the fan effect. For both true and false sentences, accuracy was negatively correlated with sentence fan. Also, accuracy was poorer for false sentences than for true sentences for all sentence fans ($ps < .05$).

Table 2: Recognition accuracy (%)

	Accuracy	
Sentence fan	True	False
3	97.5	95.5
6	95.1	91.7
9	92.1	86.3
12	82.7	77.6

Reaction time increased with sentence fan ($p < .001$) for both true and false sentences, and true sentences were judged more quickly than false ones ($p = .001$).

Table 3: Recognition reaction time (ms/char)

	True		False	
Sentence fan	Reaction time	SD	Reaction time	SD
3	59.0	18.5	64.1	19.9
6	63.6	20.0	69.3	22.6
9	74.2	21.8	86.2	31.1
12	91.3	31.5	102.5	43.6

Table 3 shows the reaction times (ms/char) for both true correct (i.e., the test sentence was true and was judged correctly) and false correct sentences of each fan.

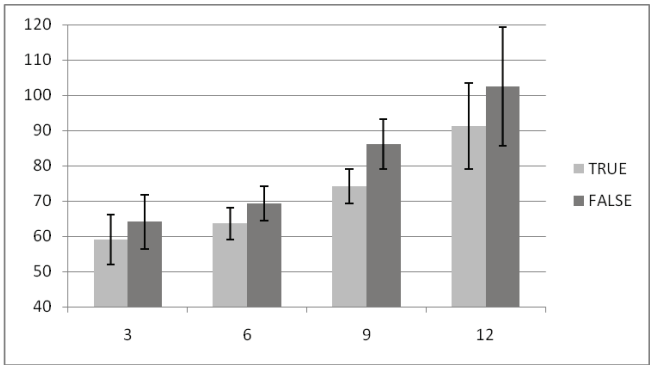

Figure 1: Recognition reaction time by sentence fan (ms/char) with confidence intervals

Figure 1 shows the mean reaction times, measured in ms/character, for both true correct and false correct sentences, for each sentence fan with confidence intervals. There was no interaction of truth and fan (p = .199). Table 4 presents the pairwise comparisons across fan using the Bonferonni adjustment.

Table 4: Pairwise comparisons for correct reaction times

Sentence fans	P (one-tail)
3 versus 6	0.129
6 versus 9	< 0.001
9 versus 12	< 0.001

In Summary

The results of the experiment confirm the fan effect as a robust phenomenon that generalizes from sentences with two content terms (Anderson, 1974) to sentences with three content terms (present research). Future work will examine whether statistically significant differences can be found in the relative contributions of the terms to the fan effect (e.g., does the *color* term contribute differently than the *thing* or *place* terms).

DSHM

The memory modeling system used to model the described experiment was Dynamically Structured Holographic Memory (DSHM) (Rutledge-Taylor & West, 2008). DSHM is based on the BEAGLE model of the lexicon (Jones & Mewhort, 2007). The details of the DSHM architecture and the similarities between BEAGLE and DSHM can be found elsewhere (Rutledge-Taylor & West, 2007).

For an account of the use of DSHM to model the classic fan effect, and a comparison to ACT-R (Anderson & Lebiere, 1998), see Rutledge-Taylor and West (2008). For those unfamiliar with DSHM, a brief introduction follows.

DSHM makes use of holographic reduced representations (HRR) to encode knowledge in memory. See Plate (1995) for a discussion of the sort of HRRs used by DSHM (and BEAGLE). A DSHM system is composed of a collection of items that are represented internally as two vectors of numbers: i) the environmental vector is static and uniquely identifies the item in the system; ii) the memory vector is dynamic and encodes all of the associations an item develops with other items. The lengths of these vectors are fixed for an instance of DSHM, but can be initially set to any positive integer which is a power of 2.

DSHM takes collections of items as input (called complex items; collections of items are items themselves). The structure of a complex item can be expressed using left and right brackets. For example the sentence "The red hat is in the garage" can be expressed [red:hat:garage]. The system can also allow items to have a hierarchical structure. Here, the context tags used to classify an item as background knowledge (false) versus experimental knowledge (true) applies to the sentence as a whole, and so is up a level in the hierarchy, expressed: [true [red:hat:garage]]. Items can bear ordered (delimited by colons) or unordered (delimited by spaces) relationships with one another.

Information is extracted from DSHM by presenting it with incomplete complex items. For example, a query for the color of an item might be expressed [true [?x:hat:garage]. Any missing items are called query items and in DSHSM syntax are always preceded with a question mark, (e.g., "?x"). A query item is like a variable that DSHM is tasked with resolving. DSHM makes use of information stored in the memory vectors of the provided items to generate a rank ordered list of candidate items for replacing the query item. Each candidate completion is accompanied by a numerical value ranging from 0.0 to 1.0 that indicates the strength of the completion. This strength is referred to as the confidence (i.e., how confident DSHM is in the completion being correct, or appropriate). It can also be thought of a context relative activation value, to use an ACT-R term.

A DSHM model is constructed by making choices about how information is represented in complex items, what vector size should be used, what training regime is used, and what sorts of queries are presented to the system.

The Model

Twenty-seven simulated participants were run (to correspond to the 27 human participants). It was found that a range of vector lengths allowed the simulated participants to produce reasonable recognition accuracy and reaction time results. However, fitting the recall data was more of a challenge. Uniformly using a vector length of 64 produced significantly poorer performance than the average for the human participants, while a vector length of 128 produced significantly better results. No value in between is possible (vector length must be powers of 2). In order to produce good average scores, nine of the simulated participants were given memory systems that made use of vector lengths of 64, while the other 18 used vector lengths of 128.

Study Phase

Prior to learning the study set, each simulated participant read 1026 background knowledge sentences, each encoded as a flat ordered list of three content terms associated with a tag 'false'; "[false [color:thing:place]]". The false sentences included either one or two of the content terms appearing in the study set. The remaining one or two terms were nonsense terms that did not occur in the study set sentences. The background knowledge was needed in order to give the simulated participants some basis for making errors. Without background knowledge there is nothing for DSHM to confuse the study set sentences with; DSHM does not make use of explicitly added noise.

The simulated participants read each sentence in the study set once or twice (to account for the differences in how well the human participants prepared themselves for the first task) prior to beginning the recall phase. Sentences from the study set were associated with a context tag representing 'true'; "[true [color:thing:place]]".

Recall Performance of the Model

Like the human participants, the simulated participants produced responses to fill-in the blank questions in the recall phase. For example, "The _____ hat is in the garage" was submitted to the DSHM participant as "[true [?x:hat:garage]]". The system outputs a list of candidate responses, in rank order. The one with the highest rank was considered to be the simulated participant's response. If the system's response item matched the correct missing term, the trial was scored as correct.

After each iteration the DSHM participant read each of the study set sentences once for every three incorrect responses on the previous iteration. The majority of human participants took the opportunity to review the study set, even after scoring perfectly on the previous iteration. Thus, the DSHM participants re-read the study set a minimum of once between trials.

Table 5: Model recall accuracy

	Iteration		
	1	2	3
Correct	11.1	14.1	14.1
SD	3.2	2.1	1.9
Percentage	69.4	88.2	88.2

Table 5 presents the recall accuracy for the simulated participants. Although, the accuracy plateaus after the 2nd iteration, there is an overall good match for accuracy and standard deviation, as demonstrated in figure 2 (only the standard deviations for the human data are shown).

Recognition Performance of the Model

The simulated participants were each tested on the same 96 test sentences as the human participants. In order to produce a truth judgment the simulated participant was presented with a query of the form "[?x [color:thing:place]]". If the system produced 'true' as its highest ranked completion candidate, the simulated participant was considered to have judged the sentence to be *true*, otherwise, the simulated participant was considered to have judged the sentence to be *false*.

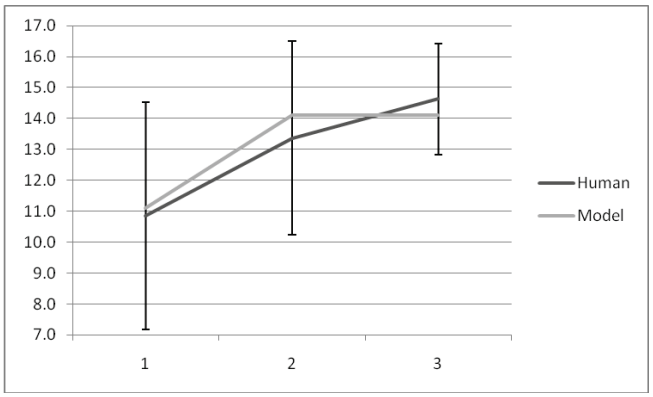

Figure 2: Recall accuracy (out of 16)

To determine the reaction time for the response, the model evaluated the degree to which the test sentence as a whole (without a context tag) ("[color:thing:place]"), resonated within the system. The sentence's resonance is produced by a built-in DSHM method, which essentially determines how closely associated the terms in the sentence are to one another. Here, the resonance value is interpreted as indicating how familiar the sentence seems to the simulated participant. Thus, if the sentence is judged to be true, a high resonance should make this decision easier. If it is lower, it should make the decision harder. The opposite is the case for judgments of false. It should be difficult to reject a sentence that seems familiar, and vice-versa.

The formula used for translating resonance to reaction time was RT = 32 / R, where R is the value provided by the memory system of the simulated participant, and RT is reaction time measured in ms per character. For true sentences R is the resonance value for the sentence. For false sentences, R is the resonance value for the sentence subtracted from an upper limit on resonance values. This upper limit was estimated to be the maximum resonance value calculated for any of the true sentences (0.64). Table 6 presents the reaction time data for the model.

Table 6: Model reaction time (ms/char)

	Reaction Time	
Sentence fan	True	False
3	61.1	67.4
6	69.0	69.5
9	79.6	77.1
12	87.8	94.5

Figure 3 presents a comparison of the human and model

data for correct trials. The solid lines correspond to the human data; the dashed lines correspond to the model data; the light lines correspond to the true data; and the dark lines correspond to the false data.

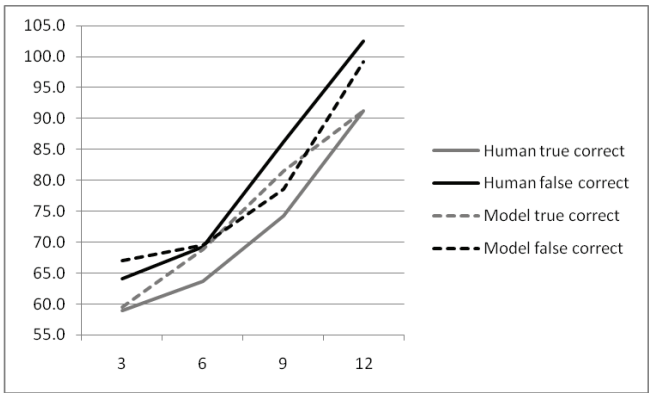

Figure 3: Recognition reaction time

In terms of judgment accuracy, the model outperformed the human participants. The simulated participants had a judgment accuracy of 100% for true sentences and 98.5% for false sentences. It is possible that this discrepancy may be due to the relatively small body of 'interfering' background knowledge in the simulated participants relative to real human participants.

In Summary

In general the model results provide a good match to the human data, in that 1) the false sentences take longer, on average, to affirm or deny than do true sentences (77.1 ms/char versus 74.4 ms/char); 2) a fan effect is observed for both true sentences and false sentences; 3) the model provided a good fit to recall performance as well as recognition performance; and 4) the formula used to convert raw model output to reaction time values is simple and provides a good fit to the recognition times using a single scaling parameter.

On-Going Work: Effect Of How Fan Is Distributed?

Part of the motivation for this experiment and model construction was to investigate whether each content word in a sentence contributes equally to the difficulty in recognizing a sentence as true (i.e., a member of study set). It was hypothesized that the color term may make a smaller contribution to the fan effect than the thing or the place. This is because the color terms are adjectives and more ubiquitous than the things or places, which are nouns. However, whether the thing or the place should carry more weight was not predicted given conflicting intuitions about why one or the other should be more influential. For example, the thing term might be the most influential because an object's type (e.g., hat) is a more intrinsic property than its location (or color). Alternately, place might be more influential: Grammatically, the color and thing share a common noun phrase, while the place does not share its prepositional phrase with any other content word.

The human data were not clear cut with regard to the influence how fan was distributed among content terms. By fan distribution, we are referring to the possible pattern of the fans of the words making up sentences with a particular fan. There are three different ways to make fan 6 sentences (color term fan = 1, thing = 1, place = 4; 1,4,1; 4,1,1), and three ways fan 9 sentences (1,4,4; 4,1,4; 4,4,1), while there is only one way to make fan 3 sentences (1,1,1) and one way to make fan 12 sentences (4,4,4).

No significant effects of fan distribution were found among fan 6 sentences. But, among sentences with a fan of 9, an ANOVA with revealed that fan distribution did have an impact on RT (p =.002). Specifically, RT was faster when either the thing or place was unique (i.e., fan 1) and slower when the color was unique. Put another way, when trying to judge whether a sentence is true (e.g., "The red hat is in the garage"), knowledge of other objects with the same color (red ball) adds less difficulty than knowledge of other items of the same type (hat) or other items in the same place (garage). Further, RTs tended to be faster when the thing type was unique rather than the location, though this trend did not reach significance.

Note: a simple variation in the representation of sentences in DSHM would be able to account for this effect because DSHM is capable of representing facts that have hierarchical structure. In fact, DSHM already leverages this capability in the current model. In the representation "[true [color:thing:place]]", the three term sentence as a whole aggregate is the hierarchical sibling of the context tag ('true'). In order to represent sentences where the thing term is dominant, the color and place need only be embedded in a list of peripheral properties as in the following representation: "[true [thing:[color:place]]".

Exploratory simulations confirm that using this type of representation predicts differences in reaction times among fan 9 sentences, where the thing fan dominates the fans of the other two terms. Similarly, for sentences with an overall fan of six and a thing fan of four are significantly slower than fan 6 sentences with a color fan of four, or a place fan of four. Additional human testing is required to gather more information about the effects of fan distribution. But it is noteworthy that such hierarchical effects could be naturally afforded by structural aspects of a DSHM architecture. This line of research is on-going.

Appendix

Relationship To the Two Term Model

Rutledge-Taylor and West (2008) presented a model of the fan effect, as described in Anderson (1974). This model provided a good match to the human data, but used a different mechanism for calculating recognition accuracy and reaction time values, than the one presented here. This mechanism, which we will refer to as the 'retrieval'

mechanism operates as described below.

Whether DSHM recognizes a sentence, or not, according to the retrieval mechanism is based on how strongly the words in the sentence are associated with one another. Specifically, if at least one of the words in the sentence (referred to as a target word) can be recovered using the other words in the sentence as cues, the sentence as a whole is recognized (as true), otherwise, it is not.

If the sentence is recognized, the reaction time is based on strengths (e.g., confidence values) of the recovered target words, which are high on average, resulting in low reaction times. If the sentence is not recognized, the reaction time is based on strengths of the words that were retrieved (but did not match the target words). On average the strengths of these retrieved words are lower, resulting in higher reaction times. Additionally, the fans of the words in the sentences affect the strengths of the retrieved words and it these strengths that are the basis for the fan effect in the DSHM model.

Using The Retrieval Mechanism In The Three Term Model

The retrieval mechanism for generating recognition and accuracy results for the DSHM model was initially tested on the current stimuli and without using background knowledge sentences, which are not necessary for this mechanism. The retrieval mechanism produced a 100% accuracy rate for identifying true sentences, but only a 36% accuracy rate for rejecting false sentences.

The retrieval mechanism produced a very good fit to the human true correct reaction times, including the characteristic exponential curve observed in the human data (for both trues and falses). However, the model results for false correct (i.e., correct rejections) reaction times were drastically different from that of the human data. See figure 4.

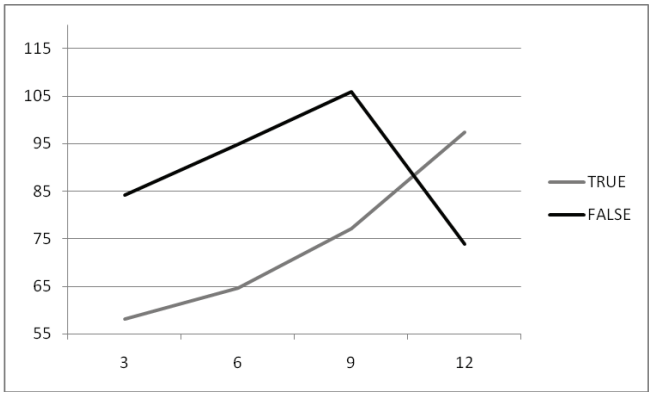

Figure 4: Reaction times (ms/char) using the retrieval mechanism

The explanation for the model's false correct data has to do with the number of true near neighbors the false sentences have. Here, 'near neighbors' are defined as two sentences that differ only by a single word. The number of near true neighbors a false sentence has is correlated with its fan. This is the result of the counter-balancing of true and false sentences. The existence of near neighbors makes little difference in the recognition results for false fan 3, 6 and 9 sentences. However, for fan 12 sentences there are true near neighbors that are retrieved (for each target word) with very high strengths. This results in low reaction times for false sentences with a fan of 12. For example, when presented with the false sentence "the black mug is in the garage", the true sentence "the grey much is in the garage" is retrieved with a high confidence value, when internally testing to see if "[?x:mug:garage]" retrieves 'black' as a candidate completion of the query term '?x'.

Due to the failure of the retrieval mechanism to provide a satisfactory account of the reaction times for correct rejections, the new mechanism described above was developed. It is the authors' belief that the retrieval mechanism ought to work for most DSHM models under most circumstances. However, in cases such as the one presented here, the new mechanism can be applied in order to generate recognition reaction times for correct rejections that are resistant to the effects of true near neighbors.

References

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

Anderson, J. R., & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and Order information in a composite holographic lexicon. *Psychological Review, 114*, 1-37.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks, 6,* 623-641.

Rutledge-Taylor, M. F. & West, R. L. (2007) MALTA: Enhancing ACT-R with a holographic persistent knowledge store. *Proceedings of the XXIV Annual Conference of the Cognitive Science Society*. Nashville, TN.

Rutledge-Taylor, M. F. & West, R. L. (2008) Modeling The fan effect using dynamically structured holographic memory. *Proceedings of the XXX Annual Conference of the Cognitive Science Society*. 385-390. Washington, DC