Author/s:
Osth, AF;Zhou, A;Lilburn, SD;Little, DR

Title:
Novelty Rejection in Episodic Memory

Date:
2023-03-13

Citation:
Osth, A. F., Zhou, A., Lilburn, S. D. & Little, D. R. (2023). Novelty Rejection in Episodic Memory. Psychological Review, 130 (3), pp.720-769. https://doi.org/10.1037/rev0000407.

Persistent Link:
https://hdl.handle.net/11343/332962

Novelty rejection in episodic memory

Adam F. Osth, Aspen Zhou, Simon D. Lilburn, & Daniel R. Little

University of Melbourne

Address correspondence to:

Adam Osth

Author Note

This work was supported by an ARC Discovery Early Career Award (DECRA) awarded to Adam Osth (DE170100106) and an ARC Discovery Project (DP160102360) to Daniel R. Little. Data, experiment code, and model code can be found on our Open Science Foundation (OSF) page: `https://osf.io/b2zyk/`. We would like to thank Simon Dennis for helpful discussions, Danièle Martinie and Ariel Goh for assistance with data collection, and Klaus Oberauer and Greg Cox for helpful comments on a previous version of this manuscript. This study was not preregistered.

Abstract

Episodic memory theories have postulated that in recognition, a probe is accepted or rejected on the basis of its global similarity to studied items. Mewhort and Johns (2000) directly tested global similarity predictions by manipulating the feature compositions of probes – novelty rejection was facilitated when probes contained novel features even when other features strongly matched, an advantage dubbed the extralist feature effect, which greatly challenged global matching models. In this work, we conducted similar experiments using continuously-valued separable- and integral-dimension stimuli. Analogs of extralist lures were constructed where one stimulus dimension contained a value that was more novel than the other dimensions while overall similarity was equated to another class of lures. Facilitated novelty rejection for lures with extralist features was only found for separable-dimension stimuli. While integral-dimension stimuli were well described by a global matching model, the model failed to account for extralist feature effects with separable-dimension stimuli. We applied global matching models – including variants of the exemplar-based linear ballistic accumulator (EB-LBA) – that employed different means of novelty rejection afforded by separable-dimension stimuli, including decisions based on the global similarity of the individual dimensions and selective attention being directed toward novel probe values (a diagnostic attention model). While these variants produced the extralist feature effect, only the diagnostic attention model succeeded in providing a sufficient account of all of the data. The model was also able to account for extralist feature effects in an experiment with discrete features similar to those from Mewhort and Johns (2000).

Keywords: recognition memory; global matching models; exemplar models; extralist feature effect

Novelty rejection in episodic memory

When we are asked if we recognize a stimulus, what makes us able to recognize items we've experienced and correctly reject novel ones? And moreover, why are some novel items easy to reject while others are extremely difficult, sometimes resulting in endorsements of new items that are nearly as strong as those of true memories (e.g. Roediger & McDermott, 1995)?

The majority of episodic memory models posit that both acceptance of studied items and rejection of novel items derive from a probe item's *global similarity* to the contents of memory, which is the basis of the recognition decision. In these models, a measure of global similarity is produced by a process of global matching (see Clark & Gronlund, 1996; Osth & Dennis, in press, for reviews), in which the probe item is matched against each item in memory and the similarity is computed. Subsequently, each of the similarities are aggregated via summation or averaging, producing a measure of global similarity that indexes the similarity between the probe cue and the stored memories.

Global similarity can be mapped to recognition decisions by comparing the similarity value to a decision criterion (e.g., Gillund & Shiffrin, 1984; Hintzman, 1988). More recent frameworks have used the global similarity to drive a noisy evidence accumulation process between "old" and "new" decisions that is capable of making predictions about both choice and response time (e.g., Cox & Shiffrin, 2017; Fox, Dennis, & Osth, 2020; Nosofsky, Little, Donkin, & Fific, 2011; Osth, Jansson, Dennis, & Heathcote, 2018). In both forms of decision-making, higher global similarity results in a higher likelihood of making an "old" decision, as the global similarity is a reflection of the likelihood that the probe item is located in the contents of memory. In this way, global matching models directly reflect the encoding specificity principle of Tulving and Thomson (1973), in which successful retrieval is a function of the similarity between the cues at the time of the retrieval and the contents of memory. The relevance of global matching as a retrieval mechanism extends beyond episodic recognition memory, as it has been proposed as a cornerstone of an integrated

theory of attention, categorization, and memory (Logan, 2002). Variants of global matching models have also been used to explain other cognitive tasks such as eyewitness identification (Clark, 2003), inductive reasoning (Hawkins, Hayes, & Heit, 2016; Heit & Hayes, 2011), lexical access (Wagenmakers et al., 2004), implicit memory (Schooler, Shiffrin, & Raaijmakers, 2001), spoken word recognition (Goldinger, 1998), and probability estimation (Dougherty, Gettys, & Ogden, 1999).

According to global matching models, a novel stimulus becomes difficult to reject when its global similarity to the contents of memory is high, as a high value of global similarity is indicative of the probe being a target item. This can result from the probe being highly similar to one stimulus in memory or being moderately similar to a number of items. Conversely, in global matching models, probes that are easy to reject have low global similarity, in that they are not similar to any of the stored memories. This account has been successful in explaining a number of findings in episodic memory research. First, global matching models are able to account for the category length effect: the finding that increases in the number of studied items from a common semantic or perceptual category increases the probability that non-studied items (lures) from the same category are endorsed (e.g., Cho & Neely, 2013; Robinson & Roediger, 1997; Shiffrin, Huber, & Marinelli, 1995; Zaki & Nosofsky, 2001). This prediction occurs because the greater number of items in memory that are similar to the probe increases the global similarity, which consequently increases the probability that they will be endorsed. A second success is the model's ability to account for the list length effect, which is the finding that increases in the number of studied items degrades performance, usually by increasing the false alarm rate. The explanation is that a larger number of items in memory results in a higher likelihood that some of the stored memories will resemble the probe (Brandt, Zaiser, & Schnuerch, 2019; Cary & Reder, 2003; Fox et al., 2020; Kahana & Sekuler, 2002; Nosofsky et al., 2011; Strong, 1912). Finally, global matching models have also been found to account for the high rates of false memory in the Deese-Roediger-McDermott paradigm (Deese,

1959; Roediger & McDermott, 1995), in which the study of a large number of associates produces a very high false alarm rate to a non-studied associate due to the non-studied associate's similarity to each of the other studied associates (Arndt & Hirshman, 1998).

A limitation of many studies focused on the consequences of similarity is that they have employed verbal stimuli. Verbal stimuli are high dimensional entities containing both perceptual and semantic properties, making it difficult to explicitly control the features that give rise to similarity. However, such a task is made easier with artificial visual stimuli – such stimuli were employed by Mewhort and Johns (2000) to more directly test the relationship between global similarity and the rejection of novel items. In their experiments, participants studied short lists of colored shapes. The critical manipulation involved the construction of the lures — the matches and mismatches of the lures to the study set items were independently manipulated to produce differing levels of global similarity. An illustration of their paradigm can be seen in Figure 1. Specifically, lures included 1:1 probes that contained exactly one match to a stimulus on one aspect and one match to a different stimulus on the other aspect (note that we use the term "aspect" here instead of dimension, as the underlying dimensional structure of the stimuli may be much greater; color, for instance, is usually represented with three or more dimensions; Ekman, 1954; Shepard, 1962.). A 1:0 probe, in contrast, is a probe that has one match to a stimulus on one aspect, while the feature of the other aspect was considered an *extralist feature* by virtue of it not being present on the study list. Likewise, a 2:0 probe pairs the extralist feature with a feature presented twice on the study list.

Global matching models predict that a novel item should become more difficult to reject as the number of matching features across memories is increased. Because global similarity is additive across memories when the aggregation consists of summing or averaging, mismatching features from some memories should merely decrease global similarity, making the novel item easier to reject. Mewhort and Johns (2000) tested this prediction by comparing the 1:1, 2:0, 1:0, and 0:0 probes in a Sternberg (1966) recognition

memory paradigm. According to global matching models, 1:1 and 2:0 probes should be equally difficult to reject by virtue of their identical number of matches to study list items (two), while 1:0 probes should be easier to reject due to their low global similarity. Contrary to this prediction, mean response time (RT) was longest for 1:1 probes and was virtually equivalent for 1:0 and 2:0 probes (RT was the relevant dependent variable as accuracy was close to ceiling), while RT was shortest for 0:0 probes. Mewhort and Johns dubbed the advantage for extralist features (i.e., the ease of rejecting 1:0 and 2:0 probes) the *extralist feature effect*. It was as if the novel information present in the extralist probes was sufficient to override the matching information; the increase in matches from 1:0 to 2:0 did not appear to make lure probes more difficult to reject, but increasing the number of extralist features to 2 (on 0:0 probes) facilitated rejection.

The rejection advantage for novel extralist probes has been replicated in subsequent work (E. E. Johns & Mewhort, 2002, 2003). E. E. Johns and Mewhort (2002) further generalized the results to novel aspects of the stimuli by simultaneously varying the color, shape, and texture of the items. They found that if each study list item shared the same feature on one aspect (e.g., all of the study items were blue), lures with novel features on that aspect (e.g., a yellow item) were rejected easier even if the other aspects (shape and texture) strongly matched the studied stimuli. In addition, Experiments 5–7 in the Mewhort and Johns (2000) article demonstrated that the effect could also be observed with word stimuli in both short- and long-term recognition paradigms, where matches and mismatches were manipulated by swapping the initial and terminal letters in the words. The combined set of results demonstrate that this is a highly replicable and generalizable phenomenon.

The experiments of Mewhort and Johns constitute what is arguably the strongest challenge to the global matching retrieval mechanism as they directly manipulated the global similarity between the lures and the study lists and failed to find the predicted equivalence between 2:0 and 1:1 lures. It is especially problematic that the findings from

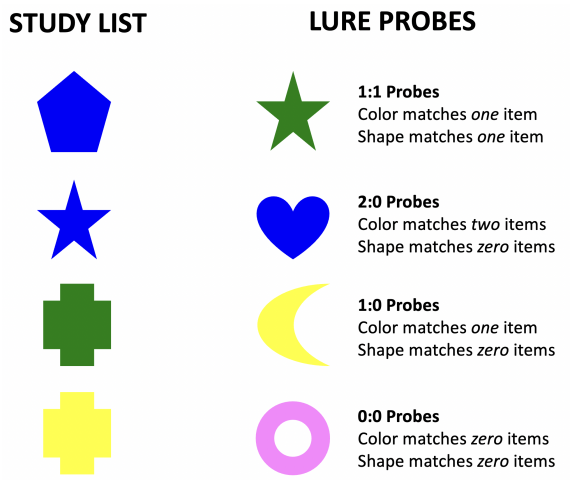

Figure 1. Illustration of the Mewhort and Johns (2000) paradigm. The members of the study set are on the left, while the right column depicts four of the probe types (1:1, 1:0, 2:0, and 0:0 probes), where the numbers refer to the number of matches to color and shape.

these studies are contrary to the predictions of virtually all global matching models. Mewhort and Johns (2005) performed simulations with two leading global matching models, namely the Minerva 2 (Hintzman, 1988) and the retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997), which were both found to be incapable of producing appropriate predictions for the Mewhort and Johns (2000) paradigm. Specifically, both models predicted equivalent performance for 2:0 and 1:1 probes along with a rejection advantage for 1:0 over 2:0 probes. The failure of the REM model to produce the extralist feature effect demonstrates just how challenging the data pattern is, given that REM iis able to capture a number of benchmark phenomena that challenged initial global matching models, including the mirror effect (Glanzer & Adams, 1985) and the null list-strength effect (Murnane & Shiffrin, 1991; Osth, Fox, McKague, Heathcote, & Dennis, 2018; Ratcliff, Clark, & Shiffrin, 1990). Since the publication of Mewhort and Johns (2000), newer global matching models have been developed such as the tensor model of Osth and Dennis (2015), the dynamic recognition model of Cox and Shiffrin (2017), and the exemplar-based random walk (EBRW) model (Nosofsky et al., 2011). However, despite

the advances of these models, none of these models have obvious explanations for the extralist feature effect.

The reason why the extralist feature effect is challenging to the entire set of global matching models concerns the global similarity computation itself. While global matching models differ considerably in the way that similarity is computed — ranging from dot products between vector representations (Hintzman, 1988; Humphreys, Bain, & Pike, 1989; Murdock, 1982; Osth & Dennis, 2015), an exponential transformation of the distance between multidimensional representations (Kahana & Sekuler, 2002; Nosofsky et al., 2011), to the likelihood ratio that the memory trace is a studied item (Cox & Shiffrin, 2017; Dennis & Humphreys, 2001; Shiffrin & Steyvers, 1997) — in the majority of models, the resulting global similarity usually arises by summing or averaging the similarity values. General predictions for the Mewhort and Johns paradigm can be illustrated if we assume the similarity between two matching features is X. Ignoring the similarity between mismatching features for the time being, a summed similarity model predicts that a 1:0 probe has a global similarity value of X due the presence of one feature match, while 2:0 and 1:1 probes both exhibit global similarity values of $2X$ due to two matching features. Substituting various similarity metrics can change the absolute values of global similarity but cannot change the differences in global similarity between 2:0 and 1:1 probes that stem from the matching features.

Given that lures containing extralist features (e.g., 2:0 lures) can yield the same global similarity as 1:1 lures, an obvious question that emerges is why novel items with extralist features are easier to reject. Does this rejection advantage reflect the fact that additional processes are occurring during recognition decisions to facilitate novelty rejection? Or is the very notion of recognition being determined by global similarity flawed? The answer to this question has far-reaching implications, as the abandonment of such models would undermine their application not just in episodic memory, but other domains such as categorization, lexical access, and eyewitness memory.

We address this question in the present article in two ways. Our first aim is to empirically evaluate the generality of the ease of rejection of novel items containing extralist features. While the extralist feature effect has been demonstrated with both words and colored shapes, in our experiments we introduce an additional stimulus manipulation that is theoretically relevant to the set of models we test, namely the distinction between *integral-* and *separable-dimension* stimuli. In the categorization literature (Algom & Fitousi, 2016; Ashby & Maddox, 1994; Burns, 2016; Garner, 1974; Nosofsky & Palmeri, 1997; Little, Nosofsky, Donkin, & Denton, 2013; Little, Wang, & Nosofsky, 2016; Shepard & Chang, 1963), integral dimensions, like the brightness and saturation of a color, refer to dimensions which are difficult to analyze independently; by contrast, separable dimensions, like color and shape, can be attended without interference from the other dimension.

As we will discuss, this is a theoretically important distinction because only separable-dimension stimuli enable the individual dimensions in a probe stimulus to be accessed differently than they are conventionally treated in global matching models, enabling other methods of novelty rejection. For instance, instead of calculating the global similarity of the entire stimulus, it might be the case that global similarity is computed for each stimulus *feature* — a probe can be rejected if any feature has sufficiently low global similarity. This would result in more frequent rejection of 2:0 lures – the extralist feature in 2:0 probes has low global similarity, enabling rejection, whereas both features in 1:1 probes have higher global similarity. A hybrid account is also possible where global similarity of the entire stimulus can be used to produce "old" decisions while rejection of novel items can be based on the global similarity of the individual stimulus features. Such a model implies that the basis for novelty rejection is fundamentally different than that of recognition of familiar stimuli. The hybrid model was suggested by Nosofsky et al. (2011) as a possible explanation for the extralist feature effect but was not implemented or fit to data.

An additional possibility is that selective attention may be directed toward features that are particularly diagnostic for the evaluation of whether a stimulus is "old" or "new".

For instance, if additional attention is directed toward the extralist feature in 2:0 probes, that feature exhibits greater weight in determining the inter-item similarities and can consequently reduce the global similarity of 2:0 probes relative to 1:1 probes. This account reflects the idea that similarity between stimuli is not fixed, but context-dependent (e.g., Tversky, 1977), and selective attention can vary systematically across conditions to optimize performance (Nosofsky, 1984, 1986). A selective attention explanation of the extralist feature effect was originally proposed by Nosofsky et al. (2011), but they did not provide a mechanism for how attention could be directed toward extralist features in this fashion.

Both of these mechanisms are only possible when stimulus dimensions are independently accessible to the participant, and thus would only be feasible for experiments where separable-dimension stimuli are employed. Thus, if the extralist feature effect is specifically restricted to separable-dimension stimuli, then it would suggest that the unique affordances from such stimuli may responsible for the ease of rejecting lures that contain extralist features.

The second aim in our work is to evaluate whether either of these proposed modifications to the global matching framework are capable of addressing the enhanced novelty rejection of lures containing extralist features. In this article, we focus on the exemplar-based linear ballistic accumulator model (EB-LBA: Donkin & Nosofsky, 2012b), which combines the global matching architecture of the generalized context model (GCM: Nosofsky, 1986, 1991) with linear ballistic accumulators (LBA: Brown & Heathcote, 2008) to produce predictions about both choice and distributions of response times. The underlying EB-LBA framework is sufficiently generalizable to explore all of the relevant considered models. The closed-form analytics of the LBA allow for generalization to both parallel and hybrid coactive-parallel decision architectures, while the selective attention components allow for constructing models where extralist features attract more attention than other features. We will elaborate more on the relevance of the GCM family of models

to the extralist feature effect in the next section.

The Exemplar-Based Random Walk (EBRW) Model and its Proposed Solutions to the Extralist Feature Effect

Despite the contrary evidence to global matching models that was presented by Mewhort and Johns (2000), a detailed global matching account of stimulus-specific effects in recognition memory was developed by Nosofsky et al. (2011) in the form of the EBRW. The EBRW is the integration of the GCM with a back-end random walk process to produce decisions, which allowed the model to make predictions about both choice and response time (RT). In the GCM family of models, the stimulus representations are specified as points in a multidimensional space — similarity between stimuli can be calculated as an exponential transformation of the distance between them. Representations of stimuli can be derived from multidimensional scaling (MDS) solutions or other proximity data, allowing for detailed, trial-by-trial predictions of summed similarity between probes and the items on the study list. Nosofsky et al. demonstrated that the model was successful in accounting for variation in both choice probability and mean RT for each individual tested item in their Experiment 1, which were patches of Munsell colors. The EBRW was also successful in accounting for many other benchmarks of the Sternberg (1966) paradigm, including set size and serial position effects.

A question remains: if the work of Mewhort and Johns (2000) demonstrated that global matching models are not able to predict the feature composition effects of study lists, how then was the EBRW able to give such a strong account of individual items in their experiment? This question is especially relevant when one considers that, according to the EBRW and other models in the GCM family, the extent to which a lure is easy or difficult to reject depends entirely on its similarity to other members of the memory set. For instance, a lure that is difficult to reject is often near the representations of the list items in the similarity space, making it confusable with the list items and resulting in high

global similarity, whereas an easy lure is often further away from the representations of the list items and consequently has low global similarity. In other words, *the success of the EBRW in accounting for the variability in the memorability of individual items can be directly attributed to its reliance on global similarity computation.*

At first glance, the successes of the EBRW are at odds with the evidence of Mewhort and Johns (2000) against a global matching retrieval mechanism. However, despite the extensive replication, these demonstrations of the extralist feature effect have often involved separable-dimension stimuli such as colored shapes or words. Nosofsky et al.'s (2011) Experiment 1 stimuli were instead integral-dimension stimuli, in which each of the dimensions combine into a holistic stimulus that cannot be disentangled into its component dimensions. Nosfosky and Palmeri (1997) claimed that the EBRW should be restricted to experiments containing integral-dimension stimuli, as separable-dimension stimuli make it possible for participants to make decisions about each dimension independently, either in serial or in parallel, rather than combining them into a single percept. In fact, subsequent investigations using systems factorial technology (SFT: Townsend & Nozawa, 1995) have confirmed that participants in categorization tasks do exactly that — strong evidence has been found for serial or parallel processing of the stimulus dimensions with separable-dimension stimuli, whereas integral-dimension stimuli employ *coactive* decision architectures, in which information from each dimension is combined into a single processing channel to make a decision (Fific, Little, & Nosofsky, 2010; Fific, Nosofsky, & Townsend, 2008; Little, Nosofsky, & Denton, 2011; Little et al., 2013; Moneer, Wang, & Little, 2016; Griffiths, Blunden, & Little, 2017). One should note that the EBRW, as well as all other global matching models, is an example of a coactive decision architecture because information from stimulus dimensions and the memory set items are pooled into a single memory strength value that drives the decision process (Fific et al., 2010).

If categorization decisions about separable-dimension stimuli have been found to be based on the individual stimulus dimensions, then it is likely that a similar strategy is

possible with recognition memory decisions as well. That is, global similarity could be calculated for each stimulus dimension and a stimulus could be rejected if any of them are unfamiliar. What is particularly relevant for our purposes is the fact that a dimension that carries a novel value has lower global similarity than a stimulus dimension which contains a familiar value, which can potentially result in a rejection advantage for stimuli that contain extralist features. We will demonstrate later that this account of novelty rejection can be tractably formalized by making recognition decisions on the basis of each dimension, producing "old" decisions if all dimensions are recognized (e.g., an exhaustive decision rule), while the stimulus is rejected as new if a "new" decision is made about any of the dimensions (e.g., a self-terminating decision rule). The EB-LBA affords construction of a hybrid account where "new" decisions are made in the same way as described, but "old" decisions are driven by the global similarity of the entire stimulus. Integral-dimension stimuli, in contrast, do not allow for parallel or serial operations over dimensions because the dimensions cannot be disentangled from each other. Thus, the above account makes the novel prediction that *no extralist feature effect should be found if integral-dimension stimuli were employed in the Mewhort and Johns paradigm.*

An additional reason why there is enhanced novelty rejection for lures containing extralist features is that selective attention is directed to the stimulus dimensions carrying the extralist feature. The GCM family of models (which includes the EBRW and EB-LBA) depart from many other global matching models because of their emphasis on selective attention to the stimulus dimensions. Specifically, attention is divided between each of the component dimensions of a stimulus and can be unequally alotted between the dimensions. Devoting additional attention to a particular dimension has the effect of expanding that dimension in the geometric space, such that two stimuli that were previously proximal can become further apart in psychological distance. In other words, additional attention to a particular dimension means that matches on that dimension are more consequential in determining similarity, whereas matches on the unattended dimension have little influence

on the resulting similarity. Consequently, consider if in Figure 1, more attention was devoted to the dimensions that comprise shape than the dimensions that comprise color during the test phase. This has the consequence that matches on the shape carry more weight in determining similarity, such that the 2:0 probe's two matches on color would carry relatively little weight in the global similarity computation. The 1:1 probe, in contrast, contains one match on the shape (with high weight) and one match on color (with little weight), resulting in higher global similarity for the 1:1 probe than the 2:0 probe.

Nosofsky et al. (2011) performed model simulations with the EBRW and demonstrated that additional attention to dimensions that carry the extralist feature is sufficient to produce enhanced novelty rejection for such lures. However, they did not provide a mechanism for how selective attention could be allocated in this fashion. While attention weight parameters are commonly fit as free parameters in applications of the GCM family of models, the stimulus aspect or dimensions that carry the extralist feature vary across trials. Thus, while shape in Figure 1 carries the extralist feature, on a later trial color could carry the extralist feature. Thus, if extra attention was devoted to shape, the model would predict the opposite of an extralist feature effect on trials where color carries the extralist feature, as higher global similarity would be predicted for 2:0 probes than 1:1 probes. Thus, the attentional account requires a specification of how attention might vary on the basis of the memory set, the probe type, or some combination thereof on a trial-by-trial basis.

In this work, we additionally explore a novel account of selective attention where the attention weight to a particular dimension is determined on the basis of how diagnostic that particular dimension is in determining whether the stimulus is old or new. Specifically, the likelihood of the value of each dimension of the probe is compared against the distribution of values from the memory set, and attention is allocated toward the least likely values. This makes it such that novel values within the probe, such as extralist features, receive extra attention due to their relative degree of novelty. A selective

attention explanation of the extralist feature effect is uniquely allowed by separable-dimension stimuli, given that individual stimulus dimensions cannot be easily disentangled in integral-dimension stimuli. There are numerous examples demonstrating that learning which dimensions to selectively attend occurs easily for separable dimensions (Nosofsky, 1986; Shepard, Hovland, & Jenkins, 1961) but not integral dimensions (Nosofsky, 1987; Nosofsky & Palmeri, 1996a). A novel theoretical implication of our work is that the extralist feature effect also belongs to the class of findings differentiating separable and integral dimensions (see Griffiths et al., 2017, for a review). An advantage of the GCM family of models — of which the EB-LBA is an example — is that it allows for implementations of attentional explanations, as selective attention to stimulus dimensions has not yet been implemented in other global matching models. [1]

The Current Experiments

Across four experiments, we test for the presence of the extralist feature effect using integral- and separable-dimension stimuli. We used a novel design where stimuli are represented using three continuously-valued dimensions. An advantage of employing continuous-dimension stimuli is that they allow us to construct study lists where the distance from the probe to the items in the memory set can be balanced in a principled fashion between items which do and do not contain an extralist feature. In addition, it allows us to further manipulate stimulus difficulty in a graded and straightforward fashion — lures that should be easy to reject are located further from the representations of the study list items in the multidimensional space. As we will demonstrate later in the paper, this allows us to test some more fine-grained predictions about whether the extralist feature effect should or should not occur. Finally, our design also allows us to construct study lists and lure probes for both integral- and separable-dimension stimuli using the

[1] While models such as the SAM model have implemented selective attention to the context and item cues (Clark & Shiffrin, 1987; Gillund & Shiffrin, 1984), we know of no application with the SAM model where selective attention is divided between the dimensions within a particular item.

same underlying spatial configurations.

An example of our experimental paradigm can be seen in Figure 2A. In Experiments 1-4, study lists consisted of three stimuli constructed from three dimensions. On each trial, the three stimuli varied along two dimensions (X and Y in the figure) while the third dimension (Z in the figure) was fixed across each stimulus. What we term "standard" lures are lures that took on novel combinations of old and new values along the two varied dimensions but contained the same value on the fixed dimension. We constructed an analog to extralist feature ("ELF") lures by constructing stimuli that take a novel value along the fixed dimension. To manipulate stimulus difficulty, both the standard and ELF lures had two levels of distance (distance 1 and 2) to the studied list items

One should note that technically speaking, the standard lures in our continuous-dimension stimuli contain values that could be considered "extralist features" in that these values are not shared by the study list items – only the fixed dimension is identical to the study list items. This highlights an important difference between discrete feature and continuously-valued stimuli – whereas the experiments of Mewhort and Johns manipulated lure difficulty by varying whether features among lure probes are studied or unstudied, experiments with continuously-valued stimuli commonly construct lures with novel values and vary the degree of distance to the studied items, with shorter distances reflecting higher difficulty due to their greater confusability with the list items (e.g., Kahana, Zhou, Geller, & Sekuler, 2007; Visscher, Kaplan, Kahana, & Sekuler, 2007; Zhou, Kahana, & Sekuler, 2004).

Where the design of our continuous-dimension experiments is conceptually analogous to that of Mewhort and Johns (2000) is that our ELF lures (a) contain a value that has a considerably greater degree of novelty than the values among the standard lures, and (b) ELF lures and standard lures exhibit the same distance on average to the study list items in the same way that 2:0 and 1:1 lures in Mewhort and Johns design have the same number of feature matches to the list items, and (c) our lure difficulty manipulation is analogous to

how Mewhort and Johns manipulated the number of matching features in lure probes (e.g., the difference between 2:0 and 1:0 probes).

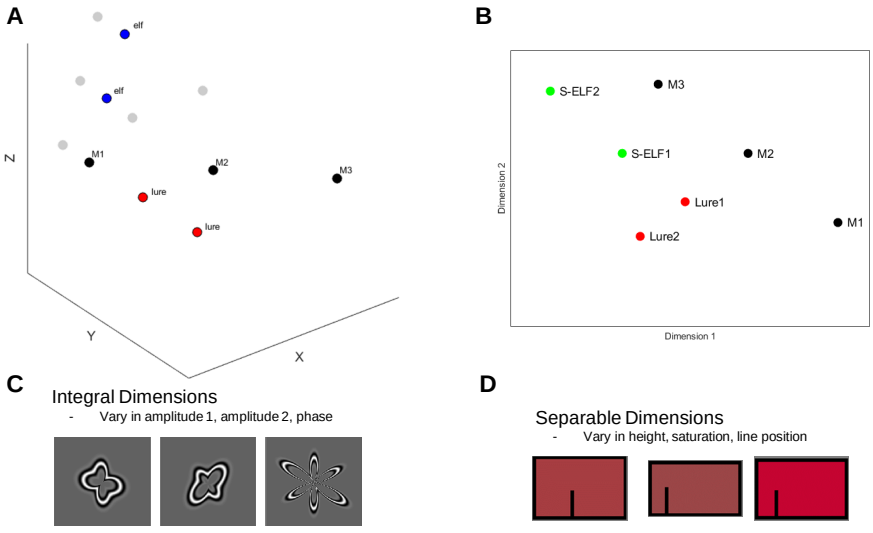

Figure 2. Examples of how continuous dimension study sets and lure probes were constructed (panel A). A two-dimensional spatial depiction of the construction of the study list items, the standard lures, and S-ELF lures (panel B). Example integral-dimension stimuli (panel C) and separable-dimension stimuli (panel D). See the text for details.

Our Experiments 1 and 2 bear a superficial resemblance to the extralist lures that were used by E. E. Johns and Mewhort (2002), where one aspect of the stimulus was fixed on the study list (such as color, shape, or texture) and extralist lures carried a novel value along that same aspect, whereas the original experiments by Mewhort and Johns (2000) created extralist features by introducing a new feature on one of the aspects (color or shape) that varied among the study list items. However, as mentioned, each of these stimulus aspects are likely represented with multiple dimensions or discrete features, so there is no guarantee that the extralist feature in the Mewhort and Johns experiments was always carried by the same psychophysical dimensions as those represented on the study list.

In addition to lures where the extralist feature is on a novel dimension (ELF lures),

in Experiments 2 and 4 we also test lures where the extralist feature is carried by one of the dimensions that varies among the study list items. We term these lures same-dimension ELF lures, or S-ELF lures for short. Specifically, for S-ELF lures, the third dimension is fixed to the same value as the fixed dimension on the study set, similar to the standard lures. However, while the standard lure probes in Experiments 2 and 4 are equidistant to the memory set items on both dimensions, the S-ELF lures exhibit values on one dimension that are proximal to the values of the dimensions of the memory set items, while the values on the other dimension are further away from the values of the dimensions of the memory set items. The S-ELF lure probes are continuous analogs to 2:0 lures in the Mewhort and Johns (2000) paradigm, because the probe dimension that is close to the memory set items represents a strong match on that dimension (similar to the twice presented feature in 2:0 probes) whereas the other dimension represents a weak match on that dimension (similar to the extralist feature in 2:0 probes). An example of an ELF trial and a S-ELF trial is shown in Figure 3 and Figure 9 while a spatial representation of S-ELF and standard lures can be seen in Figure 2B, where it can be seen that both S-ELF and standard lures lie on the same plane, contrary to the novel dimension ELF lures.

Figure 2C and 2D depicts examples of the integral-dimension (Experiments 1 and 2) and separable-dimension (Experiments 3 and 4) stimuli that we employ. One can see that the variations among the stimuli are rather subtle in comparison to the experiments of Mewhort and Johns and consequently, it would be rather difficult to construct labels for each individual stimulus. In each of our experiments the dimension that is fixed varies from trial-to-trial to minimize the probability that participants deduce the relevant dimension.

To foreshadow our results, we found no extralist feature effect in Experiments 1 and 2 with integral-dimension stimuli but found a substantial extralist feature advantage with the separable-dimension stimuli in Experiments 3 and 4 – a result which was unable to be accounted for by the EB-LBA model. We developed additional variants that employed the proposed modifications to novelty detection within the model, including parallel decision

architectures or dynamically varying attention weights, both of which are made possible with separable-dimension stimuli. The models were jointly constrained by the difference between standard and extralist lures, but additionally the difficulty of the lures (the two levels of distance) and the serial position effect — the finding that recently studied items considerably outperform older studied items (Corballis, 1967; Monsell, 1978). In addition, the models were constrained by how the complete RT distribution varies across the stimulus types and difficulty manipulations.

The usage of our continuous-dimension stimuli with psychophysically principled representations offers two advantageous constraints for our modeling. First, because the similarity space of the stimuli is known, the EB-LBA models are constrained by the degree of probe-item similarity on an individual trial basis. Second, because the number of psychophysical dimensions is known, this allows us to constrain the number of accumulators in parallel architectures, which involve a race between the stimulus dimensions of the probe. Appendix B shows the results of MDS studies we performed with a limited set of stimuli that show that both the integral- and separable-dimension stimuli we employ conform to the three-dimensional structure we use in our experiments. Furthermore, there was a very strong correspondence between the MDS estimated coordinates and the coordinates we employed in our experiments for separable-dimension stimuli. Correspondence was also found for integral-dimension stimuli but to a weaker degree.

To test the generality of our models to the original Mewhort and Johns (2000) design, we additionally performed a conceptual replication of the original Mewhort and Johns experiments with colored shapes that included 1:1, 2:0, 1:0, and 0:0 lures (Experiment 5). Because we lack psychophysically principled representations of these particular stimuli, our EB-LBA modeling relied on discrete feature matches and mismatches to each of the stimuli. Consequently, all mismatches on a particular stimulus dimension all have the same weight. We found that the models endorsed in our separable-dimension stimuli experiments similarly performed well with the colored shapres in this experiment, suggesting a

generality of the account across both continuous and discrete features.

Data from each experiment along with the experiment code and model code can be found on our OSF page (`https://osf.io/b2zyk/`) (Osth, Little, & Lilburn, 2019).

Experiments 1 and 2: Integral-dimension stimuli

Experiments 1 and 2 used the same stimuli and study list construction. The only difference is that Experiment 2 additionally tested same-dimension ELF (S-ELF) lures in addition to the novel dimension ELF lures and the standard lures. We compensated for the greater number of trial types in Experiment 2 by having participants complete one additional experimental session (four sessions in Experiment 2, as opposed to three sessions in Experiment 1).

Method

Participants. Participants were 31 members of the University of Melbourne community (15 in Experiment 1, 16 in Experiment 2), with normal or correct-to-normal vision. In Experiment 1, participants completed in three one-hour sessions, with the exception of one participant who discontinued before completion of the first session and was excluded from the analysis. In addition, three participants were not able to complete all three sessions: one participant only completed one session, while the two others completed two, but their data was included in the analysis.

In Experiment 2, participants completed four one-hour sessions. Three participants did not complete all four sessions: one participant completed only a single session while the other two completed two sessions. All three participants' data were still included in the analyses. Due to the impact of COVID-19, we were not able to test all of the participants in Experiment 2 in an isolated testing booth. Instead, several of our participants downloaded the experimental software and performed the experiment on their home computers. Our integral dimensions all involve aspects of shape and should not be affected

too severely by changes in lighting conditions or monitor position (e.g., unlike color dimensions which would vary greatly between viewing conditions).

Participants were renumerated at a rate A$15 per session. Human testing was approved by the Melbourne Human Research Ethics Committee (Approval number: 1034866).

Materials

The stimuli were blob-shaped radial frequency patterns (see Figure 2) generated by convolving three sine waves plotted in polar coordinates. One sine wave varied in amplitude and phase angle and a second varied only in phase angle. The amplitude of the second and third sine waves was fixed at .5, and the phase angle of the third sine wave was fixed at 0. Previous experiments using these stimuli have found results consistent with the notion that these dimensions are integral (Op de Beeck, Wagemans, & Vogels, 2003; Cortese & Dyre, 1996; Lewandowsky, Roberts, & Yang, 2006). For instance, Lewandowsky et al. (2006) (2006) found that similarity ratings of pairs these stimuli were better captured by a Euclidean distance metric rather than a city-block distance metric. Op de Beeck and colleagues found no evidence of dimensional learning for these dimensions consistent with integrality (Op de Beeck et al., 2003).

Study list items varied on three dimensions: Amplitude, Phase Angle 1, and Phase Angle 2. Table 1 shows the range of values that were used to generate the study list items and lures.[2] Study list generation is described in Appendix C. After setting the values of the study list, the order of study list presentation was then randomized. Ignoring

[2]To determine the psychologically differentiable step-size within each dimension (i.e., the just-noticiable difference, JND), we used the psi method (Kontsevich & Tyler, 1999; Prins, 2006) in a separate calibration study. Two participants were tested, and their JNDs were similarly found to be .065 for amplitude and .12 radians for phase angle. See Appendix A for more details. Although we varied the physical coordinates of the stimuli, we confirmed using multidimensional scaling (MDS) that these dimensions captured the psychological representation of the stimuli (see Appendix B).

presentation order, there were 697 unique study lists in Experiment 1 and 742 unique lists in Experiment 2.

Table 1

Range of values (in steps of JND) used to generate study list items and lures.

	Integral Dimensions		Separable Dimensions		
Index	Phase Angle	Amplitude	Saturation	Height	Bar Position
1	-1.92	0.13	2	12	24
2	-1.68	0.26	4	24	36
3	-1.44	0.39	6	36	48
4	-1.20	0.52	8	48	60
5	-0.96	0.65	10	60	72
6	-0.72	0.78	12	72	84
7	-0.48	0.91	14	84	96
8	-0.24	1.04	16	96	108
9	0.00	1.17	18	108	120
10	0.24	1.30	20	120	132
11	0.48	1.43	22	132	144
12	0.72	1.56	24	144	156
13	0.96	1.69	26	156	168
14	1.20	1.82			
15	1.44	1.95			

Note: Step size for the integral dimension stimuli in Experiment 2 was 4 × JND; hence, the values in this table should be multiplied by 2 for Experiment 2.

In Experiment 1, test items were either (a) old target items (sampled randomly from the study list), (b) new items in the with the same fixed dimension as the memory set

(standard lure probes), or (c) novel dimension extralist feature (ELF) lure probes. Lures were generated at two distances from the memory set: for distance 1, the lure had a minimum Euclidean distance of 2 JNDs to any list item. For distance 2, the lure had a minimum Euclidean distance of 4 JNDs to any list item.

Standard lures were generated by extending a vector perpendicular to the vector formed by the study list within the plane formed by the fixed dimension. For instance, if the variable study list dimensions had index values of [4, 7], [6, 5], and [8, 3] (see the second to last row of Table C1), the varied dimensions of the lure would have index values of [6.59, 1.59] (or at the same distance but in the other perpendicular direction, [9.41, 4.41]). These indices were linearly interpolated into the values listed in Table 1 to set the dimensional values for the lure. The fixed dimension of the standard lure was therefore equal to the fixed dimension of the study list. The length of the vector was either 2 × the JND (distance-1) or 4 × the JND (distance-2). Further representative lists are shown in Table 2.

ELF lure probes were generated by projecting a vector of the same length as the standard lures but perpendicular (normal) to the plane of the study list items. Hence, the distance of the extralist feature lures to the study list was adjusted to be equivalent to the distance of the lures to the memory set (see Table 2).

In Experiment 2, to generate the standard and S-ELF lures for each distance, we generated all possible dimensional combinations that were of a specific distance from the memory set space (i.e., that varied in the same dimensions varying in the memory set). Standard lures were chosen such that the absolute difference between the distances for dimension 1 and dimension 2 was equal (less than .05 JND). For S-ELF lures, the absolute difference between the distances for dimension 1 and dimension 2 was constrained to be asymmetric (greater than .05 JND). Hence, standard lures have roughly equivalent distance to the study list on both dimensions (and hence, roughly equivalent similarity); S-ELF lures have asymmetric similarity to the study list on both dimensions. ELF lures were found in the same way as Experiment 1, first by projecting a vector normal to the

plane of the study list and then adjusting the length of that vector using an optimization procedure to vary the fixed dimension distance, equating it to the standard and S-ELF lure distance. Examples of each of the probe types are shown in Figure 3.

Table 2

Representative lists and lure indices for Experiments 1 and 2

Integral Dimensions							
Distance	Fixed Dimension	M1	M2	M3	Lure	Type	sum(1/d)
1	1	[4, 3, 8]	[4, 5, 6]	[4, 7, 4]	[4, 6, 7]	standard	1.00
1	1	[4, 3, 8]	[4, 5, 6]	[4, 7, 4]	[4, 7, 6.4]	S-ELF	1.01
1	1	[4, 3, 8]	[4, 5, 6]	[4, 7, 4]	[5.55, 5, 6]	ELF	1.01
1	2	[4, 3, 6]	[6, 3, 8]	[8, 3, 10]	[6.8, 3, 7.8]	standard	1.01
1	2	[4, 3, 6]	[6, 3, 8]	[8, 3, 10]	[2.6, 3, 6]	S-ELF	1.01
1	2	[4, 3, 6]	[6, 3, 8]	[8, 3, 10]	[6, 4.55, 8]	ELF	1.01
1	3	[3, 4, 10]	[5, 6, 10]	[7, 8, 10]	[3.6, 7.4, 10]	standard	0.86
1	3	[3, 4, 10]	[5, 6, 10]	[7, 8, 10]	[6, 4, 10]	S-ELF	0.87
1	3	[3, 4, 10]	[5, 6, 10]	[7, 8, 10]	[7, 8, 11.725]	ELF	0.86
2	1	[7, 4, 10]	[7, 6, 8]	[7, 8, 6]	[7, 9, 5]	standard	0.77
2	1	[7, 4, 10]	[7, 6, 8]	[7, 8, 6]	[7, 4, 12]	S-ELF	0.77
2	1	[7, 4, 10]	[7, 6, 8]	[7, 8, 6]	[9.24, 6, 8]	ELF	0.77
2	2	[6, 10, 5]	[8, 10, 7]	[10, 10, 9]	[10.4, 10, 4.6]	standard	0.63
2	2	[6, 10, 5]	[8, 10, 7]	[10, 10, 9]	[10, 10, 4.2]	S-ELF	0.63
2	2	[6, 10, 5]	[8, 10, 7]	[10, 10, 9]	[10, 12.63, 4.6]	ELF	0.63
2	3	[4, 8, 6]	[6, 6, 6]	[8, 4, 6]	[3, 9, 6]	standard	0.77
2	3	[4, 8, 6]	[6, 6, 6]	[8, 4, 6]	[8, 7.8, 6]	S-ELF	0.76
2	3	[4, 8, 6]	[6, 6, 6]	[8, 4, 6]	[6, 6, 8.24]	ELF	0.77

Participants were tested in sound- and light-attenuated testing booths on Linux OS machines running Matlab and Psychophysics toolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). A mid-gray background (RGB: [128 128 128]) was used. Response times were recorded using a calibrated RT Box (Li, Liang, Kleiner, & Lu, 2010).

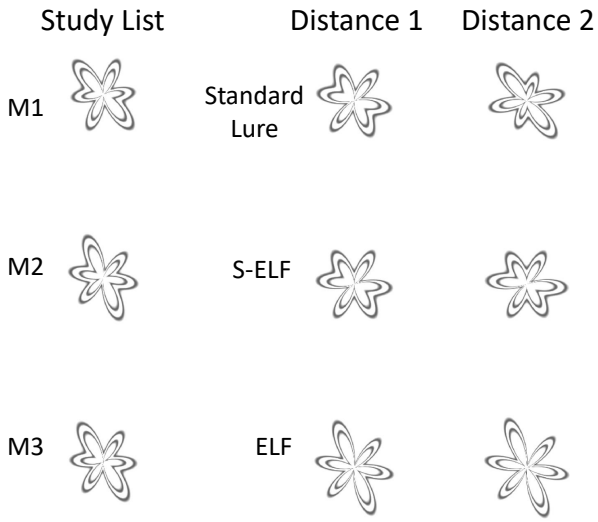

Figure 3. Illustration of the list types used in Experiments 1 and 2. The members of the study list are on the left, while the right column depict the lure types. Refer to the text for descriptions of how the lures were generated.

Procedure

On each trial, a fixation cross was presented for 1 s, followed by a blank interval of 1 s. Study items were then presented sequentially for 1.75 s each with an intervening interval of .25 s. Following the presentation of the last item, a second fixation cross was displayed for 1 s followed by the test item, presented until a response was made (or until 10 s).

In Experiment 1, each session yielded 336 trials: 168 target trials along with 42 trials of each of the four lure types (standard lures at distance 1 and 2 along with extralist feature lures at distance 1 and 2). For the target trials, there were an equal number (i.e., 56) trials in which the test item matched the first, second, or third presented study item. The fixed study set dimension was varied randomly from trial to trial with a total of 112 trials for each fixed dimension.

In Experiment 2, in each session there were 189 target trials and either 32 trials of each lure type (standard lures at distance 1 and 2, ELF lures at distance 1 and 2, and S-ELF lures at distances 1 and 2) or 31 trials of each lure type. There were 2 sessions of

each frequency making 4 sessions in total of either 381 or 375 total trials. For the target trials, there were 63 trials in which the test item matched the first, second, or third presented study item. The fixed study set dimension was varied randomly from trial to trial with a total of 127 or 125 trials for each fixed dimension depending on the session.

This study (and all other experiments in this article) was not preregistered.

Results

Experiment 1. One participant was excluded from the analysis for performing extremely close to chance levels ($d' = .017$). In addition, responses with RTs less than .2 or greater than 4 seconds were excluded, which resulted in the exclusion of 1.94% of the data.

Results for Experiment 1 can be seen in Figure 4, with the top row (A) showing the results for each target serial position and lures separated by distance. One can see that error rates are considerably higher than in many studies in the Sternberg procedure where study lists are composed of digits or letters, where accuracy is usually at ceiling (e.g., Sternberg, 1966; Monsell, 1978). However, it is important to note that when studies have been conducted with confusable continuous dimension stimuli, error rates are considerably higher and within the range found in these experiments (e.g, Kahana & Sekuler, 2002; Nosofsky & Kantner, 2006; Nosofsky et al., 2011).

A typical recency effect was found, with lower error rates found for the third and final item ($M = .169, SEM = .023$) from each study list as opposed to the second ($M = .330, SEM = .023$) or first item ($M = .422, SEM = .022$), which is consistent with many previous studies using the Sternberg paradigm (e.g., Corballis, 1967; Monsell, 1978).

We focused on responses to lure stimuli using 2 x 2 x 3 ANOVAs with lure difficulty (distance 1 or 2), lure type (standard and ELF), and fixed dimension (amplitude of sine wave 1, phase angle of sine wave 1, and phase angle of sine wave 2) as factors. The critical effects in the data were most evident in the error rates ($p(error)$). Distance 1 lures exhibited much higher false alarm rates (FAR, $M = .512, SEM = .022$) than lures at

distance 2 ($M = .269, SEM = .031$), $BF_{10} = 6.846e + 20$. No extralist feature effect was found, as roughly equivalent FAR were observed for standard lures ($M = .393, SEM = .025$) and ELF lures ($M = .389, SEM = .024$), $BF_{10} = .170$. The dimension that was fixed in the study set had little effect on the FAR, $BF_{10} = .403$, with similar FAR for phase-1 ($M = .408, SEM = .026$) and phase-2 ($M = .416, SEM = .029$) dimensions, although FAR was noticeably lower for the amplitude-1 dimension ($M = .349, SEM = .020$).

An interaction between lure type and the fixed dimension was found, as evidenced by the fact that the model that included such an interaction along with all three main effects was the preferred model, $BF_M = 18.706$. The results separated by each fixed dimension can be seen in Figure 4B. The interaction is due to the fact that there was an extralist feature effect when amplitude-1 was the fixed dimension (lure FAR $M = .458, SEM = .026$, ELF FAR $M = .242, SEM = .021$), whereas there was a *inverse* extralist feature effect for phase-1 (lure FAR $M = .341, SEM = .031$, ELF FAR $M = .475, .025$) and phase-2 (lure FAR $M = .379, SEM = .027$, ELF FAR $M = .454, SEM = .035$) were the fixed dimensions. Recall from the Introduction that extra attention to one of the stimulus dimensions produces exactly this pattern — an extralist feature effect on the attended dimension, along with an inverse extralist feature effect on the other dimensions. Below, we discuss converging evidence from both the data and the modeling that supports this possibility.

Parallel analyses were conducted on the mean RTs on correct trials. Mean RTs were longer for distance 1 ($M = 1.21$, $SEM = .078$) than distance 2 lures ($M = 1.12$, $SEM = .063$), $BF_{10} = 37.92$. Mean RTs were nearly equivalent for ELF lures ($M = 1.14$, $SEM = .065$) than standard lures ($M = 1.17$, $SEM = .073$), $BF_{10} = .276$. Mean RTs were slightly different depending on the fixed dimension ($M_{amplitude-1} = 1.10$, $SEM_{amplitude-1} = .068$, $M_{phase-1} = 1.20$, $SEM_{phase-1} = .074$, $M_{phase-2} = 1.17$, $SEM_{phase-2} = .066$), although the Bayes Factor revealed only weak evidence for this

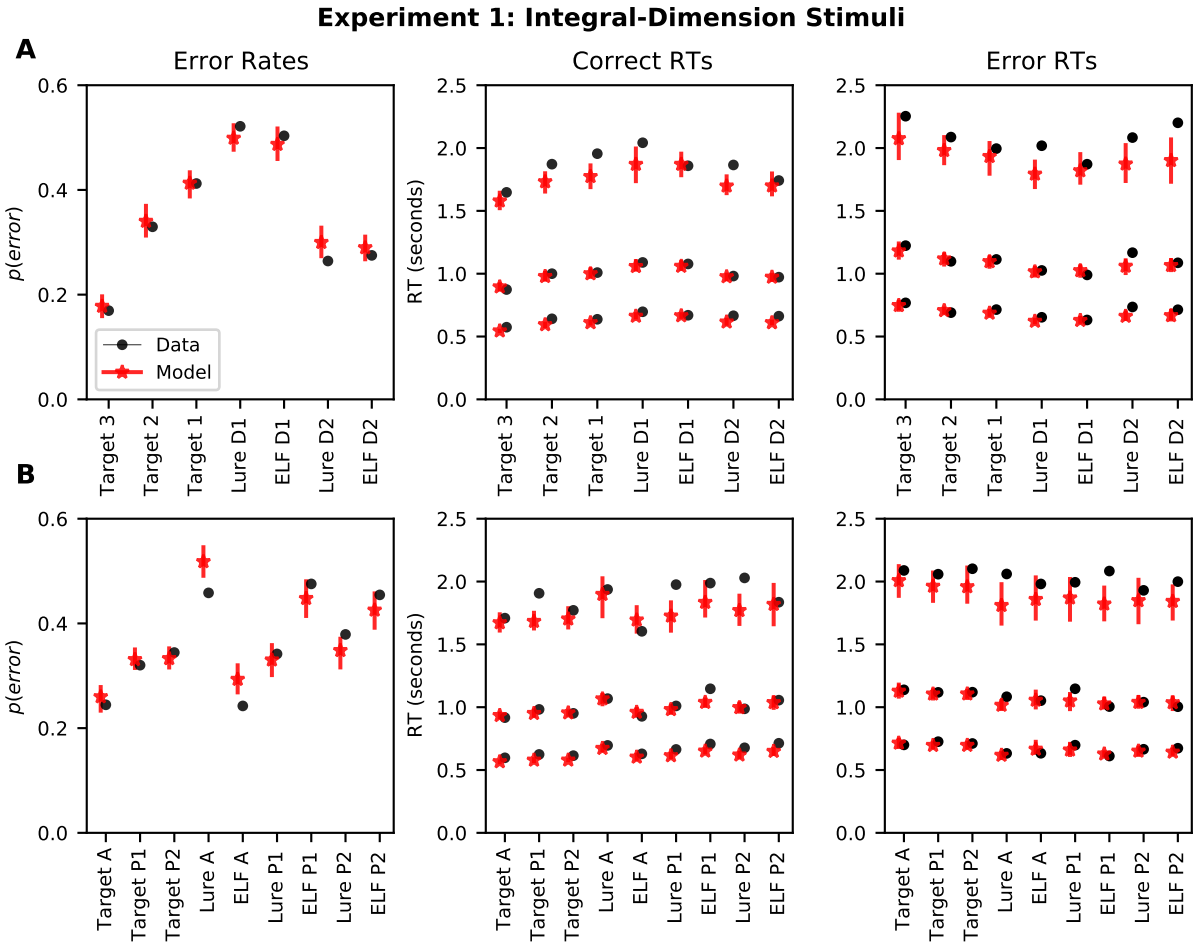

Figure 4. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from from Experiment 1 with integral-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictive distributions from the EB-LBA model. Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, A = amplitude of sine wave 1, P1 = phase angle of sine wave 1, P2 = phase angle of sine wave 2.

difference, $BF_{10} = 1.619$.

These data appear to be congruent with the predictions of global matching models. No extralist feature effect was found — instead, memory performance was well described by the global similarity of the probes to the list items, as evidenced by the very large effect on performance of the distance between the lures and the study set items. In the next section, we describe the details of the EB-LBA model and its application in order to evaluate its quantitative correspondence with the data.

Experiment 2. Due to a bug in the experimental code, on 487 trials (1.37% of the data) no probe stimulus was presented. These trial were removed prior to further anlaysis. Three participants were excluded for exhibiting a very large proportion of slow RTs (>20% of responses slower than 4 seconds) and an additional participant was excluded for exhibiting poor performance ($d' = .16$). The large proportion of slow RTs could be due to poor compliance from the participants performing the experiments on their home computers.

Results can be seen in Figure 5. FAR were considerably higher for distance 1 lures ($M = .469$, $SEM = .0239$) than distance 2 lures ($M = .214$, $SEM = .0244$), $BF_{10} = 1.30 \times 10^{28}$. Consistent with the results of Experiment 1, there were no differences in FAR between the different lure types, $BF_{10} = .670$. While the Bayes Factor did not yield decisive evidence for a null effect, inspection of Figure 5 reveals that while the FAR are approximately equal for each lure type for the distance 2 lures, there do appear to be FAR differences for the relatively difficult distance 1 lures. However, they are in the opposite direction of an extralist feature effect: compared to the standard lures ($M = .421$, $SEM = .0249$), FAR are higher for the novel-dimension ELF lures ($M = .474$, $SEM = .0248$) and the same-dimension ELF lures ($M = .513$, $SEM = .0329$). This observation is corroborated by support for an interaction between lure type and distance, which is indicated by the fact that such an interaction is included in the preferred model, which additionally includes an interaction between lure type and the fixed dimension, in

addition to main effects of lure type, distance, and dimension type ($BF_M = 23.582$ for the preferred model).

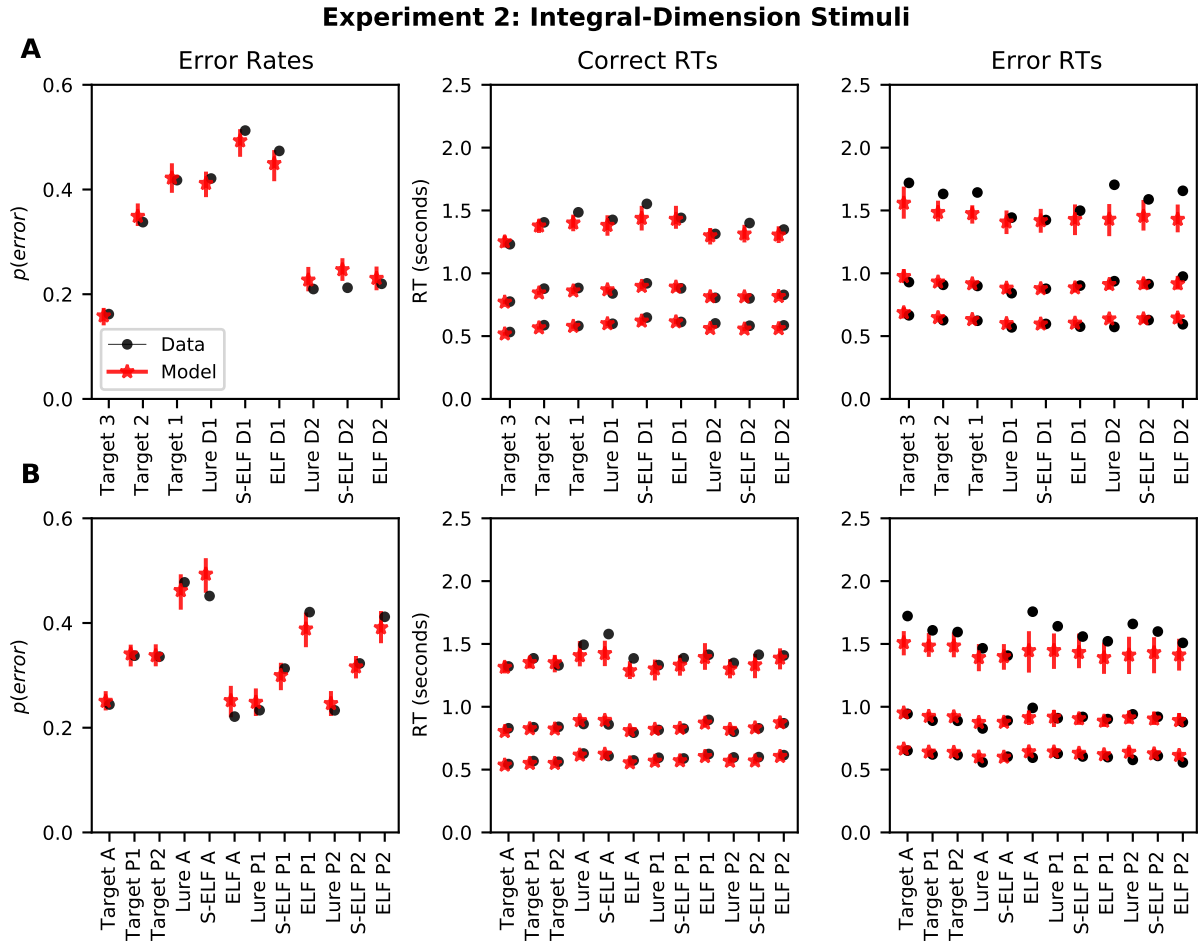

Figure 5. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from from Experiment 2 with integral-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. Model predictions are group-averaged posterior predictive distributions from the core EB-LBA model. Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, A = amplitude of sine wave 1, P1 = phase angle of sine wave 1, P2 = phase angle of sine wave 2.

The interaction between lure type and the fixed dimension is consistent with the

results of Experiment 1, where higher hit rates and false alarm rates were found for probes where the amplitude was the fixed dimension. As can be seen in Figure 5B, we found similar results here — when amplitude-1 was the fixed dimension, FAR were substantially lower for novel-dimension ELF lures ($M = .211$, $SEM = .0269$) than for same-dimension ELF stimuli lures ($M = .452$, $SEM = .0251$) and standard lures ($M = .478, SEM = .0282$). However, the opposite results are found when phase-1 or phase-2 was the fixed dimension: FAR to novel-dimension lures were highest when either phase-1 ($M = .416$, $SEM = .0273$) or phase-2 ($M = .400$, $SEM = .0228$) was the fixed dimension, relative to same-dimension ELF lures (phase-1 $M = .313$, $SEM = .0217$, phase-2 $M = .323$, $SEM = .0280$) and the standard lures (phase-1 $M = .233$, $SEM = .0228$, phase-2 $M = .233$, $SEM = .0319$). As with Experiment 1, we will demonstrate below that these results are due to a great deal of attention being placed on the amplitude-1 dimension. Evidence for this stems from the fact that hit rates are higher to targets when amplitude-1 was the fixed-dimension ($M = .756$, $SEM = .0450$) compared to the other two dimensions (phase-1 $M = .662$, $SEM = .0363$, phase-2 $M = .664$, $SEM = .0430$).

Similar to Experiment 1, analyses of the mean RTs of correct trials revealed many of the same trends as the error rates. Mean RTs were longer for distance 1 ($M = .971$, $SEM = .0375$) than distance 2 ($M = .908$, $SEM = .0418$) lures, $BF_{10} = 16,094$. Little difference was found in the mean RT to the different lure types – standard lures ($M = .914$, $SEM = .0412$), ELF lures ($M = .937$, $SEM = .0394$), and S-ELF lures ($M = .945$, $SEM = .0406$) all exhibited similar mean RTs, although the Bayes Factor did not reveal strong evidence for equivalence, $BF_{10} = .355$. Unlike Experiment 1, no differences in mean RT were found for each fixed dimension, $BF_{10} = .169$.

A strong recency effect was again found, with the lowest error rate found for targets in the third serial position ($M = .162$, $SEM = .0465$) compared to the second ($M = .338$, $SEM = .0423$) and first ($M = .418$, $SEM = .038$) position.

EB-LBA Modeling

The data from both Experiments 1 and 2 appear to be congruent with the predictions of global matching models. No extralist feature effect was found — instead, memory performance was well described by the global similarity of the probes to the list items, as evidenced by the very large effect on performance of the distance between the lures and the between the probe and the study set items. In this section, we describe the details of the EB-LBA model and its application in order to evaluate its quantitative correspondence with the data.

Description of the model. A complete illustration of each of the EB-LBA model can be found in Figure 6. In the EB-LBA model, each item on our study list is represented as a point in a three-dimensional space (a simplified two-dimensional Euclidean space is depicted in Figure 6 for the ease of visualization). In Experiments 1 and 2, these dimensions correspond to amplitude 1, phase angle 1, and phase angle 2, and the values of each of these dimensions are the same as the indices used to generate the stimuli. Each exemplar is assumed to be encoded perfectly into memory. At test, the probe stimulus i is compared to each of the exemplars stored in memory. This comparison process involves the calculation of the distance between the probe i and the exemplar in memory j, which is denoted as:

$$d_{ij} = [\sum_{k=1}^{3} w_k |x_{ik} - x_{jk}|^{\rho}]^{1/\rho} \quad (1)$$

where x_{ik} is the value of the ith item's dimension k. ρ defines the distance metric of the space, with $\rho = 2$ for Euclidean distance and $\rho = 1$ for city-block distance. We follow previous convention and employ Euclidean distance for integral-dimension stimuli and city-block distance for separable-dimension stimuli (e.g., Garner, 1974; Nosofsky, 1984; Shepard, 1964). w_k ($0 < w_k$, $\sum w_k = 1$) is the attention weight given to dimension k. Changes in the attention weights have the effect of expanding or decreasing the attention

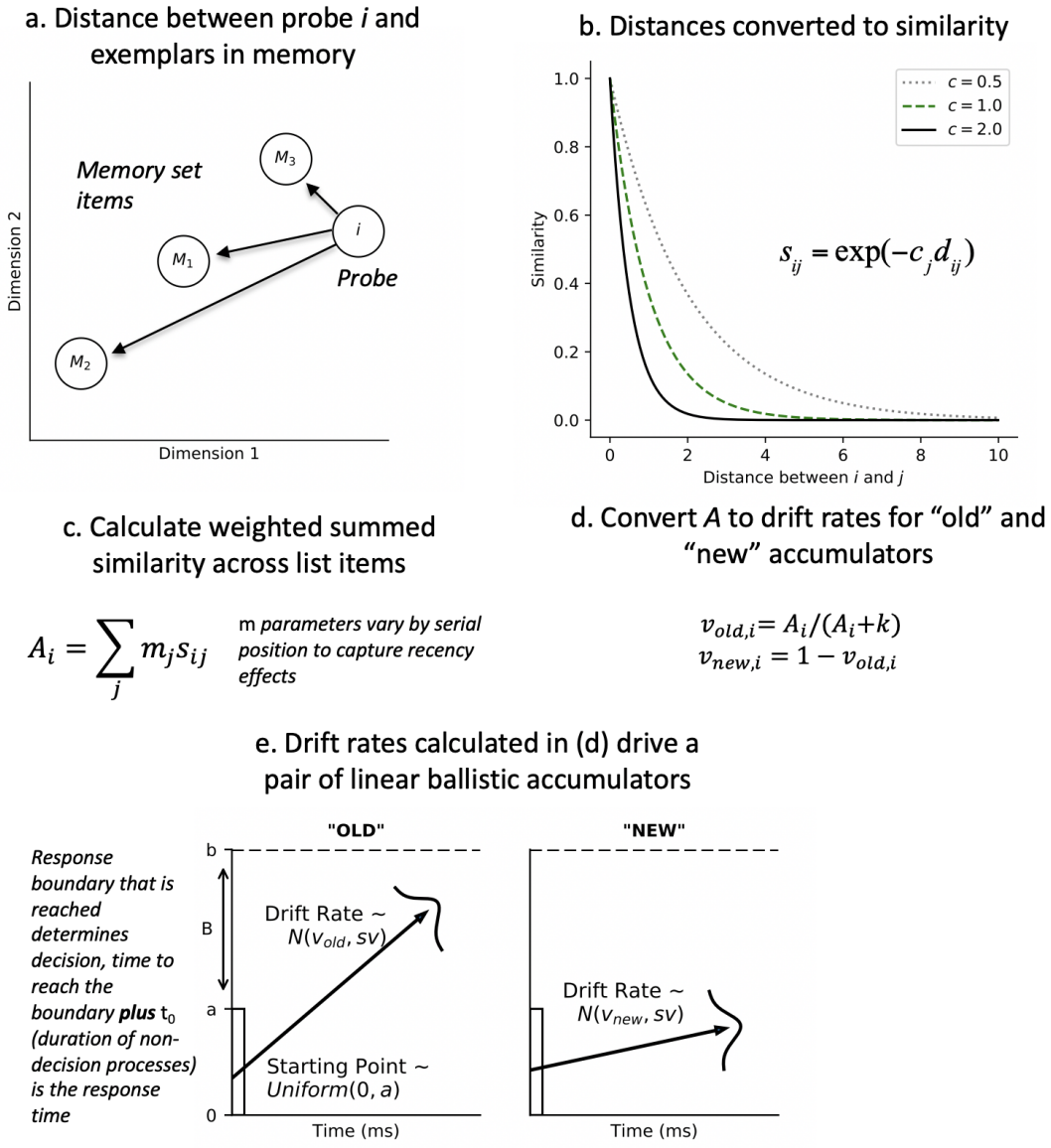

Figure 6. Illustration of the EB-LBA model using a simplified two-dimensional space. This includes (a) calculating the distance between the probe i and the members of the study set (M_1, M_2, and M_3), (b) conversion of the distances to similarities using an exponential transformation, (c) calculating a weighted summed similarity, (d) converting the weighted summed similarity values to mean drift rates, and (e) using the mean drift rates to drive a pair of linear ballistic accumulators to produce predictions about choice and response time. See the text for more details.

given to a given dimension, with a given dimension having no effect on the similarity calculation as w_k approaches zero.

The similarity of the test probe i to a studied item j is an exponentially decreasing function of their psychological distance (Shepard, 1987):

$$s_{ij} = \exp(-c_j d_{ij}) \tag{2}$$

c_j is a sensitivity parameter associated with item j. The sensitivity parameter governs the steepness of the exponential function — examples depicting three different values of c are depicted in panel B of Figure 6. As c increases, memory set items that are further from the probe stimulus yield similarity values that are closer to zero. For instance, in Figure 6B one can see that when $c = 2.0$, all distances of 2.0 or greater result in a similarity value that is close to zero. However, when $c = .5$, there are still substantial differences in similarity between distances of 2.0 ($s = .367$) and 4.0 ($s = .135$). Thus, as c increases, the summed similarity computation is dominated by exemplars that are closer to the probe. Because the distance between an item and itself is always zero, the similarity between a probe item and its own representation in memory is always one regardless of the value of c. We follow previous convention and allow c to vary with serial position (Nosofsky et al., 2011). We obtained reasonable fits with just two parameters: one for the most recent item (c_3) and another for the first two items ($c_{1,2}$)

Similarity values are subsequently converted to summed similarity A of the probe item i using the following equation:

$$A_i = \sum_j m_j s_{ij} \tag{3}$$

where m_j is the memory strength associated with serial position j. Memory strength parameters vary by serial position to capture recency effects in recognition memory tasks (e.g., McElree & Dosher, 1989; Monsell, 1978). It is conventional to fix one of the m

parameters to 1.0 – this was done for the most recent item, while m_1 and m_2 were estimated as free parameters that were constrained to be less than 1.0. The fact that information is pooled across both the dimensions and each of the memory exemplars into a single underlying decision variable (A) is why the standard implementation of the EB-LBA can be considered a coactive model.

Activation values were subsequently converted to drift rates for the "old" and "new" accumulators using the following equation:

$$v_{old,i} = A_i/(A_i + K) \tag{4}$$

$$v_{new,i} = 1 - v_{old,i} \tag{5}$$

where K is a drift criterion parameter. Increases in A or decreases in K have the effect of increasing drift rate for the "old" accumulator while decreasing the drift rate for the "new" accumulator. Thus, even though there are two independent accumulators, increases in global similarity produce increases in v_{old} and simultaneous reductions in v_{new}, producing more frequent and faster "old" decisions.

The drift rate parameters are used to drive a pair of linear ballistic accumulators (LBA: Brown & Heathcote, 2008), which are illustrated in panel E of Figure 6. In the LBA, each response option is associated with its own accumulator that is racing toward its associated response threshold, denoted by b. Whichever response's threshold is reached first results in the corresponding response being made, and the time taken to reach the threshold is the RT plus some additional time t_0 which corresponds to the time taken to encode the stimulus and make the associated motor response (nondecision time). Evidence accumulation is linear and noiseless: the slope of the line is denoted by the drift rate, which is a sample from a normal distribution with mean v and standard deviation sv. Evidence begins to accumulate at a randomly sampled point from a uniform distribution with height a (we adopt the lowercase notation to avoid confusion with the activation values in Equation 3). B corresponds to the distance between the top of the starting point

distribution and the response boundary b, and provides for more efficient parameter estimation because it prevents any cases where $a > b$. In all of our fits, separate B parameters were estimated for the "old" and "new" accumulators to allow for bias.

The combination of variability in starting point (a) and drift rate over trials (sv)[3] allows the LBA to capture the shapes of RT distributions for correct and error responses across a range of conditions, and the LBA has generally been successful in accounting for data from recognition memory paradigms (Donkin & Nosofsky, 2012b, 2012a; Osth, Bora, Dennis, & Heathcote, 2017; Rae, Heathcote, Donkin, Averell, & Brown, 2014). The EB-LBA model more specifically has been successful at accounting for data from short-term recognition memory (Donkin & Nosofsky, 2012a), categorization (Little et al., 2016), and long-term recognition memory and inductive generalization (Hawkins et al., 2016).

Applying the Model to Data. Throughout the article, we applied the model to data using hierarchical Bayesian techniques. A distinct advantage of hierarchical Bayesian implementation of cognitive models is that it allows for the simultaneous estimation of group- and participant-level parameters. This avoids averaging artifacts associated with fitting group data (e.g., Estes & Maddox, 2005; Liew, Howe, & Little, 2016) but simultaneously allows for "pooling" information across individuals, such that parameter estimates from one participant are influenced by the group-level parameters (Shiffrin, Lee, Kim, & Wagenmakers, 2008). Parameters were sampled using differential evolution Markov chain Monte Carlo (DE-MCMC) methods, which are robust to correlations among model parameters (Turner, Sederberg, Brown, & Steyvers, 2013).

The model was fit to the responses and RTs from all individual trials in the

[3]It may seem somewhat strange that there is variability in drift rate that is not generated by memory retrieval in the model. However, in a previous application of the EBRW, Nosofsky and Stanton (2006) found that additional variability outside of the variability in summed similarity across items was necessary to account for errors that are slower than correct responses. In that work, they were able to produce such variability by varying the sensitivity parameter c across trials within a mixture model. However, the usage of the ad hoc sv parameter within the LBA can accomplish the same goal in a simpler fashion.

experiment. Specifically, in each iteration of MCMC, the likelihood of each response and RT was calculated according to the LBA's likelihood function (where the mean drift rates for the LBA are determined by the equations described above), similar to maximum likelihood estimation (MLE) — closer matches between the data and model predictions are reflected in higher likelihoods. However, hierarchical Bayesian parameter estimation differs from MLE in three important ways. First, the sum of the log likelihood of each data point is summed with the log likelihood of the participant parameters under their respective group-level distributions. This is how the "pooling" occurs in hierarchical Bayesian models: higher likelihoods are assigned to participant parameters that are closer to the group-level distributions. Second, likelihoods of the parameters that define the group-level distributions are also calculated on each iteration. Third, MCMC does not result in a single set of parameter values with the highest likelihood, but instead collects a *distribution* of parameters referred to as the posterior distribution, where parameter values with higher likelihoods are represented more strongly in the posterior distribution. In addition, the width of a posterior distribution for a given parameter corresponds to the uncertainty with which the given parameter is estimated. A complete description of the technical details of applying the models to data can be found in Appendix D. Readers seeking more detail on MCMC methods should consult van Ravenzwaaij, Cassey, and Brown (2018).

The representations of both the studied items and lure probes were specified using three-dimensional representations. The coordinates were represented as values along ideal orthogonal dimensions scaled in terms of the JNDs. 12 parameters were estimated for each participant, including two memory strength parameters (m_1 and m_2), two sensitivity parameters ($c_{1,2}$ and c_3), two attention weight parameters ($w*_1$ and $w*_2$, see Appendix D for how w can be derived from the $w*$ parameters), the drift criterion K, and five LBA parameters (sv, B_{yes}, B_{no}, a, and t_0). While this may seem like a large number of parameters, *no parameters explicitly varied across the different lure types in the experiment.*

Model Fitting Results for Experiments 1 and 2. Group-averaged posterior predictives from the EB-LBA to the data from Experiments 1 and 2 can be seen in Figure 4 and 5, respectively, which not only depict predicted choice proportions for each trial type (targets from each serial position, standard lures, ELF lures, and S-ELF lures in Experiment 2) but also depicts predicted correct and error RT distributions for each trial type. Throughout the article, we summarize the RT distributions using the .1, .5, and .9 quantiles, which respectively are the 10th, 50th, and 90th percentiles. One should note that these quantiles are purely used to summarize the data and posterior predictives — the fitting procedure instead fits each individual response and RT combination from each participant. Uncertainty in the group averages is depicted using the 95% highest density interval (HDI).

Figures 4A and 5A reveal that the model appears to provide an excellent account of the choice probabilities and predicts no extralist feature effect: roughly equivalent FAR for standard lures and ELF lures are predicted at both distances due to both trial types having equal global similarity. For Experiment 2, the EB-LBA also captures the equivalent FAR between standard lures and S-ELF lures. In addition, the model is generally providing a strong account of the RT distributions for each trial type, although the model appears to underpredict the slowness of the errors for the distance 1 lures. In addition, the model capably addresses the recency effect seen in the data, demonstrating higher HR and faster RT for targets in the third serial position as opposed to targets in the first or second serial position. A minor shortcoming is that the model underpredicts the slowest correct RTs (the .9 quantile) for targets in the second and third serial position.

Analyses of Experiments 1 and 2 revealed that there was little difference in mean RT between each of the lure types. The EB-LBA's posterior predictives in Figures 4 and 5 demonstrate that the model predicts very little difference in the .1 and .5 quantiles across each of the stimulus types. Instead, the differences are most apparent in the .9 quantile, where it can be seen that distance 1 lures exhibit a slower .9 quantile than the easier

distance 2 lures, similar to what is seen in the data. Why does the model show this pattern? In the EB-LBA model, each of these stimulus classes only show differences in the drift rate. A counter-intuitive property of evidence accumulation models is that changes in drift rate show the most pronounced differences in the *slowest* RTs, with very little difference in the fastest or median RTs (e.g., Ratcliff & McKoon, 2008; White & Poldrack, 2014).

Figures 4B and 5B depict the data and model fits to the three probe types (targets, standard lures, and ELF lures) separated by each of the fixed dimensions (amplitude of sine wave 1, phase angle of sine wave 1, and phase angle of sine wave 2). This demonstrates the model is actually capable of addressing the extralist feature effect present when amplitude-1 is the fixed dimension (ELF FAR < standard lure FAR) and the inverse extralist feature effect when the other two dimensions are the fixed dimensions (ELF FAR > standard FAR). How does the model accomplish this? Analysis of the group means of the attention weight parameters reveals that attention is largely fixated on the amplitude-1 dimension in both experiments (Experiment 1: $M = .663$, 95 % HDI = [.610, .713], Experiment 2: $M = .660$, 95% HDI = [.609, .710])), while attention is almost equally divided between the other two dimensions (Experiment 1: phase-1 $M = .154$, 95% HDI = [.125, 1.83], phase-2 $M = .182$, 95% HDI = [.137, .227], Experiment 2: phase-1 $M = .167$, 95% HDI = [.143, .191], phase-2 $M = .173$, 95% HDI = [.134, .213]).

Extra attention allocated to the amplitude-1 dimension has the effect of making the ELF lures less similar to the memory set items than the standard lures when amplitude-1 is the dimension that carries the extralist feature. However, this comes at the cost of a inverse extralist feature effect on the other two dimensions, as this makes the ELF lures more similar to the memory set items than the standard lures when the other dimensions are fixed. Allocating extra attention to the amplitude dimension additionally predicts that there should be a larger hit rate for targets on trials when amplitude is the fixed dimension; Figure 4B demonstrates that this pattern is present in both the data and the

model's predictions. Thus, evidence from both the data and the modeling strongly suggest that the interaction between lure type and the fixed dimension is due to the extra attention allocated to that dimension. Readers may wonder why w is not split equally between amplitude-1, phase-1, and phase-2, since we argue that these dimensions are integral. To clarify, the point is that the dimensions are difficult to attend but not impossible (cf. Nosofsky, 1987). It is reasonable to assume that over the course of many trials and several sessions, participants are able to focus more on one of the dimensions than another. The critical distinction is that attention cannot be easily switched from one dimension to another on a trial-by-trial basis resulting in higher attention overall to one dimension. It is also possible that the amplitude-1 dimension is more salient than the other dimensions due to distortions in the psychological representation of the stimuli (see Appendix B).

Scatterplots depicting the results of individual participants against the mean posterior predictive distributions from the model can be seen in Figures 7 and 8 for Experiments 1 and 2, respectively. The EB-LBA model appears to provide an excellent account of the error rates for each item type (top rows) with the majority of r^2 values in the .87-.94 range. The model's coverage of individual participant RTs is quite strong for targets, even for the error RTs. The r^2 values for the .1 quantile are surprisingly low in Experiment 1 ($r^2 = .33$ for correct RTs for lures at distance 1) — this may be due to the fact that one participant exhibited an unusually slow .1 quantile (1 second for both lure types), whereas the majority of values are in the .5-.6 range. Coverage for the error RTs is extremely poor for the third target item in Experiment 2. However, performance was extremely high for this item, and thus there were very few errors. Overall, the model appears to be capturing differences across individual participants quite well.

Discussion

The results of Experiments 1 and 2 demonstrated that no extralist feature effect was found with integral-dimension stimuli: virtually equivalent FAR were found for both

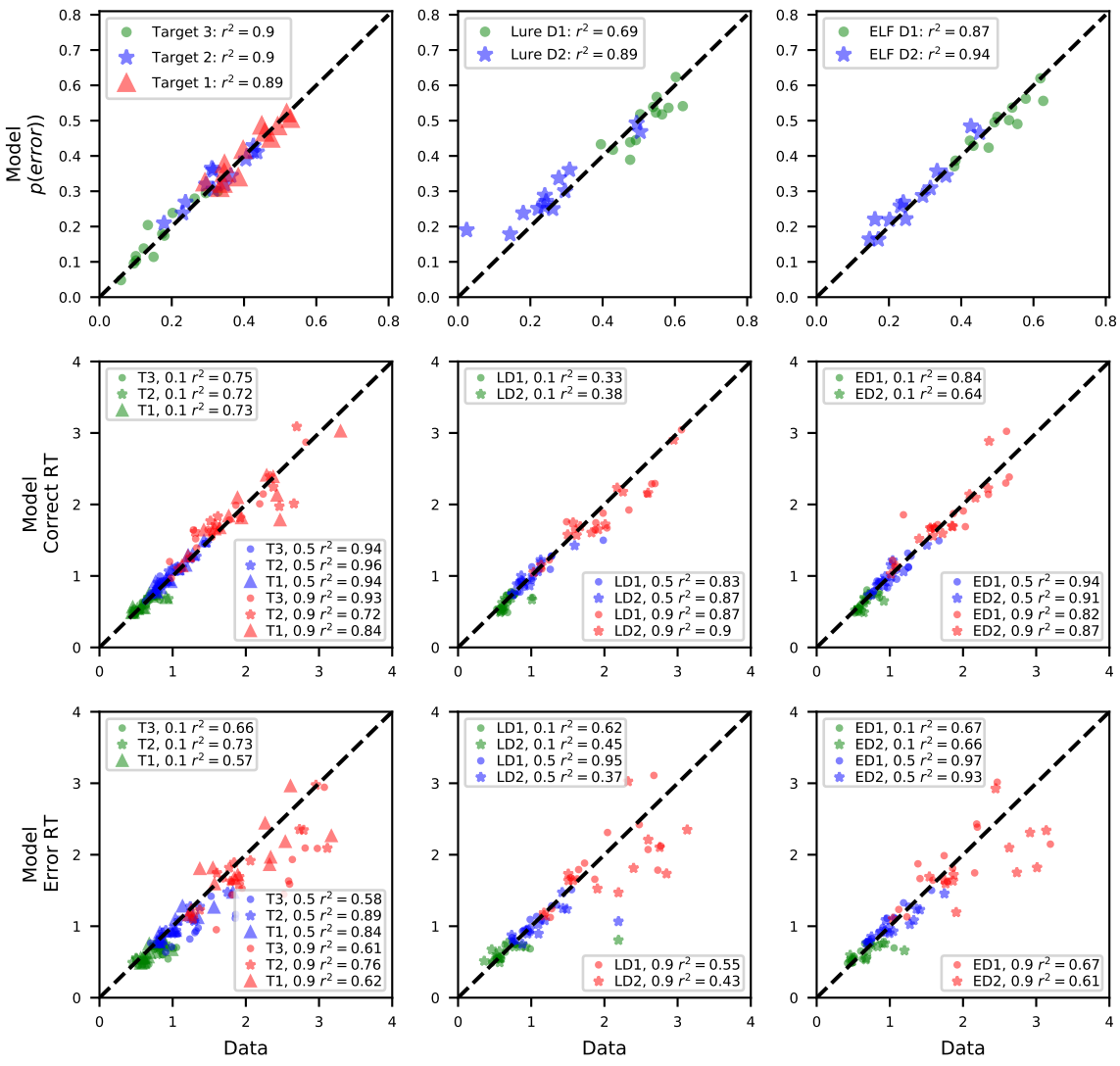

Figure 7. Scatterplots depicting the fit to individual participants from Experiment 1, including the error rates (top row), correct RTs (middle row), and error RTs (bottom row), where RT distributions are summarized using the .1, .5, and .9 quantile. Model predictions consist of the mean of the posterior predictive distribution from the EB-LBA model. Notes: T3, T2, T1 = targets at serial positions 3, 2, and 1. L = lures. E = ELF lures. D1 = distance 1. D2 = distance 2.

standard lures and ELF lures at both levels of distance. Experiment 2 replicated this trend and further demonstrated the equivalence when the extralist feature was carried on the

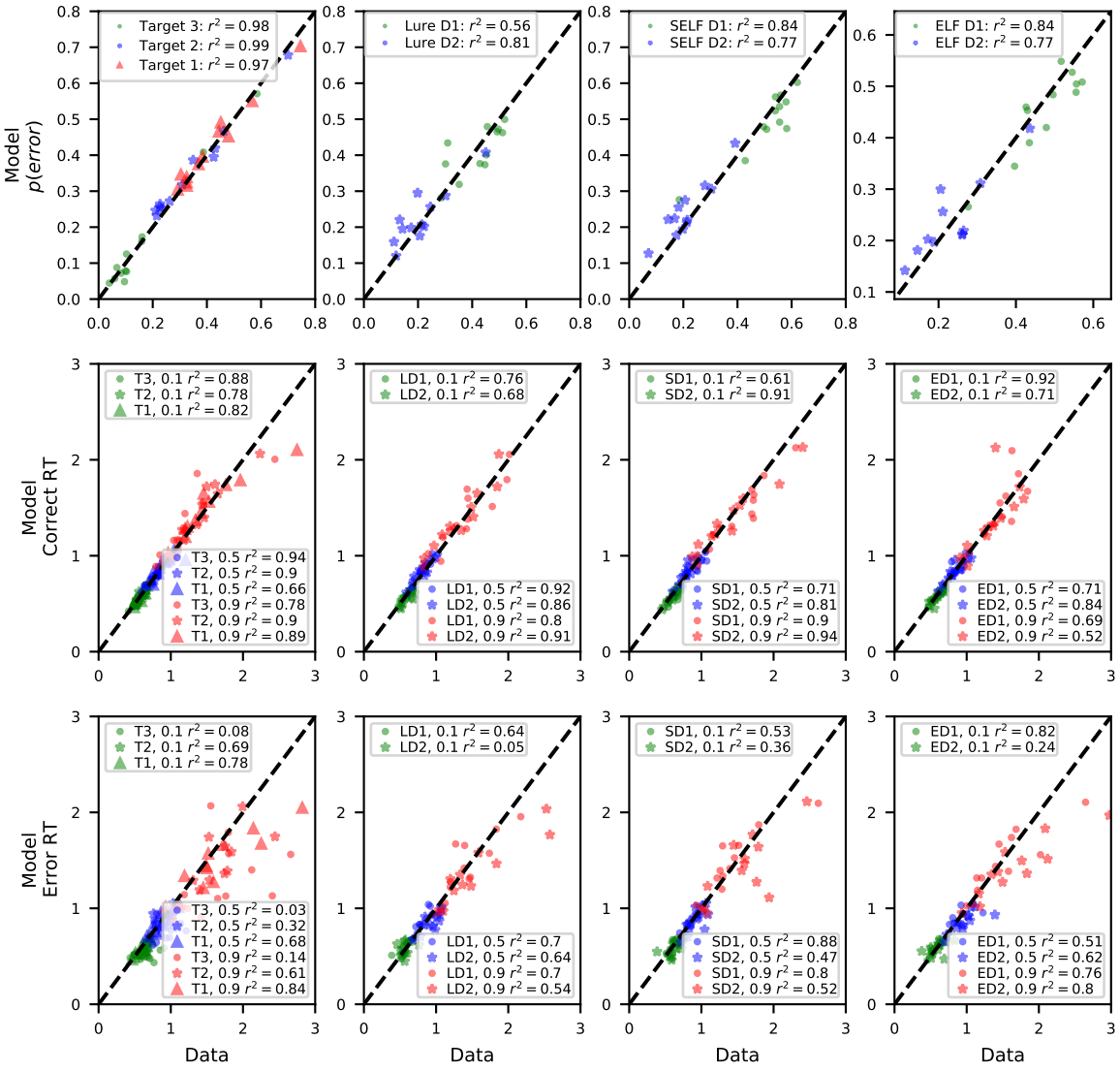

Figure 8. Scatterplots depicting the fit to individual participants from Experiment 2, including the error rates (top row), correct RTs (middle row), and error RTs (bottom row), where RT distributions are summarized using the .1, .5, and .9 quantile. Model predictions consist of the mean of the posterior predictive distribution from the core EB-LBA model. Notes: T3, T2, T1 = targets at serial positions 3, 2, and 1. L = lures. S = S-ELF lures. E = ELF lures. D1 = distance 1. D2 = distance 2.

same dimension as the standard lures. The data from both experiments were well accounted for by the standard EB-LBA model without modification. These results

undermine the generality of the extralist feature effect and suggest that global matching models are appropriate models for integral-dimension stimuli.

Experiment 3: Separable-Dimension Stimuli

The next major question we are concerned with in this article are whether we can demonstrate an extralist feature effect using continuously valued separable dimensions and whether plausible modifications to global matching models can provide a parsimonious explanation of the extralist feature effect. Thus, in the third experiment, we used a very similar paradigm to Experiment 1 but with separable-dimension stimuli that were composed of rectangles that varied in their height, color saturation, and position of a vertical bar (see Figure 2 for a visual demonstration). As mentioned previously, the ability to separately process and attend to stimulus dimensions offers some unique possibilities to the decision-maker, such as decisions based on component dimensions, or the ability to strategically allocate attention to relevant stimulus dimensions on a trial-by-trial basis. We test these ideas within a global matching framework by implemenenting (a) parallel global matching models, where strengths for each stimulus dimension are calculated and compete to be retrieved and (b) a variant of the EB-LBA where attention is allocated to dimensions that are more "diagnostic" of the stimulus being novel.

Distance between separable dimension is typically based on the city-block metric (Attneave, 1950; Nosofsky, 1992; Shepard, 1987; Shepard & Chang, 1963; Torgerson, 1958). Because the city-block distance is not rotation invariant, we balanced the quantity summed(1/distance) between lure trials and extralist feature trials by first generating the standard lure probe and using an optimization procedure on summed(1/distance) to adjust the value of the fixed dimension to generate an ELF probe with equivalent similarity. We ran an additional experiment where we equated the city-block distance directly between standard lure probes and ELF probes. This experiment found an inverse extralist feature effect – better performance on standard lure probes than extralist lure probes, which is

likely due to the fact that distance – but not similarity – was equated between the two probe types, and therefore similarity was lower on average for lure probes than for extralist feature probes. We report the results of this experiment in Supplementary Materials A and the data can also be found on our OSF page (`https://osf.io/b2zyk/`).

Unlike Experiments 1 and 2, we found a large extralist feature effect for novel-dimension ELF lures, whereas the size of the extralist feature effect with same-dimension ELF lures depended on lure distance. Because the differences between these effects were consequential for the models under consideration, we discuss the results of Experiment 3 and 4 separately, the latter of which contains the same-dimension ELF lures.

Method

Participants. Participants were 19 members of the University of Melbourne community, with normal or correct-to-normal vision, who participated in three one-hour sessions. Participants were renumerated at a rate A$15 per session. Human testing was approved by the Melbourne Human Research Ethics Committee (Approval number: 1034866).

Materials and Procedure

The stimuli were rectangles varying in the internal saturation of the color, their height, and the horizontal position of an internal bar (see Figure 2). The background color was selected from the Munsell hue 5R with a brightness value of 4. The standard xyY coordinates corresponding to the Munsell brightness and saturations (available at `http://www.cis.rit.edu/research/mcsl2/online/munsell_data/all.dat`) were converted to RGB values by converting the xyY values first to CIE XYZ color space coordinates and then to RGB values using standard transformations (Rossel, Minasny, Roudier, & McBratney, 2006; Travis, 1991). The width of the rectangle was set to 200 pixels. The bar width and outline were set to 10 pixels. The range of values used for

saturation, height, and bar position (from the internal left border of the rectangle) are shown in Table 1. [4]

For Experiment 3, study sets and test items were constructed in the same manner as Experiment 1, except that rather than constraining lures and extralist feature lures to have the same distance from the study set, we constrained lure and extralist features lists to have equivalent values for summed inverse city-block distance to each study set item. sum(1/distance) is scale invariant, meaning that multiplying all of the distances by a constant would still result in the same sum(1/distance) for each probe type. Representative list values are shown in Table 3. The procedure was also the same as in Experiment 1. Each of the lure types in Experiment 3 are shown in Figure 9. The MDS analysis reported in Appendix B shows that the psychological representation of the stimulus dimensions corresponded closely to the dimensions manipulated in the experiment.

Results

Two participants were excluded for exhibiting chance or close-to chance level performance ($d' = .15$ and $d' = -.03$). Exclusion of responses faster than .2 seconds or slower than 4.0 seconds resulted in the exclusion of less than 1% of the data.

Results can be seen in Figure 10. Similar to Experiment 1, distance 1 lures exhibited much higher FAR ($M = .437, SEM = .018$) than lures at distance 2 ($M = .213, SEM = .016$), $BF_{01} = 2.29 \times 10^{24}$. Contrary to the results of Experiment 1, an extralist feature effect was found, with ELF probes showing lower FAR ($M = .285, SEM = .013$) than standard lures ($M = .364, SEM = .020$), $BF_{10} = 35.65$. This result demonstrates that the extralist feature effect can be found even with continuously-valued stimuli that are not easily labeled.

[4] We again used the psi method as described in Appendix A in a preliminary calibration experiment with two observers to find the JND's for each dimension. The JND for saturation was near 2 steps, which is consistent with previous scaling results for saturation by Nickerson (1936). The JND's for height and bar position were both 12 pixels, which is just slightly larger than the line width used to construct the rectangles.

Table 3

Representative lists and lure indices in Experiments 3 and 4

Separable Dimensions							
Distance	Fixed Dimension	M1	M2	M3	Lure	Type	sum(1/d)
1	1	[4, 3, 8]	[4, 5, 6]	[4, 7, 4]	[4, 6, 7]	standard	1.00
1	1	[4, 3, 8]	[4, 5, 6]	[4, 7, 4]	[4, 7, 6.4]	S-ELF	1.01
1	1	[4, 3, 8]	[4, 5, 6]	[4, 7, 4]	[5.55, 5, 6]	ELF	1.01
1	2	[4, 3, 6]	[6, 3, 8]	[8, 3, 10]	[6.8, 3, 7.8]	standard	1.01
1	2	[4, 3, 6]	[6, 3, 8]	[8, 3, 10]	[2.6, 3, 6]	S-ELF	1.01
1	2	[4, 3, 6]	[6, 3, 8]	[8, 3, 10]	[6, 4.55, 8]	ELF	1.01
1	3	[3, 4, 10]	[5, 6, 10]	[7, 8, 10]	[3.6, 7.4, 10]	standard	0.86
1	3	[3, 4, 10]	[5, 6, 10]	[7, 8, 10]	[6, 4, 10]	S-ELF	0.87
1	3	[3, 4, 10]	[5, 6, 10]	[7, 8, 10]	[7, 8, 11.725]	ELF	0.86
2	1	[7, 4, 10]	[7, 6, 8]	[7, 8, 6]	[7, 9, 5]	standard	0.77
2	1	[7, 4, 10]	[7, 6, 8]	[7, 8, 6]	[7, 4, 12]	S-ELF	0.77
2	1	[7, 4, 10]	[7, 6, 8]	[7, 8, 6]	[9.24, 6, 8]	ELF	0.77
2	2	[6, 10, 5]	[8, 10, 7]	[10, 10, 9]	[10.4, 10, 4.6]	standard	0.63
2	2	[6, 10, 5]	[8, 10, 7]	[10, 10, 9]	[10, 10, 4.2]	S-ELF	0.63
2	2	[6, 10, 5]	[8, 10, 7]	[10, 10, 9]	[10, 12.63, 4.6]	ELF	0.63
2	3	[4, 8, 6]	[6, 6, 6]	[8, 4, 6]	[3, 9, 6]	standard	0.77
2	3	[4, 8, 6]	[6, 6, 6]	[8, 4, 6]	[8, 7.8, 6]	S-ELF	0.76
2	3	[4, 8, 6]	[6, 6, 6]	[8, 4, 6]	[6, 6, 8.24]	ELF	0.77

The dimension that was fixed in the study set had no effect on the FAR, $BF_{10} = .099$, with roughly equivalent FAR for each fixed dimension (saturation $M = .346, SEM = .021$, height $M = .316, SEM = .015$, bar position $M = .313, SEM = .019$). A recency effect is again evident, with much higher HR for

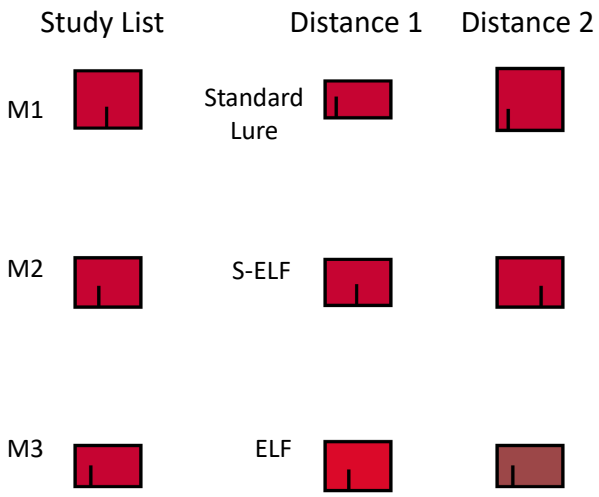

Figure 9. Illustration of the list types used in Experiments 3 and 4. The members of the study list are on the left, while the right column depicts the lure types. The study list varies in height and bar position with saturation as the fixed dimension. Refer to the text for descriptions of how the lures were generated.

targets in the third position ($M = .872, SEM = .019$) than in the second ($M = .676, SEM = .034$) and first ($M = .596, SEM = .036$) positions.

Interactions of a.) lure type and distance and b.) distance and the dimension that held the fixed dimension were both found, as the model with main effects of lure type, distance, and the two interactions was preferred, $BF_M = 34.37$. The interaction between lure type and distance is due to the fact that the extralist feature effect is larger at distance 1 (standard lure FAR $M = .496, SEM = .025$, ELF lure FAR $M = .378, SEM = .015$) than at distance 2 (standard lure FAR $M = .233, SEM = .021$, ELF lure FAR $M = .192, SEM = .014$). The second interaction is not of theoretical interest and was not analyzed further.

Parallel analyses were conducted on mean RTs for each lure type. Mean RT to distance 1 lures were longer ($M = 1.03$, $SEM = .064$) than to distance 2 lures ($M = .960$, $SEM = .051$), $BF_{10} = 43,154$. Standard lures exhibited longer latencies ($M = 1.01$,

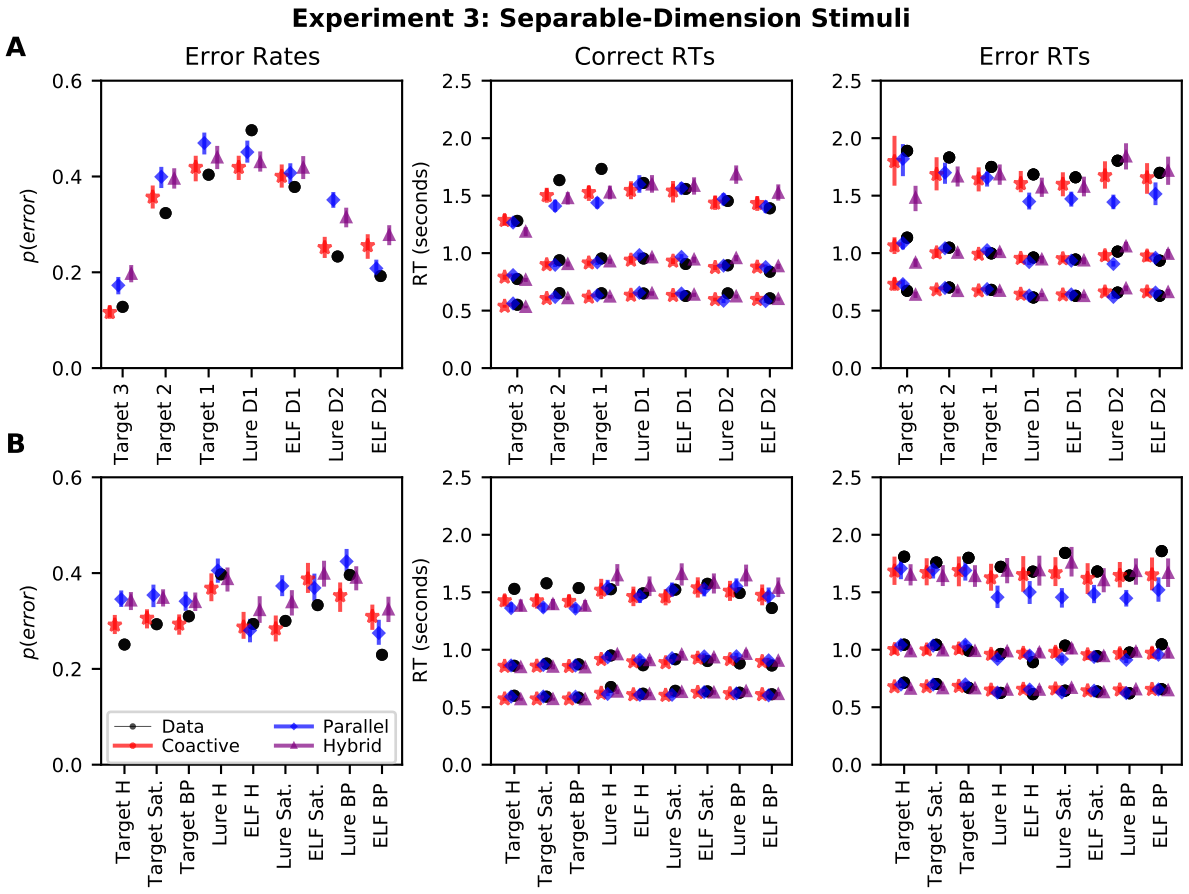

Figure 10. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 3 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictive distributions from both the coactive (red) and parallel (blue, with an exhaustive decision rule for "old" responses and self-terminating rule for "new" responses) variants of the EB-LBA model. Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

$SEM = .060$) than ELF lures ($M = .970$, $SEM = .054$), although the Bayes Factor revealed weak evidence for this difference, $BF_{10} = 2.053$. The Bayes Factor was indecisive as to whether mean RT varied by which dimension was fixed, $BF_{10} = 1.395$.

Unlike the results with integral-dimension stimuli in Experiment 1, the extralist feature effect we found in this experiment with separable-dimension stimuli are challenging to global matching models which predict equivalent FAR for the standard lure and ELF lure probes. We verify this prediction in the next section where we fit the EB-LBA model to the data. Subsequently, we introduce various modifications to the model in an attempt to capture the effect.

EB-LBA Modeling

We fit the standard EB-LBA model to the data in the same manner as in Experiment 1 with the same 12 parameters being estimated for each participant. Group-averaged posterior predictive distributions of error rates and correct and error RT distributions can be seen in red in Figure 10, where the model is referred to as the "coactive" model. It can be seen that contrary to the data, the model predicts equivalent FAR between the standard and ELF lure types — the standard EB-LBA model is not able to capture the extralist feature effect in the data. However, aside from that major limitation, the model is able to capture several of the other constraints, including the recency effect, the differences between easy (distance 2) and difficult (distance 1) lures, and is reasonably able to capture the shapes of the RT distributions across each trial type. Similar to Experiment 1, however, the model underpredicts the .9 quantile for correct RTs for targets in the second and third position.

In the coming sub-sections, we describe model variants that make allowances for different methods of novelty rejection that are afforded by separable-dimension stimuli These take the form of either a.) alternative decision architectures to a coactive architecture or b.) variants that allow for attention to the component dimensions vary on a

trial-by-trial basis. Each of these model variants vary in their number of parameters and their flexibility in their ability to capture the data. To place these models on equal footing, model variants were also compared using the widely applicable information criterion (WAIC: Watanabe, 2010). WAIC is a metric for model selection that is an approximation to leave-one-out cross validation, and similar to other information criteria it produces a value that strikes a balance between goodness-of-fit and model complexity. Because WAIC is on a deviance scale, lower values are preferred. In contrast to the deviance information criterion (DIC), WAIC is considered to be a "fully Bayesian" information criterion in the sense that its penalty term is determined by calculation of the variability in the likelihood of each data point across all of the parameters in the posterior distribution. In other words, a more complex model has more "ways" it can fit the data than a less complex model. WAIC has been recommended over DIC (Gelman et al., 2014).

WAIC values for each model variant can be seen in Table 4.

Table 4

WAIC values for each variant of the EB-LBA for the data from Experiments 3, 4, and 5. The winning model is depicted in bold. P = the number of participant parameters from the model.

	Exp 3		Exp 4		Exp 5	
Model	*P*	WAIC	*P*	WAIC	*P*	WAIC
Coactive (standard EB-LBA model)	12	28291	12	45623	11	24325
Parallel: Exhaustive "old" self-terminating "new"	12	29421	12	47663	11	24901
Parallel: Exhaustive "old" two "new"	12	29757	12	47084		n/a
Parallel: Two "old" self-terminating "new"	12	30183	12	48220		n/a
Parallel: Exhaustive "old" and "new"	12	30181	12	47212	11	25920
Hybrid Coactive-Parallel: Self-terminating "new"	17	28807	17	46415	15	24210
Hybrid Coactive-Parallel: Two "new"	17	28652	17	46859		n/a
Hybrid Coactive-Parallel: Exhaustive "new"	17	29439	17	46387	15	24778
Coactive: Diagnostic Attention	**16**	**28035**	**16**	**45076**	**14**	**24088**
Coactive: Attention to Unvarying Dimensions	16	28150	16	45174		n/a

Parallel Global Similarity Models: Decision on the Basis of Component Stimulus Dimensions. We formalize the notion of decision by component stimulus dimensions by inheriting the global matching assumptions of the GCM, but calculating the summed similarity on the basis of each component dimension separately. In other words, instead of calculating the global similarity of A across all items and dimensions, separate global similarities are calculated for each stimulus dimension, reflecting how similar the probe's value on that dimension to the values from the list items. This results in the calculation of three A values: one for each stimulus dimension (height, saturation, and position of the vertical bar). These A values are then converted to drift rates for an "old" and "new" accumulator for each stimulus dimension. As we describe below, various decision rules can be used to map the outcomes of these races to the resulting decision.

Such parallel decision architectures are not new to recognition memory. Several proposed models and investigations have used parallel or serial architectures for recognition memory decisions (e.g., Cox & Criss, 2017, 2019; Donkin & Nosofsky, 2012b; Ratcliff, 1978; Townsend & Fifić, 2004). However, an important contrast between such models and the ones we consider here is that in these investigations the race was either between the items in the memory set or between item and associative information[5]. Here, we assume that the items in the memory set are always aggregated together by virtue of the global similarity computation — it is instead the stimulus dimensions that race against each other.

One should note that because decision making is based on the stimulus dimensions and not the entire stimulus, the model can no longer be considered an "exemplar-based" model. For this reason, we refer to this model as the parallel global similarity model. The mathematics of the model are described below. They are very similar to the analytics of

[5]It is also worth mentioning that parallel races between the items in the memory set are common in free recall models (Osth & Farrell, 2019; Osth, Reed, & Farrell, 2021; Polyn, Norman, & Kahana, 2009; Sederberg, Howard, & Kahana, 2008).

the EB-LBA model, with the exception that distances and similarities are calculated for each dimension separately.

The distance between the probe i and memory exemplar j on dimension k is given by:

$$d_{ijk} = w_k(|x_{ik} - x_{jk}|) \tag{6}$$

The similarity between the probe i's value on dimension k to memory exemplar j is given by:

$$s_{ijk} = exp(-c_j d_{ijk}) \tag{7}$$

Activation values for a particular dimension are produced by summing the similarities between the probe's value on that dimension and the values on each of the list items, weighted by the memory strengths of the exemplars:

$$A_{ik} = \sum_j m_j s_{ijk} \tag{8}$$

Unlike the standard EB-LBA model, the parallel global similarity model contains six linear ballistic accumulators, three "old" and three "new" accumulators for each stimulus dimension. In other words, the decision process produces recognition decisions on each of the dimensions at the same time. The drift rates for the "old" and "new" accumulators on each dimension are given by:

$$v_{old,ik} = A_{ik}/(A_{ik} + k) \tag{9}$$

$$v_{new,ik} = 1 - v_{old,i} \tag{10}$$

In order to apply a parallel model to data, a decision rule for response termination is required. We explored a number of different decision rules, but preferred a model with an

exhaustive rule for "old" decisions as well as a self-terminating rule for "new" decisions, meaning that an accumulator for a single dimension reaching the "new" boundary is sufficient to terminate the decision. The lax rule for "new" decisions was used to capture the idea that an extralist feature in a probe will mismatch the memory set, producing a low value of A for that dimension, which results in a low drift rate for its "old" accumulator and a high drift rate for its "new" accumulator, and thus can easily trigger a "new" decision.

An advantage of the LBA is that analytic expressions for parallel models can be expressed for a wide variety of decision rules (e.g., Cox & Criss, 2019; Donkin & Nosofsky, 2012b; Eidels, Donkin, Brown, & Heathcote, 2010). Analytic expressions for the exhaustive "old" and self-terminating "new" model can be found in Appendix E. We did not pursue serial models because this requires the convolution of each searched dimension, and there are no analytic expressions for this.

The parallel global similarity model contains all the same parameters as the coactive EB-LBA model. The group-averaged posterior predictives for error rates and correct and error RT distributions for Experiment 2 can be seen in Figure 5. One can see that the model captures the extralist feature effect: FAR to ELF lures are lower than to the standard lure probes. This provides a proof-in-concept that the extralist feature effect can be produced within a global matching model if decisions are made on the basis of component dimensions, as the mismatch on the dimension that carries the extralist feature is sufficient to generate more rejections to ELF lures.

However, aside from its ability to capture the extralist feature effect, the model does not bear a close quantitative correspondence with the data. The model predicts an interaction between lure distance and lure type, but in the opposite direction as is seen in the data — a large extralist feature effect is predicted for the relatively easy distance 2 lures while a smaller effect is predicted for the distance 1 lures. In addition, the model is predicting faster error responses for lure stimuli than are seen in the data, presumably because its relatively lax decision termination rule for lure stimuli makes it such that errors

can be triggered more quickly for lures than for targets. Inspection of Table 4 reveals that the parallel model performs extremely poorly in WAIC relative to the core model (Δ WAIC = 1130 relative to the full model), reinforcing the conclusion that the model is not providing an adequate account of the data.

Because of our concern about the decision rule for "new" responses being too lax, we additionally applied three other parallel model variants: one with an exhaustive rule for "old" responses and two required "new" thresholds, one with two required "old" thresholds and a self-terminating rule for "new" responses, and a model with an exhaustive rule for both "old" and "new" responses. Table 4 reveals that none of the alternative decision rules produced any improvement over the model with exhaustive "old" and self-terminating "new" model. Group-averaged posterior predictive distributions for each of these alternative parallel models can be found in Supplementary Materials B.

A parallel model that uses stimulus dimensions as the basis for decision making should fail for other reasons. In the original Mewhort and Johns (2000) paradigm that uses discrete features, the parallel model should be unable to discriminate between 1:1 lure probes and targets because there is no sensitivity to the conjunction of the feature dimensions that are present in the target probes but not in the 1:1 lure probes. The coactive model does not share this limitation because the exponential transformation of distances between exemplars into similarity is non-linear and results in higher similarities for targets. Due to this issue with the parallel global similarity model – along with the generally poor performance in model selection – fits of the parallel global similarity model to subsequent datasets can be found in the Supplementary Materials.

Coactive-Parallel Hybrid Architecture. Nosofsky et al. (2011) proposed a hybrid decision architecture where the EBRW would use a coactive decision rule for "old" decisions using only a single accumulator where the memory strength is pooled across all of the stimulus dimensions and exemplars in memory. For "new" decisions, in contrast, a parallel decision architecture is employed, with an accumulator corresponding to each

stimulus dimension and the evidence for each dimension being driven by the similarity between the value of the dimension in the probe and the values of the corresponding dimension in each exemplar in memory. In other words, the hybrid model suggests novelty rejection operates differently than acceptance of old items in that each decision utilizes different forms of global similarity computation. The hybrid coactive-parallel EB-LBA model bears a superficial resemblance to the iterative resonance model (IRM: Mewhort & Johns, 2005) in the sense that separate sources of evidence drive "old" and "new" decisions in each model (more on the IRM can be found in the General Discussion).

In implementing the hybrid coactive-parallel model, one possible difficulty is that the larger number of accumulators for "new" responses can result in faster "new" decisions than "old" decisions. Due to the stochastic nature of the LBA, a larger number of accumulators for one decision makes it such that there is a high probability that one of the accumulators will spuriously sample a high starting point or a high drift rate from their respective sampling distributions. For this reason, we allowed several decision-related parameters to vary across the two decision architectures (sv, a, and t_0) in addition to the two distictiveness parameters ($c_{1,2}$ and c_3), resulting in a total of 17 parameters being estimated per participant. The purpose of varying the nondecision time parameter t_0 was to allow for the possibility that the coactive and parallel races might initiate at different times. Thus, the larger number of "new" accumulators can be compensated by reductions in noise (sv and a) or longer nondecision times (t_0) to produce comparable decision times for "old" and "new" responses.

We also explored each of three possible decision rules for "new" responses (self-terminating, two "new" accumulators required to hit threshold to produce a "new" decision, and exhaustive). Inspection of Table 4 reveals that contrary to the pure parallel models, the hybrid model that required two "new" responses was the best performing model of the three hybrid models, which is the model we focus on in the main text. Figures depicting the fit for the other two hybrid models can be found in Supplementary Materials

B.

Group-averaged posterior predictive distributions for choice probabilities and correct and error RT distributions for Experiment 3's data can be seen in Figure 10. Similar to the parallel models, the hybrid model is capable of producing the extralist feature effect, albeit one that is much smaller in magnitude than what is seen in the data.

However, the fact that the hybrid model is capable of addressing the critical phenomenon of interest is offset by other major shortcomings in its ability to capture the data. The model is unable to capture the large distance effect in the data and does not predict a big enough rejection advantage for the relatively easy distance 2 lures. In addition, the model performs somewhat poorly in its ability to capture the RT distributions. In particular, it predicts much faster errors to the most recent target than seen in the data. While the poor performance on the RTs for each stimulus type is likely due to the different architectures for "old" and "new" responses, we attempted to mitigate against this as much as possible by allowing a number of parameters to vary freely across the two architectures. Table 4 reveals that this model performed more poorly than the core coactive EB-LBA model ($\Delta_{WAIC} = 361$).

Diagnostic Attention Model. Due to the generally poor fits of the parallel global similarity model and the hybrid coactive-parallel EB-LBA model, we resorted back to a coactive decision architecture and pursued a model where extra attention is devoted to the fixed dimension in the study set, which carries the extralist feature. Additional attention to the fixed dimension can reduce the similarity of ELF lures to the list items. However, as previously mentioned, an important challenge that our paradigm brings to bear is the fact that the dimension that carries the extralist feature varies across trials. Thus, we require a mechanism that determines how extra attention is allocated to the fixed dimension on a trial-by-trial basis.

We developed a coactive EB-LBA variant where attention weights are allocated to dimensions that are *diagnostic* of the probe having appeared in the study set. In this

model, attentional shifts are inversely proportional to the likelihood of the value of each probe dimension under the values of the memory set. In other words, attention weights are allocated to dimensions with values that are surprising or unlikely according to the memory set items. In the model, base attention weights to each dimension are still estimated in the normal fashion. We refer to the baseline level attention to each dimension as w^{base} while the shift in attention is referred to as w^{shift}.

To carry out such a calculation, we need a distribution that describes the variation over the particular dimension on the study list. We chose a normal distribution here purely for convenience given that it is suitable for continuously valued dimensions with support over the real number line. To implement the model, first the mean μ and standard deviation σ are calculated for the values of each dimension k among the memory set items. Subsequently, we calculate the likelihood λ of each of the probe i's values on each dimension k as follows:

$$\lambda_{ik} = \phi(i_k, \mu_k, \sigma_k) \tag{11}$$

where ϕ is the probability density function of the normal distribution. Note that for the fixed dimension, the λ_{ik} will be undefined because σ_k for that dimension will be zero. For this reason, we truncate σ_k to a very small value as a minimum (10^{-10}). Attention weights are allocated to the probe's dimensions that have lower likelihoods according to the values of the study list:

$$\gamma_{ik} = \frac{1}{\lambda_{ik} + 10^{-10}} \tag{12}$$

$$w_{ik}^{shift} = \frac{\gamma_{ik} + \beta_k}{\sum_k (\gamma_{ik} + \beta_k)} \tag{13}$$

where β_k is a bias parameter for dimension k. A small value is included in the denominator of Equation 13 to prevent γ_{ik} from being undefined when values of λ_{ik} are zero.

Equation 13 states that attentional allocation to a given dimension is proportional to how unlikely the value on that dimension is. We introduced bias parameters to the model because we found that the attentional shifts determined by the model without any additional parameters were too extreme and also required variation for each dimension. The bias parameters have the effect of mitigating the attentional allocation determined by Equation 13. As β_k approach infinity, w^{shift} will be dominated by the relative values of β parameters. The psychological underpinning of the bias parameters is that participants may only have approximate knowledge of the likelihood of each probe value, similar to the idea behind "subjective" likelihood ratio models of recognition memory (e.g., Criss & McClelland, 2006; McClelland & Chappell, 1998; Osth, Dennis, & Heathcote, 2017; Shiffrin & Steyvers, 1997). Indeed, we find it extremely unlikely that participants would be able to perform such likelihood calculations in an exact manner. The separate β parameters for each dimension may reflect the fact that some dimensions may be easier to calculate likelihoods for than others.

The final attention weights are a weighted combination of the original attention weights w^{base} and the shift w^{shift}:

$$w_{ik} = (1 - p)w_{ik}^{base} + pw_{ik}^{shift} \tag{14}$$

where p is a proportion parameter bounded between zero and one that governs the size of the attentional shift. The w^{base} values are estimated in the same manner as the core EB-LBA model. This results in an additional four parameters (p, β_1, β_2, and β_3) in addition to the parameters from the core EB-LBA model.

Illustration of the Diagnostic Attention Model. An illustration of how attention shifts can produce lower similarity for ELF trials can be seen in Figure 11, which uses a simplified two dimensional space instead of the three dimensional space in this experiment. The top panel shows items from the memory set, where all of the items vary only on the y dimension but have common values on the x dimension. Depicted are a lure

probe and an ELF probe, where the ELF trial carries a novel value on the x dimension, but both the ELF and standard lure have the same similarity (summed(1/distance)) to the memory set items.

Figure 11a illustrates the likelihood calculation for both the lure and ELF probes. For each dimension, a normal distribution is depicted in gray, where the mean and standard deviation are determined by the values of the memory set on that dimension. The tick marks depict the values of the memory set items, whereas the vertical dotted lines depict the values of the lure and ELF probes on that dimension. The lure probe contains a value that is the same as the memory set items on dimension x, resulting in a very high likelihood for that value. The lure's value on the y dimension is less likely, as illustrated by the fact that it lies toward the tail of the normal distribution.

Equation 13 dictates that the probe's value on a given dimension will receive greater attention allocation (w_{shift}) if the value has a low likelihood. Thus, in this example, the lure probe will receive extra attention to the y dimension by virtue of its lower likelihood. Thus, Equation 13 results in a shift in attention to the y dimension for the lure probe. Figure 11b illustrates the consequences of that attention shift – the y dimension is expanded while the x dimension is contracted. Consequently, the lure is more distant from the memory set items, resulting in a lower similarity to them.

The ELF probe shows a similar attention shift to the opposite dimension, but in a more exaggerated fashion. Figure 11b illustrates that the ELF's value on the x dimension is completely outside of the distribution of the memory set, which has very low variability. Consequently, the likelihood of the ELF's value on this dimension is essentially zero, indicating that this value is highly surprising relative to what was present in the memory set. The value on the ELF's y dimension, in contrast, is toward the center of the normal distribution, which has high variability. This results in a high likelihood, but not as high as the lure's value on dimension x. Both of these effects in tandem result in a shift in attention to the x dimension that is significantly stronger than the lure's attention shift to

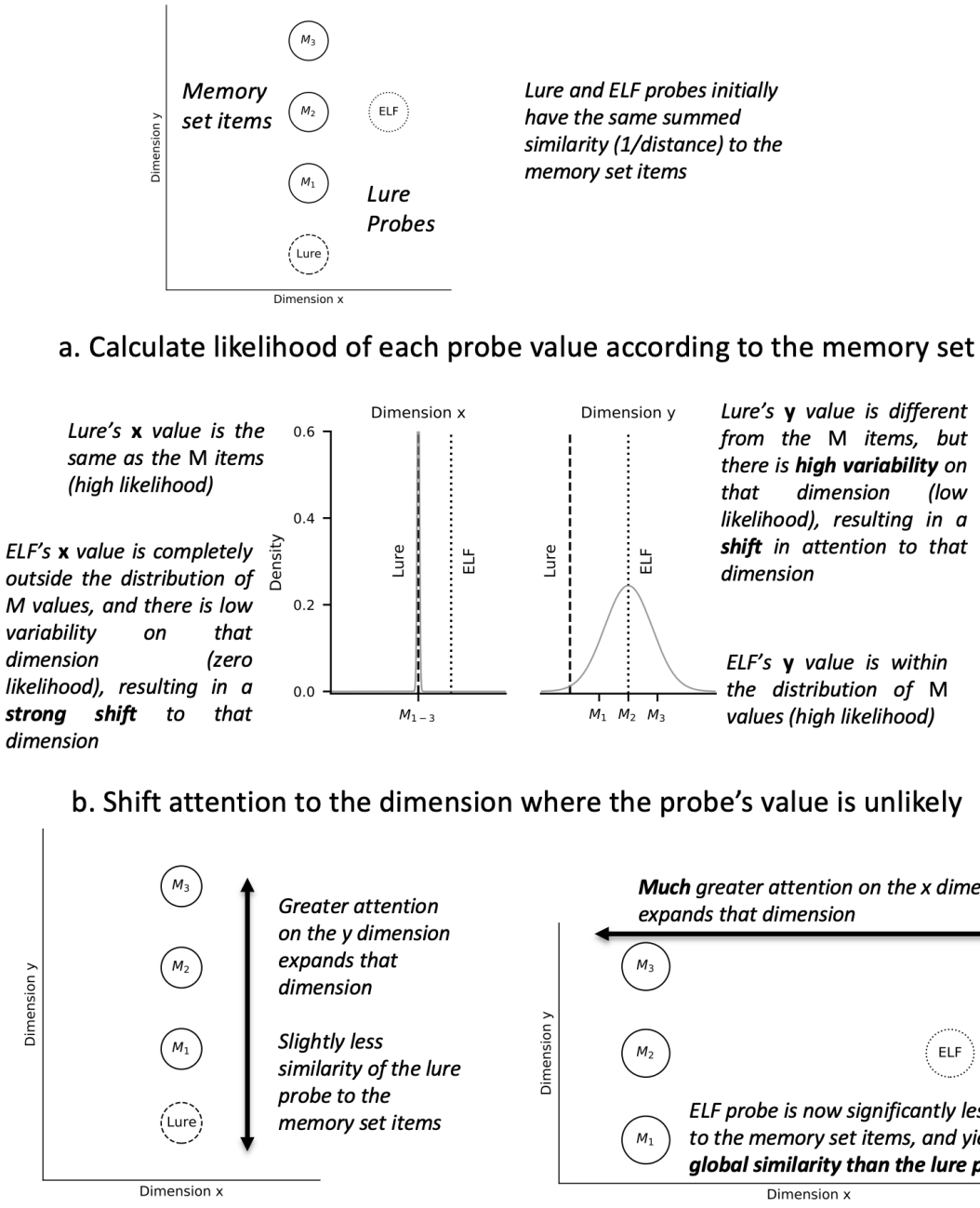

Figure 11. Illustration of attention shifts in the diagnostic attention EB-LBA model using a simplified two-dimensional space. This includes a.) calculating likelihoods of the probe's values according to the memory set (M_1, M_2, and M_3), where the distribution is depicted in gray, and b.) allocating attention to dimensions that have lower likelihoods, resulting in the expansion of dimensions that carry novel values. See the text for more details.

the y dimension. Figure 11b illustrates the expansion on the x dimension, where it can be seen that the ELF lure is pushed much further from the memory set items, resulting in a low global similarity for such a probe item.

It is important to note that this diagnostic attention variant of the EB-LBA is identical to the core model with the exception of the trial-to-trial changes in attention. Each of the probe's values are processed to evaluate which dimension should receive the most attention, with the idea behind that a dimension that carries a relatively novel or surprising value relative to what was studied in the memory set will receive the most attention; the model then proceeds with a global similarity computation where the attended dimensions carry the most weight.

Fits to Data and Estimated Attention Weights. Group-averaged posterior predictives from the diagnostic attention model can be seen in Figure 12. The model successfully accounts for the extralist feature effect in the data without compromising its ability to fit other aspects of the data. The model successfully accounts for the distance effect (higher FAR for distance 1 than distance 2 lures). The model is also able to account for the relatively small differences in RT between each of the lure types, as it predicts only small differences in the .9 quantile across both lure difficulty and lure type. In addition, Figure 12B reveals that the model also accounts for the differences across the fixed dimensions, specifically the finding that the extralist feature effect is evident when height and bar position are the fixed dimensions, but not when saturation is the fixed dimension.

The estimated values of w, w^{base}, and w^{shift} focused on the fixed dimension are depicted in Figure 13 for each trial type, calculated from the group mean parameters of the model after it had been fit to data. The panel in the top row shows the estimated values collapsed across all dimensions, while the bottom row depicts the attention on the fixed dimension conditioned on each individual dimension. On ELF lure trials, w^{shift} shows maximal attention on the fixed dimension ($w^{shift} = 1$) for all dimensions. However, the final estimates of w are much more restrained — showing extra attention to the fixed dimension

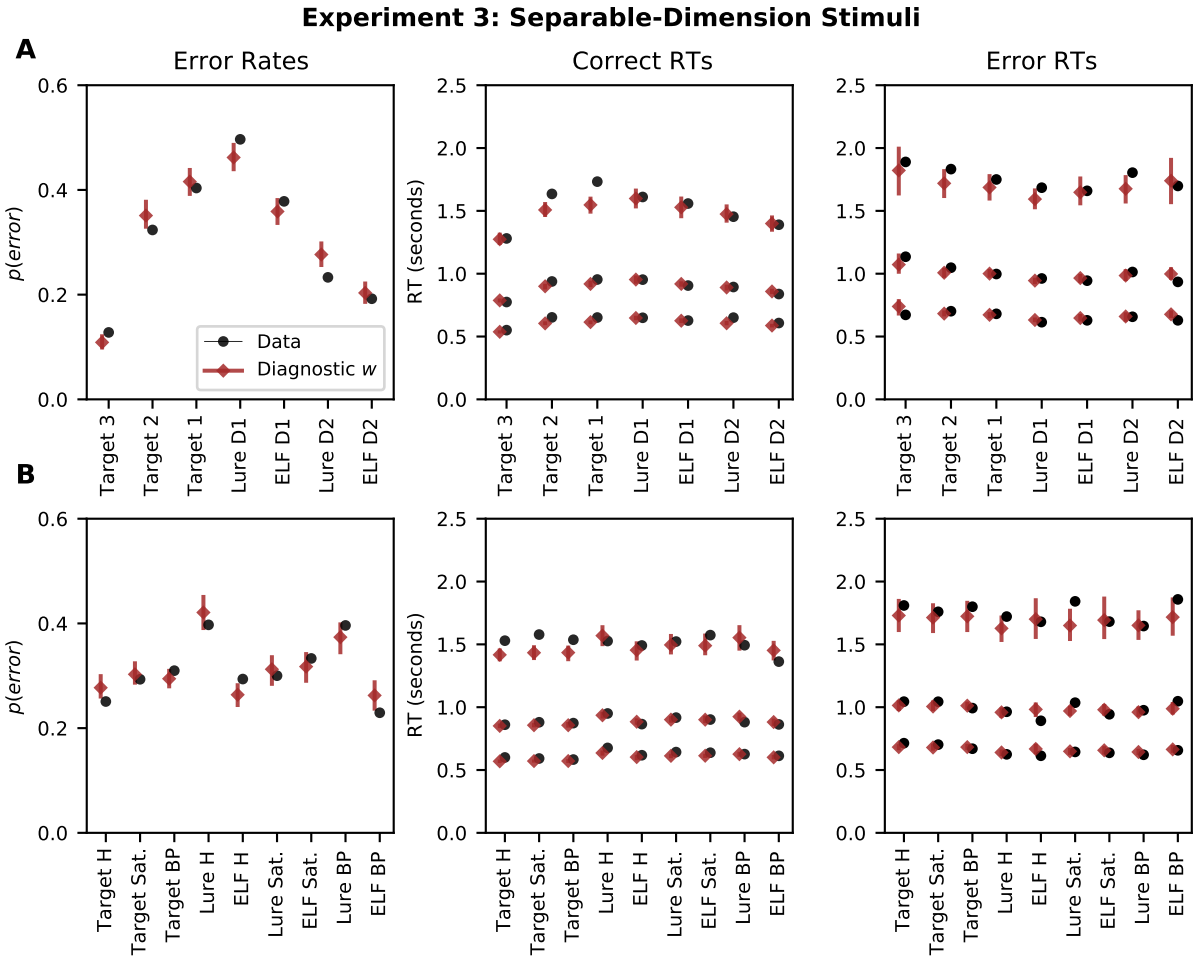

Figure 12. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 3 with separable-dimension stimuli. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lure results separated by distance. The bottom row (B) shows the results separately for each fixed dimension in the probe stimulus. Model predictions are group-averaged posterior predictive distributions from the diagnostic attention EB-LBA model. Error bars depict the 95% highest density interval (HDI). Note: D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

for ELF lures relative to standard lures, but not to an enormous degree. This is likely because the resulting estimates of p^{mu} reveal relatively small shifts in attention ($M = .175$, 95% HDI = [.014, .289]). The amount of attention devoted to the fixed dimension does not depend on the distance of the lures to the study list — the extra attention devoted to the fixed dimension is virtually equivalent for ELF lures at distances 1 and 2.

Possibly the only shortcoming of the model is that it does not account for the interaction wherein the extralist feature effect is smaller for the distance 2 than distance 1 lures. This is likely due to the fact that the attention on the fixed dimension does not vary with ELF distance, as revealed by the attention estimates in Figure 13. Inspection of Table 4 reveals that the diagnostic attention model is hugely improved over the core EB-LBA model ($\Delta_{WAIC} = -256$), suggesting that the improvement in fit more than justifies the additional complexity offered by the four additional model parameters.

Scatterplots depicting the fit to individual participants from the diagnostic attention model can be seen in Figure 14. The model yields an excellent account of the error rates on target trials ($r^2 \sim .9 - .97$) while yielding a very good account of the FAR ($r^2 \sim .78 - .88$). The point of systematic misfit is that for the standard lures, FAR are overpredicted for the distance 2 lures, while they are underpredicted for the distance 1 lures. A similar degree of systematic misfit is present for the ELF lures, albeit to a lesser degree. Similar to the core model, the account of the RT distributions is again excellent. The fact that the diagnostic attention model can capture the extralist feature effect while yielding a strong account of the individual differences across participants is impressive.

Attention to Unvarying Dimensions Model. The previously presented diagnostic attention model allocates attention to dimensions based on the relationship between the probe's values and the values in the memory set. A reviewer inquired as to whether it would be possible for selective attention to be based entirely on the values of the memory set. For instance, if attention was placed on the dimensions that do not vary, then attention will be directed toward the fixed dimension, which is the dimension that carries

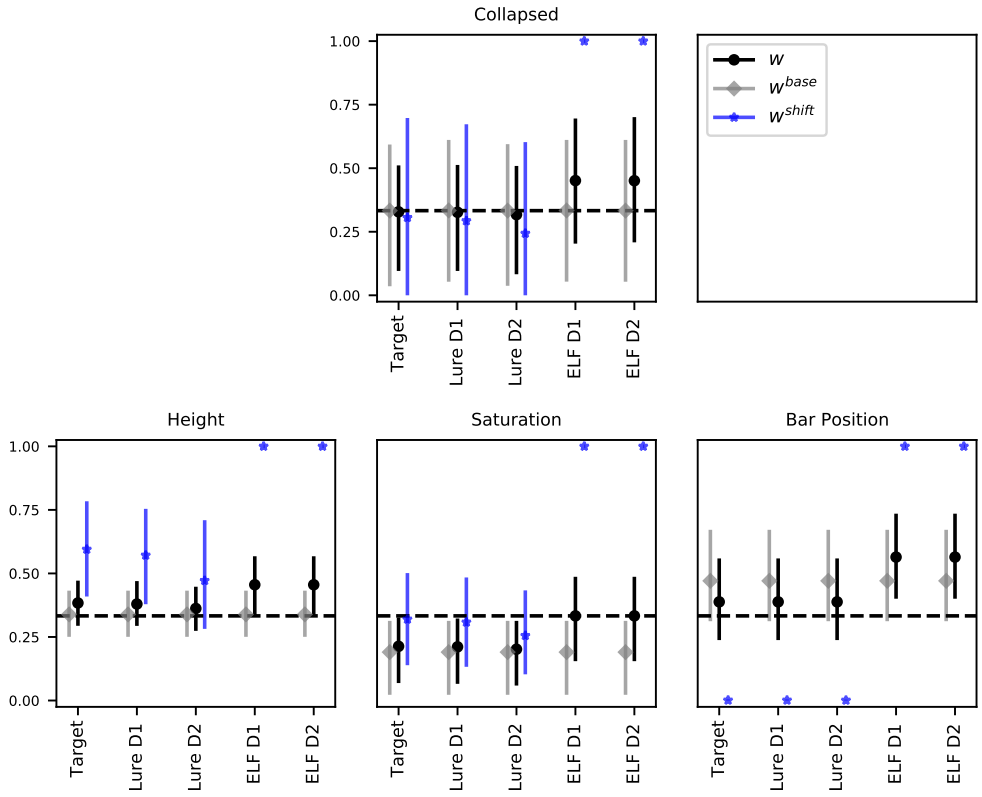

Figure 13. Mean attention weights w on the fixed dimension for each trial type from the diagnostic attention model from the fits to Experiment 3. The first panel collapses across all dimensions, while the second, third, and fourth panel condition on the particular dimension (height, saturation, or bar position). Error bars represent the 95% highest density interval (HDI).

the extralist feature on ELF trials.

We formalized such a model to evaluate its ability to capture the data that is very similar to the diagnostic attention model, with the exception that attention is directed toward non-varying dimensions. For each probe i, we calculate the standard deviation on dimension k. We similarly enforce a minimum value of σ_k (10^{-10}). Attentional allocation w_{shift} is guided toward dimensions with low variability:

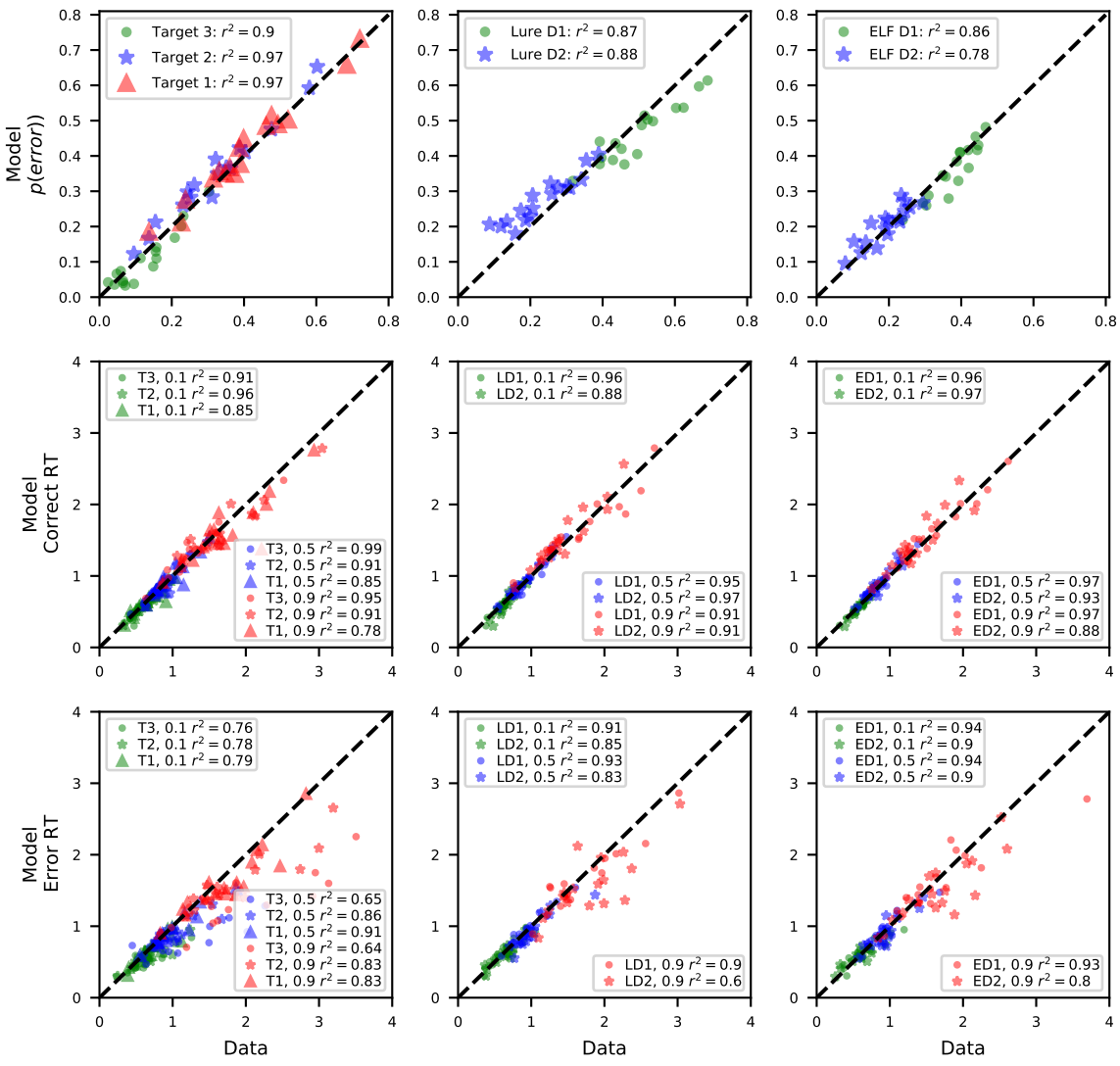

Figure 14. Scatterplots depicting the fit to individual participants from Experiment 3, including the error rates (top row), correct RTs (middle row), and error RTs (bottom row), where RT distributions are summarized using the .1, .5, and .9 quantile. Model predictions consist of the mean of the posterior predictive distribution from the diagnostic attention EB-LBA model variant. Notes: T3, T2, T1 = targets at serial positions 3, 2, and 1. L = lures. E = ELF lures. D1 = distance 1. D2 = distance 2.

$$\rho_{ik} = \frac{1}{\sigma_k} \tag{15}$$

$$w_{ik}^{shift} = \frac{\rho_{ik} + \beta_k}{\sum_k (\rho_{ik} + \beta_k)} \tag{16}$$

The final attention weights are determined by Equation 14. The model has the same number of parameters as the diagnostic attention model (16). Table 4 indicates that the model outperforms the core model, but performs considerably worse than the diagnostic attention model ($\Delta_{WAIC} = 115$). A comprehensive evaluation of the fit of the model can be seen in Supplementary Materials D, where it can be seen that the model is sufficiently able to capture the extralist feature effect in the data. An important distinction from the diagnostic attention model is the fact that attention is determined entirely by the memory set, and therefore does not vary across probe types. The addition of S-ELF lures in Experiment 4 is particularly constraining for this model. As the model always shifts attention to the fixed dimension, it is not able to allocate attention to an extralist feature that is on a different dimension than the fixed dimension.

Discussion

In Experiment 3, we found a sizeable extralist feature effect using separable-dimension stimuli, contrary to Experiments 1 and 2 in which no extralist feature effect was found with integral-dimension stimuli. Both experiments used continuously-valued stimuli where the study lists and lure probes were constructed in as similar of a fashion as possible, and overall levels of performance were quite similar across the two experiments.

The data from Experiment 3 served as an important test for a variety of mechanisms for capturing the extralist feature effect that are uniquely afforded by separable-dimension stimuli. Taking influence from the categorization literature, we initially pursued models where decision-making is based on component dimensions rather than the entire stimulus. These models took the form of a.) parallel models, in which there is a decision as to whether each probe value was seen before, and b.) hybrid coactive-parallel models, in which "old" decisions are based on the entire stimulus but "new" decisions are based on whether the values of each probe dimension were seen before. Both models could capture the extralist feature effect in fits to data because the extralist feature is not aggregated

with the other features – the novelty of the extralist feature is sufficient to produce a "new" decision. Unfortunately, these models fared poorly in capturing several other aspects of the data.

The inadequacies of these models motivated us to pursue a model within the coactive decision architecture where extra attention is devoted to the dimension that carries the extralist feature. Extra attention to such a dimension has the consequence of making a mismatch on that dimension more consequential in determining the similarity between that probe and the study list items, which can consequently reduce the global similarity for ELF lures relative to standard lures. In this work, we developed a model called the diagnostic attention model where, prior to making a recognition decision, attention is shifted to dimensions that may be diagnostic of the probe being a lure to facilitate novelty rejection. Specifically, each of the values of the probe's stimulus dimensions are subject to a pre-processing stage where the likelihood of their values are calculated according to the distribution of the values in the memory set. Attention weights are determined to be inversely proportional to such likelihood values, with the idea being that the most attention is allocated to the dimension where the probe's value might be considered surprising to the participant. The model not only succeeded in capturing the extralist feature effect, but also achieved a very good fit to the data overall. The model succeeded in capturing the recency effect, the difficulty manipulation (reduced FAR for lures at distance 2), and was able to capture the variability across participants in both choice probabilities and summary statistics of the RT distributions.

In Experiment 4, we retained the same design as the previous experiment, with the exception that S-ELF lures were additionally included at both levels of lure distance (similar to Experiment 2). Due the inclusion of an additional trial type, four sessions of data were collected per participant instead of three.

Experiment 4: Separable-Dimension Stimuli

Experiment 4 was virtually identical to Experiment 2 in that it contained both novel-dimension and same-dimension ELF lures, with the exception that it was applied to the separable dimension stimuli tested in Experiment 3. Experiment 3 demonstrated an extralist feature effect wherein novel-dimension ELF lure FAR were lower than the FAR to standard lures. However, can we similarly expect reduced FAR to same-dimension ELF lures?

Mewhort and Johns (2000) gave many compelling demonstrations of an extralist feature effect, wherein a novel shape or color was sufficient to reduce FAR even if the other shape or color was well represented in the study set. However, an important difference between their stimulus manipulation and ours is that it is unclear whether a novel shape or color constitutes a novel value on a single represented dimension. To the contrary, it is very likely that both shape and color comprise *several* dimensions. For instance, Lee & Navarro (2002) showed that shape and color can be represented by four dimensions (two dimensions for each).Thus, we cannot conclude from the Mewhort and Johns findings that we should expect to see an extralist feature effect for same-dimension ELF lures.

For the current paradigm, does the diagnostic attention model predict a large decrease in FAR for the same-dimension ELF lures relative to the standard lures, just as the model predicts with novel-dimension ELF lures? Not necessarily – unlike with novel-dimension ELF lures, with same-dimension ELF lures *the attention placed on the stimulus dimension carrying the novel extralist feature depends on lure distance.* Figure 15 illustrates the likelihood calculation of the probe values on S-ELF trials: one at distance 1 (top row) and the other at distance 2 (bottom row). For both trials, the probe shares the same value on the fixed dimension (dimension 3 in the trial depicted in the top row, dimension 2 in the trial in the bottom row) as the memory set items, meaning the likelihood for that dimension will be very high, resulting in w^{shift} estimates being shifted away from the fixed dimension. Of the two remaining probe values, one is very close to the

center of the memory set's distribution (dimension 1 for both lures), resulting in a shift in attention away from that dimension.

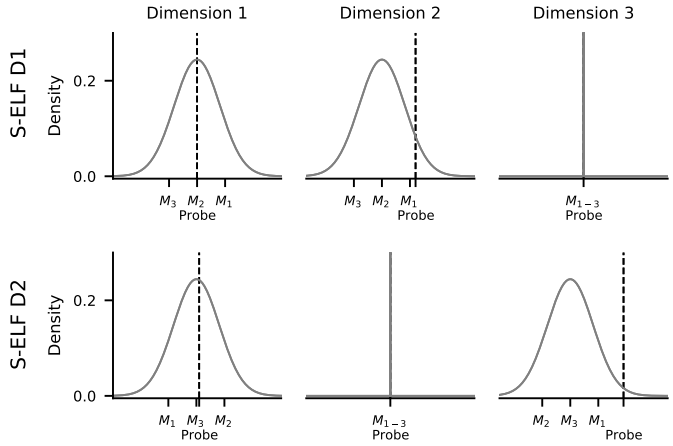

Figure 15. Illustration of the likelihood calculation for each value in the probe for a S-ELF trial at distance 1 (top row) and a S-ELF lure trial at distance 2 (bottom row). For each dimension, the probability density (depicted on the y-axis) of the probe's value (dashed vertical line) is calculated via the distribution comprising the memory set values (M_1, M_2 and M_3). Note that for the fixed dimension (Dimension 3 in the top row, Dimension 2 in the bottom row), all memory set items have the same value.

The remaining dimension is the relatively surprising dimension (dimension 2 for the S-ELF D1 trial in the top row, and dimension 3 for the S-ELF D2 trial in the bottom row), which consequently receive the bulk of the attentional weight. Recall that for the novel-dimension ELF lures, the novel value on the fixed dimension is sufficiently surprising that it elicits a probability density close to zero, thus eliciting w^{shift} values that are entirely focused on the fixed dimension. However, for the same-dimension ELF lures, the surprising probe value is still within the distribution of the memory set items, indicating that these novel values are considerably less surprising than those on novel-dimension ELF trials. In addition, a further difference from the novel-dimension ELF trials is that there is a meaningful difference in probability density from the distance 1 and distance 2 lures. In

sum, a.) *S-ELF lures are less surprising than ELF lures due to the variability across the memory set items on the dimension that carries the extralist feature, and thus should be less likely to show an extralist feature effect* and b) *the S-ELF distance 2 lures are more surprising than S-ELF distance 1 lures, making them more likely to benefit from an attentional shift to the novel dimension and show reduced FAR relative to standard lures.* We explore these predictions in the current experiment.

A further advantage for the addition of the S-ELF lures is that it provides additional constraint on the EB-LBA variant where attention is allocated to dimensions in the study set that do not vary. As mentioned previously, because the values of the probe do not change how attention is allocated, the fixed dimension will generally receive the most attention for all probe types. Because the fixed dimension does not carry the extralist feature on S-ELF trials, the model has no mechanism for producing reduced FAR on S-ELF trials.

A spatial depiction of the construction of S-ELF lures can be seen in Figure 2B. Examples of each lure type are shown in Figure 9.

Method

Participants. Participants were 21 members of the University of Melbourne community, with normal or correct-to-normal vision, who participated in four one-hour sessions. Participants were remunerated at a rate AUD$15 per session. Three participants did not complete all four sessions: one participant completed only a single session while the other two completed two sessions. All three participants' data were still included in the analyses. Human testing was approved by the Melbourne Human Research Ethics Committee (Approval number: 1034866).

Materials and Procedure. The stimuli were identical to the stimuli from Experiment 3. We generated standard, S-ELF and ELF lures using the same method as in Experiment 2 but equated on summed(1/distance). Lure condition trial numbers were

identical to Experiment 2.

Results

Three participants were excluded for exhibiting chance or close-to-chance level performance ($d' = .197$, $-.050$, and .187). Exclusion of responses faster than .2 seconds or slower than 4.0 seconds resulted in the exclusion of 1.4% of the data.

Results can be seen in Figure 16. As with Experiments 1-3, distance 1 lures exhibited much higher FAR ($M = .451$, $SEM = .026$) than lures at distance 2 ($M = .280$, $SEM = .016$), $BF_{10} = 4.87 \times 10^{27}$. Additionally, FAR differences were found between the lure types, $BF_{10} = 28.84$, with standard lures exhibiting the highest FAR ($M = .401$, $SEM = .022$), S-ELF probes showing moderate FAR ($M = .366$, $SEM = .020$), and ELF probes showing the lowest FAR ($M = .330$, $SEM = .019$).

Inspection of Figure 16 suggests that the S-ELF FAR depends on the level of distance between the probe and the study set. Post hoc analyses with Bayesian t-tests found virtually equivalent FAR for standard lure probes ($M = .465$, $SEM = .033$) and the S-ELF lure probes ($M = .468$, $SEM = .0265$) at distance 1, $BF_{10} = .246$. However, at distance 2, there was substantial evidence for an FAR difference, $BF_{10} = 30.77$, with S-ELF lure probes showing reduced FAR ($M = .264$, $SEM = .020$) relative to standard lure probes ($M = .336$, $SEM = .016$). Thus, the results confirmed the predictions of the diagnostic attention model that extralist feature effects should be more likely at for the relatively easy distance 2 lures due to them being more surprising than the distance 1 lures.

It was unclear whether the dimension that was fixed had any effect on the FAR, $BF_{10} = 1.067$. Similar FAR were found for each fixed dimension, with slightly higher FAR found for the saturation dimension ($M = .396$, $SEM = .023$) than for the other two dimensions (height $M = .352$, $SEM = .020$, bar position $M = .348$, $SEM = .022$). Similar to Experiment 2, an interaction was found between lure type and the fixed dimension, as evidenced by the fact that the model that included this interaction (along with main effects

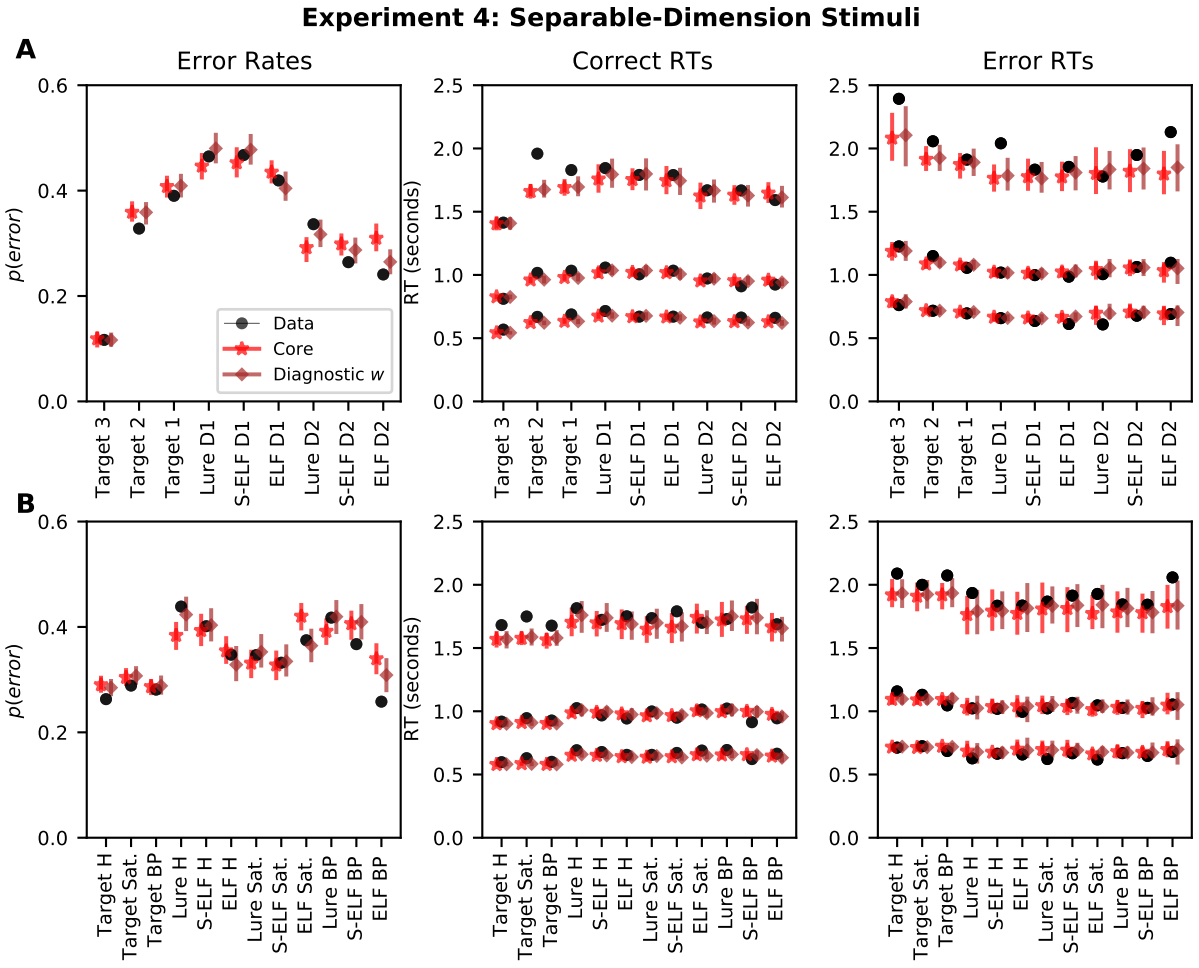

Figure 16. Group-averaged error rates (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 4. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows targets separated by serial position and lures separated by distance. The bottom row (B) shows results separately for each fixed dimension in the probe. Model predictions are group-averaged posterior predictives from the standard coactive EB-LBA model (labeled as "core") and the diagnostic attention EB-LBA (labeled as "diagnostic w"). Error bars depict the 95% highest density interval (HDI). Note: S-ELF = same-dimension ELF lures, D1 = distance 1, D2 = distance 2, H = height, Sat. = saturation, BP = bar position.

of distance, lure type, and fixed dimension) was the preferred model ($BF_M = 6.106$). This interaction was not of theoretical interest and was not analyzed further.

As with the previous experiment, differences in FAR did not necessarily translate to differences in mean RT. Lures at distance 1 did exhibited longer mean RTs ($M = 1.15$, $SEM = .078$) than lures at distance 2 ($M = 1.06$, $SEM = .068$), $BF_{10} = 2.62x10^7$. However, mean RT did not vary between standard lures ($M = 1.13$, $SEM = .082$), ELF lures ($M = 1.09$, $SEM = .074$), and S-ELF lures ($M = 1.09$, $SEM = .067$), $BF_{10} = .281$. Mean RT did not depend on which dimension was fixed, $BF_{10} = .040$.

A strong recency effect was again found, with the lowest error rate found for targets in serial position 3 ($M = .117$, $SEM = .016$). Error rates were higher for serial position 2 ($M = .228$, $SEM = .027$) than in the first serial position ($M = .390$, $SEM = .030$).

EB-LBA Modeling

We fit each of the EB-LBA variants to the data using the same methods as in the preceding experiments. No new parameters are required for the fourth experiment as the predictions for the additional trial type (S-ELF lures) depend on the distances between the probe and the memory set members, which are already specified by the stimulus space. However, for the sake of space, we focus consideration on the core coactive EB-LBA model and the diagnostic attention coactive variant.

Figure 16 shows the predictions of the core EB-LBA model along with the diagnostic attention model. What is perhaps not surprising is that the core EB-LBA model predicts virtually equivalent FAR for all three lure types. Aside from that, the model exhibits the same strengths as in previous experiments: it successfully captures the recency effect, the distance effect, and it does a reasonable job of capturing the RT distributions for each stimulus type. One of the largest departures of the model's predictions from the data is the fact that the slowest RTs for targets in serial position 2 are much slower than the model predicts. However, it is also somewhat strange that the .9 quantile for the second serial

position is even slower than the first serial position, despite the fact that targets in the second position exhibited a higher hit rate. Such a pattern was not found in Experiment 2.

The diagnostic attention variant, in contrast, yields an impressive account of the data. For the distance 1 lures, the model predicts equivalent FAR for the standard lures and same-dimension ELF lures, while predicting reduced FAR for the novel-dimension ELF lures. However, for the distance 2 lures, the model predicts reduced FAR for same-dimension ELF lures relative to standard lures, and even lower FAR for novel-dimension ELF lures, although the predicted effects are somewhat smaller than seen in the data. Similar to Experiment 3, Figure 16B reveals that the diagnostic attention model is also capable of addressing the extralist feature effect when height and bar position are the fixed dimensions along with the lack of the extralist feature effect when saturation is the fixed dimension.

These results suggest that with the same-dimension ELF stimuli, the variability across the memory set items on the dimension that carries the novel value makes it such that it takes larger differences in probe values in order for the probe's value to be considered surprising or unlikely, and consequently allocate attention to that value. For the distance 1 lures, the novel values are sufficiently close to the memory set items to prevent much additional attention to be allocated to the dimension that carries the novel value. However, for the distance 2 lures, the novel values are sufficiently distinct from the memory set to trigger attentional re-allocation. For the novel-dimension ELF probes, the variability in the memory set on the fixed dimension is essentially zero, so virtually any novel value will be considered novel and trigger a shift in attention, which is why the model does not predict that the extralist feature effect should depend on lure distance for such probes.

The diagnostic attention model similarly captures the recency effect, the distance effect, and does a good job of capturing the RTs for each stimulus type. Just as with the core model, the model's main shortcoming is its inability to capture the slow RTs of targets in the second serial position. Table 4 shows the WAIC scores of each model, where it can

be seen that the diagnostic attention model wins decisively over the core EB-LBA model ($\Delta_{WAIC} = -547$).

Scatterplots depicting the fit of the diagnostic attention model to individual participants can be seen in Figure 17. The model again shows impressive coverage of the hit rates for each serial position across participants ($r^2 \sim .92 - .98$). For the FAR to the different lure types, the model does an impressive job with the relatively difficult distance 1 lures for all trial types ($r^2 \sim .87 - .89$) but performs somewhat poorer with the easier distance 2 lures ($r^2 \sim .6 - .8$). Nonetheless, the deviations from the individual participants appear somewhat more random than in previous experiments, rather than exhibiting a systematic pattern. For the response times, the model does an impressive job in capturing the correct RT distributions across participants for all trial types ($r^2 \sim .76 - .99$).

We have focused on the core EB-LBA model and the diagnostic attention EB-LBA variant in this section, both of which rely on a coactive decision architecture. The results for the alternative parallel and hybrid coactive-parallel architectures can be found in Supplementary Materials C whereas the results for the attention to unvarying dimensions model can be found in Supplementary Materials D.

Experiment 5: Colored Shapes

Experiments 3 and 4 established that the diagnostic attention variant of the EB-LBA model can capture the range of extralist feature effects with our continuous separable-dimension stimuli. Nonetheless, an important remaining question concerns whether the model can account for the effect with stimuli from the original Mewhort and Johns (2000) paradigm. In this paradigm, participants studied colored shapes and were tested on lures where the number of matches and mismatches on each aspect (color or shape) were manipulated. As mentioned previously, a 1:1 lure contains one match on each aspect, whereas a 2:0 lure contains matches to two study list items on one aspect while matching none of the items on the other aspect. An example of each of the lure types was

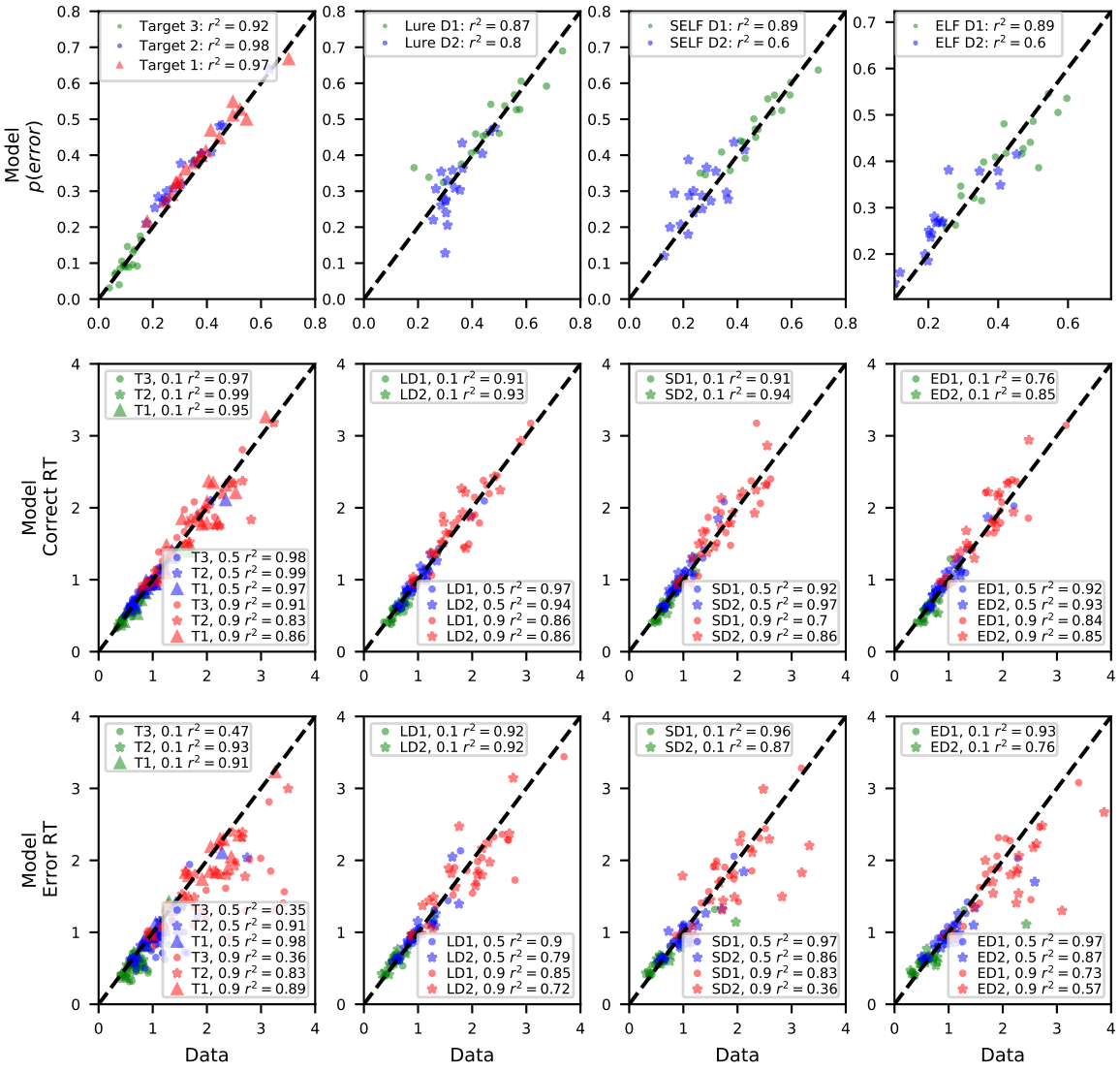

Figure 17. Scatterplots depicting the fit to individual participants from Experiment 4, including the error rates (top row), correct RTs (middle row), and error RTs (bottom row), where RT distributions are summarized using the .1, .5, and .9 quantile. Model predictions consist of the mean of the posterior predictive distribution from the diagnostic attention EB-LBA model. Notes: T3, T2, T1 = targets at serial positions 3, 2, and 1. L = lures. S = S-ELF lures. E = ELF lures. D1 = distance 1. D2 = distance 2.

presented earlier in Figure 1.

We conducted a partial replication of the Mewhort and Johns' Experiments 1-3 in

order to test the same models we used in Experiments 3 and 4. We did not conduct an exact replication of their experiments because each of the relevant lure types, namely the 1:1, 2:0, 1:0, and 0:0 lures, were never tested within the same experiment. We found it desirable to have all of these lure types tested within the same experiment and same participants to exert the maximum constraint on the model given that we have fit these models to individual participant data. In addition, while Mewhort and Johns had participants give confidence ratings (participants were asked to give a "sure" or "unsure" rating after their initial decision), we only collected old/new responses, which not only is consistent with our previous experiments, but circumvents difficulties in simultaneously modeling confidence and choice decisions.

A change from our previous experiments was an expansion of the study list length to four items. In the original Mewhort and Johns experiments that contained 2:0 lures and tested colored shapes (namely Experiments 2 and 3), on each study list, one shape and one color were presented twice without exactly repeating any of the target items, resulting in the presentation of three shapes and three colors across the memory set. Consistent with our previous experiments, participants completed four sessions in order to maximize the amount of data at the participant level.

As mentioned previously, the key findings from Mewhort and Johns were that a.) performance was higher for 2:0 probes than 1:1 probes, despite the equal number of feature matches to the study list, b.) equivalent performance for 2:0 and 1:0 lures, despite the higher number of matches in the 2:0 lures, and c.) better performance for 0:0 lures, which contain two extralist features, than lures which contain only a single extralist feature (1:0 and 2:0 lures). Based on these findings, Mewhort and Johns concluded that only the number of extralist features determines performance – the number of matching features does not exert an influence. These conclusions were also based on comparisons of mean RT to correct trials. We accompany analyses of mean correct RT with analyses of error rates, not only to be consistent with our previous experiments, but also because our

computational modeling jointly considers choice proportions and RTs.

The lack of psychophysically principled representations of these stimuli required a different modeling approach than in our previous experiments. Rather than use the GCM front-end, which calculates similarity as an exponential transformation of distance in a continuous multidimensional space, we followed Nosofsky et al. (2011) and used the context model as our front-end memory model (Medin & Schaffer, 1978). Specifically, the similarity between a probe and given study item was calculated as matches or mismatches on each aspect (color or shape). Each match was counted as a 1 where each mismatch was represented by a similarity parameter S that was constant for all mismatches. The overall probe-item similarity was the product of matches and/or mismatches between the probe and the memory set item.

Due to the impact of COVID-19, the experiment was administered online and participants completed it on their home computers.

Method

Participants. Participants were 19 members of the University of Melbourne community who participated in four online one-hour sessions on their home computers. Participants were remunerated at a rate AUD$15 per session. All participants completed all four sessions. Human testing was approved by the Melbourne Human Research Ethics Committee (Approval number: 12033).

Materials. Stimuli consisted of the eight colors and eight shapes originally used by Mewhort and Johns (2000) in their Experiments 1-3. The colors included purple, yellow, blue, red, gray, pink, green, and peach. The shapes included a star, heart, donut, cross, diamond, triangle, pentagon, and moon. Each combination of color and shape was 96 x 96 pixels. Each of the color-shape combination images can be found in our OSF repository (`https://osf.io/b2zyk/`).

Procedure. The procedural details are very similar to the original Mewhort and Johns (2000) experiments. Each study list began with the presentation of a fixation cross for 1,000 ms. Following that presentation, four colored shapes were presented for 1,750 ms each followed by a 250 ms ISI. Each study list consisted of three colors and three shapes, with one color and one shape presented twice with the provision that they were never repeated within the same stimulus (no exact repetitions of targets were allowed). One should note that this design ensures that each studied item comprises a twice-presented aspect along with a once-presented aspect.

Upon completion of the study list, a fixation cross was presented for 1,000 ms. After its disappearance, a probe item was presented along with the text "Have you seen this object before (from the previous study phase)?" and in the line below was presented a reminder of the response keys ("1 = YES, 0 = NO"). On half of the trials, the probe was one of the target items from the study list and on the other half the probe was a lure. The lure types included trials where the probe contained a once-presented color and a once-presented shape (1:1 lures), a twice-presented color or shape and a non-presented color or shape (2:0 lures), a once-presented color or shape and a non-presented color or shape (1:0 lures), or probes where neither the color or shape were presented on the study list (0:0 lures).

A total of 224 trials were presented in each session (112 targets and lures). Each of the different trial types were counterbalanced. That is, there were an equal number of targets from each serial position (1, 2, 3, or 4) and an equal number of each lure type. In addition, 2:0 and 1:0 trials were counterbalanced such that there were an equal number of trials where the color was presented and the shape was the extralist feature and visa versa.

To prevent slow guesses, a time limit of 8,000 ms was imposed. If the participant had not responded with the appropriate response keys by that time point, the trial timed out with a text display that read "TOO SLOW." To prevent fast guesses, responses faster than 250 ms were presented with a text display that read "TOO FAST." In both cases, the

feedback was presented for 1,500 ms. Trials in either case were excluded from the analysis.

The experiment was programmed in Javascript using jsPsych (de Leeuw, 2015).

Results

One participant was excluded from the analysis for showing poor discrimination between targets and 1:1 lures ($d' = .17$). Another participant was excluded for focusing almost exclusively on color – when lures employed shape as an extralist feature along with a presented color (1:0 or 2:0 trials), FAR were extremely high (98.2%). Likewise, when 1:0 or 2:0 trials employed color as an extralist feature, the same participant's FAR was very low (1.78%). In addition, responses faster than 250 ms or slower than 4 seconds were excluded from the analysis, resulting in the exclusion of 1.63% of the data.

Results can be seen in Figure 18. The top row (Figure 18A) depicts results for targets broken down by serial position along with lure 1:1, 2:0, 1:0, and 0:0 probes. Because accuracy was close to perfect for 0:0 trials, error RTs were not presented. The bottom row (Figure 18B) shows results broken down by the frequency of each aspect (color or shape). Target trials always comprised a twice-presented shape or color along with a once-presented shape or color – Figure 18B shows decomposes results for targets depending on which aspect (color or shape) was once-presented. Figure 18B also shows results for 1:0 and 2:0 probes depending on which aspect was the extralist feature. Because there were very few errors on 1:0 and 2:0 probes when these trial types were decomposed by the aspect carrying the extralist feature, only the median error RT is depicted.

Accuracy was somewhat lower in our experiment than in the original Mewhort and Johns (2000) experiments. For instance, Mewhort and Johns reported hit rates of 81.9%, 86.5%, and 81.3% for their colored shapes in Experiments 1-3, respectively, which are higher than the hit rate we have observed in this experiment ($M = .713$, $SEM = .037$). Part of this discrepancy may be due to faster responding overall in our experiment. While mean RTs to correct target trials was 1.13 seconds in our experiment, Mewhort and Johns

reported mean RTs 1.40, 1.52, and 1.63 seconds in their Experiments 1-3 for the same trial type. It may be that Mewhort and Johns' requirement to report whether participants are sure or unsure after each decision may have caused participants to raise their decision threshold, trading speed for accuracy. The overall proportion of correct responses in our experiment was 78.9%.

Our analyses of FAR revealed a sizeable extralist feature effect in error rates - FAR were much higher for 1:1 lures ($M = .303$, $SEM = .029$) than 2:0 lures ($M = .129$, $SEM = .016$, $BF_{10} = 109,745$, despite the fact that both lure types contained the same number of matches. FAR were nearly perfect for 0:0 lures ($M = .029$, $SEM = .010$), reproducing the finding that lure performance was affected by the number of extralist features. FAR to 0:0 lures was lower than both 2:0 ($BF_{10} = 24,671$) and 1:0 lures ($BF_{10} = 2,750$).

To compare 2:0 and 1:0 lures, we conducted a 2x2 ANOVA with number of matches (2 vs. 1) and the aspect carrying the extralist feature (color vs. shape) as factors. FAR were considerably higher when shape was the extralist feature ($M = .141$, $SEM = .019$) as opposed to color ($M = .074$, $SEM = .016$), $BF_{10} = 387.78$. FAR were slightly higher for 2:0 ($M = .129$, $SEM = .016$) than 1:0 lures ($M = .086$, $SEM = .013$), but the Bayes Factor revealed only weak evidence for this difference ($BF_{10} = 3.56$).

While this might appear consistent with the finding from Mewhort and Johns that performance was unaffected by the number of matching features in lures, we found an interaction between the number of matches and the stimulus aspect carrying the extralist feature, as evidenced by the fact that the model including the interaction was the preferred model ($BF_M = 3.616$). Post hoc analyses revealed little difference in FAR between 2:0 ($M = .084$, $SEM = .018$) and 1:0 lures ($M = .065$, $SEM = .016$) when color was the extralist feature, with the Bayes Factor being agnostic between the null and alternative hypotheses ($BF_{10} = 1.049$). However, when shape was the extralist feature, the Bayes Factor revealed strong evidence for an FAR difference between 2:0 ($M = .174$,

$SEM = .023$) and 1:0 lures ($M = .110$, $SEM = .017$), $BF_{10} = 84.70$. We will demonstrate later that this interaction is consistent with selective attention being focused on the color aspect, which has the effect of lowering FAR when color is the extralist feature and also makes matches on the shape aspect more consequential in determining performance.

Analyses of mean RT on correct trials in our experiment did not reflect the interaction between the number of matches and the aspect carrying the extralist feature, as evidenced by the fact that the model containing an interaction between the two factors was not the preferred model ($BF_M = .795$). While there were shorter latencies on 2:0 and 1:0 trials when color was the extralist feature ($M = 1.06$, $SEM = .088$) as opposed to shape ($M = 1.18$, $SEM = .100$), $BF_{10} = 2,515$, the number of matches did not exhibit much of an influence on RT – RT for 2:0 probes ($M = 1.14$, $SEM = .098$) were very similar to 1:0 probes ($M = 1.10$, $SEM = .088$). However, the Bayes Factor did not reveal decisive evidence in favor of the null or alternative hypothesis ($BF_{10} = .628$).

Aside from this difference, the mean RTs largely conformed to the accuracy results. 1:1 probes exhibited slower RTs ($M = 1.35$, $SEM = .118$) than 2:0 probes, $BF_{10} = 61.23$. 1:1 probes were also slower to reject than 1:0 probes, $BF_{10} = 158.75$. 0:0 probes exhibited the shortest latencies overall ($M = .979$, $SEM = .077$) and were quicker to reject than 1:0 probes ($BF_{10} = 695.58$) and 2:0 probes ($BF_{10} = 2761.74$).

While serial position effects were not analyzed by Mewhort and Johns (2000), we again found a recency effect, as evidenced by strong evidence for a main effect of serial position of the target, $BF_{10} = 2.46 * 10^{10}$. Error rates were lowest for the final item ($M = .098$, $SEM = .020$), lower for the penultimate item ($M = .275$, $SEM = .038$), and were comparable for the first two items ($M_1 = .384$, $SEM_1 = .049$, $M_2 = .391$, $SEM_2 = .050$).

EB-LBA Modeling

As mentioned previously, we adopt the representational assumptions of the context model (Medin & Schaffer, 1978) due to the lack of principled representations of the colored shapes. Unlike in Experiments 3 and 4, we did not apply the attention to unvarying dimensions model in Experiment 5 for two reasons. First, lacking the precise values of the stimulus dimensions prevents us from calculating the variability across the memory set items on the underlying dimensions. Second, even a discrete analog of this model that allocates attention to aspects that carry fewer features would result in equal attention to each aspect – both the shape and color aspects each carry three features across the memory set.

The remaining model variants were compared using WAIC. WAIC values for each model variant can be seen in Table 4. The parallel and hybrid coactive-parallel models can be found in Supplementary Materials E.

Core EB-LBA Model. The usage of the context model as a front-end leads to some differences in the similarity calculation for the core EB-LBA model. We refer to the value of the match or mismatch on aspect k (color or shape) as τ. τ takes the value of 1 if the probe i's aspect matches that of the study list item j's aspect, whereas it takes the value of S if there is a mismatch, where S is bounded between 0 and 1.0.

The probe-item similarity s_{ij} can be expressed as:

$$s_{ij} = \prod_k \tau_k^w \tag{17}$$

where w is an attention weight parameter for the particular aspect. Because there are only two aspects in these stimuli, only a single attention weight parameter is required to be estimated. Specifically, we estimated the attention weight for color as $w*$ whereas the attention weight for shape was $1 - w*$. Attention weights are implemented as powers rather than scalar coefficients due to the usage of the product rule to combine the

similarities across color and shape (e.g., Gillund & Shiffrin, 1984).

The usage of the product rule in Equation 17 is rather critical for the present design. If additive combination was used instead, the global similarity for 1:1 lures would be approximately equal to that for targets because the 1:1 lure's two partial matches would be equivalent to the target match[6]. The usage of multiplicative combination ensures that the probe-item similarity for a target exceeds that of two partial matches to the study list items (e.g., Medin & Schaffer, 1978; Shiffrin & Steyvers, 1997).

The remaining aspects of the EB-LBA are identical to the version we applied to continuous stimuli. The global similarity A for a particular probe item was calculated according to Equation 3 and used to drive a pair of linear ballistic accumulators. Because there were four items in the study set, we used a total of three memory strength parameters to capture the serial position effects in the data – m_1, m_2, and m_3 – while the memory strength for the final item (m_4) was fixed to 1.

In total, the core EB-LBA model contains a total of 11 parameters – three memory strength parameters (m_1, m_2, and m_3), the similarity parameter for mismatches S, the attention weight parameter $w*$, the drift criterion K, and the five LBA parameters shared with the other models (t_0, a, sv, b_{old}, and b_{new}).

Group-averaged posterior predictives of the core EB-LBA model can be seen along with the data in Figure 18. The most obvious point of misfit is that the model fails to capture the extralist feature effect – approximately equal FAR are predicted for 2:0 and 1:1 lures. This challenge is also shared in the model's account of the RTs, where the model has difficulty in capturing the very slow RTs to 1:1 probe relative to 2:0 probes. In addition, the model overpredicts the FAR difference between 2:0 and 1:0 lures. These shortcomings are shared with previous explorations of global matching models by Mewhort and Johns (2005), namely the REM and Minerva 2 models.

[6]One should note that the Minerva 2 model (Hintzman, 1988) circumvents this problem due to an additional non-linearity, namely the cubing of probe-item similarities prior to calculation of the global similarity.

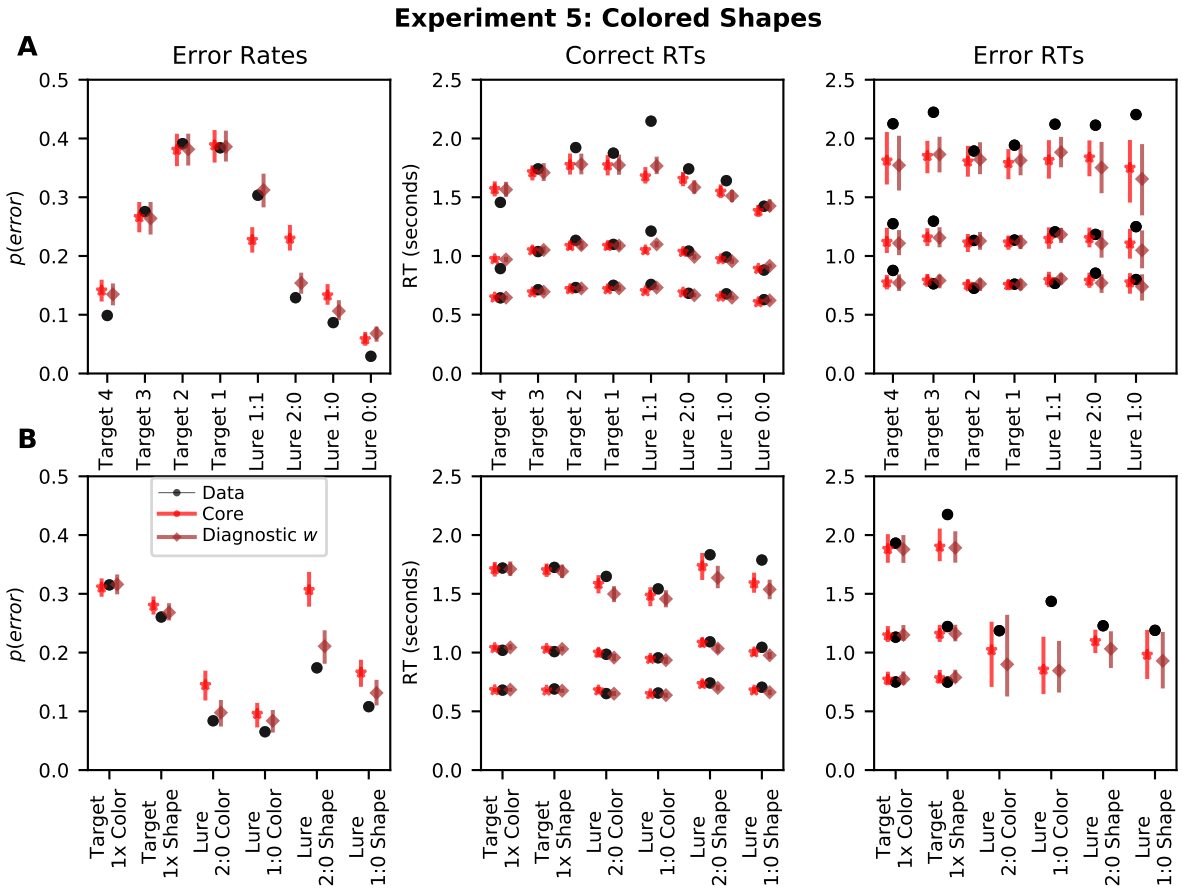

Figure 18. Group-averaged choice probabilities (left panel) and correct and error RT distributions (middle and right panels, respectively) for the data from Experiment 5 with colored shapes as stimuli along with posterior predictives from the core and diagnostic attention EB-LBA models. RT distributions are summarized using the .1, .5, and .9 quantiles. The top row (A) shows results for target (broken down by serial position) and lure (1:1, 2:0, 1:0, and 0:0) trials. The bottom row (B) breaks down results for targets depending on whether the shape or color is the once-presented feature along with 1:0 and 2:0 probes when color or shape is the extralist feature. Error bars depict the 95% highest density interval (HDI).

Nonetheless, the model succeeds in capturing the low FAR to 0:0 probes. Despite the model not possessing a mechanism for capturing the extralist feature effect, there is a

simple reason why the model is able to capture this pattern - 0:0 probes exhibit the lowest global similarity to the study list. The model also produces a higher error rate for targets when color is the once-presented feature, and also produces higher FAR for lures when color is the extralist feature. Both patterns are likely due to the fact that greater attention is placed on the color aspect, as revealed by the estimates of the group mean of w ($w^{\mu} = .595$, $95\%HDI = [.566, 0.624]$). Greater attention to color results in lower FAR when color is missing from a lure probe, but also results in higher error rates for targets when color is the once-presented feature, as less attention is devoted to the twice-presented feature in such stimuli.

Diagnostic Attention Model. In our previous applications of the diagnostic attention EB-LBA model, the shifts in attention to a particular stimulus dimension were proportional to how unlikely the probe's value on that dimension were. Because the dimensions were continuous, we could calculate the likelihoods using a normal distribution, where the mean and standard deviation were determined by the values along the study set.

Due to the discrete nature of our stimuli, we instead adopted a discrete likelihood that only considers the number of occurrences h of the feature in the study set. Specifically, the likelihood λ of feature k in probe i is:

$$\lambda_{ik} = h/3 \tag{18}$$

where 3 is present in the denominator because there are always three shapes or colors in each study set. Higher likelihoods are assigned to twice-presented features. The attentional shifts are calculated according to Equations 12-14.

Because the model employs a coactive decision architecture, there are only two accumulators ("old" and "new"). The model contains a total of 14 parameters – it inherits the 11 parameters of the core EB-LBA model, but uses an additional three parameters for the attention shift (β_{color}, β_{shape}, and p).

Group-averaged posterior predictives from the diagnostic attention variant can be

seen in Figure 18. The model yields an even stronger account of the data than the hybrid model. It predicts a large extralist feature effect in that FAR to 2:0 lures is much lower than for 1:1 lures. This is because on trials with 2:0 lures, attention is shifted to the feature that contains the extralist feature, such that it carries more weight in computing similarity. The attention shift is evident in the estimated attention weights, which we depict in Figure 19, which depicts w^{base}, w^{shift} and w to the color aspect. Results are separated by trials where both shape and color are relevant (left panel, 1:1 and 0:0 trials) and trials where color or shape are the less frequently presented feature (middle and right panels, respectively). One can see that attention to color is higher when 1:0 and 2:0 probes contain color as an extralist feature, but is substantially less when shape is the extralist feature.

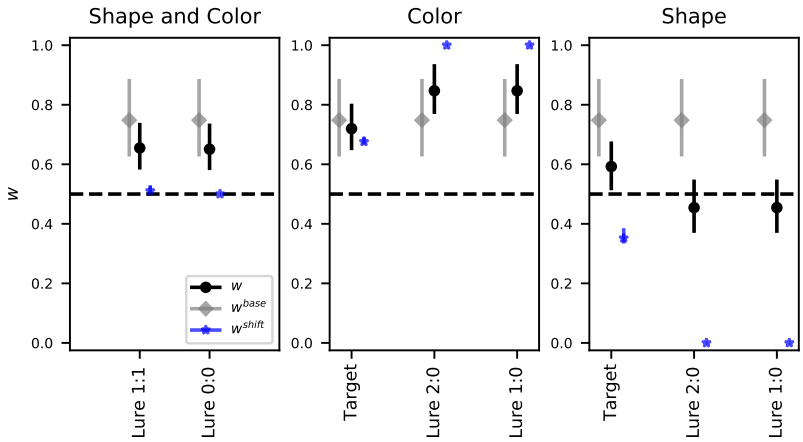

Figure 19. Mean attention weight w to color from the diagnostic attention model based on the fits to Experiment 5. The first panel depicts trials where both shape and color are relevant (0:0 and 1:1 trials). The second and third panels show target and lure trials where color and shape are the less frequently presented feature. Error bars represent the 95% highest density interval (HDI).

Another consequence of the model's attention weights is that it is able to predict a relatively small FAR difference between 2:0 and 1:0 lures that is similar in size to the data. This is likely for the same reason – while the 2:0 lures should in principle have double the

global similarity of 1:0 lures, the extra match in 2:0 probes occurs on a dimension that is not the focus of attention, and therefore carries less weight in the overall global similarity calculation. In addition, Figure 19B reveals the interaction between the lure's number of matches (2 vs. 0) and the stimulus aspect that carries the extralist feature. Larger FAR are predicted when shape carries the extralist feature, whereas only a small difference is predicted when color carries the extralist feature. This is due to the fact that attention is biased toward color, such that the matches on shape on 1:0 and 2:0 trials are more consequential, resulting in a larger FAR difference between such trials.

It may seem somewhat unusual that Figure 18 reveals roughly equivalent attention to the extralist feature on 2:0 and 1:0 trials given that Equation 18 dictates that the presented feature has a higher likelihood in 2:0 trials, implying that more attention should be given to the extralist feature on such trials. However, because the extralist feature has a likelihood of zero in both trial types, when the likelihoods are inverted in Equation 12 in the calculation of γ, this results in extremely high values of γ for the extralist feature, which consequently result in maximal attention to the extralist feature when the ratio of γ values is calculated in Equation 13. In other words, the extralist feature is sufficiently surprising such that the higher frequency of the presented feature in 2:0 trials does not drive attention away from the extralist feature. This is similar to the logic of why in Experiments 3 and 4, the fixed dimension in novel-dimension ELF trials receives maximal attention for both distance-1 and distance-2 lures – the extralist feature is sufficiently surprising that the stronger matches to studied items on distance-1 lures are insufficient to drive attention away from the extralist feature.

Nonetheless, one noticeable aspect of misfit is that the model falls short of capturing the slow rejections to 1:1 lures. This is the most noticeable advantage of the hybrid coactive-parallel model's fit to the data over the diagnostic attention model, which can be seen in Supplementary Materials E. However, the model selection results still decisively favor the diagnostic attention model, with large advantages over both the core EB-LBA

model ($\Delta_{WAIC} = -237.29$) and the hybrid model ($\Delta_{WAIC} = -122.17$). This suggests the quantitative advantage of the diagnostic attention model is still superior for many comparisons to outweigh the hybrid model's advantage in capturing the correct RT distributions of 1:1 lures.

Figure 20 shows scatterplots depicting the individual participant error rate and RT quantiles. The figure reveals that the model does an excellent job in accounting for individual partipant error rates. The worst r^2 value was for the 0:0 lures, as the model appears to over-predict the FAR to such probes. However, other r^2 values are as high as .71-.8 for lures and .84-.99 for targets, indicating an excellent coverage of individual participant variation. Similar to the other experiments, the ability to account for individual participant RT distributions is somewhat worse, but still yields a strong account of the RT distributions for correct responses.

Discussion

We conducted an experiment with discrete feature stimuli that bore a strong conceptual resemblance to the experiments of Mewhort and Johns (2000) with the exception that it included all four lure probe types in the same design (1:1, 2:0, 1:0, and 0:0 probes) to offer the most constraint on the computational models we considered. Results largely replicated those of Mewhort and Johns (2000), with two important exceptions. The first is that the relevant trends, including the extralist feature effect, were clearly evident in error rates and not just in mean RTs to correct trials. The second is that we did not reproduce the equivalence between 2:0 and 1:0 probes. Specifically, when shape was the extralist feature, FAR to 2:0 probes was noticeably higher than to 1:0 probes. While we did not find this pattern in mean RTs, we did not find strong evidence for the null hypothesis in this comparison either.

These results are contrary to those from Mewhort and Johns (2000), and suggests that the number of matching features does influence performance even when the two

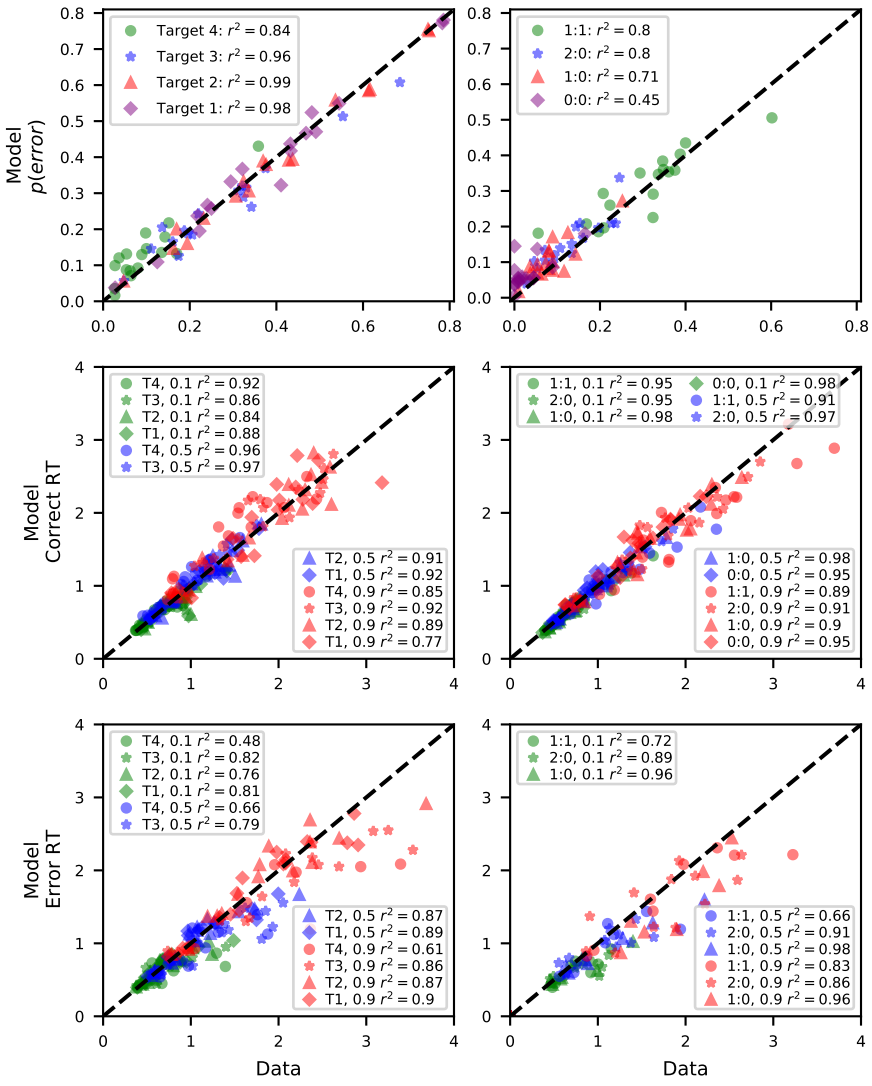

Figure 20. Scatterplots depicting the fit to individual participants from Experiment 5, including error rates (top row), correct RTs (middle row), and error RTs (bottom row). Targets and lures are depicted in the left and right columns, respectively. RT distributions are summarized using the .1, .5, and .9 quantile. Model predictions consist of the mean of the posterior predictive distribution from the diagnostic attention EB-LBA model. Due to the infrequency of errors, individual participant RT quantiles were only plotted if they had at least 10 errors. Notes: T4, T3, T2, T1 = targets at serial positions 4, 3, 2, and 1.

probes both share the same number of extralist features. While this might suggest a failed replication of the original experiments, we would like to emphasize three important differences from their data and analyses. First, while Mewhort and Johns (2000) found no difference in accuracy between 2:0 and 1:0 probes in their Experiment 2, they did not statistically analyze the difference in accuracy in their Experiment 3. Table 5 in their report demonstrates higher accuracy for 2:0 probes (93.4%) than 1:0 probes (81.6%). While it is unclear whether this difference was reliable, this accuracy difference was larger than in our experiment. Second, the analyses of Mewhort and Johns collapsed across trials where color and shape were the extralist feature. Third, the analyses of Mewhort and Johns relied on null hypothesis significance testing, which is unable to find evidence for the null hypothesis. Thus, it is unclear whether the analyses in their report were "true" null effects or whether they had inadequate power to detect a difference (Wagenmakers, 2007).

The results of Experiment 5 were very well captured by the diagnostic attention EB-LBA model that successfully explained the extralist feature effects with continuous-dimension stimuli in Experiments 3 and 4. A discrete version of the likelihood calculation enables an attention shift to the stimulus aspect that carries the extralist feature, making it such that lures containing extralist features have lower global similarity overall, and thus the model produced lower FAR to 2:0 and 1:0 lures than to 1:1 lures. Additionally, the model also captured the interaction between the number of lure matches and the stimulus aspect that carries the extralist feature – the model produced only a small increase in FAR from 1:0 to 2:0 lures when color carried the extralist feature, but produced a larger FAR difference when shape carried the extralist feature. The model was able to account for this interaction because there was a higher baseline degree of attention to color. Thus, when shape carried the extralist feature, there was less attention to shape on such trials than there was to color when color carried the extralist feature, making it such that the additional match on the shape aspect on 2:0 trials was more consequential in determining the global similarity.

A potential concern with the diagnostic attention model in each of our experiments is that it might invariably fit data better than the standard EB-LBA in all comparisons with an improvement in fit that vastly outweighs the complexity costs used by the WAIC model selection metric. If this is the case, then the model should be preferred in model selection even when the extra attention mechanisms are not warranted. To investigate this possibility, we fit the diagnostic attention variant to Experiments 1 and 2, which did not show evidence of extralist feature advantages due to the usage of integral-dimension stimuli. In contrast to Experiments 3-5, model comparison favored the core EB-LBA model, with WAIC penalties found for both Experiment 1 ($\Delta_{WAIC} = 23.90$) and Experiment 2 ($\Delta_{WAIC} = 26.28$). These results accord with intuitions that selective attention to diagnostic dimensions of the probe is uniquely afforded by separable-dimension stimuli.

General Discussion

Over twenty years ago, global similarity predictions were directly tested in the studies of Mewhort and Johns (2000), who discovered that novel features were sufficient to facilitate novelty rejection despite other probe features being well represented on the study list. Such findings were contrary to the predictions of the majority of global matching models, which should yield equivalent levels of global similarity for both lure types. In the present work, we explored the extent to which one aspect of stimuli may predict the presence of the extralist feature effect, namely the integrality of the stimulus. Four experiments with continuous-dimension stimuli confirmed that extralist feature effects do not occur with integral-dimension stimuli (Experiments 1 and 2), while they can be found with separable-dimension stimuli (Experiments 3 and 4).

To explore the extent to which these data were challenging for global matching models, we applied the exemplar-based linear ballistic accumulator (EB-LBA: Donkin & Nosofsky, 2012b) model, a variant of the EBRW model, to all experiments, including a fifth experiment that used discrete feature stimuli similar to those from the Mewhort and Johns

experiments. The standard EB-LBA implementation — like virtually all other global matching models of recognition memory — relies on a coactive decision architecture, in that information from both stimulus dimensions and memory exemplars is pooled into a single global similarity value that is used to produce a decision. We found that the data using integral-dimension stimuli (Experiments 1 and 2) were well described by this model. The data using separable-dimension continuous-dimension stimuli (Experiments 3 and 4) and the discrete feature stimuli (Experiment 5), in contrast, were challenging for this model in that it could not predict an advantage for extralist features.

A great deal of evidence from the categorization literature that separable-dimension stimuli employ parallel or serial decision architectures (Fific et al., 2010, 2008; Little et al., 2013; Moneer et al., 2016). This motivated development of parallel global similarity models along with a hybrid-coactive parallel EB-LBA, wherein entire probes could be rejected if only one of its values or features was considered "new." While each of these models succeeded in producing extralist feature effects, they unfortunately performed quite poorly on other aspects of the data and often were unable to capture the constraints imposed by RT distributions, to the point where they were often performed worse in model selection than the core EB-LBA model. One exception was that the hybrid model showed a noticeable improvement over the core model with the discrete feature stimuli in Experiment 5. However, even in this experiment the hybrid model was not the preferred model.

The most successful model in our model comparison was the diagnostic attention EB-LBA model developed in this article. In this model, novelty rejection is facilitated by shifting attention to values or features in the probe that are unlikely according to the memory set. While this model employed the same coactive architecture as the core EB-LBA model, the crucial deviation is that selective attention the stimulus dimensions varies on a trial-by-trial basis. The model succeeded not only in producing the extralist feature effect but was able to capture virtually all other qualitative trends in the data, including lure difficulty, serial position effects, biases in stimulus dimensions or aspects,

and the shapes of the RT distributions across these comparisons. The model succeeded in all of the right places, in that it was the preferred model for separable-dimension stimuli – including both continuously-valued (Experiments 3 and 4) and those composed of discrete feature (Experiment 5) – in that separable-dimension stimuli uniquely afford rapid changes in attention (e.g, Nosofsky & Palmeri, 1996b). The model's attention shift mechanism was unnecessary in the right places as well, in that it provided no advantage for integral-dimension stimuli (Experiments 1 and 2) where such shifts in attention should not be possible.

What was additionally impressive was that the diagnostic attention model also made predictions about when the extralist feature effect should *not* occur. In Experiment 4, the model made the prediction that the same-dimension ELF lures should not show an advantage when they were close to the memory set (distance 1) but should show an advantage when they are further away (distance 2). This was a stark contrast from the novel-dimension ELF lures, where the predictions did not depend on lure distance. The reason for the discrepancy concerns the differences in diagnosticity. With novel-dimension lures, the extralist feature was always carried by a novel dimension that was fixed across the probe items. This makes it such that any difference between the probe's values and the memory set values are extremely novel and result in shifts in attention. For the same-dimension ELF lures, in contrast, the extralist feature is carried by a dimension that varies across the memory set items. Thus, when extralist features are at close distance to the memory set items, they are considerably less novel because they are within the distribution of the memory set items. However, when the extralist feature is at a further distance from the memory set, it is taken outside the distribution of the memory set. This results in the prediction of an extralist feature effect at distance 2 but not at distance 1, which was confirmed in our data.

Implications of the Diagnostic Attention Model

The attention shift in the diagnostic attention model does not need to be unique to the EB-LBA — a similar mechanism could be implemented in other recognition memory models, such as REM (Shiffrin & Steyvers, 1997), Minerva 2 (Hintzman, 1988), the Osth and Dennis (2015) model, or the Cox and Shiffrin (2017) model. As mentioned previously, we chose the EB-LBA because it is one of the few recognition memory models that has formalized a selective attention process to stimulus dimensions. While other models have employed selective attention to item and context cues (e.g, Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981), the GCM family of models – of which the EB-LBA is a member – have enjoyed great success in accounting for both categorization, recognition, and inductive reasoning of individual items with realistic stimulus representations (e.g., Hawkins et al., 2016; Heit & Hayes, 2011; Nosofsky, 1986; Nosfosky & Palmeri, 1997; Nosofsky et al., 2011).

Part of the reason why many recognition memory models have not employed selective attention to stimulus dimensions is that they are often applied to experiments that use single words as stimuli, where representations are harder to define and models instead often use randomly generated representations. Nonethless, an increasing focus in episodic memory modeling is employing more realistic representations for words derived from semantic space models (B. T. Johns, Jones, & Mewhort, 2012; Kimball, Smith, & Kahana, 2007; Monaco, Abbott, & Kahana, 2007; Morton & Polyn, 2016; Osth, Shabahang, Mewhort, & Heathcote, 2020; Polyn et al., 2009). Incorporation of such representaitons allows models to produce similarity effects and false memory phenomena "for free", similar to the EB-LBA modeling performed in this article.

A potential stumbling block for this line of research is the possibility that extralist feature effects could occur when orthographic, phonological, or semantic features that are not represented on the list are present in a probe word. Indeed, in their Experiments 5-7 Mewhort and Johns (2000) observed the extralist feature effect with word stimuli,

demonstrating that letters in the first or last position of a word that were not present on the list promoted rejection advantages even if other letters strongly overlapped with those from the memory set. Similar extralist feature effects may be able to be found with semantic features by constructing analogous paradigms where semantic features are manipulated using the interpretable dimensions provided by Hollis and Westbury (2016), or by the usage of semantic feature norms (McRae, Cree, Seidenberg, & McNorgan, 2005).

Memory models with word representations would likely benefit from a diagnostic attention mechanism similar to the one we present in this work, which can facilitate novelty rejection for lures that contain stimulus features that were not present in the memory set. A difficulty in formalizing selective attention mechanisms with semantic representations is that semantic space models employ a very large number of dimensions, often 300 or more (e.g., Landauer & Dumais, 1997). However, Hollis and Westbury (2016) demonstrated that performing principal components analysis on the semantic vectors can produce a smaller number of dimensions that are interpretable, which would allow models to specify attention to the semantic dimensions of the words.

A potential concern with the diagnostic attention mechanism we propose is the knowledge it requires on the part of the participants. Specifically, in order to calculate the likelihood of each probe value, participants have to possess knowledge of the means and standard deviations of each probe dimension in the memory set. A potential criticism is that it's unlikely that participants have such exact knowledge of the characteristics of the memory set. This is a sensible criticism — we have implemented the model using the true statistics of the memory set to carry out the likelihood computation merely because it was the simplest way to implement and test the model. It is much more likely that participants only possess approximate knowledge of the characteristics of the memory set, similar to the operation of subjective likelihood models of recognition memory (e.g., Dennis & Humphreys, 2001; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In our experiments, the key requirement is that participants appreciate that a.) some dimensions

are fixed, making novel values on such dimensions surprising (novel dimension ELF probes) and b.) although other dimensions vary, they sometimes contain relatively surprising novel values (distance-2 same-dimension ELF probes).

A reviewer inquired as to whether the two-stage assumption of the diagnostic attention model is necessary to account for the extralist feature effect. The model currently assumes that the diagnosticity of each value within the probe is calculated, and attention is shifted to the stimulus dimension that is most diagnostic being a novel stimulus. A possible single-stage variant could be formulated by weighing the difference of each probe's value to the studied items by the diagnosticity of its respective dimension. Because this involves a multiplicative combination of diagnosticity and distance, this is mathematically equivalent to attention setting, which involves a multiplication of the attention weight parameter w by the difference between the probe's value and each studied item's value on that given dimension. We do not deny the possibility of a single-stage account of the extralist feature effect. However, we have a preference for the diagnostic attention model as it is rooted in the underlying philosophy of the GCM, which is that the context-dependence of similarity between stimuli is explained by the selective attention distributed to each stimulus dimension (Nosofsky, 1986, 1987).

The Role of Attention in Episodic Memory

Our diagnostic attention EB-LBA model is within the GCM family of models, which was originally developed as a model of categorization. One of the important contributions of the GCM was how differences in selective attention to stimulus dimensions can be used to understand the differences between categorization and identification, which previously had been argued to reflect different processes (Nosofsky, 1986). Since then, the crucial goal of category learning has been cast as learning how to weight the stimulus dimensions in a manner that provides optimal separation between the categories (e.g., Kruschke, 1992; Nosofsky, Palmeri, & McKinley, 1994). Our diagnostic attention variant of the EB-LBA is

similar in that it focuses attention on the stimulus dimensions that best divide studied from unstudied items, albeit within a single trial rather than over the course of an experimental session.

A recent investigation by Logan, Cox, Annis, and Lindsey (2021) made more direct comparisons between episodic memory and the attention literature. Their studies adapted the Eriksen flanker paradigm (Eriksen & Eriksen, 1974) to episodic memory. In traditional studies of the flanker task, participants are presented with a cue, such as a left or right arrow, and required to press the associated response. The focal target is accompanied by arrows referred to as flankers that point either in the same or opposite direction. A common finding is that performance is worse when the flankers point in the opposite direction. To account for these findings, Eriksen and colleagues posited that selective attention is a spotlight that initially focuses on the entire display (targets and flankers), but increasingly becomes more focused on the target item and less focused on the flankers.

Logan et al. (2021) conducted experiments analogous to the Eriksen and Eriksen (1974) studies in the context of episodic memory. Specifically, participants studied a short list of items and were presented with an item and position cue and asked whether the item was presented in that particular serial position (e.g., Oberauer, 2003). Comparable effects were found to the original studies – performance benefited when the flankers were in the same order as the studied items were presented in. Based on these parallels, Logan et al. (2021) concluded that memory retrieval can be thought of as attention driven inward. This description parallels the way that memory cues have been described as spotlights on regions in memory (Polyn et al., 2009). Our diagnostic attention EB-LBA also accords with the spotlight metaphor – the guidance of attention toward relevant stimulus dimensions can be thought of as guiding the spotlight toward relevant regions in memory to best exclude novel items.

Comparisons Between Extralist Feature Effects in Discrete Feature and Continuous-Dimension Stimuli

We have argued in this work that our experiments with continuous dimension stimuli are analogous to those from the original experiments by Mewhort and Johns, which have discrete mismatching features. If this is the case, then why was the extralist feature effect found with same-dimension ELF lures in Experiment 4 weaker and less evident than with the novel-dimension ELF lures? We would like to note that there are two important differences from the original experiments that need to be considered. The first is that the manipulation of shape and color in the original experiments by Mewhort and Johns is unlikely to reflect manipulation of the "true" psychophysical dimensions, which may instead comprise a combination of multiple continuous dimensions and possibly even discrete features. For instance, while the digits 0-9 might intuitively be represented purely along a magnitude dimension, Navarro and Lee (2003) instead showed that such numbers are better represented by a richer set of complex dimensions, comprising both discrete features (e.g., powers of 2, powers of 3, Tenenbaum, 1996) and continuous values (e.g., magnitude). Color is typically represented with at least three dimensions: hue, brightness, and saturation (Indow, 1988; Indow & Kanazawa, 1960; Shepard, 1978). For these reasons, it may be the case that the extralist features were *not* carried by the same dimensions as the memory set items (as in the S-ELF probes in our Experiments 2 and 4) but were actually carried along novel dimensions that were not represented in the memory set, making them more likely to attract attention.

The second consideration is that because the similarity between the lure types and the memory set items was not controlled using a continuous space in the original experiments, it is unclear how similar or dissimilar the extralist feature lures were to the memory set items. As we have demonstrated earlier in the article, the extent to which the extralist feature attracts attention depends on its similarity to the memory set items, so it is possible that the extralist features in their experiments were sufficiently distinct from the

memory set items to consistently attract attention.

With that being said, while the work of Mewhort and Johns found extralist feature effects consistently across their experiments, there was important variation in the size of the extralist feature effect that partly reflects our findings with same-dimension ELF lures in Experiment 4. Specifically, E. E. Johns and Mewhort (2002) manipulated the number of study set feature alternatives to the extralist feature. They found that as the number of colors or shapes that varied among the study set items was reduced, the extralist feature effect was larger. A common explanation may apply to both stimulus classes – when there is less variation among the study set items, an extralist feature effect is more noticeable. Indeed, this is the explanation from the diagnostic attention model's account of the S-ELF lures from Experiment 4, where the dimension that carries the extralist feature is more novel and surprising in novel dimension ELF lures than in the same dimension ELF lures. Our work takes this further and demonstrates that the extralist feature effect depends not only on the variability of the dimension that carries the extralist feature, but also depends on the continuous distance of the probe's value or feature from the study list items.

An additional difference between our experiments and the work of Mewhort and Johns is that their work found that only the number of extralist features determined performance for lures – the number of matching features exerted no influence on lure rejection. This was based on the finding that rejection latency did not significantly differ between 2:0 and 1:0 lures, whereas rejection latency was much lower for 0:0 lures, which contain two extralist features. While we similarly did not find reliable differences in latency between 1:0 and 2:0 lures in our Experiment 5 which used colored shapes, we did find reliably higher FAR to 2:0 lures when shape carried the extralist feature. However, this difference was well captured by our diagnostic attention model – higher FAR was predicted to 2:0 lures because there was more attention overall to color. When shape was the extralist feature, while there was additional attention to the shape aspect, there remained a nontrivial degree of attention to color. This made it such that the additional match to

colors from 2:0 probes was more consequential in determining the global similarity of the probe and higher FAR were predicted for 2:0 lures. While we did reproduce the rejection advantage for 0:0 lures, it is important to note here that 0:0 lures exhibit substantially lower global similarity than 1:0 and 2:0 lures. In fact, the rejection advantage for 0:0 lures is well captured by the core EB-LBA model which is not able to capture the extralist feature effect.

Relations Between the Extralist Feature Effect and Distinctiveness Effects in Recognition Memory

There are parallels between the extralist feature effect – observed both with discrete feature and continuous dimension stimuli – and distinctiveness effects in recognition memory. The focus on distinctiveness effects within an exemplar model framework was prompted by the face recognition studies of Busey and Tunnicliff (1999). They defined distinctive faces as face stimuli that arise from isolated regions of an MDS space, which exhibited superior recognition memory performance to typical faces, which come from a dense region of the MDS space. What was particularly challenging for the GCM was the finding that distinctive faces exhibited both higher hit rates and lower false alarm rates than typical faces. The GCM predicts that probe items sampled from isolated regions of the MDS space are more likely to be dissimilar to study list items – this results in lower global similarity for both targets and lures, resulting in lower false alarm rates than probe items sampled from dense regions of the MDS space. However, the GCM does not predict a hit rate advantage for distinctive targets, as the global similarity for targets is dominated by the match of the target to itself, and the similarity of a target item to its own representation is always equal to one.

Two important follow-up studies were conducted to evaluate how constraining distinctive stimuli are to models such as the GCM. Zaki and Nosofsky (Nosofsky & Zaki, 2003; Zaki & Nosofsky, 2001) argued that the faces from isolated regions of the MDS space

used in the studies by Busey and Tunnicliff (1999) also possessed unusual features that were not captured by the MDS space, such as beards. Thus, it is unclear to what extent the previously observed distinctiveness effects were due to the unusual features or due to their isolation within the MDS space. Their initial studies focused on the consequences of dense versus isolated regions of the MDS space using stimuli that exhibited a stronger correspondence to MDS representations, namely Munsell color patches (Zaki & Nosofsky, 2001). The data strongly accorded with predictions of the GCM and found that distinctive colors from isolated regions of the MDS space exhibited lower hit and false alarm rates than typical colors from dense regions of the MDS space.

Additional studies by Nosofsky and Zaki (2003) tested recognition memory for stimuli that included discrete features that are not represented by the MDS space. Specifically, they tested standard Munsell color patches, but additionally constructed distinctive stimuli that included features such as symbols, plus signs, or letters. The distinctive stimuli exhibited higher hit rates and lower false alarm rates, which paralleled the results of Busey and Tunnicliff (1999). The fact that the discrete features were not part of the continuous MDS space took them "outside the similarity space of the other items" (Nosofsky & Zaki, 2003, 1204). For this reason, they accounted for these results with a hybrid-similarity model, where similarity is a joint function of the the similarity within the continuous space along with matches and mismatches to the additional discrete features. Matches on the discrete features boost the self-similarity, resulting in higher hit rates, while mismatches on the discrete features reduce inter-item similarity, resulting in a lower false alarm rates. The hybrid similarity model was also able to account for the distinctive faces originally used by Busey and Tunnliciff (Knapp, Nosofsky, & Busey, 2006).

The extralist feature effect can be considered a distinctiveness effect, in the sense that a lure containing an extralist feature effect is made more distinctive than the remaining study list items. However, it differs importantly from previously studied distinctiveness effects in that the probes containing extralist features are made distinctive by virtue of the

study set and not by the features or probe values themselves. That is, an extralist feature is not specific to a particular color or shape in the studies of Mewhort and Johns (2000) or our Experiment 5, as the test stimuli are counterbalanced such that each color and shape can be an extralist feature. With our experiments with continuous dimension stimuli, the dimension that carries the extralist feature is similarly not specific to any region with the similarity space but is defined relative to the values of the study set items. Thus, there is no particular reason to believe that such values are outside the range of the similarity space to any further degree than the values contained in the study list items.

We do not deny that a hybrid-similarity model could similarly account for the extralist feature effect. If the extralist feature in the probe is defined as a discrete feature that mismatches the study list items, it would decrease the global similarity of lures containing extralist features and reduce the false alarm rate. However, the important challenge for such an account would be understanding how a probe containing an extralist feature would result in the creation of a discrete mismatching feature.

The Iterative Resonance Model (IRM) Account of Extralist Feature Effects

As mentioned previously, we have focused on the EB-LBA to implement various mechanisms that can produce the extralist feature effect. However, an additional model that has been used to simulate the extralist feature effect is the iterative resonance model (IRM) of Mewhort and Johns (2005). The IRM bears a strong resemblance to the Minerva 2 model (Hintzman, 1988). In the IRM, items are represented as vectors containing randomly generated binary values (0 or 1). Within each vector, a number of elements are used to represent each of the features of the studied items (shape or color) and an additional set of vector elements are used to represent the unique conjunction of features. Retrieval proceeds through a number of iterations – during each iteration, the probe is used to construct an "echo" vector which is comprised of a superposition of the vectors in memory in which each trace's contribution is weighted by a nonlinear function of the

similarity between the probe and each trace. Specifically, similarity is measured by the cosine between the probe and trace vector raised to the power of the number of iterations that have taken place. This implies that retrieval is initiated with a linear similarity function that becomes progressively more non-linear as retrieval proceeds, akin to increasing the c parameter in the EB-LBA as retrieval proceeds.

During each iteration, the echo vector is compared against the probe vector. Positive evidence is measured by the cosine between the probe and echo vectors, which is akin to a global similarity calculation. The negative evidence is measured by the number of elements in the echo that mismatch the probe. A decision is made if the positive evidence exceeds an "old" criterion or if the negative evidence exceeds a "new" criterion, otherwise retrieval proceeds through an additional iteration. Response times (RTs) are proportional to the number of iterations to produce a decision. Mewhort and Johns (2005) demonstrated that the model predicted slower RTs for 1:1 probes than 2:0 probes, and roughly equivalent RTs for 1:0 and 2:0 probes, as was seen in the data. In addition, the model was capable of other benchmarks of short-term recognition memory tasks, including increasing RTs with set size (Sternberg, 1966) and serial position effects (Corballis, 1967; McElree & Dosher, 1989; Monsell, 1978). Similar principles, including global similarity calculations for "old" responses and discrete feature mismatches for "new" responses, along with dynamically changing evidence through the trial, were also parts of the recognition by semantic synchronization (RSS) model (B. T. Johns et al., 2012), which was capable of addressing a number of benchmarks in long-term recognition memory including semantic similarity effects.

The separate calculations of positive evidence and negative evidence bears a superficial resemblance to the hybrid coactive-parallel model we considered, where global similarity of the exemplars drives "old" decisions while global similarity of the stimulus dimensions (Experiments 1-4) or aspects (Experiment 5) drives "new" decisions. In this model, mismatches on an probe's value on a dimension or a probe's feature are sufficient to

produce a "new" decision due to the self-terminating nature of the decision architecture. However, it is unclear the extent to which the IRM can produce the appropriate shapes of RT distributions for correct and error responses as only mean RT predictions were shown in the original paper. In addition, the model was not fit to data from individual participants. This may be due to the fact that the model is extremely simulation intensive — not only are large numbers of Monte Carlo simulations required to produce stable predictions (typically thousands of simulations), but individual timesteps *within* the trial additionally have to be simulated. This can make the model extremely laborious to simulate.

The EB-LBA, in contrast, is capable of addressing correct and error RT distributions across a wide range of conditions and is a tractable model. Since the IRM was not quantitatively fit to data, it is unclear whether the correspondence with the phenomena is purely qualitative. In our investigation, we found that the purely parallel and hybrid coactive-parallel variants of the EB-LBA were all capable of producing the extralist feature effect, but struggled otherwise in our experiments with continuous-dimension stimuli (Experiments 3 and 4). In these experiments, the hybrid-coactive parallel model that resembles the IRM performed even more poorly in model selection than the core EB-LBA model despite the fact that the core model was unable to predict the extralist feature effect. While the hybrid model outperformed the core model in our Experiment 5 which used discrete features, it was still outperformed by the diagnostic attention model. The success of the IRM in producing the extralist feature effect does not guarantee that it can reproduce many other aspects of the data that are of interest.

Conclusions

The experiments and modeling in the present manuscript were focused on addressing a lingering challenge for global matching models of recognition memory — the finding that novelty rejection is facilitated when lure probes contain features not present on the study list (the extralist feature effect). However, we have found that this challenge mainly applies

to experiments consisting of separable-dimension stimuli, as we did not find this rejection advantage for novel feature values when integral-dimension stimuli were employed. Our modeling work has demonstrated that the extralist feature effect is likely due to the fact that participants shift their attention to dimensions that carry values that are relatively surprising according to the memory set, facilitating novelty rejection. It is highly beneficial for future work to focus on the extent to which similar mechanisms might be operating with word stimuli.

References

Algom, D., & Fitousi, D. (2016). Half a century of research on Garner interference and the separability–integrality distinction. *Psychological Bulletin*, *142*, 1352.

Arndt, J., & Hirshman, E. (1998). True and False Recognition in MINERVA2: Explanations from a Global Matching Perspective. *Journal of Memory and Language*, *39*, 371–391.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology*, *38*, 423–466.

Attneave, F. (1950). Dimensions of similarity. *The American journal of psychology*, 516-556.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, *10*, 433–436.

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.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Burns, D. M. (2016). Garner interference is not solely driven by stimulus uncertainty. *Psychonomic Bulletin & Review*, *23*, 1846–1853.

Busey, T. A., & Tunnicliff, J. L. (1999). Accounts of blending, distinctiveness, and typicality in the false recognition of faces. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1210–1235.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*(2), 231–248.

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. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, *17*, 629-654.

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.

Clark, S. E., & Shiffrin, R. M. (1987). Recognition of Multiple-Item Probes. *Memory & Cognition*, *15*(5), 367–378.

Corballis, M. C. (1967). Serial order in recognition and recall. *Journal of Experimental Psychology*, *74*, 99-105.

Cortese, J. M., & Dyre, B. P. (1996). Perceptual similarity of shapes generated from Fourier descriptors. *Journal of Experimental Psychology: Human Perception & Performance*, *22*, 133-143.

Cox, G. E., & Criss, A. H. (2017). Parallel interactive retrieval of item and associative information from event memory. *Cognitive Psychology*, *97*, 31-61.

Cox, G. E., & Criss, A. H. (2019). Parametric supplements to systems factorial analysis: Identifying interactive parallel processing using systems of accumulators. *Journal of Mathematical Psychology*, *92*.

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

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.

Deese, J. (1959). On the prediction of the occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*, 17–22.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*, 1-12.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word

recognition. *Psychological Review*, *108*(2), 452–478.

Donkin, C., & Nosofsky, R. M. (2012a). A power-law model of psychological memory strength in short- and long-term recognition. *Psychological Science*, *23*(6), 625–634.

Donkin, C., & Nosofsky, R. M. (2012b). The structure of short-term memory scanning: an investigation using response time distribution models. *Psychonomic Bulletin & Review*, *19*, 363–394.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgements of likelihood. *Psychological Review*, *106*(1), 180–209.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*, *17*, 763–771.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, *38*(2), 467–474.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identication of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143-149.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, *12*(3), 403–408.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, *117*(2), 309–348.

Fific, M., Nosofsky, R. M., & Townsend, J. T. (2008). Information-processing architectures in multidimensional classification: A validation test of the Systems Factorial Technology. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 356–375.

Fox, J., Dennis, S., & Osth, A. F. (2020). Accounting for the build-up of proactive interference across lists in a list length paradigm reveals a dominance of item-noise in recognition memory. *Journal of Memory and Language*, *110*.

Garner, W. R. (1974). *The processing of information and structure*. Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Aki, V., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd ed.). CRC Press.

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.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Griffiths, D. W., Blunden, A. G., & Little, D. R. (2017). Logical-rule based models of categorization: Using systems factorial technology to understand feature and dimensional processing. In D. R. Little, N. Aliteri, M. Fifić, & C.-T. Yang (Eds.), *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms* (p. 245-269). Academic Press.

Hawkins, G. E., Hayes, B. K., & Heit, E. (2016). A dynamic model of reasoning and memory. *Journal of Experimental Psychology: General*, *145*(2), 155–180.

Heit, E., & Hayes, B. K. (2011). Predicting reasoning from memory. *Journal of Experimental Psychology: General*, *140*(1), 76-101.

Hintzman, D. L. (1988). Judgments of Frequency and Recognition Memory in a Multiple-Trace Memory Model. *Psychological Review*, *95*(4), 528–551.

Hollis, G., & Westbury, C. (2016). The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, *23*, 1744-1756.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different Ways to Cue a Coherent Memory System: A Theory for Episodic, Semantic, and Procedural Tasks. *Psychological Review*, *96*(2), 208–233.

Indow, T. (1988). Multidimensional studies of munsell color solid. *Psychological Review*, *95*, 456.

Indow, T., & Kanazawa, K. (1960). Multidimensional mapping of munsell colors varying in hue, chroma, and value. *Journal of experimental psychology*, *59*, 330.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*(4), 486–518.

Johns, E. E., & Mewhort, D. J. K. (2002). What information underlies correct rejections in short-term recognition memory? *Memory & Cognition*, *30*(1), 46–59.

Johns, E. E., & Mewhort, D. J. K. (2003). The effect of feature frequency on short-term reconigiton memory. *Memory & Cognition*, *31*(2), 285–296.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: a noisy exemplar approach. *Vision Research*, *42*(18), 2177–2192.

Kahana, M. J., Zhou, F., Geller, A. S., & Sekuler, R. (2007). Lure similarity affects visual episodic recognition: Detailed tests of a noisy exemplar model. *Memory & Cognition*, *35*, 1222–1232.

Kimball, D. R., Smith, T. A., & Kahana, M. J. (2007). The fSAM model of false recall. *Psychological Review*, *114*, 954–993.

Kleiner, M., Brainard, D. H., Pelli, D. G., Ingling, A., Murray, R., Broussard, C., et al. (2007). What's new in psychtoolbox-3. *Perception*, *36*(14), 1.

Knapp, B. R., Nosofsky, R. M., & Busey, T. A. (2006). Recognizing distinctive faces: A hybrid-similarity account. *Memory & Cognition*, *34*, 877-889.

Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, *39*(16), 2729–2737. Retrieved 2019-03-18, from `http://www.sciencedirect.com/science/article/pii/S0042698998002855` doi: 10.1016/S0042-6989(98)00285-5

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

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.

Lewandowsky, S., Roberts, L., & Yang, L.-X. (2006). Knowledge partitioning in categorization: Boundary conditions. *Memory & Cognition*, *34*, 1676-1688.

Li, X., Liang, Z., Kleiner, M., & Lu, Z.-L. (2010). Rtbox: A device for highly accurate response time measurements. *Behavior Research Methods*, *42*(1), 212-225.

Liew, S. X., Howe, P. D. L., & Little, D. R. (2016). The appropriacy of averaging in the study of context effects. *Psychonomic Bulletin & Review*, *23*, 1639–1646.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1.

Little, D. R., Nosofsky, R. M., Donkin, C., & Denton, S. E. (2013). Logical rules and the classification of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 801–820.

Little, D. R., Wang, T., & Nosofsky, R. M. (2016). Sequence-sensitive exemplar and decision-bound accounts of speeded-classification performance in a modified Garner-tasks paradigm. *Cognitive Psychology*, *89*, 1–38.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*, 376–400.

Logan, G. D., Cox, G. E., Annis, J., & Lindsey, D. R. B. (2021). The episodic flanker effect: Memory retrieval as attention turned inward. *Psychological Review*, *128*(3), 397-445.

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.

McElree, B., & Dosher, A. (1989). Serial position and set size in short-term memory: the time course of recognition. *Journal of Experimental Psychology: General*, *118*, 346–373.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature

production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547-559.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: Evidence against item matching in short-term recognition memory. *Journal of Experimental psychology: General*, *129*(2), 262–284.

Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative-resonance model for short-term recognition memory. *Memory*, *13*, 300–307.

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.

Moneer, S., Wang, T., & Little, D. R. (2016). The processing architectures of whole-object features. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 1443–1465.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology*, *10*, 465–501.

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. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*(6), 609–626.

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 855–874.

Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (p. 67-74). Cambridge, MA: MIT Press.

Nickerson, D. (1936). The specification of color tolerances. *Textile Research*, *6*, 505–514.

Nosfosky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *10*, 104–114.

Nosofsky, R. M. (1986). Attention, similarity and the idenitification-categorization relationship. *Journal of Experimental psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87.

Nosofsky, R. M. (1991). Typicality in logically defined categories: exemplar-similarity versus rule instantiation. *Memory and Cognition*, *19*, 131–50.

Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honour of William K. Estes Vol. 1.* Hillsdale, NJ: Lawrence Erlbaum.

Nosofsky, R. M., & Kantner, J. (2006). Exemplar similarity, study list homogeneity, and short-term perceptual recognition. *Memory & Cognition*, *34*, 112-124.

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.

Nosofsky, R. M., & Palmeri, T. J. (1996a). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review*, *3*, 222–226.

Nosofsky, R. M., & Palmeri, T. J. (1996b). Learning to classify integral-dimension stimuli. *Psychonomic Bulletin & Review*, *3*, 222-226.

Nosofsky, R. M., & Palmeri, T. J. (1997). Comparing exemplar-retrieval and decision-bound models of speeded perceptual classification. *Perception & Psychophysics*, *59*, 1027–1048.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus exception model of classicification learning. *Psychological Review*, *101*, 53-79.

Nosofsky, R. M., & Stanton, R. D. (2006). Speeded old-new recognition of multidimensional perceptual stimuli: Modeling performance at the individual-participant and individual-item levels. *Journal of Experimental Psychology: Human Perception and Performance*, 314–334.

Nosofsky, 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.

Oberauer, K. (2003). Understanding serial position effects in short term recognition and recall. *Journal of Memory and Language*, *49*, 469–483.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General*, *132*, 491-511.

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. (in press). Global matching models of recognition memory. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of Human Memory.*

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101–126.

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.

Osth, A. F., Little, D. R., & Lilburn, S. L. (2019, February). Novelty rejection in episodic memory . Retrieved from osf.io/b2zyk/.

Osth, A. F., Reed, A., & Farrell, S. (2021). How do recall requirements affect decision-making in free recall initiation? A linear ballistic accumulator approach. *Memory & Cognition*.

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*.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*, 437–442.

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.

Prins, N. (2006). The psi-marginal adaptive method: How to give nuisance parameters the attention they deserve (no more, no less). *Journal of Vision*, *13*, 1-17.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of Associative Memory. *Psychological Review*, *88*(2), 93–134.

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., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect : I. Data and discussion. *Journal of Experimental Psychology: Learning Memory and Cognition*, *16*(2), 163–178.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Robinson, K. J., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, *8*, 231–237.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not present in lists. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 803–814.

Rossel, R. A. V., Minasny, B., Roudier, P., & McBratney, A. B. (2006). Colour space models for soil science. *Geoderma*, *133*, 320-337.

Schooler, L., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A model for implicit effects in perceptual identification. *Psychological Review*, *108*, 257–272.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, *115*(4), 893–912.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*, 54–87.

Shepard, R. N. (1978). The circumplex and related topological manifolds in the study of perception. *Theory construction and data analysis in the behavioral sciences*, 29–80.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shepard, R. N., & Chang, J.-J. (1963). Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, *65*(1), 94.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of

classifications. *Psychological Monographs*, *13*.

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., Lee, M. D., Kim, W., & Wagenmakers, E. J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM - retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652–654.

Strong, E. K. J. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*, 447–462.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In *Advances in neural information processing systems* (pp. 3–9).

Torgerson, W. S. (1958). *Theory and methods of scaling.* Wiley.

Townsend, J. T., & Fifić, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Perception & Psychophysics*, *66*, 953-962.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and co-active theories. *Journal of Mathematical Psychology*, *39*, 321–359.

Travis, D. (1991). *Effective color displays: Theory into practice.* Academic Press.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

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.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, *25*, 143–154.

Visscher, K. M., Kaplan, E., Kahana, M. J., & Sekuler, R. (2007). Auditory short-term memory behaves like visual short-term memory. *PLoS Biology*, *5*.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin and Review*, *14*, 779–804.

Wagenmakers, E.-J., Steyvers, M., Raaijmakers, J. G. W., Shiffrin, R. M., van Rijn, H., & Zeelenberg, R. (2004). A model for evidence accumulation in the lexical decision task. *Cognitive Psychology*, *48*(3), 332–367.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

White, C. N., & Poldrack, R. A. (2014). Decomposing bias in different types of simple decisions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 385–398.

Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1022–1041.

Zhou, F., Kahana, M. J., & Sekuler, R. (2004). Short-term episodic memory for visual textures: A roving probe gathers some memory. *Psychological Science*, *15*(2), 112-118.

Appendix A

Determination of JNDs

In order to obtain an estimate of the scale at which observers would be sensitive to changes in the dimensions of the radial frequency stimuli, a small pilot experiment was conducted with two observers (the authors AZ and SL) detected whether two stimulus intervals contained the same stimuli, with four out of five trials containing different stimuli of increasingly smaller differences. The psi method of (Kontsevich & Tyler, 1999) was used to generate the differences for each of the dimensions the in stimulus space for testing during the calibration experiment. This process was repeated for five different starting points in the stimulus space, each for one hundred trials. The same procedure was used for the separable dimension stimuli (authors SL and DL were observers).

Two simple psychometric models, each modeling the change with a cumulative Gaussian psychometric function space with an estimated guess rate, were then tested to characterize performance across the range of changes and starting points: a model where the slope of the psychometric function for each dimension was constant across the entire stimulus range and a model in which the slope of the psychometric function could change as a linear function of the position in stimulus space. A qualitative inspection of the data and model fits indicated no systematic change in the slope as a function of the position in the stimulus range, so constant increments to stimulus dimensions (75% JNDs) were used to construct the stimuli used in the main experiment.

Appendix B

Similarity Scaling Studies

We sought to confirm that the psychological representation of our integral- and separable-dimension stimuli corresponded to the representation assumed by the experiment and modelling analysis (i.e., the ideal representation). Our recognition experiments sampled over 3000 unique items. Collecting pairwise similarity ratings for the entire set would have required a prohibitively large number of comparisons (around 4.5 million). Instead, for both the integral and separable dimension stimuli, we conducted a similarity rating study using a smaller $3 \times 3 \times 3$ set of items sampled from the larger space used in our recognition experiments. Specific values used in each study are shown in Table B1.

Table B1

Values used in the similarity rating study. All values were crossed factorially creating a 3 × 3 × 3 stimulus space.

		Integral Stimuli	
Value	Amplitude	Phase Angle 1	Phase Angle 2
1	1.3	-1.92	-1.92
2	2.08	-0.48	-0.48
3	2.86	0.96	0.96
		Separable Stimuli	
Value	Saturation	Height (pixels)	Bar Position (pixels)
1	8	48	60
2	14	84	96
3	20	120	132

Note: Color was determined using Munsell Hue 5R, Brightness 4, and varying Saturation as shown. Bar Position is measured in pixels from the left border of the rectangle.

Similarity ratings were obtained either from the students as part of the Melbourne School of Psychological Sciences Research Experience Program (Integral stimuli, N = 15) or via a Human Intelligence Task (HIT) placed on Amazon Mechanical Turk (Integral

stimuli, N = 29; Separable stimuli, N = 34). REP participants received subject credit for completion of a one hour session. Mechanical Turk participants received $15-$16 USD for completion of the study. Prior to completing the ratings, participants were shown all 27 stimuli once in random order and were shown examples regarding the least and most similar pairs. Participants were instructed to use the entire response scale. Each of the 351 unique pairs was presented twice for each participant. The order of presentation and the left-right presentation of each pair was randomized. The study was self-paced.

For the integral-dimension stimuli, we removed one participant whose ratings had a correlation with the distances between the stimuli in the ideal representation with $r < .20$. We removed another six participants who did not use the full response scale (i.e., responding with mostly one or two responses). For the separable-dimension stimuli, seven participants were removed for not using the full response scale.

For both sets of stimuli, we first averaged the similarity ratings across observers. We determined the dimensionality of the scaling solution which provided the best fit to the average similarity by fitting scaling models of different dimensionality and minimizing the sum of squared deviations (SSD) between the 351 observed dissimilarities and the predicted stimulus distances (using a Euclidean metric for the integral stimuli and city-block distance for the separable stimuli). A Nelder-Mead SIMPLEX algorithm was used to estimate the best fitting coordinate for each stimulus. We used 100 different starting points for the model and to ensure convergence to the global minimum.

The coordinate of the first stimulus was fixed to the origin; all of the remaining coordinates were free to vary. The number of parameters in each model is given in Table B2 along with the SSD. To facilitate model selection, we computed the Bayesian Information Criterion (Schwarz, 1978) values for each model and each observer by:

$$BIC = n \log\left(\frac{SSD}{N_{datapoints}}\right) + N_{parameters} \log\left(N_{datapoints}\right). \tag{19}$$

For both the integral and separable dimensions, the preferred scaling solution has

three dimensions (see Table B2). We conducted two further tests for each stimulus set. We first fit a scaling solution with three dimensions where the values were constrained to be monotonically increasing in each row and column (*monotonic space*). Second, we fit a scaling solution in which the values of each of the stimuli in a row or column were constrained to be equal (*grid space*). A linear scaling of the predicted distances to the similarities was also introduced for these analyses. Neither of these models were preferred for the integral dimension stimuli (monotonic: SSD = 39.21, BIC = -300.21; grid: SSD = 154.92, BIC = -240.18). The monotonic model fit slightly better than the full model (monotonic: SSD=48.98, BIC = -222.36); but due to the relatively smaller number of parameters (grid: $N_{parameters} = 8$), the grid model was preferred for the separable stimuli (grid: SSD=59.36, BIC = -576.89). Consequently, the estimated coordinates are from the full model for the integral dimension stimuli (see Figure B1) and the grid model for the separable dimension stimuli (see Figure B2).

Table B2

Number of parameters, Sum of Square Deviations (SSD), and Bayesian Information Criterion (BIC) values for scaling solutions of different dimensionality.

	Integral Stimuli		
Dimensions	$N_{parameters}$	SSD	BIC
1	26	325.08	125.45
2	52	73.78	-242.71
3	78	36.68	-335.59
4	104	26.73	-294.31
	Separable Stimuli		
Dimensions	$N_{parameters}$	SSD	BIC
1	26	550.02	310.04
2	52	173.36	57.16
3	78	49.62	-229.56
4	104	32.79	-222.55

For the separable dimensions, the coordinates correspond almost exactly to the ideal coordinates used in the recognition experiment. The correlation between the estimated and ideal coordinates is $r = .96$. For the integral stimuli, there is more confusability between the stimuli particularly at the more extreme values of amplitude. Nevertheless, the dimensions revealed by the MDS analysis seem to correspond to amplitude, phase angle 1 and phase angle 2. The correlation between the estimated coordinates and the ideal coordinates for the integral-dimension stimuli is $r = .77, p < .01$.

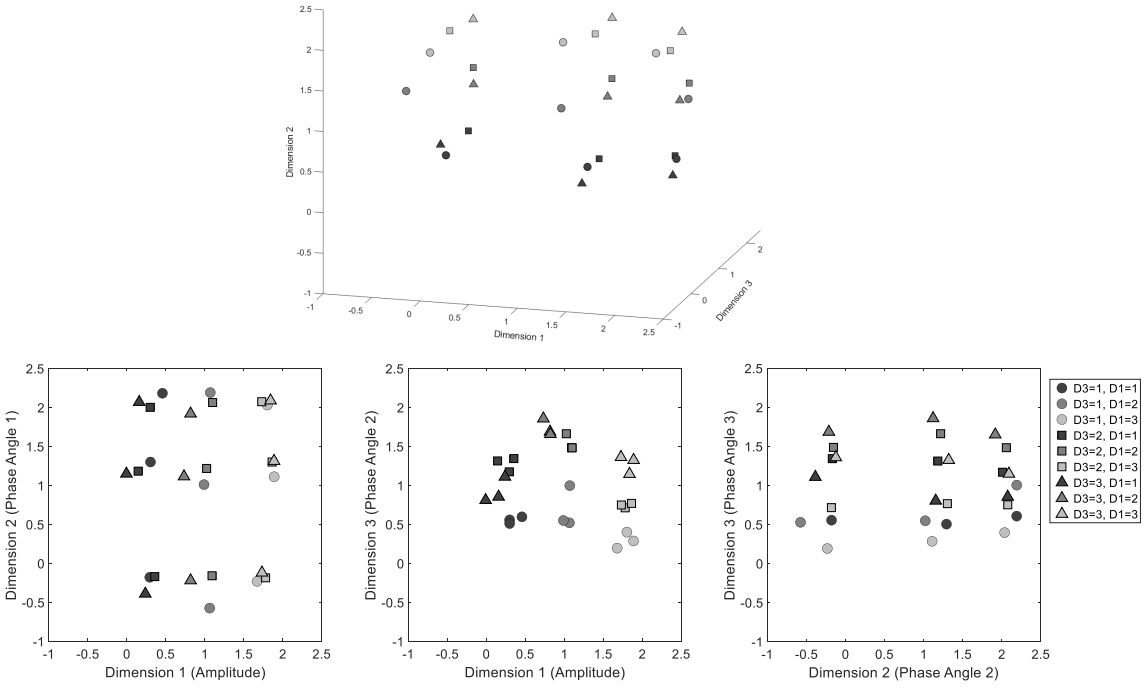

Figure B1. Three-dimensional scaling solution for the integral dimension stimuli.

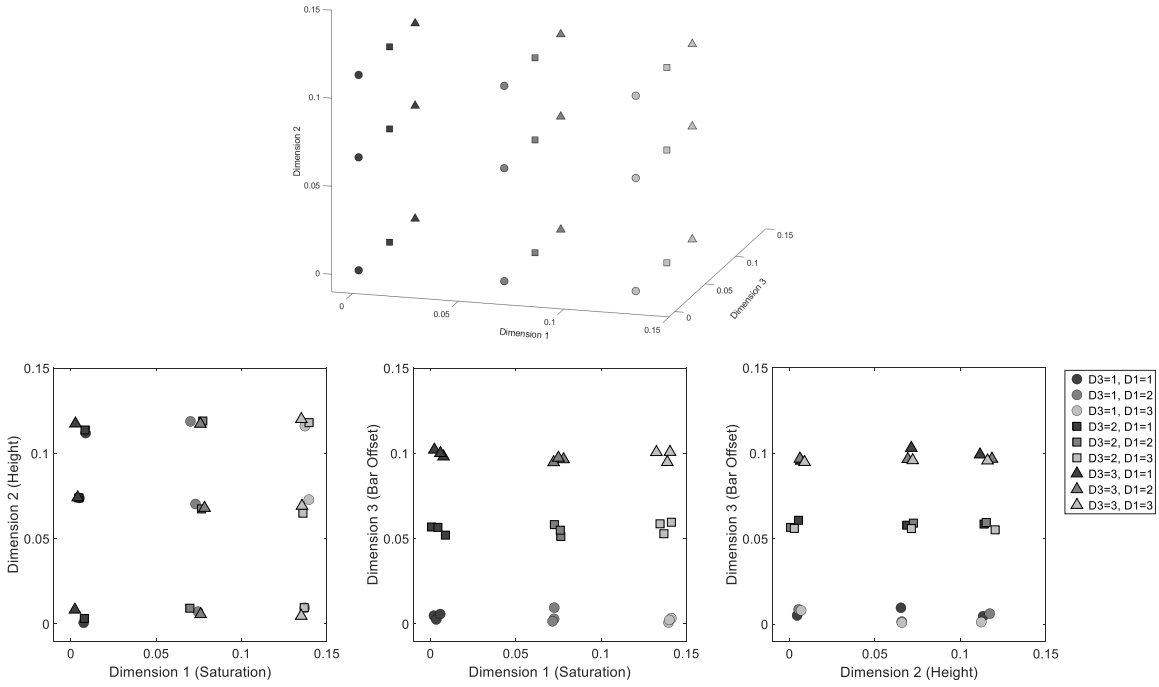

Figure B2. Three-dimensional scaling solution for the separable dimension stimuli. For the lower panels a small amount of random jitter was added to prevent overlap.

Appendix C

Study List Construction

To generate a study list, a dimension was chosen randomly and fixed to a random index value in the range 3 to 10 (see Table 1, Index column). Indices for the values of the remaining two variable dimensions were selected for all three items by selecting a row at random from Table C1. Each row represents the variable dimensions across all three study list items with a minimum step size of two JNDs between each study list item. Value pairs were allocated randomly to dimensions, and the indices from Table C1 were used to set values based on the numbered rows in Table 1. A subset of the full range was used for the study list items to allow the similarity of the lures to be adjusted appropriately outside of the study list range. After setting the values of the study list, the order of study list presentation was then randomized.

Table C1

Indices for study list items.

Value Pair 1		Value Pair 2		Value Pair 3	
3	3	5	5	7	7
3	4	5	6	7	8
3	5	5	7	7	9
3	6	5	8	7	10
3	7	5	5	7	3
3	8	5	6	7	4
3	9	5	7	7	5
3	10	5	8	7	6
4	3	6	5	8	7
4	4	6	6	8	8
4	5	6	7	8	9
4	6	6	8	8	10
4	7	6	5	8	3
4	8	6	6	8	4
4	9	6	7	8	5
4	10	6	8	8	6
5	3	7	5	9	7
5	4	7	6	9	8
5	5	7	7	9	9
5	6	7	8	9	10
5	7	7	5	9	3
5	8	7	6	9	4
5	9	7	7	9	5
5	10	7	8	9	6
6	3	8	5	10	7
6	4	8	6	10	8
6	5	8	7	10	9
6	6	8	8	10	10
6	7	8	5	10	3
6	8	8	6	10	4
6	9	8	7	10	5
6	10	8	8	10	6

Appendix D

Details on Hierarchical Bayesian Models

Attention Weight Parameterization

In models such as the GCM or EB-LBA, for n dimensions a set of $n-1$ attention weights need to be estimated. A difficulty in parameter estimation for three dimensions is ensuring that the two parameters sum to less than 1. We addressed this in Experiments 1-4 by sampling parameters $w*_1$ and $w*_2$. We accomplished this by partitioning the area of a normal distribution into three regions:

$$w_1 = \Phi(w*_1) \tag{20}$$

$$w_2 = \Phi(w*_1 + w*_2) - w_1 \tag{21}$$

$$w_3 = 1.0 - w_2 - w_1 \tag{22}$$

where $w*_1$ ranges from $-\inf$ to inf, $w*_2$ is always positive, and Φ is the cumulative distribution function of the standard normal distribution.

Prior Distributions on Model Parameters

Many of the individual paramaters are sampled from group-level distributions with a defined mean μ and standard deviation σ. Because the β parameters could sometimes take large values and were strictly positive, these parameters were sampled on a log scale. Superscripts are used to denote the group-level distributions:

$$
\begin{aligned}
c &\sim Normal_{(0,\text{inf})}(m^{\mu}, m^{\sigma}) \\
m &\sim Normal_{(0,\text{inf})}(c^{\mu}, c^{\sigma}) \\
k &\sim Normal_{(0,\text{inf})}(k^{\mu}, k^{\sigma}) \\
w*_1 &\sim Normal(w*_1^{\mu}, w*_1^{\sigma}) \\
w*_2 &\sim Normal_{(0,\text{inf})}(w*_2^{\mu}, w*_2^{\sigma}) \\
S &\sim Normal_{(0,1)}(S^{\mu}, S^{\sigma}) \\
a &\sim Normal_{(0,\text{inf})}(a^{\mu}, a^{\sigma}) \\
B &\sim Normal_{(0,\text{inf})}(B^{\mu}, B^{\sigma}) \\
sv &\sim Normal_{(0,\text{inf})}(sv^{\mu}, sv^{\sigma}) \\
t_0 &\sim Normal_{(0,\text{inf})}(t_0^{\mu}, t_0^{\sigma}) \\
log(\beta) &\sim Normal(log(\beta)^{\mu}, log(\beta)^{\sigma}) \\
p &\sim Normal_{(0,1)}(.5, .5)
\end{aligned}
$$

We used only weakly informative prior distributions on the group-level parameters that were similar to those used in a previous hierarchical Bayesian implementation of the EB-LBA (Hawkins et al., 2016):

$$
\begin{aligned}
c^{\mu} &\sim Normal_{(0,\text{inf})}(2,1) \\
m^{\mu}, k^{\mu}, a^{\mu}, B^{\mu}, sv^{\mu}, t_0^{\mu} &\sim Normal_{(0,\text{inf})}(.5,.5) \\
w*_1^{\mu} &\sim Normal(0,2) \\
w*_2^{\mu} &\sim Normal_{(0,\text{inf})}(.5,2) \\
S^{\mu}, p^{\mu} &\sim Normal_{(0,1)}(.5,.5) \\
log(\beta)^{\mu} &\sim Normal(0,10) \\
c^{\sigma}, m^{\sigma}, k^{\sigma}, w*_1^{\sigma}, w*_2^{\sigma}, a^{\sigma}, B^{\sigma}, sv^{\sigma}, t_0^{\sigma} &\sim Gamma(1,3) \\
log(\beta)^{\sigma}, p^{\sigma} &\sim Gamma(1,1)
\end{aligned}
\tag{23}
$$

Details on DE-MCMC Parameter Estimation and Simulation

For each model, the number of chains was set equal to three times the number of participant level parameters. The purely coactive and parallel model architectures used 10,000 burn-in iterations followed by 15,000 iterations. The chains were heavily thinned such that 1 in 10 samples were accepted, resulting in 1,500 accepted samples per chain. The hybrid coactive-parallel required more iterations to converge. For these models, we utilized 30,000 burnin iterations followed by 100,000 iterations where 1 in 40 samples were accepted, resulting in 2,500 accepted samples per chain. A model was considered converged if its Gelman-Rubin (G-R) diagnostic was below 1.2 for all participant and group-level parameters. This criterion was satisfied for all models, with the G-R statistic close to 1.0 in many cases. Individual participant posterior predictives were generated for each posterior sample by simulating the model using the same number of trials as in the original data using 1 in 40 posterior samples.

Appendix E

Analytic Expressions for Parallel Models

Analytics for the parallel global similarity models are constructed from the probability density function (PDF) and cumulative density function (CDF) of an individual LBA accumulator (analytics can be found in Brown & Heathcote, 2008) with parameter vector θ. We denote the PDF and CDF as f and F, respectively.

The likelihood of "old" and "new" responses in a coactive decision architecture can be described as a simple race equation:

$$L(old, t) = f(t, \theta_{old})[1 - F(t, \theta_{new})] \tag{24}$$

$$L(new, t) = f(t, \theta_{old})[1 - F(t, \theta_{new})] \tag{25}$$

The parallel models consist of three accumulators for each response option ("old" and "new") corresponding to each dimension i. The first parallel model we adopted assumed an exhaustive rule for "old" responses and a self-terminating rule for "new" responses:

$$L(old, t) = \sum_i \left[f(t, \theta_{old,i}) \prod_{j, j \neq i} F(t, \theta_{yes,j}) \prod_k \Big[1 - F(t, \theta_{no,k})\Big] \right] \tag{26}$$

$$L(new, t) = \sum_i \left[f(t, \theta_{new,i}) \prod_{j, j \neq i} \Big[1 - F(t, \theta_{no,j})\Big] (1 - \prod_k F(t, \theta_{yes,k})) \right] \tag{27}$$

When the model is exhaustive for both "old" and "new decisions, the likelihoods are as follows:

$$L(old, t) = \sum_i \left[f(t, \theta_{old,i}) \prod_{j, j \neq i} F(t, \theta_{old,j}) (1 - \prod_k F(t, \theta_{new,k})) \right] \tag{28}$$

$$L(new, t) = \sum_i \left[f(t, \theta_{new,i}) \prod_{j, j \neq i} F(t, \theta_{new,j}) (1 - \prod_k F(t, \theta_{old,k})) \right] \tag{29}$$