Author/s:
Osth, AF;Shabahang, KD;Mewhort, DJK;Heathcote, A

Title:
Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model

Date:
2020-04-01

Citation:
Osth, A. F., Shabahang, K. D., Mewhort, D. J. K. & Heathcote, A. (2020). Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model. Journal of Memory and Language, 111, https://doi.org/10.1016/j.jml.2019.104071.

Journal of Memory and Language 111 (2020) 104071

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Global semantic similarity effects in recognition memory: Insights from BEAGLE representations and the diffusion decision model☆

Adam F. Osth[a,*], Kevin D. Shabahang[a], Douglas J.K. Mewhort[b], Andrew Heathcote[c]

[a] University of Melbourne, Australia
[b] Queens University, Canada
[c] University of Tasmania, Australia

ARTICLE INFO

Keywords:
Recognition memory
Semantic similarity
Semantic space models
Diffusion decision model

ABSTRACT

Recognition memory models posit that false alarm rates increase as the global similarity between the probe cue and the contents of memory is increased. Global similarity predictions have been commonly tested using category length designs where it has been found that false alarm rates increase as the number of studied items from a common category is increased. In this work, we explored global similarity predictions within unstructured lists of words using representations from the BEAGLE model (Jones & Mewhort, 2007). BEAGLE differs from traditional semantic space models in that it contains two types of representations: item vectors, which encode unordered co-occurrence, and order vectors, in which words are similar to the extent to which they are share neighboring words in the same relative positions. Global similarity among item and order vectors was regressed onto drift rates in the diffusion decision model (DDM: Ratcliff, 1978), which unifies both response times and accuracy. We implemented this model in a hierarchical Bayesian framework across seven datasets with lists composed of unrelated words. Results indicated clear deficits due to global similarity among item vectors, but only a minimal impact of global similarity among the order vectors. We also found evidence for a linear relationship between global similarity and drift rate and did not find any evidence that global similarity differentially affected performance in speed vs. accuracy emphasis conditions. In addition, we found that global semantic similarity could only partially account for the word frequency effect, suggesting that other factors besides semantic similarity may be responsible.

Many models characterize recognition memory as a process of *global matching*, in which memory strength is assessed by determining the similarity between the probe and each representation in memory. Because each non-target representation contributes interference at retrieval, two predictions emerge: (a.) increases in the number of representations in memory should degrade performance because each representation contributes noise to the total memory strength and (b.) increases in the similarity between lure probes and the representations in memory should hurt performance (Clark & Gronlund, 1996). Both predictions are commonly tested using category length designs, which manipulate the number of items from a particular category on a study list. It is generally found that discriminability worsens as category length increases, a finding attributed to a greater number of similar items in memory (e.g., Robinson & Roediger, 1997; Shiffrin, Huber, & Marinelli, 1995). Such *global similarity effects* have been argued to support models where the primary source of interference is from the study list items (i.e., item-noise); as more items from a given category are studied there are more features shared between the probe and the memory traces, which increases the likelihood of falsely recognising lures from studied categories (Criss & Shiffrin, 2004).

Here, we aim to test the predictions of global similarity memory models without manipulation of the lengths of study lists or categories, as such manipulations have been criticised for introducing additional confounds (e.g., Dennis, Lee, & Kinnell, 2008). Instead, our tests are based on measuring variation in global similarity across items using *semantic space models* (see Jones, Willits, & Dennis, 2015, for a review). Semantic space models represent words as points in a high dimensional space; these representations are learned in an unsupervised fashion based on a large corpus of natural language. Similarities between words can be calculated from the vector cosine between two words' vector

☆ We would like to thank Amy Criss and Asli Kiliç for generously sharing their data and Brendan Johns for providing the datasets of similarity and relatedness judgments. This work was supported by an ARC Discovery Early Career Research Award (DE170100106) awarded to Adam Osth. BEAGLE vectors, datasets, and model code can be found on https://osf.io/gtdqf/.

* Corresponding author.
E-mail address: adamosth@gmail.com (A.F. Osth).

https://doi.org/10.1016/j.jml.2019.104071
Received 1 April 2019; Received in revised form 7 November 2019; Accepted 10 November 2019
Available online 28 November 2019

representations, yielding a continuous measure of similarity between −1 and 1; values closer to 1 indicate higher similarity between the words. These representations are useful here because they enable computation of global similarity measures – one can easily compute the similarity between the probe word's semantic representation and each of the study list word's semantic representations.

Across a number of recognition memory datasets, we used the Bound Encoding of the Aggregate Language Environment (BEAGLE) semantic space model (Jones & Mewhort, 2007) to calculate global similarity between the probe word and each of the study list words. These global similarity values were regressed onto drift rates in the diffusion decision model (DDM: Ratcliff, 1978). The DDM relies on the accumulation of evidence to choose between two alternatives; the time it takes for evidence to reach a response boundary determines the response time (RT). The rate at which evidence is accumulated, which is referred to as the *drift rate*, is a latent measure of performance that provides a simultaneous account of accuracy and response time (RT). Thus, our analysis is able to provide a unified picture of the relationship between semantic similarity and recognition memory performance.

We first summarise explanations of similarity effects by different memory models, and discuss how such models have previously been tested in designs manipulating the length of categories. We follow this discussion with descriptions of BEAGLE and the DDM, then report a hierarchical Bayesian analysis of the relationship between BEAGLE representations and performance in memory experiments as characterised by the DDM.

Similarity effects in global matching models

In global matching models, as the similarity between the studied items and the probe increases, global similarity increases and there is an greater likelihood of saying "yes" to the probe item. This can result in increased false alarms to lure items, but can also produce an increase in hits to target items, as the target items receive a boost from the similar, non-target representations in memory.

In category length (CL) designs, the increase in the false alarm rate (FAR) with increasing CL is robust across a wide range of studies (e.g., Cho & Neely, 2013; Criss & Shiffrin, 2004; Dennis & Chapman, 2010; Neely & Tse, 2009; Robinson & Roediger, 1997; Shiffrin et al., 1995). The effects on the hit rate (HR) are somewhat less consistent. Some studies have found increases in the HR with increasing CL (e.g., Cho & Neely, 2013; Maguire, Humphreys, Dennis, & Lee, 2010; Neely & Tse, 2009), which is consistent with the predictions of global matching models. Others have found no effect (e.g., Dennis & Chapman, 2010), even within the same article. For instance, Shiffrin et al. (1995) manipulated category length between 2 and 9 items and found no effect on HR in their Experiment 1, but found an increase in the HR in Experiments 2 and 4.

Although the lack of an improvement in HR with increasing category length superficially seems at odds with global matching models, there are other mechanisms that can mitigate the effect of increasing similarity for target items. These mechanisms imply that the most similar item can dominate the global similarity computation, and since on target trials the most similar item is its own representation the non-target items exert a minimal effect on the global similarity. A mechanism in some models is a non-linear transformation of the similarities, such as a cubing of each similarity value (Hintzman, 1988) or an exponential transformation of distance (Nosofsky, 1988). These transformations exaggerate the differences between the strongest and weakest matches, and since the target item is generally the strongest match it will exert a disproportionately large influence on the global similarity. A related idea is that in some models the similarity between probes and memory traces is based on the likelihood ratio that the memory is the same as the target item (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In such models, the target often has such an extremely high likelihood ratio that the presence of other items has a minimal impact on the resulting global similarity. Both mechanisms have the potential to explain how increases in CL might have no effect on HR for targets. Nonetheless, such models cannot predict decreasing hit rates with increasing similarity without appeal to additional mechanisms.

Although this article focuses on the effects of semantic similarity, interference in item-noise models can stem from perceptual similarity between the probe and the items in memory, even for word stimuli. In fact, when categories are constructed from words that are orthographically or phonologically similar to each other, impairments with increasing category length are also observed (e.g., Criss & Shiffrin, 2004; Heathcote, 2003; Shiffrin et al., 1995; Steyvers, 2000).

Nonetheless, there is another class of models referred to as context-noise models that assume that similarity has no direct effect on memory performance. Context-noise models contend that study-list items contribute no interference at all, and therefore manipulations of the number of items on the study list and the similarity among the items does not affect performance. These models instead claim that interference stems from prior contexts in which a probe had been experienced (Anderson & Bower, 1972; Dennis & Humphreys, 2001). Suppose a word that occurs on a study list (e.g., "dog") has appeared in many contexts prior to the recognition memory experiment. Context-noise models claim that the goal of recognition memory is ascertaining whether the study list context is among the many contexts in which "dog" was experienced. They predict no effects of item interference because they make the counter-intuitive claim that item representations are orthogonal to each other, implying that adding words from a similar semantic category cannot produce interference. An argument in support of context-noise models is the fact that manipulations of list length do not harm performance when a number of confounds are controlled (e.g., Dennis et al., 2008; Kinnell & Dennis, 2011), although list length effects can be found under the circumstances when the buildup of interference across study lists is taken into account (Brandt, Zaiser, & Schnuerch, 2019; Fox, Osth, & Dennis, 2020).

How can context-noise models account for the large impairments of similarity on recognition memory performance? Dennis and Humphreys (2001) appealed to the generation of implicit associative responses (IARs) during the study phase. That is, during study of a list of words, participants may think of other words from the same category and bind them to the study list context, increasing the probability that they will false alarm to such items during the test phase. More recently, context-noise models have made an appeal to a strategic explanation of category length effects in terms of using the categories as cues in conjunction with the probe items (Dennis & Chapman, 2010; Osth & Dennis, 2015). That is, when a longer semantic category is presented, participants may notice that a number of items come from the same category (e.g., "dog", "cat", and "mouse" all belong to the category of "animals"). This may cause a change in strategy during the test phase such that when a participant is presented with a test word such as "fox", instead of asking themselves "Was this word one of the words I studied?" they may be asking "Was this word among the *categories* I studied?"

Such concerns about category cuing are not new. It has been known for some time that presenting all of the items from a category in succession produces larger similarity effects than when the categorized exemplars are distributed randomly throughout the list (D'Agostino, 1971; Dennis & Chapman, 2010; Mather, Henkel, & Johnson, 1997; Neely & Tse, 2009). Although there are some models that are sensitive to the sequential order of items at study (e.g., Howard & Kahana, 2002a), to our knowledge there has been no demonstration within a computational model of how massed categories produce larger similarity effects. One possibility is that participants notice the structure of the categories and alter their decision or cuing strategies during the test phase. This possible influence was noted by Shiffrin et al. (1995) who attempted to control for this confound by distributing all of their categorized items within a very large study list. Another possibility that

we note in the General Discussion is that encoding of items into memory may be affected by the similarity between items on the study list.

Insights from semantic space models

Given the possible strategies that participants can employ in category length designs, such as IAR generation and category cuing, how then can the influence of semantic similarity on recognition memory be better investigated? One potential remedy is to use study lists composed of unrelated words, so there is no categorical structure for participants to infer, and measure variation in global semantic similarity using word representations from semantic space models. Analysis of episodic memory data from experiments using unrelated words with semantic space models was pioneered by Howard and Kahana (2002b) in the domain of free recall by re-analzying archival datasets using word representations derived from latent semantic analysis (LSA: Landauer & Dumais, 1997).

In LSA, word representations are created from a large corpus of natural text by constructing a word-by-document matrix. The vector representation is simply its row in the matrix and similarities between words reflects their co-occurrence in the same documents. Singular value decomposition (SVD) is performed on the matrix and it is reduced to the 300 dimensions with the highest singular values – this has the effect of bringing similar documents closer together. After the dimensionality reduction, words that do not directly co-occur with each other will become similar to each other by virtue of having co-occurred with the same words. For example, two synonyms such as "swordsman" and "fencer" may not co-occur with each other because speakers might use one word or the other but have no need to use both. Nonetheless, both words will occur with the same kinds of words, such as "blade" and "duel."

Vectors from LSA enable similarity computation by calculating the cosine of the angle between pairs of vectors. Howard and Kahana used LSA to analyze participants' recall sequences and found that recall of a word was more likely to be followed by another list item that was semantically similar to the just recalled word, even when the lists were composed entirely of unrelated words with no obvious categorical structure. Specifically, the higher the semantic similarity between two words (as measured by the cosine of their LSA vectors), the more likely it is that the two words will be recalled adjacent to each other in the recall sequence.

In the present work, we adopt a similar approach in the domain of recognition memory by leveraging semantic space models to calculate measures of global semantic similarity. However, instead of using representations derived from LSA, we use the BEAGLE model (Jones & Mewhort, 2007). BEAGLE departs from LSA in that it produces two distinct representations for each word. These include *item representations* which are derived from unordered co-occurrence between words in text. Similarities among item representations bear a strong resemblance to those from LSA, which also learns semantic representations using unordered co-occurrence in text (Jones, Kintsch, & Mewhort, 2006; Jones & Mewhort, 2007).

Where BEAGLE extends beyond LSA is with its set of *order representations*, where words are similar to each other to the extent to which they share the same neighboring words in the same relative positions. Order representations in BEAGLE have been shown to bear a resemblance to other "moving window" type approaches to developing semantic representations, such as the hyperspace analog to language (HAL: Lund & Burgess, 1996) model (Jones et al., 2006). In HAL, a moving window is slid across a text corpus and associations are learned between words that are proportional to the lag between them. While HAL also leverages the co-occurrence between words, it introduces an ordered element to their representations as words will be similar to the extent to which they are adjacent to each other or are adjacent to the same words. Below, we will both demonstrate and discuss how techniques that leverage positional similarity yield different semantic representations than co-occurrence alone.

The vector representations in BEAGLE enable us to calculate the similarity between each probe word and the words on the study list as vector cosines[1] to construct a measure of global similarity between the probe and the contents of the study list. The fact that both measures can be simultaneously obtained on an item-by-item basis is a distinct advantage of the approach, as investigations using category length manipulations have not often been able to control for the similarity *between* semantic categories. For instance, in Shiffrin et al. (1995) there is a prototype word "gambler" that comes from a category of gambling related words. However, this word may also bear some non-negligible similarity to words from a separate category, such as "comedian" and "clown." Through the usage of a semantic space model, we can calculate the similarity between the probe and *all* of the items on the study list.

Most experiments manipulating CL have measured the influence of semantic similarity on accuracy but not RT, potentially ignoring an important constraint. Rather than separately investigating the effects of semantic similarity on accuracy and RTs, we jointly modeled the influence of semantic similarity on both measures by functionally relating semantic similarity to drift rates in the diffusion decision model (DDM: Ratcliff, 1978; Ratcliff & McKoon, 2008). The DDM is a model of binary choice in which decisions are made by a process of noisy accumulation of evidence toward one of two response boundaries. Each response boundary corresponds to a choice alternative (e.g., "YES" and "NO" in a two-choice recognition memory experiment). RT is determined by the time taken to reach a boundary. The rate at which evidence accumulates is called the drift rate, with faster rates of accumulation producing a greater proportion of correct responses and shorter RTs. The DDM has been very successful in accounting for both accuracy and the shapes of RT distributions across correct and error responses in a wide range of conditions in recognition memory experiments (Ratcliff & Smith, 2004; Ratcliff, Thapar, & McKoon, 2004; Starns, Ratcliff, & McKoon, 2012) and has also been successfully integrated with process models of recognition memory (Osth, Jansson, Dennis, & Heathcote, 2018).

Combining BEAGLE representations and the DDM enables us to address several aims:

- First, we are able to test predictions from item-noise and context-noise models. Item-noise models predict that increasing global semantic similarity among BEAGLE's item and order vectors should increase drift rates for lures in recognition memory. While item-noise models are able to predict no effect of increasing global similarity on target items, they are likely unable to predict impairments on target items without additional mechanisms. Context-noise models, in contrast, predict no effects of global similarity a priori. Although they can appeal to the use of a category cuing strategy when participants are able to notice categories, each of our datasets contain lists of unrelated words with no obvious structure, making this much less likely.
- The two types of BEAGLE representations, item and order vectors, enable us to address what types of semantic representations underlie semantic similarity effects in recognition memory. Semantic representations can be similar to each other to the extent that they co-occur with each other or occupy similar positions relative to other words. BEAGLE's two representations more cleanly distinguish between these two types of similarity than prior models and employs the same vector architecture for each.
- Semantic space models provide another advantage to memory models by their ability to explain word frequency effects without

[1] One should note that other similarity metrics, such as dot products or inverse distance, can also be used to calculate similarity between semantic vectors, but the cosine metric is the most commonly employed.

any additional mechanisms. It has been well demonstrated that low frequency (LF) words outperform high frequency (HF) words in recognition memory (e.g., Glanzer & Adams, 1985, 1990; Hemmer & Criss, 2013). In some semantic space models, HF words are more similar to each other than LF words, which is consistent with the explanation of word frequency effects offered by the retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997). Two models using semantic space representations have demonstrated that they were able to account for the word frequency effect using semantic representations alone (Johns, Jones, & Mewhort, 2012; Monaco, Abbott, & Kahana, 2007). However, neither of these models have been fit to data so it remains unclear whether such representations can capture the size of the word frequency effect, especially across a range of different study lists and word sets. We explore the extent to which global similarity effects alone can produce word frequency effects in the subsection entitled "Can the Global Similarity of BEAGLE's Representations Account for Word Frequency Effects?"

- One of the distinct advantages of the DDM is that it enables a principled account of the speed-accuracy tradeoff – a shallower response boundary requires less evidence for responses, enabling faster responses, but comes at the cost of more errors due to the noise in accumulation. Recent evidence has indicated that the drift rate is also reduced in conditions of speed emphasis (Rae, Heathcote, Donkin, Averell, & Brown, 2014), but it remains unclear what impairs drift rates under such conditions. Our approach enables us to investigate whether this is due to differences in global similarity across the speed and accuracy emphasis conditions, which might arise if participants emphasize encoding of different cues early vs. late in the test trial.
- Psychological representations for word stimuli allow us to test different proposed mechanisms for calculating global similarity. For instance, the retrieving effectively from memory (REM: Shiffrin & Steyvers, 1997) model averages the similarities among all of the studied items. In contrast, the subjective likelihood in memory (SLiM) model (McClelland & Chappell, 1998) uses the maximum similarity as the basis of the recognition decision. Although Criss and McClelland (2006) found that these two methods of aggregation produced equivalent levels of performance, their simulations were performed with random vectors for each stimulus; the results are not guaranteed to generalize to cases with more realistic representations (e.g., Johns & Jones, 2010).[2] Indeed, we will demonstrate later that while both mean and maximum similarity are strongly correlated, that correlation is not perfect and each method performed differently in our fits to real data.

Constructing word representations with BEAGLE

BEAGLE generates two representations for each word, an item representation denoted by **m**, and an order representation denoted by **o**. Each of these representations is constructed from text and begins with an environmental representation for each word, which we denote using **e**. Unlike the item and order representations, the environmental representations do not change with experience. Each environmental vector is sampled from a zero-centered Gaussian distribution with standard deviation $1/\sqrt{D}$ where D is the vector dimensionality. Consistent with prior applications of the model, we used vectors of 1024 elements (Jones et al., 2006; Mewhort, Shabahang, & Franklin, 2018; Recchia, Sahlgren, Kanerva, & Jones, 2015).

[2] A reviewer pointed out that the equivalence of the results may be due to the fact that in both models the global similarity computations are dominated by the highest matching trace. This is indeed very likely, however, realistic assumptions about semantic similarity may introduce more discriminating cases, such as having two or more strongly matching traces in memory.

Item representations for a given word are constructed from the environmental representations that co-occur with the word in a sentence. For a word i in sentence k, we can construct a context vector c that is the sum of each of the environmental vectors of the other words j in the sentence:

$$c_{ik} = \sum_{j=1, j\neq i}^{n} e_j \quad (1)$$

A word's item representation m is the sum of all context vectors constructed from the corpus:

$$m_i = \sum_{k=1}^{N} c_{ik} \quad (2)$$

This process is conducted for each word in the corpus. Because each word's item representation is composed entirely of other words that accompanied the given word, two words' item representations will be similar to the extent to which they accompanied similar words in text. To give an example, consider if a single sentence such as "Cats eat mice" was presented to the model. This would result in the following item vectors:

$$m_{cats} = e_{eat} + e_{mice}$$
$$m_{eat} = e_{cats} + e_{mice}$$
$$m_{mice} = e_{cats} + e_{eat}$$

The item vectors for each of the words would bear some similarity to each other by virtue of sharing some of the same environmental vectors in their vectors. For instance, both m_{cats} and m_{mice} share e_{eat} in their sum.

BEAGLE's learning algorithm enables its item representations to extend beyond direct co-occurrence and capture latent semantic relationships. Consider if a sentence such as "My dog chases mice" was also learned. The vector m_{dogs} from this sentence would be $e_{my} + e_{chases} + e_{mice}$. Consequently, the vector m_{dogs} would bear some similarity to m_{cats} by virtue of both words having accompanied the word "mice" even though the two words did not occur together. In short, words will be similar to each other if they are accompanied by the same words even if they do not co-occur. LSA also accomplishes this, but it does so via the SVD.

One kind of information that is absent from BEAGLE's item representations is word order – one could present a corpus where the words within each sentence are completely scrambled and identical item representations would be produced (a similar problem is present in the LSA model). In BEAGLE, a separate type of representation – order representations – are influenced by the order of words within sentences. In the original BEAGLE implementation, order representations are built by binding words together in an ordered fashion using vector convolution. However, an improved method for representing order information comes from the usage of vector permutations (e.g., Sahlgren, Holst, & Kanerva, 2008). Vector permutations can represent the relative position of a word by shuffling the elements of the environmental vectors, where different shuffling operations are defined for each position.

Possibly the simplest way to represent position with permutations is to rotate each vector element by n slots to represent the nth position relative to a target word. When elements reach the slots at the end of a vector, they are rotated to the other end of the vector. To use a concrete example, assume we have an environmental vector of [1, 2, 3, 4, 5] and we want to represent it being two positions before the target word ($n = -2$). A simple way to implement the permutation method is to rotate all elements by two ([3, 4, 5, 1, 2]). We will use the operator Π^n to refer to the permutation of a vector by n slots. While the permutation method proposed by Sahlgren et al. (2008) employed binary vectors of very high dimensionality, Recchia et al. (2015) created a modified version of BEAGLE that used the permutation method for order representations by using the same environmental vectors and found it

outperformed order vectors constructed from the convolution method.

Returning to our example, we can construct an order representation for the word "eat" in the sentence "Cats eat mice" as $o_{eat} = \Pi^{-1}e_{cats} + \Pi^{1}e_{mice}$. BEAGLE's order representations are similar to the extent that they are flanked by the same words in the same positions. To understand this, let's consider if we substituted other verbs into the above example such as "kill" ("cats kill mice") or "like" ("cats like mice"). The order representations of these verbs for these sentences would all be identical to each other by virtue of sharing the same words in the same relative positions. Thus, order representations differ from item representations in that words are similar to the extent to which they occupy similar roles in sentences. We used permutations to encode relative order for up to five positions on either side of the target word.

Our BEAGLE representations were constructed from a corpus of novels, which has been used in previous publications (Johns & Jamieson, 2018; Johns, Jones, & Mewhort, 2019; Mewhort et al., 2018). The corpus contained 39,076 unique words in 10,238,600 sentences. More details on how BEAGLE is trained on text corpora can be found in Recchia et al. (2015), and the BEAGLE vectors can be found on our OSF page (https://osf.io/gtdqf/). In Supplementary Materials A ("Comparisons Between Convolution and Permutation Methods on Human Data"), we compared both the convolution and permutation methods across our novels corpus and the more traditionally used TASA corpus on a set of benchmarks, including the Test of English as a Foreign Language Exam (TOEFL) and five datasets containing similarity ratings from human observers. The permutation method outperformed the convolution method in every case. In addition, the vectors from the novels corpus outperformed those from the TASA corpus. For this reason, the remainder of the text uses order vectors derived from the permutation method applied to the novels corpus.

What kinds of meaning are captured by item and order representations?

We have described that BEAGLE has two different means of constructing semantic representations. But what kinds of meaning are captured by each of its representations? Table 1 shows the *nearest neighbors* of a set of words on both the item and order spaces. Nearest neighbors are constructed by calculating the vector cosine between the target word and all of the words in the lexicon and then selecting *n* words with the highest cosine values. On first inspection of the table, one may notice that some unusual choices of nearest neighbors from the model, as it's not immediately clear how the word "smart" is similar to the words "kid" or "strong." These occur for a few reasons. First, semantic representations in models such as BEAGLE and LSA depend on both the size and kind of the corpus on which they are trained. As mentioned previously, item representations are learned from co-occurrence, so if "smart kid" appears frequently in the corpus there will be a high similarity among the item vectors. Likewise, order representations are similar to the extent to which they are flanked by the same words. The high similarity of the order vectors for "smart" and "strong" might arise from a high frequency of phrases such as "she's a strong woman" and "she's a smart woman."

The different types of information underlying each space provides some insight into the nearest neighbors that are produced. While the item space produces a high frequency of synonymous neighbors, there are also instances of words that are associated with the target word but not similar in meaning. For instance, "boat" is similar to "dock" and "musician" is similar to "jazz." In addition, sometimes the nearest neighbors come from a different grammatical class, such as "sword" being similar to "thrust."

The order vectors contrast sharply with the item vectors because there is a strong respect for grammatical class among its nearest neighbors. Jones and Mewhort (2007) demonstrated how, among BEAGLE's order vectors, nouns are similar to other nouns, verbs are similar to other verbs, and adjectives are similar to other adjectives. This is likely because of the constraints imposed by positional learning. In the sentence "cats eat mice," the word "eat" is flanked by two nouns. It is extremely unlikely for a noun or an adjective to be both followed and preceded by nouns, but this can be much more common for verbs. In addition, there is some respect for similar kinds in the nearest neighbors. "Boat" is similar not just to other sea-related words, but it is also similar to other vehicles such as "ship", "plane", and "bus" and the word "cat" is similar to other animals. This is, however, not as strong of a regularity as with grammatical class. For instance, the word "sing" is similar to other verbs of artistic expression such as "write" and "perform", but some less obvious verbs appear in its nearest neighbors such as "eat", "survive", and "feed." The HAL model also demonstrates similar clustering by grammatical class and has been demonstrated to produce similar structure as those from simple recurrent networks (Burgess & Lund, 2000). This is sensible when one considers that the similarity relationships among the hidden units in a simple recurrent network also cluster very strongly by grammatical class (Elman, 1990).

Table 1
Eight nearest neighbors to a set of words on BEAGLE's item and order spaces.

Word	Space	Neighbors
boat	Item	boats, shore, ship, raft, dock, lake, swim, island
	Order	ship, plane, bus, raft, van, truck, bridge, river
bicycle	Item	bike, horseback, car, bicycles, motorcycle, vehicle, riding, driving
	Order	motorcycle, donkey, horse, flashlight, clipboard, broomstick, camel, crutch
cat	Item	dog, mouse, big, saw, little, huge, instead, back
	Order	dog, bird, wolf, lion, snake, spider, tiger, dragon
sword	Item	dagger, weapon, blade, rapier, knife, scimitar, thrust, axe
	Order	rifle, horse, pipe, spear, pistol, gun, knife, guitar
musician	Item	music, musicians, jazz, band, thought, kind, always, singer
	Order	politician, scientist, businessman, jew, painter, writer, scholar, fighter
fight	Item	fighting, battle, anyway, kill, whatever, lose, unless, start
	Order	swim, play, visit, break, fly, start, hunt, cook
sing	Item	song, sang, singing, sung, songs, lyrics, sings, melody
	Order	play, burn, eat, write, perform, survive, draw, feed
smart	Item	nice, kind, probably, stupid, well, kid, actually, clever
	Order	clever, strong, brave, stupid, dumb, foolish, tough, stubborn
janet	Item	michael, sarah, peter, someone, sam, alex, told, anna
	Order	sarah, amy, paul, jack, peter, david, charlie, alex
doug	Item	sam, peter, tom, paul, alex, andy, after, matt
	Order	paul, michael, jack, peter, alex, mike, sam, charlie

Table 2
Summary of the datasets fit by the model.

Dataset	N	Obs.	LL	Manipulation
Rae et al. (2014)	47	756.96	28	Speed-accuracy emphasis, WF (LF/HF, within-list)
Osth et al. (2017)	35	1076.80	28	Speed-accuracy emphasis, WF (LF/HF, within-list)
Criss (2010, E2)	16	1412.88	50	Presentations (1×/5×, cross-list), WF (LF/HF, cross-list)
Kiliç et al. (2017, E1)	30	1736.76	150	Encoding task (deep vs. shallow, cross-list)
Kiliç et al. (2017, E2)	26	1737.11	150	Encoding task (deep vs. shallow, mixed list)
Kiliç et al. (2017, E3)	24	1770.00	150	Encoding task (deep vs. shallow, mixed list)
Kiliç et al. (2017, E4)	39	857.80	150	Encoding task (deep vs. shallow, mixed list)

Notes: E = experiment, N = number of participants, Obs. = mean number of observations per participant, LL = study list length. WF = word frequency, LF = low frequency, HF = high frequency.

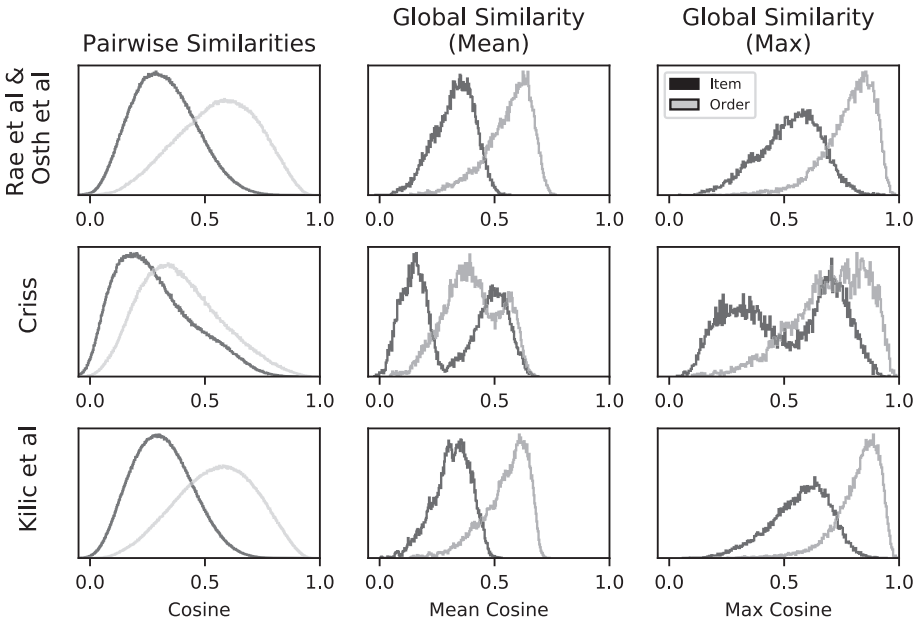

Fig. 1. Standardized similarities from the BEAGLE model, including pairwise similarities (left column), mean similarities between probe words and study list items (middle column), and maximum similarities between the probe word and study list words (right column) for each word set.

Global semantic similarity in recognition memory

Using BEAGLE's item and order representations, we can calculate a probe word's global semantic similarity and use it to predict recognition memory performance on an item-by-item basis. Global semantic similarity can be constructed using either the mean cosine similarity between the probe word and each study list word or from the maximum cosine similarity value. An additional possibility that is often employed is summed similarity (e.g., Gillund & Shiffrin, 1984; Nosofsky, Little, Donkin, & Fific, 2011). However, summed similarity produces identical predictions to mean similarity if the global similarity is rescaled.

Table 3
Correlations between each of BEAGLE's global similarity measures and log word frequency from SUBTLEX for each word set.

Dataset	Measure	1	2	3	4	5
Rae et al. & Osth et al.	1. Mean Item Similarity					
	2. Mean Order Similarity	0.2				
	3. Max Item Similarity	0.9	0.14			
	4. Max Order Similarity	0.32	0.77	0.28		
	5. Log WF	0.68	0.04	0.66	0.16	
Criss E2	1. Mean Item Similarity					
	2. Mean Order Similarity	0.62				
	3. Max Item Similarity	0.95	0.6			
	4. Max Order Similarity	0.63	0.78	0.63		
	5. Log WF	0.88	0.59	0.83	0.59	
Kiliç et al.	1. Mean Item Similarity					
	2. Mean Order Similarity	0.23				
	3. Max Item Similarity	0.85	0.19			
	4. Max Order Similarity	0.41	0.68	0.4		
	5. Log WF	0.66	0.18	0.63	0.31	

For targets, we omitted the self-similarity since the cosine between a word vector and its own representation is always 1.0. Thus, on target trials, for a list of length L the global similarity is calculated among the $L-1$ non-target words. Global similarity values for targets can be considered as a measure which evaluates the extent to which similarity to non-target representations influences memory. This is also analogous to investigating the effects of category length manipulations on the hit rate – the target is always in memory even when there are no categories (CL = 0).

Example pairwise similarities and global similarities calculated from each of the datasets we analyze in this article (details of each dataset can be seen in Table 2) can be seen in Fig. 1. We selected our datasets on the grounds that they (a.) recored the actual words that were studied and tested, (b.) used lists of unrelated words (i.e., had no semantic categories or category length manipulations), and (c.) had sufficient numbers of observations per participant to constrain the DDM parameters. Two of the datasets, namely the data of Rae et al. (2014) and Osth, Bora, Dennis, and Heathcote (2017), contained a manipulation of speed-accuracy emphasis, which enables us to test the hypothesis that

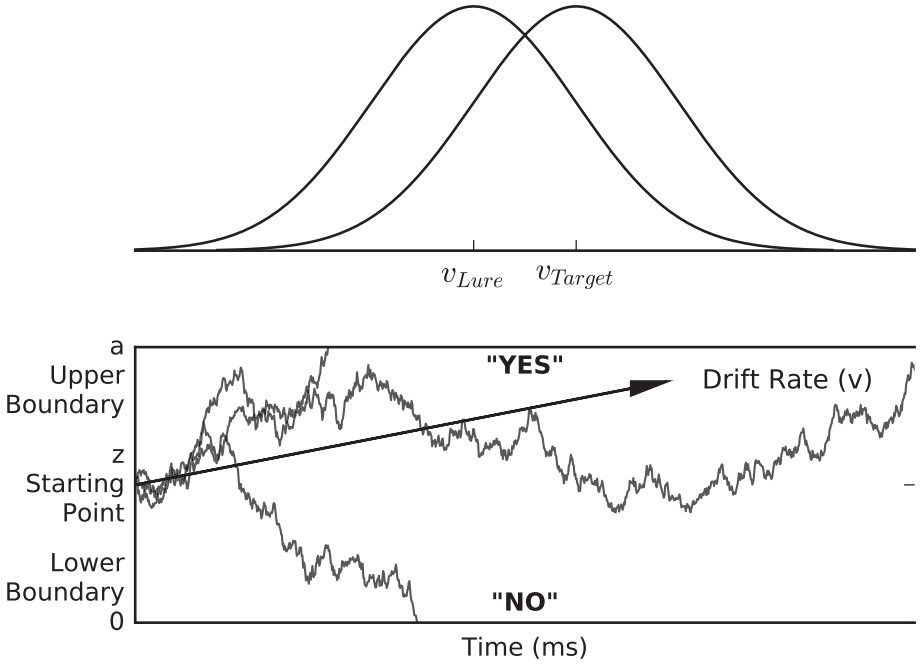

Fig. 2. Basic description of the diffusion decision model (DDM). The top panel shows example drift rate distributions for targets and lures. The bottom panel shows three example evidence accumulation trajectories with the same sampled drift rate, with the drift rate being denoted by the angle of the arrow. See the text for more details.

global similarity effects might change during the course of retrieval. Because the Rae et al. and Osth et al. datasets both used the same word sets, these results are shown together. Each of the datasets we employ can be found online on our OSF page (https://osf.io/gtdqf/), where the dataset files additionally contain trial-by-trial measures of each of the global similarity measures discussed.

What is perhaps most obvious from Fig. 1 is that order similarities are higher than item similarities. For this reason, we standardized the global similarity scales using z-scores to place item and order similarities on the same scale in the remainder of the article. Although pairwise similarity values show a slight right-skew in their distribution (Johns & Jones, 2010), global similarity values actually show the opposite pattern. The central limit theorem guarantees that global similarity should approach normality as the length of a study list is increased. Given that, it is rather surprising that in the datasets of Kiliç et al., where study lists were 150 items in length, still showed the left skew in the global similarity distributions. The bimodality in the global item similarity in the Criss dataset not only reflects the strong separation between high and low frequency words in that dataset, but also how study lists were composed entirely of a single word frequency category (no mixing of HF and LF words within a study list). These factors can produce bimodality because similarity in co-occurrence models is correlated with word frequency (Landauer & Dumais, 1997; Monaco et al., 2007).

Correlations between BEAGLE's global similarity measures for each of the word sets along with the log word frequency (WF) from SUBTLEX (Brysbaert & New, 2009) can be seen in Table 3. Correlations between global item and order similarity range range from low ($r \sim 0.2$ in the Rae et al. and Kiliç et al. experiments) to moderate ($r \sim 0.6$ in the Criss dataset). In addition, correlations between global item similarity and log word frequency are rather strong, ranging from 0.69 to 0.9. Correlations between global similarity on the order space and word frequency are weak in two of the word sets (Rae et al./Osth et al. & Kiliç et al.) and are somewhat stronger in the Criss dataset (0.59). Later in the article, we investigate the extent to which global semantic similarity alone can account for performance differences between HF and LF words in the section "Can the Global Similarity of BEAGLE's Representations Account for Word Frequency Effects?" Finally, it is perhaps unsurprising that there are large correlations between mean and max similarity in each dataset, although the correlations are not perfect.

One should be careful to note that, although BEAGLE's representations allow for the calculation of global semantic similarity, this remains an incomplete picture of global similarity in recognition memory. In addition to semantic features, representations formed during presentations of the list items likely also include phonological and orthographic features in addition to features that represent the temporal context of each item's occurrence. Although in principle it would be possible to incorporate measures of each of these dimensions, doing so goes considerably beyond the scope of the present work.

Relating global semantic similarity to recognition performance via the Diffusion Decision Model (DDM)

We related global similarity to both choice responses and RTs using the diffusion decision model (DDM). In the DDM (a diagram can be found in Fig. 2), evidence begins to accumulate in a noisy fashion at the starting point z toward one of two response boundaries: an upper boundary at a corresponding to a "YES" response, and a lower boundary at zero corresponding to a "NO" response. The time taken for the process to reach a boundary is the RT plus additional non-decision time processes denoted by t_0, which correspond to the time taken to encode the probe cue and produce the response. The drift rate v corresponds to memory strength[3] and determines the rate of evidence accumulation; as drift rate increases, evidence accumulates faster and is more likely to reach the "YES" boundary, increasing the proportion of "YES" responses. Because evidence accumulation is noisy, a positive drift rate corresponding to a target item can still erroneously terminate at the "NO" boundary, leading to an error. Manipulations that affect memory performance such as study time, word frequency, or study-test delay, exert their largest influences on the drift rate parameter. For this reason, the drift rate parameter is analogous to memory strength in signal detection theory (SDT).

The starting point z is a bias parameter; increases in z increase the likelihood of "YES" responses even if drift rates are negative and tend toward the lower boundary. If the response boundary is increased, it is

[3] The term "memory strength" has been used in a variety of different senses throughout the literature. We use the term agnostically here to represent the decision variable itself.

less likely that the process will erroneously terminate at an incorrect boundary, but it also increases the time it takes for accumulation to reach a boundary, which is the basis of the speed-accuracy tradeoff in the model. The a parameter corresponds to response caution and is varied across conditions that differ in speed-accuracy emphasis. Because both z and a can be set by the participant according to the perceived difficulty of the test list, they can vary across different list conditions (e.g., cross-list manipulations), but are conventionally fixed across different conditions that vary unpredictably from trial to trial. For instance, if high and low frequency words were intermixed on a test list, the same value of z and a should be used for both classes because the participant would have to know they were being tested on a high or low frequency word to alter their starting point or response boundary.

Drift rate and boundary separation can also be dissociated by their differential effects on RT distributions. While boundary separation and bias have profound effects on the fastest RTs in the RT distribution (the "leading edge"), drift rate exerts its most pronounced effects on the slowest RTs in the upper tail (Ratcliff & McKoon, 2008). This implies that conditions that vary only in terms of their drift rates will exhibit relatively constant leading edges across conditions while their slowest RTs will show the most pronounced differences, with longer upper tails of the RT distributions found in more poorly performing conditions.

Finally, the DDM includes cross-trial variability in drift rate (which is normally distributed with standard deviation η), starting point (uniformly distributed with range s_z), and non-decision time (uniformly distributed with range s_t). Cross-trial variability in drift rate is analogous to variability in memory strength in SDT and enables the model to produce errors that are slower than correct responses, while variability in starting point enables the model to produce errors that are faster than correct responses. In recognition memory, errors are typically slower than correct responses and a fast error pattern has not usually been found (Ratcliff & Smith, 2004) and thus prior work suggests that starting point variability is unnecessary in recognition memory (Osth et al., 2017). Thus, in our applications we fix $s_z = 0$ in all cases. In addition, it has also been found that drift rate variability for targets is larger than for lures (e.g., Osth, Dennis, & Heathcote, 2017; Starns, 2014; Starns, Ratcliff, & McKoon, et al., 2012; Starns & Ratcliff, 2014), which converges with analyses from receiver operating characteristics (ROCs: Heathcote, 2003; Ratcliff, Sheu, & Gronlund, 1992; Wixted, 2007). Thus, in each of our models we allowed drift rate variability η to vary across targets and lures.

Prior investigations in recognition memory with the DDM have often allowed the mean drift rate v to vary over conditions affecting performance but have not generally allowed for systematic differences in mean drift rates among items in the experiment (but see Cox, Hemmer, Aue, & Criss, 2018). In our investigations, in contrast, we allow the global similarity of each probe word's item and order representations to the study list items to influence the mean drift rate v on an item-by-item basis, and hence a trial-by-trial basis as items are only tested once. Specifically, the mean drift rate v for subject i in condition j on trial k is:

$$v_{ijk} = V_{ij} + \alpha_{ij} I_{ijk} + \beta_{ij} O_{ijk} \quad (3)$$

where V is an intercept parameter that can be construed as a mean drift rate for an experimental condition, I is the global similarity between the probe on trial k and each study list item's item representations, O is the global similarity between the probe on trial k and each study list item's order representations, and α and β are weights on the global similarities for item and order vectors, respectively. In our investigation, the mean drift rate intercept V varies across experimental conditions such as number of presentations, depth of processing, speed-accuracy emphasis, and word frequency. The α and β parameters only vary across targets and lures. The additive combination of I and O is consistent with prior applications of BEAGLE, where lexical vectors are formed by additively combining the item and order vectors (Jones et al., 2006; Jones & Mewhort, 2007; Mewhort et al., 2018). We assumed global similarity only affects the mean drift rate parameter v and not accumulator-related parameters, such as the starting point z and response boundary a, because global similarity varies considerably across the words on a test list, making it extremely unlikely that participants could use each word's global similarity to alter their bias or level of response caution.

Note that the systematic cross-trial variability that stems from the inclusion of global similarity is not the only source of drift rate variability in the model. In addition, we also estimate the random drift rate variability parameter η. Random drift rate variability was included for two reasons. First, other sources of variability outside of global similarity are likely to affect performance, such as variability in encoding strength and the phonological and orthographic similarities between the probe and test items – neither factor is present in the current models. In addition, our approach is consistent with previous integrated diffusion models, which generally combine systematic and random variability (e.g., Ratcliff & McKoon, 2018; Ratcliff & Rouder, 1998). Later, in the sub-section entitled "Drift Rate Variability Estimates" we examine how the inclusion of systematic drift rate variability affects the magnitude of η.

An example illustration of a five item study list along with item and order similarities between two probe words ("hat" and "karma") can be seen in the upper part of Fig. 3. The segment below the study list illustrates how drift rates for "hat" and "karma" are calculated according to Eq. 3 for both mean and max similarity rules.

An alternative to our DDM approach would be to regress semantic similarity scores directly onto manifest variables such as responses and RTs across items (e.g., Hutchison, Balota, Cortese, & Watson, 2008; Mandera, Keuleers, & Brysbaert, 2017). Our approach offers two distinct advantages. First, response probabilities and RTs for each item are unified into a single latent variable in the DDM, namely drift rates. In addition, both RTs and choice probabilities can vary according to a number of factors such as bias or speed-accuracy emphasis, which would contribute noise to a regression analysis on manifest variables. The DDM enables the disentangling of these influences by describing effects on both RT and accuracy in terms of drift rates, starting points (characterising bias), and thresholds (characterising speed-accuracy). This in turn enables our measures of global similarity to only influence the parameters corresponding to memory strength, namely the drift rate parameter.

The DDM we are applying is *not* a complete process model of recognition memory. Process models such as Minerva 2 and REM describe processes of encoding and sometimes assume more sophisticated match computations, such as the likelihood ratio of a trace in memory being identical to a probe word (e.g., Glanzer, Adams, Iverson, & Kim, 1993; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Our model is mute with respect to a number of non-semantic factors that are manipulated in our datasets, and which a process model can usually capture with an encoding strength parameter, such as depth of processing and number of repetitions. In our model, these performance differences are captured by differences in the mean drift rate intercept V across such conditions. Similar to the traditional DDM approach or SDT, our model is agnostic as to how these differences in performance arise. Nonetheless, an advantage of our approach is that our model allows us to measure the direction and magnitude of the relationship between global similarity and performance. A process model, in contrast, imposes a qualitative relationship *a priori* – similarity between the probe and the representations in memory should increase global similarity and make "yes" responses more likely.

Applying the model to data produces estimates of α and β that reflect the influence of global item and order similarity. In each of our models, we allow α and β to vary across targets and lures because, although previous work has found robust effects of semantic similarity on false alarm rates (FAR), results on the hit rates (HR) are mixed, with some studies finding increases in the HR with increasing category length (e.g., Cho & Neely, 2013; Maguire et al., 2010; Neely & Tse,

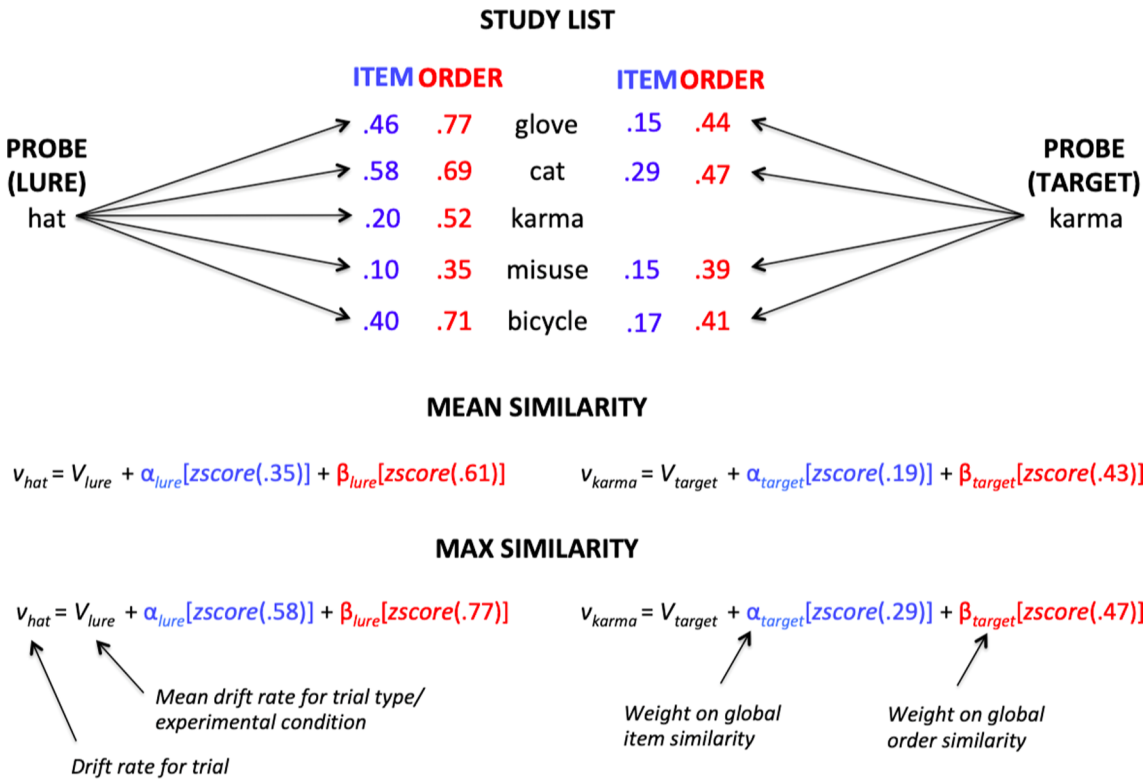

Fig. 3. Illustration of global similarity calculation with a five-word study list (top center) and two probe words "hat" (top left) and "karma" (top right). Item and order similarities are calculated between the probe word and each of the study list items. For the word "karma", the self-similarity is omitted because the cosine between a vector and itself is always 1.0. The bottom half of the figure shows how drift rates for each word are calculated using either the mean or max similarity for the two probe words.

2009) and some finding no effect (e.g., Shiffrin et al., 1995). Positive values of α and β reflect higher drift rates with increasing global similarity, resulting in an increasing tendency to say "YES."

The model fit

We fit the model to a total of seven datasets to measure the influence of global item (α) and order (β) similarity on mean drift rates in the DDM for both targets and lures. A summary of the datasets can be seen in Table 2. In addition to spanning a range of study list lengths, the experiments also contain manipulations such as the number of presentations (Criss, 2010), encoding tasks (Kiliç, Criss, Malmberg, & Shiffrin, 2017), and speed-accuracy emphasis (Rae et al., 2014; Osth et al., 2017). Within-list manipulations (either within study or test lists) mix levels within a list, such as having high frequency (HF) and low frequency (LF) words randomly ordered within both the study and test lists. Cross-list manipulations vary levels across different study-test cycles. In the DDM, parameters for bias (z) and response boundary (a) are generally fixed for within-list manipulations, but can vary for cross-list manipulations, as such parameters are theorized to be based on the instructions and metacognitive factors such as the perceived difficulty of the test list. Data and participant exclusions along with a complete description of which DDM parameters were allowed to vary across these manipulations can be found in Appendix A.

Models were applied to data using hierarchical Bayesian methods (see Lee, 2011; Rouder & Lu, 2005, for introductions), which simultaneously estimate both group and participant level parameters. An advantage of such methods is that the group level distribution allows for "pooling" across individuals while still avoiding the distortions that arise from fitting to group-averaged data. Throughout the article, we will use the μ superscript to refer to the mean of a group level distribution for a particular model parameter (α^{μ} and β^{μ} refer to the group means of α and β). Approximately non-informative prior distributions were used for the group level parameters to impose minimal constraints on their estimated values. Model parameters were estimated using differential evolution Markov chain Monte Carlo (DE-MCMC: Turner, Sederberg, Brown, & Steyvers, 2013), a method of posterior sampling that is robust to parameter correlations. Details on prior distributions and MCMC methods can be found in Appendix B.

In the following sub-section, we describe model selection results, where we compare simple models that lack one or more similarity components to more complex models that include each of the similarity components, and additionally compare mean and max similarity computations. In the sub-section "Parameter Estimates" we analyze and compare the resulting estimates of α and β across datasets and place these values in context by comparing them to drift-rate intercept estimates for manipulations in our experiments, such as word frequency and depth of processing. The sub-section "Hit Rates, False Alarm Rates, and Response Times for Varying Levels of Global Similarity" demonstrates how measures of HR, FAR, and RT vary over the measures of global similarity and additionally show the fit of the DDM to the data. Finally, the sub-section "Drift Rate Variability Estimates" compares the estimates of systematic drift rate variability derived from the BEAGLE representations to the random drift rate variability estimated by the η parameter.

Model selection

Incorporation of semantic similarities from BEAGLE with weights as free parameters necessarily introduces complexity into the DDM. To evaluate whether such an inclusion is warrented, we compared different nested models using the deviance information criterion (DIC: Spiegelhalter, Best, Carlin, & van der Linde, 2002), a metric which evaluates the tradeoff between a model's goodness of fit and its complexity. DIC is similar to other information criteria such as AIC and BIC in that it is based on log likelihoods, so differences that are greater than

Table 4
Δ DIC values and DIC weights for each model variant in each dataset. The DIC weights are depicted in parentheses and the winning model for a given dataset is highlighted in bold.

Dataset	No Sim.	Mean Similarity			Max Similarity		
		Item	Order	Both	Item	Order	Both
Rae et al.	104 (0)	**0 (0.88)**	101 (0)	14 (0)	4 (0.12)	105 (0)	11 (0)
Osth et al.	95 (0)	**0 (0.93)**	108 (0)	12 (0)	6 (0.05)	105 (0)	8 (0.02)
Criss	241 (0)	29 (0)	220 (0)	**0 (1.0)**	42 (0)	254 (0)	44 (0)
Kilic E1	343 (0)	21 (0)	348 (0)	**0 (1.0)**	65 (0)	342 (0)	67 (0)
Kilic E2	251 (0)	2 (0.16)	263 (0)	1 (0.26)	**0 (0.43)**	256 (0)	2 (0.16)
Kilic E3	180 (0)	16 (0)	186 (0)	16 (0)	1 (0.38)	180 (0)	**0 (0.62)**
Kilic E4	85 (0)	**0 (0.92)**	89 (0)	5 (0.08)	11 (0)	81 (0)	21 (0)

Notes: Sim. = similarity.

ten are conventionally considered "large." In addition, we also report DIC weights, in which a ratio of transformed DIC values divided by their sum (Wagenmakers & Farrell, 2004). Under certain assumptions, DIC weights can be interpreted as the probability that a given model is the data generating model for a given dataset among the set of models under consideration.

We compared nested models where there was (a.) no similarity contribution ($\alpha = 0, \beta = 0$), (b.) only similarities among item vectors ($\beta = 0$), (c.) only similarities among order vectors ($\alpha = 0$), and (d.) incorporated both item and order vector similarities. In each of the models where α or β were estimated, each parameter was allowed to vary across targets and lures.

A second factor in the model comparison was whether global similarity was measured as the average similarity between the probe and the list items or the maximum similarity. These two factors – in addition to models lacking any semantic similarity component – collectively resulted in seven models being applied to each dataset.

Model selection results for each dataset and model can be seen in Table 4. Results are depicted using Δ DIC, which is the difference between the model's DIC and the winning model's DIC (which receives a score of zero). Two trends emerge from the table. The first is that models that include semantic similarity exhibit a clear and large advantage over the models that lack any relations between global similarity and drift rate ($\alpha = 0, \beta = 0$). This suggests that BEAGLE's representations are capturing meaningful variability in the data that justify the additional complexity incurred by estimating additional weight parameters.

Second, mean global similarity outperforms maximum similarity for five of the seven datasets despite the high correlation between the two measures. In Kiliç et al.'s Experiments 2 and 3 a max model was preferred, but the advantage in one of these datasets (Experiment 2) was relatively small. In the cases where the mean similarity model win, in contrast, the DIC advantages were larger, with marginal DIC weights are between 0.88 and 1.0. For this reason, the remainder of the article focuses on mean similarity as the measure of global similarity computation. However, we would like to emphasize that these results are likely conditional on the similarity operation we employed (vector cosines). Other measures, such as the likelihood ratio that a trace matches the probe (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), produce highly skewed similarity values in contrast to the vector cosine.

An inconsistency among the datasets concerns the inclusion of whether global similarity of the order vectors, as indicated by the β parameter, is necessary, as five of the seven datasets preferred only the inclusion of the global item similarity parameter α. The poor performance of global similarity among the order vectors is evident by the fact that models that include only the β parameter show very large model selection penalties. The coming sections provide insight into this discrepancy by analyzing both the resulting parameter estimates and evaluating how global similarity influences performance measures in each of the datasets. The sub-section "Drift Rate Variability Estimates" compares estimates of η (standard deviation of the drift rate distribution) to the standard deviation of the systematic cross-trial variability estimates derived from BEAGLE.

Model selection results for order vectors derived using the convolution method can be found in Supplementary Materials B ("Global Similarity with Convolution Order Vectors"). These results are largely congruent with the model selection results in Table 4, suggesting that our results generalize across different methods of encoding order information in BEAGLE.

Parameter estimates: influences of global item and order similarity in each dataset

We investigated the relative magnitudes of global item and order similarity by comparing the means and 95% highest density intervals (HDIs) of α^{μ} and β^{μ} for each dataset, which can be seen in Fig. 4. Positive estimates of α and β indicate an increasing tendency to give "YES" responses as global similarity increases, negative estimates indicate the opposite, while estimates of zero indicate no influence.

For the global similarity among BEAGLE's item vectors parameter α, effects are most pronounced for lures, where it can be seen that α^{μ} is positive in every dataset, indicating an increase in memory strength for lures and an increased likelihood of false alarms as lures' global item similarity increases. In contrast, for lures the global order similarity parameter β^{μ}, is close to zero in almost every dataset, indicating that increases in global order similarity have very little influence on false alarms.

For target stimuli, a different picture emerges. Recall that for targets, the self-match was omitted from the global similarity calculation. For this reason, global similarity values for targets reflects the extent to which the match between the probe and *other* items on the list influences performance on target items. For targets, α^{μ} is close to zero for several of the datasets from Kiliç et al., but shows pronounced negative estimates for the datasets from Rae et al., Osth et al., as well as the Criss dataset. These negative estimates imply that increases in global item similarity *decreases* memory strength for targets, decreasing the likelihood of hits. For the β^{μ} parameter, the opposite is the case, with positive estimates of β present for the Criss dataset as well as the datasets from Kiliç et al., indicating that increases in global order similarity increase the likelihood of hits, although the β^{μ} estimates for targets are again quite close to zero.

To place these parameter estimates in context, Table 5 shows estimates of the drift rate intercept, V^{μ}, for each condition in every dataset. For the data of Rae et al., the word frequency effect for lures can be measured as the marginal difference in V^{μ} between HF and LF words, which was around 0.3. The α^{μ} parameter for lures was 0.22, meaning that one and a half standard deviations of global item similarity was roughly equivalent to the difference in memory strength between HF and LF words. In the Criss dataset, items were presented either once (weak lists) or five times (strong lists), and FAR were considerably lower in the strong list condition (mean FAR in the weak list = 0.16,

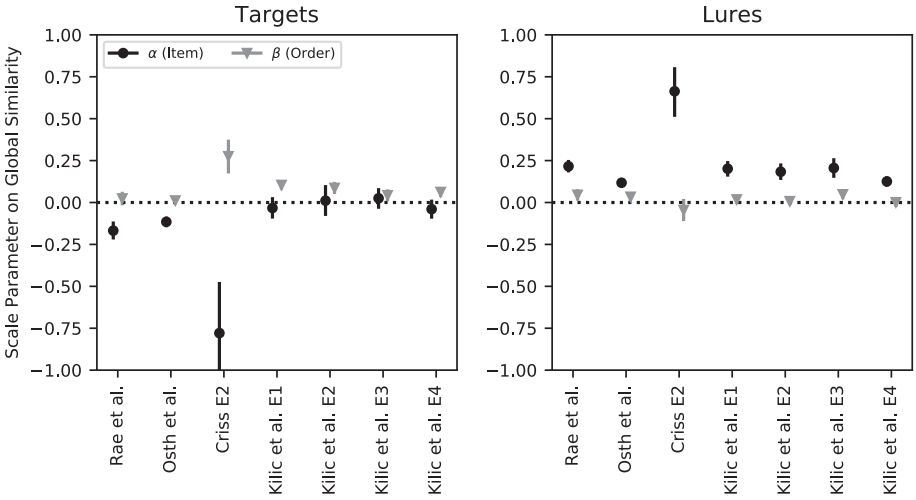

Fig. 4. Estimates of α^{μ} (weight on global item similarity) and β^{μ} (weight on global order similarity) along with the 95% HDIs for each dataset.

Table 5
Estimates of V^{μ} (mean drift rate intercept) along with the 95% highest density interval (HDI) in brackets for each condition in each dataset.

Dataset	Cond.	Targets	Lures
Rae et al.	HF Speed	1.78 [1.54, 2.02]	−1.42 [−1.7, −1.14]
	LF Speed	2.15 [1.9, 2.4]	−1.69 [−2.03, −1.37]
	HF Accuracy	1.66 [1.46, 1.86]	−1.97 [−2.22, −1.71]
	LF Accuracy	2.32 [2.07, 2.57]	−2.19 [−2.44, −1.94]
Osth et al.	HF Speed	1.03 [0.83, 1.23]	−0.7 [−0.94, −0.46]
	LF Speed	1.21 [0.95, 1.47]	−0.81 [−1.08, −0.54]
	HF Accuracy	1.07 [0.88, 1.27]	−1.22 [−1.41, −1.04]
	LF Accuracy	1.57 [1.32, 1.81]	−1.35 [−1.55, −1.15]
Criss	HF Weak	1.27 [0.71, 1.84]	−2.41 [−2.86, −1.95]
	LF Weak	1.17 [0.64, 1.7]	−1.74 [−2.18, −1.3]
	HF Strong	2.65 [1.89, 3.41]	−3.11 [−3.6, −2.61]
	LF Strong	2.7 [1.95, 3.45]	−2.84 [−3.36, −2.34]
Kilic et al. E1	Weak	0.27 [−0.0, 0.54]	−0.87 [−1.07, −0.67]
	Strong	1.52 [1.04, 1.99]	−1.56 [−1.82, −1.29]
Kilic et al. E2	Weak	0.16 [−0.13, 0.45]	−1.37 [−1.63, −1.11]
	Strong	1.65 [1.28, 2.02]	
Kilic et al. E3	Weak	0.17 [−0.13, 0.47]	−1.07 [−1.33, −0.83]
	Strong	1.55 [1.05, 2.04]	−1.19 [−1.47, −0.92]
Kilic et al. E4	Weak	0.01 [−0.21, 0.23]	−1.04 [−1.22, −0.86]
	Strong	1.12 [0.83, 1.4]	−1.25 [−1.48, −1.02]

mean FAR in the strong list = 0.10). The marginal effect of list strength on V^{μ} was around 0.9, while the α^{μ} parameter for lures was 0.66, meaning that one and a half standard deviations of global item similarity is roughly equivalent to the difference in memory strength between a list of once presented items and five times presented items.

In Supplementary Materials B, we demonstrate very similar parameters when order vectors were constructed from the convolution method instead of the permutation method that is used in the main text. However, weight parameters were highly sensitive to the corpus used to train the BEAGLE vectors. In Supplementary Materials C ("Global Similarity with the TASA Corpus"), we found that scale parameters were much weaker when BEAGLE vectors were trained using the TASA corpus. We believe that using vectors from the TASA corpus is undesirable for two reasons. First, they performed more poorly on our benchmarks than vectors derived from the novels corpus (see Supplementary Materials A). Second, a large proportion of the words in our experiments were not represented in the TASA corpus, which resulted in the omission of considerable portions of the recognition memory data.

Non-linear relationships between global similarity and drift rate

Eq. 3 assumes a linear relationship between global similarity and drift rate. One possible reason why global order similarity often showed a weak relationship to memory strength may be because the underlying relationship is non-monotonic, which can lead to a relatively flat linear approximation. Some evidence that the relationship may be non-monotonic comes from Neely and Tse (2009) who found non-linear relationships between category length (CL) manipulations of taxonomic categories and recognition memory performance. In both a literature review and several experiments of their own, the authors reported that recognition memory performance actually improves when CL is increased from 1 to 4 items, decreases from 4 to 7 items to roughly the level of the CL 1 condition, and subsequently declines further.

The investigation of Neely and Tse (2009) prompted us to explore an additional model that allows for quadratic relationships between global similarity and drift rate. The models and model selection results are discussed in Supplementary Materials D. However, the quadratic model was decisively rejected in six of the seven datasets of our model selection procedure, with one dataset showing ambiguous support between the linear or quadratic function (DIC weight of ~0.5 for both models). These results suggest that any benefit produced by the quadratic function was small compared to its increase in model complexity.

Hit rates, false alarm rates, and response times for varying levels of global similarity

Although the parameter estimates in the previous section describe how drift rates for targets and lures change with increases in global similarity, it is helpful to understand how this translates into measures of performance such as choice probabilities and response times. To depict how performance varies with changes in global similarity, we performed median splits on global item similarity and global order similarity for each dataset and then subsequently split the data into four bins: (a.) high global similarity on both item and order dimensions, (b.) high item similarity and low order similarity, (c.) high order similarity and low item similarity, and (d.) low global similarity on both item and order dimensions. Because global similarity was generally higher for high frequency (HF) words than low frequency words (LF), this analysis was performed separately for HF and LF words in each of the datasets that contained a word frequency manipulation (Rae et al., Osth et al., and Criss datasets). We collapsed across all other manipulations (item strength, speed-accuracy emphasis, etc.). One should note that due to correlations between the item and order similarities, similarity is not equated across the bins (e.g., the mean item similarity in the high item, high order similarity bin is often different than the mean item similarity in the high item, low order similarity bin).

In each bin, we calculated the mean hit rates (HR), false alarm rates

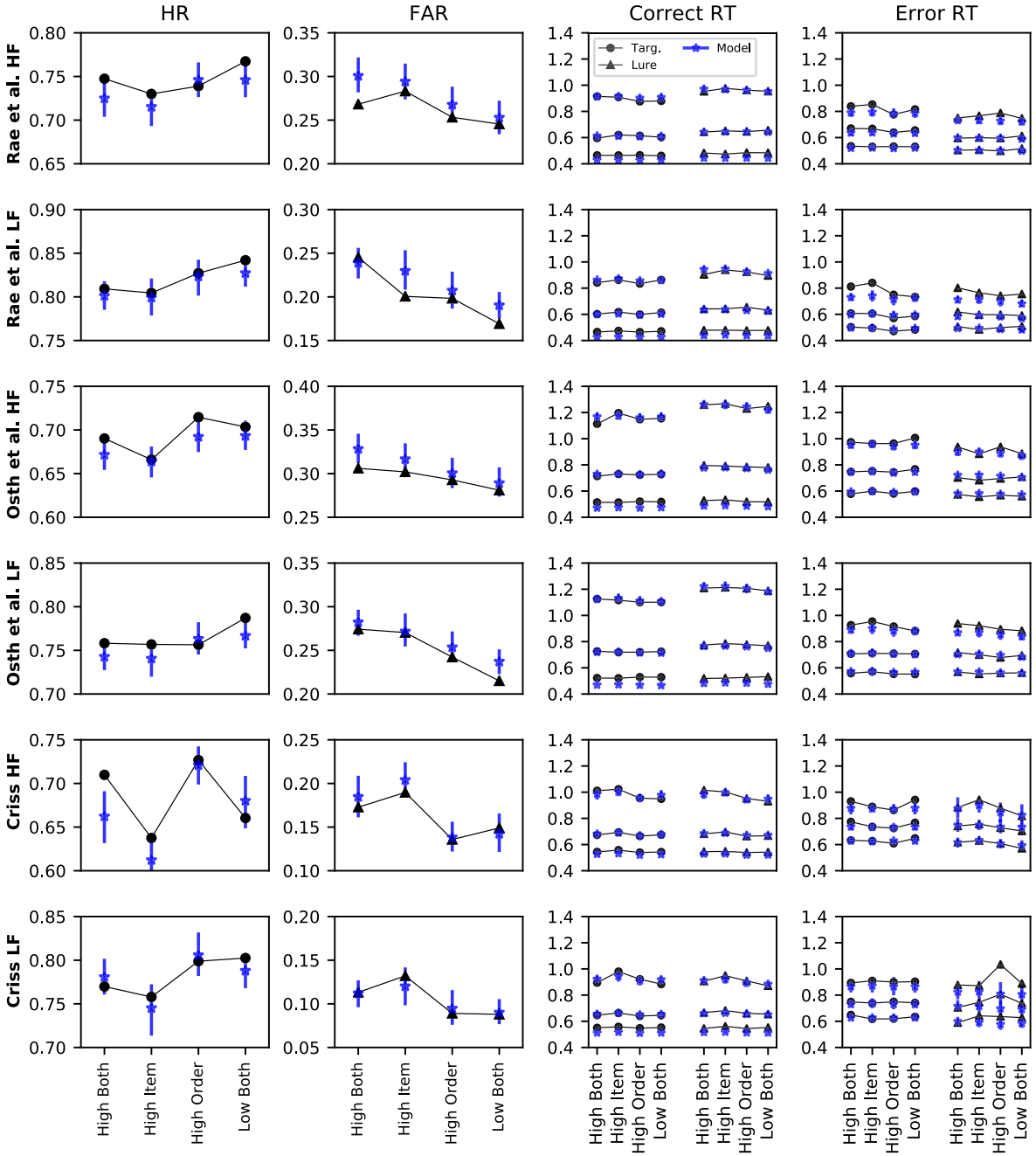

Fig. 5. Hit rates (left column), false alarm rates (second column), correct response time (RT) quantiles (third column) and error RT quantiles (fourth column) for the four semantic similarity bins (high global item and order similarity, high item similarity and low order similarity, high order similarity and low item similarity, and low global similarity on both dimensions) for HF and LF words for the data of Rae et al., Osth et al., and Criss datasets along with the mean and 95% highest density interval (HDI) of the DDM's posterior predictive distribution.

(FAR), and response time (RT) quantiles for correct and error responses. Correct responses were depicted using the 10th, 50th, and 90th percentiles of the RT distribution. Due to the relative infrequency of error responses, we depicted the error RT distribution using the 25th, 50th, and 75th percentiles of the error RT distribution. We additionally calculated the same summary statistics on the posterior predictive distributions of the DDM that employs mean global similarity of the item and order vectors. Data and model predictions along with the 95% HDIs from the model can be seen in Figs. 5 (for the Rae et al., Osth et al., and Criss datasets) and 6 for the experiments from Kiliç et al. (2017).

It is apparent from the figures that semantic similarity effects are larger for lures than for targets, with changes in the FAR being as large as 0.05–0.10 across the similarity bins. For context, Shiffrin et al.'s (1995) Experiment 1 reported increases in the FAR from 0.085 for a category length (CL) of 2 to 0.195 for CL of 9, and Experiment 2 reported increases in FAR from 0.108 to 0.196 for the same CLs. Thus, the changes in performance we observed across the global similarity bins constructed from lists of unrelated words were about half than for experiments that manipulated category length, which would be expected given that unrelated lists lack the structure of categorized lists.

Some of the effects of global semantic similarity reflect the parameter estimates depicted in Fig. 1. Estimates of α showed that increases in global item similarity generally hurt performance for both targets and lures. Although estimates of β were often close to zero, when they deviated from zero they indicated an increase in HR for targets. This pattern is present in the figures - performance is generally lowest when

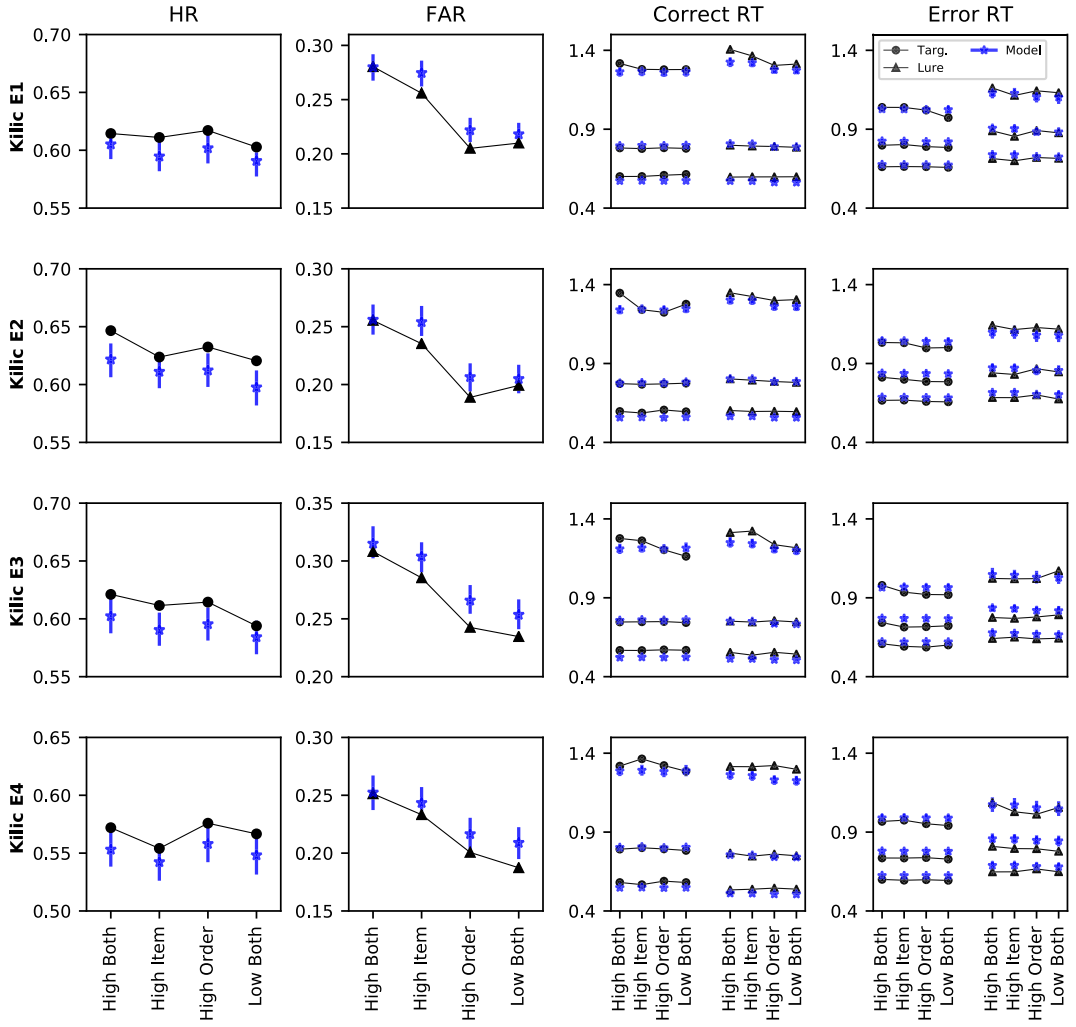

Fig. 6. Hit rates (left column), false alarm rates (second column), correct response time (RT) quantiles (third column) and error RT quantiles (fourth column) for the four semantic similarity bins (high global item and order similarity, high item similarity and low order similarity, high order similarity and low item similarity, and low global similarity on both dimensions) for the datasets from Kiliç et al. along with the mean and 95% highest density interval (HDI) of the DDM's posterior predictive distribution.

there is high item similarity, while high similarity on both dimensions does not appear to significantly impair performance more than high item similarity alone. The data from Rae et al. were an exception, with FAR being highest in the high item, high order similarity bin, but this was a dataset where the β estimate for lures was positive, albeit small.

The effects of semantic similarity on RTs appear to be relatively small. Both the lowest quantile and median of the RT distribution are relatively stationary across the semantic similarity bins while the upper quantiles show small differences. The fact that semantic similarity affects the RTs of the slowest responses but not the fastest is consistent with drift rate predictions of the diffusion model (Ratcliff & McKoon, 2008).

Can the global similarity of BEAGLE's representations account for word frequency effects?

In the previous models, we allowed the drift rate intercept parameter *V* to vary across high and low frequency words in the Rae et al., Osth et al., and Criss Experiment 2 datasets. However, in other modeling attempts, similarity measures derived from semantic space models were sufficient to produce poorer performance of HF words without any additional parameters, as HF words are more similar to each other than LF words, as postulated by the REM model (Shiffrin & Steyvers, 1997). The vectors for HF words are often more similar to each other than LF words in models such as LSA and BEAGLE because HF words are more likely to co-occur with other words and are especially more likely to co-occur with other HF words. Recognition memory models that capitalized on this include the neural network model of Monaco et al. (2007), which was trained on representations derived from word association spaces (Steyvers, Shiffrin, & Nelson, 2004), along with the recognition by semantic synchronization model (RSS: Johns et al., 2012), which is a holographic vector model that uses sparse lexical co-occurrence vectors for words. Both of these modeling efforts demonstrated that semantic representations can reproduce the *qualitative* pattern of word frequency effects. However, in neither case were the respective models fit to real data. Thus, it remains to be seen whether such semantic representations can capture the quantitative differences in performance between HF and LF words.

In order to test whether global similarity among the BEAGLE vectors was sufficient to capture the word frequency effect, we fit additional models that used the same mean drift rate intercept (*V*) for both HF and LF words. In these models, the global similarity among the item and order vectors is the *only* means for capturing performance differences between HF and LF words. For this reason, we refer to these models as the "Sim Only" models, which we applied to the three datasets that included a word frequency manipulation (Rae et al., Osth et al., and Criss). The omission of the mean drift rate intercept (*V*) parameters that vary by word frequency resulted in the removal of four parameters for

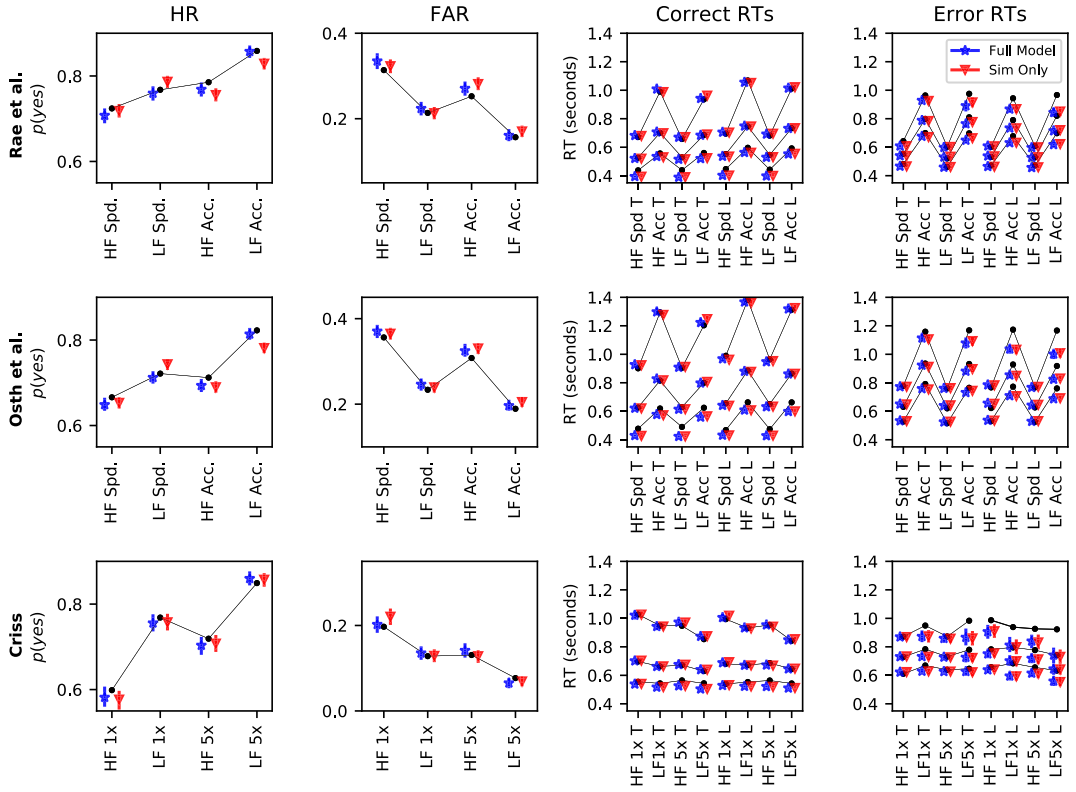

Fig. 7. Hit rates (left column), false alarm rates (second column), correct response time (RT) quantiles (0.1, 0.5, and 0.9, third column) and error RT quantiles (0.25, 0.5, and 0.75, fourth column) for HF and LF words for the data of Rae et al., Osth et al., and Criss datasets along with the mean and 95% highest density interval (HDI) of the full model (where mean drift rate *V* varies by word frequency) and Sim Only model (where *V* is fixed across HF and LF words). In the Rae et al. and Osth et al. datasets, performance was separated by the speed (spd.) and accuracy (acc.) emphasis conditions, whereas in the Criss Experiment 2 dataset performance was separated into the conditions where words were presented once (1×) or five times (5×).

each dataset relative to the "full" model. *V* continued to vary across targets and lures, speed-accuracy emphasis (Rae et al. and Osth et al.), or strength conditions (Criss).

The Sim Only models incurred substantial model selection penalties in each dataset (21, 18, and 119 DIC points respectively for the Rae et al., Osth et al., and Criss datasets). To understand why, we compared the group averaged posterior predictives from both the full and Sim Only models for HF and LF words in each dataset – these are depicted in Fig. 7. The Sim Only model captures the qualitative trends in the data and predicts poorer performance for HF words due to higher global similarity among the BEAGLE vectors. Nonetheless, the model does not perform as well as the full model which includes separate mean drift rate parameters for LF and HF words. The apparent misfit of the Sim Only model is that it overpredicts the word frequency effect in some circumstances. In both the Rae et al. and Osth et al. datasets, it overpredicts the HR differences between LF and HF words in the speed emphasis condition. In the Criss dataset, the Sim. Only model overpredicts the FAR difference between HF and LF words in the condition where words are only presented once.

The fact that global similarity among the semantic vectors falls somewhat short of capturing the performance differences between HF and LF words may be indicative of other factors contributing to the word frequency effect. HF words contain less distinctive letters than LF words (Malmberg, Steyvers, Stephens, & Shiffrin, 2002), are more perceptually similar to each other (Landauer & Streeter, 1973; Yarkoni, Balota, & Yap, 2008), and are also encountered in a wider variety of semantic contexts (Steyvers & Malmberg, 2003). In addition, models of recognition memory contain other mechanisms for WF effects, such as greater encoding variability for HF words (McClelland & Chappell, 1998), more interference from prior contexts in which HF words have been experienced (Dennis & Humphreys, 2001; Reder et al., 2000), and HF words attracting less attention than LF words (Glanzer et al., 1993). It's possible that semantic vectors could be augmented either with orthographic representations or with any of the above theoretical mechanisms to improve the ability to capture the word-frequency effect.

The correlation between global similarity and word frequency introduces a different question – are the global similarity effects we observed merely word frequency effects in disguise? In order to evaluate this, we ran additional models that included the influence of standardized log word frequency from the SUBTLEX database on the mean drift rate in addition to the global similarity among item and order vectors. In Supplementary Materials E ("Inclusion of Word Frequency as a Covariate") we additionally fit two models that allow relationships between log word frequency and drift rate, specifically a linear and a quadratic model. We investigated a quadratic function based on previous reports of U-shaped (Hemmer & Criss, 2013) and inverted U-shaped (Zechmeister, Curt, & Sebastian, 1978) word frequency effects. Quadratic functions of word frequency were strongly favored for six of the seven datasets.

Although the inclusion of word frequency reduced the influence of global item similarity, the qualitative results were very similar – high global item similarity continued to impair lures, and some datasets showed an appreciable impairment of targets as well. The fact that such results persisted when word frequency was included as a covariate suggests that the impairments due to global item similarity are unlikely to be due to other factors that correlate with word frequency. While global item similarity is correlated with word frequency, an important way in which it differs is that global semantic similarity is sensitive to the contents of the study list. That is, a given probe word's value of global semantic similarity depends on the words that are present on the study list, and thus different occurrences of the same word across different study lists can yield different values of global similarity.

Are the effects of global similarity modulated by speed/accuracy emphasis?

In our model, the effect of similarity between the probe cue and the test items is assumed to be constant through the duration of the test trial. However, recently there has been evidence that memory strength is reduced when participants are asked to respond quickly. The traditional explanation from the DDM for the speed-accuracy threshold is that participants reduce their response thresholds (the a parameter), resulting in faster decisions that are more susceptible to noise in the decision process (e.g., Ratcliff & Rouder, 1998). However, in some circumstances the drift rate is also impaired under conditions of speed emphasis in recognition memory, although the cause remains unclear (Rae et al., 2014; Starns, Ratcliff, & McKoon, et al., 2012). One possibility is that similarities between words changes as a function of processing time.

Hendrickson, Navarro, and Donkin (2015) presented participants with a probe word along with two words. One of the presented alternatives bore a thematic or associative relationship to the probe word, while the other word bore a taxonomic relationship. Participants were instructed to focus on one type of relationship only and select the appropriate word. An example trial might be a probe word such as "dog" presented along with "bone" and "cat" - if the participant was asked to select the associatively related word, "bone" is the correct answer. In addition, speed-accuracy emphasis was manipulated across two levels. Under the slower accuracy-emphasis condition, participants were equally able to focus on the associative or taxonomic relationship among the words. However, under speed emphasis, participants were able to select the associatively related word but were differentially impaired in selecting the taxonomically related word. In a similar vein, a recent model of recognition memory, the model of Cox and Shiffrin (2017), posits a dynamic process of recognition where probe features are slowly assembled as processing time unfolds, such that later decisions will result in different global similarity values than earlier decisions. If impairments due to global similarity are reduced in accuracy emphasis conditions – which could be manifested as estimates of the global similarity coefficients α or β that are closer to zero – this would result in improved drift rates in accuracy emphasis conditions.

Although most of our datasets do not control for the time in which participants respond, two (Rae et al. and Osth et al.) contain a speed-accuracy (SA) emphasis manipulation. We fit two additional models to each of these datasets. In the first model (Model 1), global similarity parameters corresponding to item (α) and order (β) similarity varied across the SA conditions, introducing four additional parameters. However, as in standard models, the drift rate intercept V varied across the speed-accuracy emphasis conditions. Thus, in the second model (Model 2), we fixed V across the speed-accuracy emphasis conditions such that *only* changes in α and β can determine the drift rate impairments under speed emphasis. Fixing V across the two conditions resulted in the omission of four parameters from the model, thus the resulting model has the same number of parameters as the standard model.

Model selection results can be seen in Table 6. For both datasets, Models 1 and 2 incurred substantial model selection penalties. The standard model, where the influence of global semantic similarity is constant through the test trial, exhibited a DIC weight of 1.0 in both datasets. These results suggest that any benefit in goodness of fit from allowing the coefficients to vary across the speed and accuracy emphasis conditions is less than the complexity penalty that such parameters incur. Model 2 suffered large penalties for fixing the mean drift rate V across SA conditions, which suggests that varying the influence of global semantic similarity across SA conditions is insufficient to capture the differences in drift rate.

Table 6
DIC values for the models that allow global similarity parameters to vary across the speed-accuracy (SA) emphasis conditions (Model 1 and Model 2) relative to the standard model where a single value of α and β are estimated. In Model 2, the mean drift rate V is fixed across the SA conditions. DIC weights are given in parentheses.

Dataset	Standard	Model 1	Model 2
Rae et al.	**−10008 (1.0)**	−9990 (0)	−9862 (0)
Osth et al.	**41973 (1.0)**	41993 (0)	42579 (0)

Fig. 8 compares the estimates of α^{μ} and β^{μ} for the standard model and Model 1. In the SA model, α^{μ} and β^{μ} do not vary drastically across speed-accuracy conditions. In the Rae et al. dataset, there is a little evidence for greater values of α^{μ} and β^{μ} in the accuracy emphasis condition. However, the 95% highest density intervals (HDIs) are large and overlap each other. In addition, these trends are not shared in the Osth et al. dataset which uses a very similar design. To conclude, there is little evidence from this analysis that the patterns of similarity effects change across speed and accuracy emphasis conditions. This resembles patterns found by Cox and Shiffrin (2017), who found no changes in the word frequency effect across speed and accuracy emphasis conditions.

Drift rate variability estimates

Inclusion of trial-by-trial global similarity estimates from the BEAGLE model allow for systematic cross-trial variability in the diffusion model, as opposed to the random drift rate variability in conventional applications of the model that is represented by the η parameter. But how much of a contribution do the global similarity estimates from BEAGLE make relative to the random variability estimates? Fig. 9 compares η^{μ} estimates from the model with no global similarity contributions (the no-similarity model: $\alpha = 0$ and $\beta = 0$) and η^{μ} estimates from the model with global similarity estimates (both α and β are estimated), along with the standard deviation of drift rates of global similarity measures calculated according to Eq. 3 from each of the datasets using the group mean parameters (α^{μ} and β^{μ}).

It is clear that the variability derived from the BEAGLE representations is rather small compared to the random variability estimated using the η parameter. This suggests that there are other forms of variability not captured by the present model. For one, despite the successes of the BEAGLE model, BEAGLE representations are not the "true" semantic representations, and a single corpus is not representative of the variability in language experience from different participants. There is also variability in global similarity among perceptual representations, such as ones constructed from orthographic and phonological features (Freeman, Heathcote, Chalmers, & Hockley, 2010), and variability in the strength of learning the list items (Hintzman, 1988; Wixted, 2007). In addition, there are various properties of words that have been found to account for variability across items in recognition memory such as word length and neighborhood size (Cortese, Khanna, & Hacker, 2010; Cortese, McCarty, & Schock, 2015). Investigating all of these possibilities would go considerably beyond the scope of the present work.

However, even in cases where the global similarity estimates from BEAGLE are largest (such as the Criss dataset), it is interesting to note that inclusion of such variability does not appear to have any effect on the estimates of the η parameter. Intuitively, one might expect η to reduce as a model accounts for more systematic cross-trial variability in drift rates. One potential reason why the estimates do not decrease is because η is responsible for the prediction of errors that are slower than correct responses. η has this effect because if there is no variability in drift rate, a counter-intuitive prediction of the diffusion model is that it predicts equal RTs for correct and error responses. When variability in drift rate is included, correct responses tend to come from high drift rates (which are fast) while error responses tend to arise from low drift rates (which are slow); when averages are calculated correct RTs are slower than error RTs (Ratcliff & McKoon, 2008). If all of the drift rate variability in a model is systematic, this would imply that error

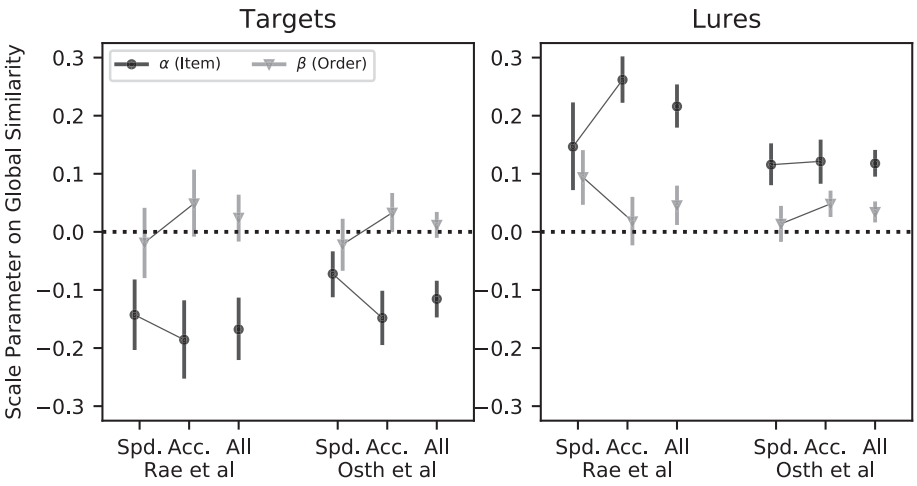

Fig. 8. Mean estimates of α^{μ} and β^{μ} along with their 95% HDIs for the speed and accuracy emphasis conditions of Rae et al. and Osth et al. along with the estimates from the models where α and β were not allowed to vary across conditions (indicated by "all"). Notes: Spd. = speed emphasis and Acc. = accuracy emphasis.

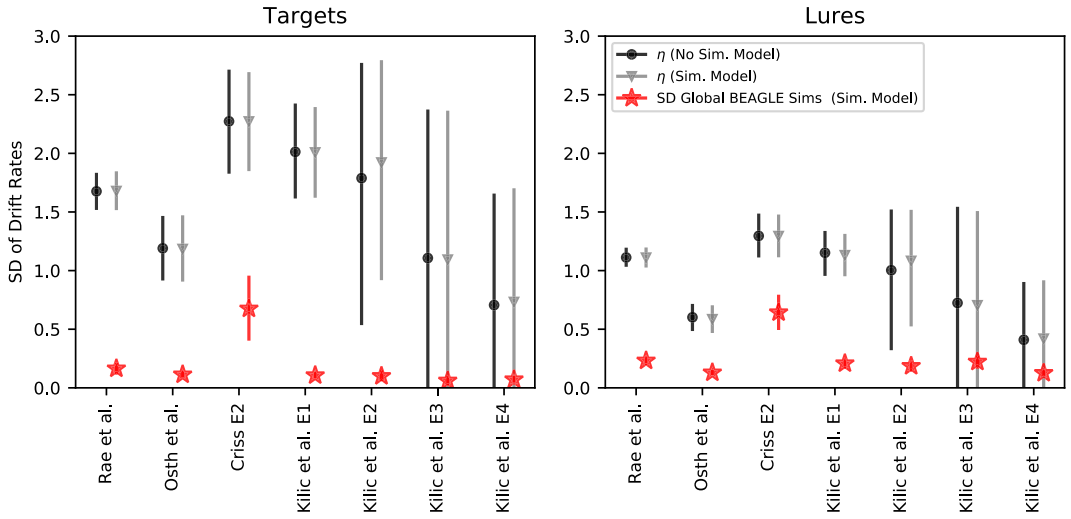

Fig. 9. Mean and 95% HDI estimates of the η^{μ} parameter (random drift rate variability) from models that exclude and include global similarity measures, along with standard deviations of global similarity calculated from the BEAGLE representations for each dataset. Error bars represent the 95% highest density interval (HDI).

responses should only be slower than correct responses when collapsed across items – for a given item, correct and error responses would be predicted to have the same speed.

General discussion

Global matching models predict that increases in global similarity should harm performance by making it more likely that participants erroneously false alarm to a probe cue. These predictions are commonly tested using category length designs, but this approach is subject to potential confounds. Instead, we investigated these effect in lists of unrelated words using the BEAGLE model, which allow us to calculate the similarity between the probe word and each of the study list words. BEAGLE differs from other semantic space models such as LSA in that it contains two types of representations: item representations, which are built from unordered co-occurrence between words in a corpus, and order representations, in which two words are similar to the extent to which they share neighboring words in the same relative positions.

We regressed individual trial global similarity measures onto drift rates in the DDM to systematically account for cross-trial variability in drift rates in recognition memory. The DDM allows for the unification of accuracy and RT into a single measure, namely drift rates. In addition, the DDM separates drift rate from other components of processing that can influence accuracy and RT, such as the time to encode the stimulus and execute the response (non-decision time), response bias, and the amount of evidence required to make a decision. Without the DDM, each of these factors would contribute noise to a regression of semantic similarity against manifest variables such as choice proportions or RTs for individual items. Instead, the DDM insures our global similarity measures are directly regressed against an estimate of memory strength, namely the drift rate parameter.

Across seven recognition memory datasets, we found pronounced impairments of performance based on global similarity calculated from BEAGLE's item vectors, especially for lures. Performance impairments due to global similarity calculated from order vectors, in contrast, were rather minimal, and in some cases model selection only favored the inclusion of global item similarity. These results suggest that semantic representations constructed on the basis of co-occurrence can be best used to understand impairments with increasing semantic similarity in recognition memory. Our analysis extends beyond previous investigations on global similarity effects because the BEAGLE model enabled us to jointly measure the contributions of both similarity measures on an item-by-item basis, and incapsulates similarity not just to other exemplars of a given semantic category but to the study list as a whole. Using such measures, we also found that global similarity measures constructed from the mean similarity generally outperformed measures

based on maximum similarity as a method for aggregating similarities across list items.

Implications for item noise and context noise models

A current debate in the recognition memory literature concerns the degree to which interference stems from the items on the study list or from the past contexts in which an item has been experienced (e.g., Criss & Shiffrin, 2004; Dennis & Chapman, 2010; Osth & Dennis, 2015). Global similarity effects in recognition memory have been traditionally used as an argument for item-noise models. To explain such effects, context-noise models appeal to additional mechanisms such as the generation of implicit associative responses (IARs) during the study list or the strategic usage of category cues when subjects notice the categorical structure on the study list. In the latter case, participants may notice that a number of study items belong to the same category (e.g., after seeing "dog", "cat", and "zebra" the participant notices the "animal" category), especially when each exemplar is presented in succession. At test, the participant may use the studied category labels as cues in conjunction with the probe items. Such a strategy resulted in increased HR and FAR with increasing category length within the BCDMEM model, a pure context-noise model (Dennis & Chapman, 2010).

Our investigation focuses on lists of unrelated words, which should minimize the possibility of both IAR generation and the abstraction of category cues. The fact that we have continued to find impairments with increasing global similarity supports the predictions of item-noise models. A pure context-noise model might be able to account for the results by assuming that IAR generation or category cuing still occurs in cases with unrelated words. However, this is unappealing for a crucial reason. A strong and counter-intuitive prediction of context-noise models is that they predict no impairment with increasing list length in recognition memory when lists are composed of unrelated words. Dennis and colleagues found support for this hypothesis by noting that previous findings of poorer performance in long lists were due to a number of confounds, and when such confounds are controlled, no list length effect occurs (e.g., Dennis & Humphreys, 2001; Dennis et al., 2008; Kinnell & Dennis, 2011). If mechanisms such as IAR generation and category cuing are still occurring in lists of unrelated words, however, longer lists should afford the generation of more IARs (Criss & Shiffrin, 2004) or the abstraction of more categories, and thus poorer performance on long lists would be predicted. At that stage, it is unclear what unique predictions a pure context-noise model generates over a pure item-noise model.

Recent investigations have also found that a pure context-noise model no longer provides a tenable account of results from the list length paradigm. Brandt et al. (2019) reproduced the null list length effects found by Dennis and colleagues using a within-subjects design where participants were tested on both a short and long list. However, Brandt et al. also found pronounced list length effects when results were restricted to the first list the participants were tested on (a between subjects analysis). Their explanation was that the first study-test cycle is most representative of differences in item-noise between the two lists, as the second study test-cycle counterbalances the number of items in memory. To understand how, consider if a participant studied a short list of 20 items followed by a list of 80 items, resulting in 100 items in memory, while another participant with the opposite order would have a different number of items in the first cycle (80) while the short list in the second cycle would result in 100 items in memory in the second cycle, the same as the other participant.

Fox et al. (2020) replicated the findings of Brandt et al. (2019) with word stimuli and modeled the increase in interference in lists across a session using the Osth and Dennis model, which is a global matching model that is able to decompose performance into the contributions of item-noise and noise from pre-experimental sources, namely context-noise and background-noise. Previous investigations using the model have found a dominance of pre-experimental interference in recognition memory (Osth & Dennis, 2015; Osth, Fox, McKague, Heathcote, & Dennis, 2018; Osth et al., 2018). However, none of these investigations considered the buildup of interference within a session, meaning there was no distinction made between short lists that were tested first (which are advantaged over long lists) or short lists that were tested after long lists (which perform comparably to longer lists). When interference buildup within a session was accounted for, Fox et al. (2020) found a dominance of item-noise over pre-experimental interference sources, as such a model was able to jointly account for the buildup of proactive interference and the finding that the list length effect depended on the length of the study list prior to the current list.

Although our results are more compatible with item-noise models, a novel finding in our analysis is that increases in global item similarity *decreased* memory strength for target items in some datasets, such as the data from Criss's Experiment 2. The fact that global item similarity had opposite effects on targets and lures (lures exhibited an increase in memory strength with increases in global item similarity) indicates a mirror effect (e.g., Glanzer & Adams, 1985). Previous investigations of category length manipulations have not generally found such effects, reporting either constant hit rates with increases in category length or increases in the hit rate (Cho & Neely, 2013; Maguire et al., 2010; Robinson & Roediger, 1997; Shiffrin et al., 1995). In models such as REM, increases in the proportion of shared features between the memory traces and the probe can increase both hits and false alarms, but can sometimes produce only negligible effects on the hit rate if the target item dominates the global similarity computation (Criss & Shiffrin, 2004; Shiffrin & Steyvers, 1997).

However, item-noise models could be modified to predict decreases in the HR with increasing global similarity. A number of memory models contain a mechanism whereby the learning rate on a given trial is proportional to the similarity between the presented item and the contents of memory (e.g., Farrell & Lewandowsky, 2002; Heath & Fulham, 1988; Mewhort et al., 2018; Murdock & Lamon, 1988). This implies that items that are similar to already encoded items will be encoded more poorly, decreasing their hit rate. This mechanism also leads to some novel predictions, such as similar items encoded later on the study list performing worse than earlier items. Studies of within-category serial position effects are consistent with this prediction, demonstrating that later items from a semantic category often perform worse than earlier items (Carey & Lockhart, 1973; Jakab & Raaijmakers, 2009; Neely & Tse, 2009). The utility of such a mechanism is not limited to similarity effects – such a mechanism can also predict advantages of spaced over massed repetitions (Murdock, 2003) along with primacy and Von Restorff effects (Elhalal, Davelaar, & Usher, 2014).[4] While recognition memory models have not generally employed such a learning algorithm, their architectures would allow for such an inclusion.

Integrating representations from semantic space models with global matching models

A number of recognition memory models have integrated realistic representations for perceptual stimuli in global matching models such as the generalized context model (Nosofsky, 1988). Such models have used representations of stimuli such as sinusoidal gratings (Kahana & Sekuler, 2002), faces (Lacroix, Murre, Postma, & van den Herik, 2006; Nosofsky & Zaki, 2003), Munsell colors (Nosofsky et al., 2011), and

[4] Recent modeling of response time distributions in free recall suggest that such a prediction error mechanism would provide an incomplete picture of primacy effects. Osth and Farrell (2019) found that primacy effects were best captured by a mixture of primacy and recency gradients, which is consistent with models where the primacy effect arises from a separate cue for the start-of-list items (e.g., Metcalfe & Murdock, 1981).

Fourier blobs and colored rectangles (Osth, Zhou, Lilburn, & Little, 2019). In several of these models, item representations are developed from multidimensional scaling solutions to a set of similarity ratings between the experimental stimuli. An advantage of such an approach is the ability to account for performance on individual items rather than merely matching performance at the level of experimental conditions.

With regard to word stimuli, however, most of the successes in integrating semantic representations within a process model have been in recall models, including the context maintenance and retrieval model (CMR: Polyn, Norman, & Kahana, 2009) which uses LSA representations, the fSAM model (Kimball, Smith, & Kahana, 2007) which employs representations from word association spaces (Steyvers et al., 2004), as well as the holographic model of Mewhort et al. (2018) which uses BEAGLE representations. Semantic representations have enabled the models to make transitions between similar words during recall, false memory intrusions, and the buildup and release from proactive interference, with each of the phenomena emerging from the interaction of the retrieval mechanisms and the similarity structure of the representations in memory.

In recognition memory, however, there are fewer examples. In fact, more recently developed models, such as the models of Osth and Dennis (2015) and Cox and Shiffrin (2017), continue to employ random representations. One exception is the work by Steyvers (2000), which used representations derived from word association spaces (Steyvers et al., 2004) in the REM model. In the traditional REM model, effects of category length are predicted by varying the proportion of shared elements between vectors. However, in the Steyvers work, category length manipulations can be predicted without any additional parameters as a natural consequence of the overlap between the semantic representations.

Another example is the recognition by semantic synchronization model (RSS: Johns et al., 2012). The RSS model employed relatively high dimensional co-occurrence vectors derived from a corpus of text. These vectors are constructed from a word by document matrix where each cell takes a value of one if a word occurred in a particular document or zero otherwise - a word representation is constructed by using rows from the matrix. Because words only occupy a relatively small proportion of documents, these representations are sparse and tend to produce little overlap with other words. Johns et al. (2012) demonstrated that the model was successful in its ability to capture a set of benchmarks such as the shapes of receiver operating characteristics, list length, and list strength effects, while also being able to capture DRM intrusions and word frequency effects.

Neither the Steyvers (2000) model or RSS have been fit to data from individual participants, which may reveal insufficiencies in their underlying structures or representations. In the RSS, semantic representations are sparse co-occurrence vectors, which resemble vectors from LSA prior to conducting the SVD. Such representations are unappealing because there can be some cases where the vectors are highly dissimilar despite a high apparent similarity between the words. Synonyms, for instance, often have low similarity in such representations using corpora such as ours because they do not frequently co-occur with each other in natural language.[5] LSA and BEAGLE do not have this problem because in such models words are sensitive to latent semantic structure (words will be similar if they co-occur with the same words). Thus, RSS's representations may be too impoverished to capture similarity effects in real data.

A related point is that in models such as the RSS, word frequency effects can be derived from the semantic representations alone. This is because high frequency (HF) words are more similar to other words than low frequency (LF) words, which makes it such that HF words suffer more interference at retrieval. We tested the extent to which the global similarity among BEAGLE's representations alone were sufficient to predict word frequency effects without any additional parameters and compared performance to a model where drift rate intercepts varied across LF and HF words. The model was capable of qualitatively capturing the word frequency effect when similarity alone produced differences between HF and LF words, but it quantitatively missed the performance differences between HF and LF words and was outperformed by the model where drift rates varied in addition to the global similarity differences. Thus, computational models may benefit from additional mechanisms for the word frequency effect, such as differences in perceptual representations or interference from pre-experimental memories.

Nonetheless, a reviewer suggested an additional note of caution in such modeling enterprises. Representations in semantic space models are highly dependent on the choice of corpus (Johns et al., 2019; Mandera et al., 2017) – this is illustrated in Supplementary Materials A where comparison between BEAGLE on two different corpora reveals different abilities to predict human performance in semantic ratings and synonym selection, and in Supplementary Materials C where we demonstrate different weight parameters of the global similarity metrics when the TASA corpus is used instead of our novels corpus. In addition, Morton and Polyn (2016) demonstrated how results in free recall are contingent both on the choice of semantic space model and retrieval model. They compared the CMR model with different choices of semantic representation and different retrieval mechanisms that determine how semantic representations are used to cue retrieval – each of these choices resulted in very different degrees of ability to capture performance in fits to free recall data.

Conclusion

In this work, we have explored global similarity effects with two types of semantic representations from the BEAGLE model and how they relate to drift rates in the diffusion decision model. Although our work has demonstrated the viability in exploring variability in performance across items using semantic space models, these representations are hardly comprehensive when it comes to accounting for variability across words. Future work will be needed to integrate other factors, such perceptual dimensions of word stimuli, in order to produce a more comprehensive representation.

Appendix A. Additional details of each dataset

Data exclusions

For each datset, we used sensible exclusions to omit fast or slow guess responses. For the Rae et al. (2014), Osth et al. (2017), and Criss (2010) datasets, responses faster than 0.2 s or slower than 2.5 s were excluded. For the Kiliç et al. (2017) datasets, responses faster than 0.3 s or slower than 3.0 s were excluded.

For the datasets of Kiliç et al., there were a number of participants who were either at chance or had a high proportion of responses with RTs less than 250 ms. For Experiment 1, one participant was excluded for having $d' < 0$ (#1), while two participants were excluded for having 58.9% and 22.2% of fast responses (#27 and 33, respectively). For Experiment 2, one participant was exluded for having $d' < 0$ (#30) and two participants were excluded for having 57.3% and 68.3% of fast responses (#26 and 29). For Experiment 3, two participants were excluded for having $d' < 0$ (#12 and

[5] This pattern does not hold in very large corpora such as a complete Wikipedia corpus (Recchia & Jones, 2009).

31, the latter of which had 76.5% fast responses) and five additional participants were excluded for having 57%, 24.2%, 92.4%, 31.1%, and 49.4% of fast responses (#1, 4, 10, 25, and 28).

For each dataset, there were additionally some words that were not present in the corpus. For the Rae et al. and Osth et al. word lists, the words "barb", "cause", "chink", "clink", "cornice", "dais", "dotage", "edict", "frill", "juggler", "larch", "lorry", "mien", "mote", "nadir", "nave", "offal", "pampas", "peal", "phial", "plait", "purser", "shed", and "simper" were not present. For the Criss dataset, the word "winless" was not present. For the Kiliç et al. datasets, the words "cause", "chandler", "cooper", "faro", "garth", "hart", "linden", "murphy", "regulus", "savoy", "sesame", "shed", "skiff", "spencer", "vita", and "wheeler" were not present.

We omitted trials that used any of these words as probe words as it is impossible to calculate global similarity for such words. If such words were present on the study list but the probe word was present in the corpus, we omitted such words from the global similarity calculation, as exclusion of all trials that contained any such words in the study list resulted in the exclusion of as much as 60% of the data for the Kiliç et al. datasets. Using our criteria, we excluded 1.37% of the data from Rae et al., 2.2% of the data from Osth et al., 0.45% of the data from Criss, and 12.30%, 12.81%, 26.90%, and 4.01% of the data from each of the experiments of Kiliç et al.

DDM parameterizations

Model parameterizations for each dataset can be seen in Table A.1, where each cell depicts the factors for which parameters were varied. A "1" indicates that a single parameter was estimated for the entire dataset. A letter such as "E" indicates speed-accuracy emphasis and implies that a separate parameter was allocated for the speed and accuracy conditions. The Rae et al. (2014) and Osth et al. (2017) datasets were combined into a single column as these datasets used extremely similar designs and identical parameterizations. The only difference between these datasets was that the Osth et al. dataset used a two-stage confidence procedure where "high" and "low" confidence responses were collected after the initial yes/no response. These confidence responses, however, were not addressed in the present modeling. While models of two-stage confidence have been developed with the DDM (Pleskac & Busemeyer, 2010; Moran, Teodorescu, & Usher, 2015), they result in intractable likelihoods when variability in model parameters is introduced. For each dataset, we follow prior precedent and allow drift rates, boundary separation, bias, and nondecision time to vary across the speed-accuracy emphasis conditions (Osth et al., 2017; Rae et al., 2014).

For the datasets that used strength manipulations, namely the data of Criss (2010) and Kiliç et al. (2017), we allowed z/a to vary across *cross-list* strength manipulations, as prior investigations have found higher z/a values in conditions of higher list strength (Criss, 2010; Starns, Ratcliff, & White, 2012). In the experiments of Kiliç et al., we additionally found large improvements in fit when a was allowed to vary across the list strength conditions, with larger response boundaries found in strong conditions. Experiments 2–4 of Kiliç et al. used mixed lists of strong (deeply encoded) and weak (shallowly encoded) words. However, in Experiments 3 and 4 participants were only tested on one strength class (strong or weak items) and were informed of which class of items they were tested on, which gives the participant the required information to adjust bias or response thresholds. For this reason, we used the same model parameterizations for Experiments 1, 2, and 4, where both a, z/a, and V_{lure} were allowed to vary across the weak and strong list conditions, while only a single value of each parameter was allocated for Experiment 2, in which no information was given to participants as to which items they would be tested on.

Finally, none of the datasets allowed η or τ to vary across conditions. This is because previous investigations have found that estimates of such parameters do not change across conditions such as word frequency, strength, or speed-accuracy emphasis (Osth et al., 2017; Starns, 2014; Starns & Ratcliff, 2014).

Table A.1
Factors over which each model parameter varies for each dataset. Notes: E = speed/accuracy emphasis, W = word frequency, S = strength, 1 = single parameter allocated.

Param.	Rae/Osth	Criss	Kiliç E1/E3/E4	Kiliç E2
t_0	E	1	1	1
s_t	E	1	1	1
a	E	1	S	1
z/a	E	S	S	1
V_{target}	E,W	W,S	S	S
V_{lure}	E,W	W,S	S	1
η_{lure}	1	1	1	1
τ	1	1	1	1

Appendix B. Additional details on the hierarchical Bayesian modeling

Prior distributions on model parameters

In the hierarchical Bayesian models we employ, participant parameters are sampled from group level distributions. Parameter of the group level distribution are denoted using μ and σ superscripts. Normal distributions are used wherever possible, however truncated normal (TN) distributions are used for distributions that are bounded, such as t_0, η, τ, s_t, and a, which are bounded on the (0,1) interval, along with z/a, which is bounded on the (0,1) interval:

$$
\begin{aligned}
z/a &\sim TN(z/a^{\mu}, z/a^{\sigma}, 0, 1)\\
a &\sim TN(a^{\mu}, a^{\sigma}, 0, \infty)\\
t_0 &\sim TN(t_0^{\mu}, t_0^{\sigma}, 0, \infty)\\
s_t &\sim TN(s_t^{\mu}, s_t^{\sigma}, 0, \infty)\\
V_{target} &\sim Normal(V_{target}^{\mu}, V_{target}^{\sigma})\\
V_{lure} &\sim Normal(V_{lure}^{\mu}, V_{lure}^{\sigma})\\
\eta_{lure} &\sim TN(\eta_{lure}^{\mu}, \eta_{lure}^{\sigma}, 0, \infty)\\
\tau &\sim TN(\tau^{\mu}, \tau^{\sigma}, 0, \infty)\\
\alpha &\sim Normal(\alpha^{\mu}, \alpha^{\sigma})\\
\beta &\sim Normal(\beta^{\mu}, \beta^{\sigma})\\
q &\sim Normal(q^{\mu}, q^{\sigma})
\end{aligned}
$$

where τ is the ratio of target drift rate variability to lure drift rate variability ($\eta_{target}/\eta_{lure}$).

For the μ parameters, we used mildly informative priors from previous applications of hierarchical Bayesian implementations of the DDM using DE-MCMC (Osth et al., 2017; Osth et al., 2017):

$$
\begin{aligned}
s_t^{\mu} &\sim TN(0.25, 0.25, 0, \infty)\\
t_0^{\mu} &\sim TN(0.5, 0.5, 0, \infty)\\
a^{\mu} &\sim TN(2, 2, 0, \infty)\\
tau^{\mu}, \eta_{lure}^{\mu} &\sim TN(1, 1, 0, \infty)\\
V_{target}^{\mu} &\sim Normal(2, 2)\\
V_{lure}^{\mu} &\sim Normal(-2, 2)
\end{aligned}
$$

Because we have considerably less information about the relations between global similarity calculated from BEAGLE's representations, we use less informative priors for the relations between global item similarity α and global order similarity β:

$$\alpha^{\mu} \sim Normal(0, 10) \alpha^{\mu} \sim Normal(0, 10)$$

We also used mildly informative priors on the σ parameters:

$$
\begin{aligned}
a^{\sigma}, \eta_{lure}^{\sigma}, V_{target}^{\sigma}, V_{lure}^{\sigma}, \alpha^{\sigma}, \beta^{\sigma}, q^{\sigma} &\sim \Gamma(1, 1)\\
z/a^{\sigma}, s_t^{\sigma}, t_0^{\sigma} &\sim \Gamma(1, 3)
\end{aligned}
$$

Details on MCMC estimation and posterior predictive simulations

For each model, the number of chains was set to three times the number of participant parameters. After 6000 burn-in iterations, chains were heavily thinned such that only one in every 20 iterations was accepted. This process continued until a total of 1500 MCMC samples were accepted for each model. Chains were considered converged if the Gelman-Rubin statistic was below 1.10. This criterion was satisfied for all models; in most cases the statistic was very close to 1.0.

To generate posterior predictive distributions, one in every 50 posterior samples was used to simulate a dataset of the same size as the original dataset.

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jml.2019.104071.

References

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*(2), 97–123.

Brandt, M., Zaiser, A., & Schnuerch, M. (2019). Homogeneity of item material boosts the list length effect in recognition memory: A global matching perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 834–850.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In Dietrich, & Markman (Eds.). *Cognitive dynamics: Conceptual change in humans and machines* (pp. 117–156). LAwrence: Erlbaum Associates.

Carey, S. T., & Lockhart, R. S. (1973). Encoding differences in recognition and recall. *Memory & Cognition, 1*(3), 297–300.

Cho, K. W., & Neely, J. H. (2013). Null category-length and target-lure relatedness effects in episodic recognition: A constraint on item-noise interference models. *The Quarterly Journal of Experimental Psychology, 66*(7), 1331–1355.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin and Review, 3*(1), 37–60.

Cortese, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory, 18*, 595–609.

Cortese, M. J., McCarty, D. P., & Schock, J. (2015). A mega recognition memory study of 2,897 disyllabic words. *Quarterly Journal of Experimental Psychology, 68*, 1489–1501.

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General, 147*, 545–590.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review, 124*(6), 795–860.

Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(2), 484–499.

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55*, 447–460.

Criss, A. H., & Shiffrin, R. M. (2004). Context-noise and item-noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review, 111*, 800–807.

D'Agostino, P. (1971). The blocked-random effect in recall and recognition. *Journal of Verbal Learning and Verbal Behavior, 8*, 815–820.

Dennis, S., & Chapman, A. (2010). The inverse list length effect: A challenge for pure exemplar models of recognition memory. *Journal of Memory and Language, 63*, 416–424.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian Analysis of Recognition Memory: The Case of the List-Length Effect. *Journal of Memory and Language, 59*, 361–376.

Elhalal, A., Davelaar, E. J., & Usher, M. (2014). The role of the frontal cortex in memory: an investigation of the Von Restorff effect. *Frontiers in Human Neuroscience, 8*, 1–20.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in

serial recall. *Psychonomic Bulletin and Review, 9*, 59–79.
Fox, J., Osth, A. F., & Dennis, S. (2020). Accounting for the buildup of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language. 00*.
Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. E. (2010). Item effects in recognition memory for words. *Journal of Memory and Language, 62*, 1–18.
Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*(1), 1–67.
Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition, 13*(1), 8–20.
Glanzer, M., & Adams, J. K. (1990). The Mirror Effect in Recognition Memory: Data and Theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(1), 5–16.
Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*(3), 546–567.
Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1210–1230.
Heath, R. A., & Fulham, R. (1988). An adaptive filter model for recognition memory. *British Journal of Mathematical and Statistical Psychology, 41*, 119–144.
Hemmer, P., & Criss, A. H. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 1947–1952.
Hendrickson, A. T., Navarro, D. J., & Donkin, C. (2015). Quantifying the time course of similarity. In D. C. Noelle, (Ed.). *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 908–913). Cognitive Science Society.
Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. *Psychological Review, 95*(4), 528–551.
Howard, M. W., & Kahana, M. J. (2002a). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 268–299.
Howard, M. W., & Kahana, M. J. (2002b). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*(1), 85–98.
Hutchison, K. A., Balota, D. A., Cortese, M. J., & Watson, J. M. (2008). Predicting semantic priming at the item level. *Quarterly Journal of Experimental Psychology, 61*, 1036–1066.
Jakab, E., & Raaijmakers, J. G. W. (2009). The role of item strength in retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 607–617.
Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science, 42*, 1360–1374.
Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review, 17*(5), 662–672.
Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology, 65*(4), 486–518.
Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review, 26*, 103–126.
Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55*(4), 534–552.
Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1–37.
Jones, M. N., Willits, J. A., & Dennis, S. (2015). Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology* (pp. 232–254). OUP.
Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: a noisy exemplar approach. *Vision Research, 42*(18), 2177–2192.
Kiliç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology, 92*, 65–86.
Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review, 114*, 954–993.
Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition, 39*, 348–363.
Lacroix, J. P. W., Murre, J. M. J., Postma, E. O., & van den Herik, H. J. (2006). Modeling recognition memory using the similarity structure of natural input. *Cognitive Science, 30*, 121–145.
Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.
Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior, 12*, 119–131.
Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55*, 1–7.
Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods Instruments & Computers, 28*(2), 203–208.
Maguire, A., Humphreys, M. S., Dennis, S., & Lee, M. D. (2010). Global similarity accounts of embedded-category designs: Tests of the global matching models. *Journal of Memory and Language, 63*, 131–148.
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition, 30*(4), 607–613.
Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and countin: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.
Mather, M., Henkel, L. A., & Johnson, M. K. (1997). Evaluating characteristics of false memories: Remember/know judgments and memory characteristics questionnaire compared. *Memory & Cognition, 25*(6), 826–837.
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760.
Metcalfe, J., & Murdock, B. B. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 20*, 161–189.
Mewhort, D. J. K., Shabahang, K. D., & Franklin, D. R. J. (2018). Release from PI: An analysis and a model. *Psychonomic Bulletin & Review, 25*, 932–950.
Monaco, J. D., Abbott, L. F., & Kahana, M. J. (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning & Memory, 14*, 204–213.
Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology, 78*, 99–147.
Morton, N. W., & Polyn, S. M. (2016). A predictive framework for evaluating models of semantic organization in free recall. *Journal of Memory and Language, 86*, 119–140.
Murdock, B. B. (2003). The mirror effect and the spacing effect. *Psychonomic Bulletin & Review, 10*(3), 570–588.
Murdock, B. B., & Lamon, M. (1988). The Replacement Effect - Repeating Some Items While Replacing Others. *Memory & Cognition, 16*(2), 91–101.
Neely, J. H., & Tse, C. S. (2009). Category length produces an inverted-U discriminability function in episodic recognition memory. *The Quarterly Journal of Experimental Psychology, 62*(6), 1141–1172.
Nosfosky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1194–1209.
Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 700–708.
Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 118*(2), 280–315.
Osth, A. F., Bora, B., Dennis, S., & Heathcote, A. (2017). Diffusion versus linear ballistic accumulation: Different models, different conclusions about the slope of the zROC in recognition memory. *Journal of Memory and Language, 96*, 36–61.
Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260–311.
Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology, 92*, 101–126.
Osth, A. F., Zhou, A., Lilburn, S., & Little, D. R. (2019). The extralist feature effect revisited: A challenge for global matching models of recognition memory. PsyArxiv Preprint.
Osth, A. F., & Farrell, S. (2019). Using response time distributions and race models to characterize primacy and recency effects in free recall initiation. *Psychological Review, 126*, 578–609.
Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory: Data and a global matching model. *Journal of Memory and Language, 103*, 91–113.
Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology, 104*, 106–142.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117*(3), 864–901.
Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*(1), 129–156.
Rae, B., Heathcote, A., Donkin, C., Averell, L., & Brown, S. (2014). The Hare and the Tortoise: Emphasizing Speed Can Change the Evidence Used to Make Decisions. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*(5), 1226–1243.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*, 59–108.
Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*, 873–922.
Ratcliff, R., & McKoon, G. (2018). Modeling numeracy representation with an integrated diffusion model. *Psychological Review, 125*, 183–217.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*(5), 347–356.
Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518–535.
Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential-sampling models for two choice reaction time. *Psychological Review, 111*, 333–367.
Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*, 408–424.
Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods, 41*, 647–656.
Recchia, G., Sahlgren, M., Kanerva, P., & Jones, M. N. (2015). Encoding sequential information in semantic space models: Comparing holographic reduced representations and random permutation. *Computational Intelligence and Neuroscience, 2015*, 1–15.
Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember-know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 294–320.
Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science, 8*, 231–237.
Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*(4), 573–604.
Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.). *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.
Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 267–287.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B,*

64, 583–639.
Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & Cognition, 42*, 1357–1372.
Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language, 70*, 36–52.
Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of the zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64*, 1–34.
Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1137–1151.
Steyvers, M. (2000). *Modeling semantic and orthographic similarity effects on memory for indvidual words.* Ph.D.Indiana University.
Steyvers, M., & Malmberg, K. J. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 760–766.
Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004). Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A. Healy (Ed.). *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. Washington, DC: American Psychological Association.
Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods, 18*(3), 368–384.
Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192–196.
Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–176.
Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*, 971–979.
Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a U-shaped function of word frequency. *Bulletin of the Psychonomic Society, 11*, 371–373.