Journal of Memory and Language 139 (2024) 104556

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Stimulus duration and recognition memory: An attentional subsetting account

Jeremy B. Caplan *, Dominic Guitard

Department of Psychology and Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Alberta, Canada
School of Psychology, Cardiff University, Cardiff, United Kingdom

ARTICLE INFO

Keywords:
Stimulus duration
Presentation rate
Selective attention
List-strength effect
Recognition memory
Mirror effect
Matched filter model

ABSTRACT

Attentional subsetting theory (Caplan, 2023) posits that only a small subset of item features are attended in episodic recognition tasks. This explained a pivotal finding for the development of recognition models: the near-null list-strength effect, where encoding strength influences recognition similarly in mixed-strength lists and pure-strength lists. Most research uses spaced repetition to manipulate encoding strength. However, the origin of the null list-strength effect was a more unusual manipulation of stimulus duration (1 s versus 2 s) — and reported an inverted list-strength effect. We present an attentional subsetting theory of duration that produces inversions — and explains why they are uncommon: Earlier-attended features dwell within a lower-dimensional feature subspace, which participants can sometimes disregard during test trials of pure-strong lists, giving strong-pure items an extra advantage. The model previously only solved for d'. We extend it to generate realistic hit and false-alarm rates by deriving the criterion from attention to each probe. Supporting the theory, two pre-registered experimental manipulations of stimulus-duration reproduced robust inverted list-strength effects, suggesting this type of finding is unlikely due to sampling error. This account of stimulus-duration, explaining inverted, as well as upright and null, list-strength effects, could be incorporated in most models with vector representations

Introduction

At the core of research on episodic memory is the nature of our working representations of items (such as words). Episodic old/new recognition distills this question. Having studied a list of items, discriminating which probe items were on the list (targets) versus those that were not (lures) is in large part interrogating the similarity of working representations to one another. High similarity between list items and lures makes the task more challenging, whereas distinctiveness makes the task easier. Because similarity drives memory behaviour in more complex tasks, a firm understanding of episodic recognition has implications far beyond recognition behaviour, itself.

The development of models of old/new episodic recognition has been substantially driven by two highly replicated findings, the null list-strength effect and the strength-based mirror effect. Both regard what happens to recognition performance when encoding strength is manipulated. Most of the research in this tradition have manipulated strength through spaced repetitions (and sometimes levels of processing; e.g., Ensor et al., 2021; Kiliç et al., 2017; Ratcliff et al., 1990). Data from spaced repetition studies also drove attentional subsetting theory (Caplan, 2023), the theory we expand upon here. However, as we elaborate below, the impetus for that line of research started with a manipulation of stimulus duration – namely, "weak" items: 1 s versus "strong" items: 2 s of study time per word – which produced a list-strength effect that differed from the standard findings. We saw this as potentially offering an interesting boundary condition on theory, and sought to develop a theory of stimulus duration that can explain why duration might produce such different results. Our main focus in this manuscript is therefore to directly apply attentional subsetting theory to manipulations of stimulus duration during the study phase.

List-strength effects. "Strength" experiments start with an experimental manipulation that is thought to modulate encoding strength, resulting in better recognition of a strong-encoded item than a weak-encoded item. The *null list-strength effect* refers to the finding that recognition of a strong item is better than recognition of a weak item — but that strength benefit is about the same size when items are mixed in the same list versus segregated to different lists, namely, pure lists of only strong items or only weak items. This was surprising because one would expect the strong items to have an advantage in mixed lists, because they should experience less competition from the half of the items that were weak; conversely, weak items should suffer in mixed, compared

* Corresponding author.
E-mail address: jcaplan@ualberta.ca (J.B. Caplan).

https://doi.org/10.1016/j.jml.2024.104556
Received 2 February 2024; Received in revised form 18 June 2024; Accepted 13 August 2024
Available online 3 September 2024

to pure lists, due to additional competition from the strong items that are present on mixed lists.

The finding was first noted by Ratcliff et al. (1990), who quantified it with their ratio-of-ratios (RoR) measure,

$$\text{RoR} = \frac{d'(\text{D mixed})/d'(\text{S mixed})}{d'(\text{D pure})/d'(\text{S pure})}, \quad (1)$$

where we use "D" to denote the strong condition (e.g., long stimulus duration) and "S" to denote the weak condition (e.g., short stimulus duration) with reference to deep and shallow levels of processing (Craik & Lockhart, 1972). This is meant to emphasize our contention, which shall become clear shortly, that weaker strength conditions often result primarily in processing (and encoding) of shallow features whereas stronger conditions result in processing and encoding of additional deeper features. The typical measure of memory, d', is the difference in hit rate minus false-alarm rate after z-transforming each. A person with no memory would make hits and false-alarms at the same rate, so $d' = 0$, whereas person who can discriminate targets from lures would make hits at a greater rather than false alarms, making d' positive. Thus, a list-strength effect would produce RoR > 1.

The specific way in which RoR > 1 is predicted is model-dependent, but one of the easiest ways to understand this is to consider the assumption that recognition is inversely proportional to the amount of competition from other list-items. Consider a fixed list length L, and pure lists of D items or S items, with a strength effect such that d'(D pure) > d'(S pure). Each D item is subject to strength-based competition from $(L-1)$ D items, whereas each S item is subject to competition from $(L-1)$ S items. If competition is indeed dependent on strength, then for pure lists, each D probe is subject to more competition than each S probe. In a mixed list, each D item is subject to competition from $L/2$ S items and $(L/2-1)$ D items and each S item is subject to competition from $L/2$ D items and $(L/2-1)$ S items. The total strength-based competition is thus greater in pure D lists than in mixed lists, because $(L-1)D > (L/2)S+(L/2-1)D$. The opposite holds for S items: $(L-1)S < (L/2)D+(L/2-1)S$. As long as false alarms do not neutralize these effects, d'(D pure) < d'(D mixed) but d'(S pure) < d'(S mixed). Thus, Ratcliff and colleagues found it curious that their experiments produced RoR values close to 1. These findings presented challenges to existing models, both local-trace models, where a separate image (usually a vector) is stored for each element of a list (e.g., a word or a pair) and global-matching memory models, where memories are summated within a single memory structure. This inspired the development new models including a particular class of local-trace models that incorporated differentiation (Shiffrin et al., 1990; Shiffrin & Steyvers, 1997) and other models assuming strict orthogonality of item representations, elaborated below.

Inverted list-strength effects. Ratcliff et al. (1990) noted that their first experiment (1 s versus 2 s duration) in fact produced a significant *inverted*[1] list-strength effect, with RoR < 1. Ratcliff and colleagues acknowledged this, but they were most struck by the absence of an "upright" list-strength effect. The inversion could be due an underlying null list-strength effect, with sampling error accounting for the apparent inversion. This was also understandable given that their second experiment, using longer durations, produced a RoR slightly above 1, but still smaller than one intuitively would have expected. Ratcliff et al. (1994) in fact found another statistically significant inversion (RoR = 0.7) in a manipulation of duration (although very short durations: 50 ms versus 200 ms) but this was one of nine RoRs across the experiments they reported in that paper (see their Table 7). These RoRs could indeed have reflected measurement variability around an underlying RoR = 1. Also worth mentioning, Sahakyan (2019) found inverted (but non-significant) list-strength effects with repeated presentations and 1.25 s/item, although the method was unique in that it compared massed ("weak") and spaced ("strong") repetition. Given the occasional inverted list-strength effect produced by attentional subsetting theory (Caplan, 2023), we wondered if their results were not simply due to random noise but a legitimate inverted list-strength effect. In that case, a formal account of inverted list-strength effects at short durations might also tell us why the inverted list-strength effect might go away when all durations are longer.

Strength-based mirror effects. The strength-based mirror effect refers to the often replicated finding that while hits (calling a target item old) increase in pure-strong versus pure-weak lists, this is accompanied by a comparable decrease in false alarms (calling a lure item old), of similar magnitude (Kim & Glanzer, 1993; Stretch & Wixted, 1998). A mirror effect did not emerge from older models, and thus demanded additional mechanisms, such as differentiation or a variable strength threshold (response criterion), which we describe below. Attentional subsetting theory dealt only with d' (Caplan, 2023). Because of the importance of the strength-based mirror effect to theory, we develop the theory further to estimate hit rate and false-alarm rate. Attentional subsetting is compatible with models that incorporate either existing approach (differentiation or criterion adjusted based on knowledge of the statistics of encoding strengths) to produce mirror effects. But in extending the theory, we found a third approach afforded by the subsetting concept. Although we present proof of principle and do not support our approach over the other two, we suggest this approach be considered as a possible way of side-stepping the ongoing debate about the cause of strength-based mirror effects.

Attentional subsetting theory. Addressing only d', Caplan (2023) proposed a novel continuum account, capable of explaining near-null as well as positive and inverted list-strength effects. The key assumptions were (Fig. 1):

(1) *Subsetting:* (a) Only a small *subset* of the features of a stimulus are attended during the study phase — and thus, encoded. (b) Just like during the study phase of an experiment, at test, only a small subset of the features of the probe stimulus are attended — and thus, available to be compared to memory.
(2) *Item-specificity of subsetting:* Due to prior knowledge, each item has its own idiosyncratic subset of features that tend to be attended (although the feature-subset may be modulated by factors like task set and proximal stimuli).
(3) *Reiteration:* When a stimulus is encountered a second time, it is highly likely that the same or similar subset of features will be attended at both times (assuming task set and contextual factors have not changed too much). Consequently, for probe items, often the same subset of features are attended at test (and compared to memory).

For example, when viewing the word CHEESE, a participant might think of a yellow wedge of Swiss cheese with holes in it, about the size of one's hand— a handful of features (assumption 1) that are item-specific (assumption 2). When encountering the word CHEESE a second time, such as a recognition probe, it is likely that the participant will think again of the same features: yellow, wedge-shaped, containing holes and hand-sized (assumption 3). Some support for assumptions 2 and 3, that features are relatively (albeit not perfectly) stable across encounters comes from experiments that asked participants to overtly generate features of stimuli. Wu and Barsalou (2009) found reliable item-specific influences of task-set on generated features and Medin

[1] Note that Ratcliff, Shiffrin and others have used the terms "positive" and "negative" describing list-strength effects corresponding to $RoR > 1$ and $RoR < 1$, respectively. This terminology would directly describe $\log(RoR)$. But because some people use "negative" to describe non-significant statistical outcomes, we prefer the terms "upright" and "inverted," respectively. The latter terminology is theoretically loaded, but by design. It reflects the perspective Ratcliff et al. (1990) had going into their list-strength studies, where $RoR > 1$ was expected.

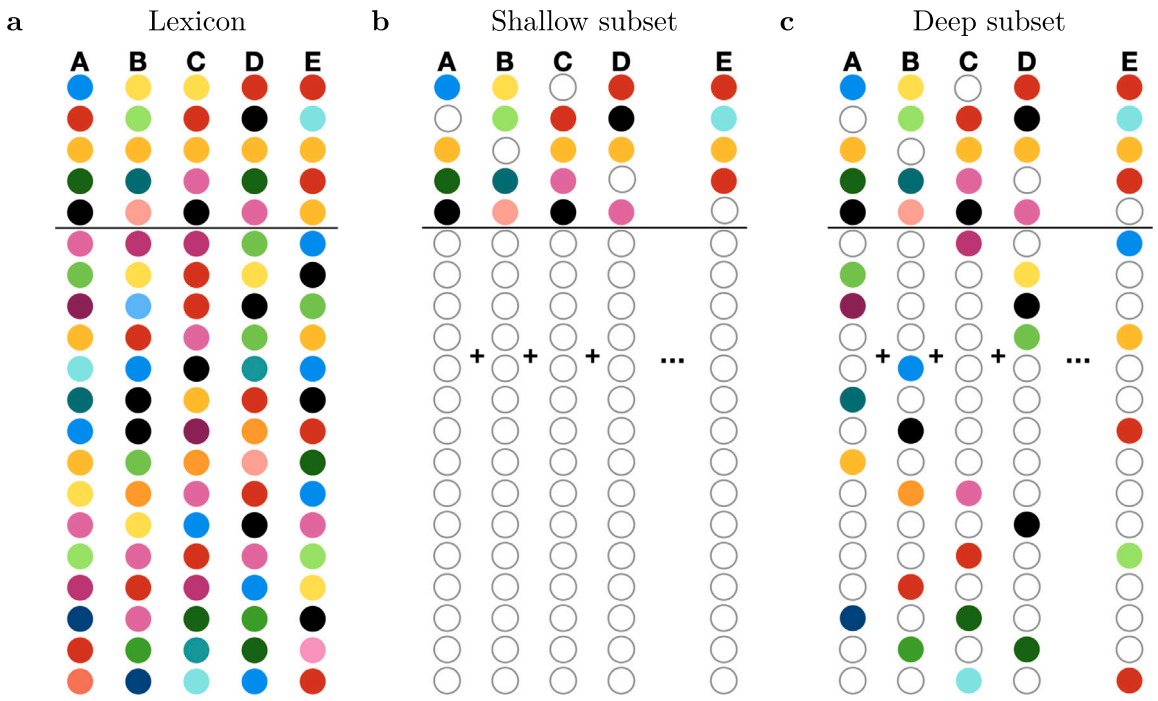

Fig. 1. Schematic depiction of how attentional subsetting could work in a model of stimulus duration, where the "shallow" condition is nested within the "deep" condition. Grey unfilled circles denote features that are not attended (and thus not encoded). We assume that the shallow features are dense, not sparsely subsetted, whereas the deep features are sparsely subsetted. The horizontal line separates the shallow feature-subspace (above) from the deep feature-subspace (below). (a) The full vector representation of five items (i.e., lexicon or knowledge). (b) The attended subset of features when studied in the shallow condition. The example list here consists of items A through D, where the memory is their sum, A+B+C+D. E is an example of a lure probe, also shallowly attended, as assumed for pure shallow lists. (c) The same as (b) but for the deep condition; in addition to the shallow features attended during both study and test, additional sparsely subsetted features from the deep subspace. Sparseness would be more pronounced if the deep subspace were greater; it is kept small here for illustration only.

and Shoben (1988) found reliable item-specific effects of task-context information on judgements of prototypicality and similarity. If these effects are reliable across participants, it is plausible that they would also be fairly stable across trials within an experimental session.

Among other things, subsetting can be seen as a way to deal with the paradox of similarity: typical experimental stimuli are really almost entirely composed of common features (for words: the font, size, colour of the text, etc.) so clearly participants are successful in disregarding most of these. If one turns this around and assumes that only a small number (a handful) of features are attended, and those are particular to each item, one can obtain representations with far fewer common features. When the subset is sparse (a small number of attended features within a high-dimensional feature-space), there is almost no confusion due to common features across items. This enabled even the very simple matched filter model (Anderson, 1970), which is just a sum of item vectors evaluated with the dot product as a measure of the strength of match to memory, to produce a near-null list-strength effect (Caplan, 2023). Given the simplicity of the matched filter model, this also suggested that attentional subsetting could have similar effects in many, if not all, models that assume a vector representation of items.

As a continuum account, the theory could also explain why other experimental manipulations do not produce a null list-strength effect. For example, the production effect (reading aloud or typing produces a memory advantage over reading silently) exhibits a list-strength effect, a bigger advantage for produced over non-produced words in mixed lists than in pure lists (Bodner et al., 2016; MacLeod et al., 2010). Attentional subsetting theory provides a simple account of production effects (Caplan, 2023), as elaborated by Caplan and Guitard (2024). Production strengthens items through additional processing of phonological features, which are presumed not to be sparsely subsetted, so those features produce substantial overlapping features across list items, producing sizeable list-strength effects. As we describe shortly, the theory also produced inverted list-strength effects, suggesting they should be expected under certain conditions.

Objectives

Attentional subsetting theory thus far provides an alternative theoretical account of list-strength effects as quantified by d', including predicting legitimate (not due to sampling error) inversions of the list-strength effect. Here we test if such inversions of the list-strength effect can be confirmed, given their scarceness in published research. We present two experiments with this as their primary goal.

But first, it is important to go beyond d'. Caplan (2023) only derived the model to solve for d', because it can be derived based on the forms of the expected distribution of matching strengths for old items and the distribution for new items. One wonders whether the model can even produce realistic hit and false-alarm rates. The next question is whether the model could produce a strength-based mirror effect, given that the mirror effect has been a contentious area of debate between groups of modellers (differentiation accounts versus variable criterion), as we elaborate below. The idea of subsetting could be incorporated in any model with a vector representation of items. For this reason, the theory could piggyback on a model like REM and produce mirror effects based on differentiation, or it could piggyback on models that incorporate processes for variable criterion and produce mirror effects for that reason. But attentional subsetting makes possible a third account: We add to attentional subsetting theory a mathematically simple way in which the model can derive a criterion that is reasonably close to optimal, *based on immediate processing of the probe item, itself, during the test trial*— in other words, setting the criterion based on task-relevant attentional processing of the current probe item. Different than current variable-criterion accounts, this does not require the participant to have any knowledge of the statistical properties of encoded items. In

the remainder of the introduction, we describe attentional subsetting theory as applied to stimulus duration, specifically. We then review list-strength effect findings and theories, delineating how our own theory differs. We describe theories of the strength-based mirror effect and describe how our new extension of attentional subsetting theory can produce a mirror effect but with a new mechanism. We then present a replication attempt of Experiment 1 of Ratcliff et al. (1990), which in fact produced both an inverted list-strength effect, where weak items were studied for 1000 ms and strong items, for 2000 ms. Because that procedure produced very small effects of duration (as in the original study), we followed this with a slight modification of the experiment, where weak items were presented for 500 ms. Our hope was to produce a bigger difference in performance for the strong and weak condition and reduce the number of participants who nominally exhibited the reverse effect, which is problematic for interpreting the results. After reporting some additional exploratory findings that speak to the theory, we conclude with a discussion of implications for experimental manipulations of strength, and for models of recognition memory.

A theory of stimulus-duration

Attentional subsetting theory can be adapted specifically to model the special case of stimulus duration with the following assumptions, illustrated in Fig. 1:

1. *Feature types.* We distinguish two classes of stimulus features (see Craik & Tulving, 1975 for a related view about the importance of such distinctions). First, *shallow* features, such as the phonology or orthography of a word, are considered to be relatively small in number, so these features will repeat a lot across stimuli and introduce considerable similarity-based confusion. Second, *deeper* features, such as those related to the meaning of a word (semantic features) or related to imagery, are considered to dwell within a very large feature space.[2] The attended subset of deeper features will tend to be *sparse* and introduce very little similarity-based confusion across items.
2. *Short duration.* For such "weak" items, shallow features are attended. These features are drawn from a low-dimensional subspace, which cannot be sparsely subsetted.
3. *Long duration.* For such "strong" items, the shallow features are also processed, but as study time continues to unfold, additional deeper features will be processed. Because deeper features are drawn from a high-dimensional subspace, they are sparsely subsetted. Thus, strong and weak items include shallow features that introduce similarity-based confusion but strong items also have features that largely avoid such confusion.
4. *Disregarding.* Finally, in some cases participants may be able to *disregard the shallow features* during the test phase. Specifically, when tested on a pure-strong list, shallow features have little diagnosticity compared to the deeper features, if these are plentiful in the strong condition. Thus, metacognitive knowledge of the list composition may enable participants to disregard shallow features in this condition (just as participants evidently can disregard other non-diagnostic features such as the fact that all stimuli are words, printed in the same font, etc., as noted earlier). Importantly, this would not be feasible in mixed lists where one does not know the strength-status of a probe.

The idea of distinguishing perceptual from semantic features has many precedents (e.g., Burgess & Hitch, 1999; Seidenberg & McClelland, 1989), and Malmberg and Nelson (2003) and Criss and Malmberg (2008) proposed further that perceptual features tend to be processed earlier than semantic features and features accessed through controlled processing. The inversion of the list-strength effect (illustrated in Fig. 2) was found by Caplan (2023) when disregarding shallow features during tests of pure-strong lists, was feasible, as we elaborate below.

[2] These distinctions, such as perceptual versus semantic, are only meant to make the point; the dimensionality of the feature space is more important to the argument, as we elaborate in the General Discussion.

Theories of the list-strength effect in recognition

The null list-strength effect in recognition memory has been replicated numerous times, with strength usually operationalized with a manipulation of the number of spaced repetitions of an item, but sometimes stimulus duration, as we consider here and occasionally levels of processing (e.g., Ensor et al., 2020, 2021; Ratcliff et al., 1990). Ratcliff et al. (1990) and Shiffrin et al. (1990) viewed the lack of a sizeable positive list-strength effect (RoR>1) as face-value evidence that items do not compete with one another, which was especially puzzling due to the stable finding of reduced recognition as list length increases.

Differentiation accounts of the null list-strength effect. This led Ratcliff et al. (1990) and Shiffrin et al. (1990) to propose that each item is stored in its own local memory trace and the match of a probe to memory would be computed for each trace individually, before aggregating matching-evidence across traces. This motivated the development of influential local-trace, differentiation-based models (e.g., McClelland & Chappell, 1998; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997). A strong item will have more encoded features. A strong item will provide more evidence of having been on the list but each strong-item trace will also provide more evidence of mismatching a lure probe. Mixed lists include more weak items than pure-strong lists, so mixed lists will produce a higher false-alarm rate than pure-strong lists. Similarly, compared to pure-weak lists, mixed lists include more strong items, which will lead to a lower false-alarm rate. The net effect can be a greater difference in d' between pure lists than between strong and weak items on mixed lists. The frequent near-null list-strength effect arises in REM because of an approximate balance between an underlying positive list-strength effect and this cause of an inverted list-strength effect.

Other accounts. Murdock and Kahana (1993) produced near-null list-strength effects by assuming competition accumulates over multiple lists and thus saturates after the first few lists in an experimental session. Still other modellers viewed the null list-strength effect as indicating that item representations are approximately orthogonal. If items are orthogonal, they will not be confused with one another, so they would be evaluated with little influence of other studied item. By design, some models therefore constructed item representations deliberately to be orthogonal to one another (e.g., Chappell & Humphreys, 1994; Dennis & Humphreys, 2001), appealing to item–context associations as the cause of list-length effects.

Caplan (2023) noted that null list-strength effects may not be as general as suggested, including upright list-strength effects with the production effect, mentioned above, and the observation that experiments manipulating duration, spaced repetition or levels of processing typically do show RoR values above 1, just not significantly so, including the second experiment reported by Ratcliff et al. (1990).

Attentional subsetting account. Caplan (2023) showed how nearly null list-strength effects could be produced without assuming local traces and without assuming *strict* orthogonality of item representations. Briefly, it was assumed that the participant attends only to a small subset (a handful) of features of a given item (illustrated in Fig. 1), but those will often be the same small subset of features attended upon a repeat presentation as happens during the test phase. When viewing the word Squirrel, the participant may think of a bushy tail and the chattering voice. When Squirrel is repeated (for example, as a recognition probe), by virtue of those features having come to mind rapidly and with little effort, the participant is likely to think again of

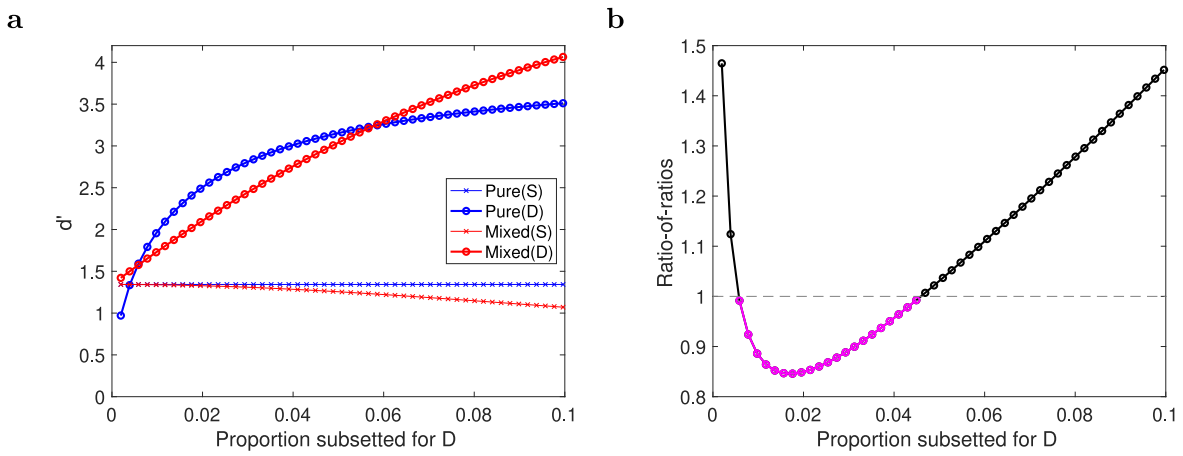

Fig. 2. The attentional subsetting account of list-strength effects. The plots show the output of an example model of a manipulation of stimulus duration such as Experiment 1 of Ratcliff et al. (1990). Here, condition S=short duration and condition D = long duration. List length, $L = 32$. $n_s = 64$ "superficial" features. $n_S = 16$ features subsetted per item and is fixed. $n_d = 512$ "deep" features. n_D, the proportion subsetted per long-duration item, varies parametrically. (a) d' as a function of item type, list type and n_D. (b) Ratio-of-ratios (RoR) as a function of n_D. The dashed line denote a null list-strength effect (RoR = 1) and pink denotes where the list-strength effect is inverted. Bear in mind that this figure is designed to illustrate the sensitivity of list-strength effect to the number of deep features while holding constant the dimensionality of the deep feature subspace. Consequently, in this graph, the left-hand values correspond to the sparse regime for and the right-hand values correspond to the dense regime, respectively. In contrast, the way we think about strengthening by adding sparse features (such as with stimulus duration, the focus of this manuscript) versus adding dense features (such as with production, investigated in Caplan & Guitard, 2024), we assume the dimensionality of the feature subspace is what primarily drives the difference.

the bushy tail and the chattering voice. As elaborated by Caplan (2023), task demands and contextual factors could modulate those subsetted features in interesting ways. The reiteration of the attentional subset does not need to be perfect; the assumption is simply that, aside from factors that modulate attention or feature-relevance, the subset will tend to be similar during repeated exposures to an item.

There are usually a very large number of "deep" features such as many features related to meaning and imagery. If one assumes only a handful of features are attended on an item, when this item-specific attentional subsetting approximates sparse representations, explaining why items did not seem to compete with one another. Sparse vectors (mostly zeros) produce very little overlap-based confusion between themselves, so the strengths of other items within a list will exert very little influence on judging a probe item. If the stronger condition adds sparsely subsetted features, a RoR very close to 1 is obtained (Fig. 2, as derived by Caplan, 2023).

In this view, orthogonality is not an invariant feature of item representations (cf. Chappell & Humphreys, 1994) but can be approximated by sparse subsetting. Because the sparse subset of one item can consist of different features than the sparse subset of another item, even items that are extremely similar to one another (same values of a given feature) can be functionally dissimilar. This account also suggests situations that might deviate from orthogonality. For "shallower" features like phonological or orthographic features, the feature space is smaller and more compact; phonemes and letters recur across words at a high rate. When features are drawn from a compact feature space, attentional subsetting cannot be sparse (corresponding to the regime towards the right of Fig. 2) and list-strength effects become pronounced, as when memory is improved by reading aloud (the production effect, MacLeod et al., 2010 and see Caplan & Guitard, 2024). In this non-sparse regime, two items with similar features will most likely have some of those similar features attended on both items.

Attentional subsetting account of inverted list-strength effects. Finally, the list-strength effect inverts is as follows. Start with a regime in which the weak condition (short duration) has only enough time to attend and encode shallow features, whereas the strong condition (long duration) has additional time to process and encode deep, sparsely subsetted features. Pure-weak lists contain items that only have shallow features, which introduce overlap-based confusion. In mixed lists, shallow features cannot be disregarded (because some probes might be weak), so the weak items are susceptible to interference due to the shallow features from the strong items as well as the weak items. Strong items fare better than weak items in mixed lists, because they benefit from having additional, sparsely subsetted features that are more distinctive than the shallow features. But in pure-strong lists, if shallow features can be safely disregarded, judgements will be based on fewer features, but those features will be the more diagnostic, sparse features rather than the more confusing, densely subsetted shallow features. This comes at a cost of reduced functional vector length. So if the strong condition is not strong enough, in that strong items do not have very many additional sparsely subsetted features attended, there may no longer be a net benefit for pure-strong items.

Theories of the strength-based mirror effect

Glanzer and Adams (1985) reported what they termed the Mirror Effect, where items that are better recognized as belonging to a studied list are also better ruled out when they appear as lures. This pattern was found for a large number of manipulations of stimulus characteristics (most notably, word frequency), and was robust across various experimental conditions (Glanzer & Adams, 1985, 1990). They posed this as a challenge to existing models of item recognition. In most models, studying items would tend to increase their strengths, right-shifting their distribution (and often increasing their variance). However, this would have no effect on the distribution of matching strengths to lure items that were not encoded in memory[3]; the "new" distribution would remain where it was. As the hit rate increases, there is no *a priori* reason to expect any effect on lure items.

However, mirror effects produced by comparing two different stimulus pools are hard to interpret, as one can never run out of hypothetical properties that might not be equated between the stimuli. When one takes concerted efforts to control such characteristics, the mirror effect can go away; when disentangled, Neath et al. (2021) found that more pure manipulations of a stimulus property largely affected either hit rate or false alarm rate (associative recognition produced similar

[3] In models with vector representations of items, this is the case if vectors are mean-centred, which is typically done. Otherwise, in some models, the strength of the lure distribution can even increase, producing the opposite of a mirror effect.

dissociations; MacMillan et al., 2022). Consistent with this, Cortese et al. (2010) and Cortese et al. (2017) failed to find mirror effects at the level of individual words, in item analyses we shall follow up. Stimulus-based mirror effects might be explained by dual-factor (or more) accounts rather than a single factor simultaneously increasing hit rate while decreasing false alarms.

The *Strength-Based* Mirror Effect (Criss, 2006, 2010; Hockley & Niewiadomski, 2007; Kim & Glanzer, 1993; Starns et al., 2012; Stretch & Wixted, 1998) manipulates encoding conditions between sets of stimuli drawn from the same overall stimulus pool, avoiding all possible stimulus-characteristic confounds. As already noted, strength is most commonly manipulated with spaced repetitions (but also stimulus duration and levels of processing). Because a manipulation of strength is at least unitary as an experimental factor (unlike stimulus-based mirror effects), the *strength-based* mirror effect does remain a challenge to models of the form proposed by Glanzer and Adams (1985).[4]

Differentiation accounts of the strength-based mirror effect. In differentiation models, strengthening a particular local memory trace both increases the subsequent match of a target probe to the trace and increases the degree to which lure probes would mismatch the trace (e.g., Criss, 2006, 2009, 2010; McClelland & Chappell, 1998; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997). For example, Retrieving Effectively from Memory (REM; Shiffrin & Steyvers, 1997) not only often produces a null list-strength effect, it can also produce a mirror effect. Different than earlier local-trace models (such as Hintzman, 1988), REM uses the number of features that match between the probe and each encoded trace, but also the number of mismatching features to estimate the likelihood that the item was studied. If a strength manipulation results in more features (accurately) encoded into the trace, the strengthened trace will lead to higher (log) likelihood that the item was studied, but also higher likelihood that a lure item was not studied. The (correct) mismatches increase as the matches increase, comprising a mirror effect. Elegantly, a mirror effect emerges from the core calculations of the model, without needing further assumptions.

Criterion-shift accounts of the strength-based mirror effect. The major alternative theoretical account of the strength-based mirror effect is the one proposed by Glanzer and Adams (1985) and their descendants, including Cary and Reder (2003) and Stretch and Wixted (1998). They proposed that a model could compute a log-likelihood of an item with a given strength having been produced by the strong versus weak expected strength distribution (Hirshman, 1995, proposed the idea of using the range of expected strengths to derive a criterion without any log-likelihood calculation). Having studied a pure-strong list, the model could thus safely increase the criterion with little cost to the hit rate, since the strong items will easily exceed a higher strength threshold, but with the advantage that lure items with chance strengths that are somewhat higher would be rejected, reducing the number of false alarms. In other words, if the expected strength distributions are known, the criterion could be optimally adjusted. One asset of the criterion-shift account is that it can be attached to any model that produces strength distributions, with or without local traces. Some support for such adjustments in criterion have been found (e.g., Starns et al., 2012). The main weak point of such models is that they arguably demand too much knowledge on the part of the participant about the expected strength distributions (but see Dubé et al., 2019; Tong & Dubé, 2022a, 2022b; Tong et al., 2019 for evidence in defence of people having this kind of knowledge). Koop et al. (2019) showed that the mirror effect is found under conditions in which they argued criterion-shifts are not plausible, after either very few test trials or in conditions in which the need to change criterion would not be blatantly obvious to participants.

In the next section, we extend attentional subsetting theory to produce separate hit and false-alarm rates. In doing so, we propose a principle by which participants could derive a good criterion based purely on immediate processing of the current probe item, influenced only by meta-knowledge of the task. We check if the model can produce realistic values of hit and false-alarm rates, as well as being able to produce mirror effects.

Attentional subsetting theory

The basic idea of feature-subsetting has been around for a while. The original log-likelihood/criterion-based account dates to Glanzer and Adams (1985) who proposed two conditions might differ in the number of features stored, and the number of features extracted at test. This meta-knowledge can be used to evaluate the match and could produce a mirror effect, computing likelihood ratios in their attention/likelihood theory (Glanzer & Adams, 1990). Similar to the later Glanzer et al. (1993) model, they assumed that a subset of features are "marked" and those features also have values. The main differences are that in our account, attentional subsets are in many circumstances quite *sparse*, and respectively, sparseness is practical only because we also assume the same subset will tend to reiterate itself similarly at test (in contrast to Glanzer et al., 1993 who assumed a random re-sampling upon each exposure). Also, Glanzer et al. (1993) assumed the set of marked features is evaluated, rather than their values; here we assume the values, themselves, are compared, and the markedness (which we call attentional subset) gates which values propagate through the comparison process. Introducing some notation for attentional subsetting, let $n_{C,i}$ denote the small number of attended features of a given item, where C can denote a particular experimental condition and i denotes a given item (Caplan, 2023; Caplan et al., 2022). When implemented in the matched filter model (Anderson, 1970),[5] the memory is a simple sum over the L list items,[6]

$$\mathbf{m} = \sum_{i=1}^{L} \mathbf{w}_{C,i} \otimes \mathbf{f}_i, \tag{2}$$

[4] There are accounts of mirror effects based on signal-detection theory (DeCarlo, 2007, 2010), which may be instructive here. Such models focus on characterizing the forms of the distributions of strength values. They do not explain where those strengths come from, but for example, DeCarlo (2007) proposes participants encode items in several discrete ways, each of which is associated with a mean encoding strength plus some variance (where the variance is equivalent across these distributions). This kind of mixture model can produce net strength distributions that resemble unequal-variance models with just two strength distributions (one for unstudied items and one for studied items). Our model does not directly include encoding-strength variability, but the number of encoded features has consequences similar encoding strength. As can be seen in Eqs. (A.3) and (A.4) and shown by Caplan (2023), the variance of target strengths is greater than that for lure strengths (although the variance come closer as the list length increases). It would be interesting to explore a mixture model where the $n_D = n_S$ but attentional subsetting has a higher probability of (all-or-none) succeeding both at study and test in the D than the S condition, or some more complex mixture.

[5] It is important to note our choice to formulate the idea within the matched-filter model is chiefly for clarity of exposition and to build our intuition. We are not suggesting this is a complete model of recognition. The matched-filter model stores a list of items as a sum of the corresponding item vectors, and probe items are evaluated by computing dot products of the probe vector with the memory vector. The model has serious limitations, but its simplicity allows us to see how attentional subsetting may function in a model. Because of its simplicity, it is also easy to see how the same principles could be embedded within well developed memory models that have been able to address problems with the matched-filter model. We are in no way endorsing the matched-filter model as a "best" or "complete" model of recognition memory, although it may be instructive to note that such a simple model is sufficient to produce the phenomena of interest here.

[6] We use the term "item" loosely, but it always refers to a putative vector in a knowledge "lexicon" corresponding to one discrete stimulus such as a single word.

where (column) vectors are denoted in boldface, i indexes distinct items, items are n-dimensional, with independent (aside from when we consider similarity across items), identically distributed values drawn from $N(0, 1/\sqrt{n})$ (approximately, but not strictly, normalized and mean-centred), and $\otimes$ denotes elementwise multiplication. As discussed in Caplan (2023), n could be arbitrarily large, we assume that the extremely large number of task-irrelevant features are disregarded, so n can be thought of as the full potential functional feature-space in a given task-setting. The $\mathbf{w}_{C,i}$ vectors are attentional masks, with value 1 for attended features and 0 otherwise. Thus, selective attention zeros-out any unattended features and lets the attended features pass through. Importantly, $\mathbf{w}$ are indexed by both item and condition, expressing the idea that the specific subset of features attended will tend to vary across items, i, but be relatively stable across presentations of a given item, although they can be substantially modulated by task-conditions, C.

Since this is meant to correspond closely to actual attention, it is plausible that the participant has a good estimate of $n_{C,i}$ for a given item at the time the stimulus is presented as a probe (perhaps even as a consciously accessible count of numbers of features processed). At first we assume for a target item, i, the attentional subset applied at test is the same as that applied at study. But in general, the subsets can differ, and we will introduce one particular such deviation shortly. For now, the judgement is based on the matching strength of a probe item, x, to the stored memory $\mathbf{m}$, their dot product

$$s_x = (\mathbf{w}_{C,x} \otimes \mathbf{f}_x) \cdot \mathbf{m}, \tag{3}$$

If $\mathbf{f}_x$ were stored (attentionally masked) in $\mathbf{m}$, then s_x will tend to be greater than if the item were not stored. Thus, the response is based on whether or not s_x exceeds a criterion, θ:

$$\begin{cases} \text{“Old”} & s_x > \theta_{C,i} \\ \text{“New”} & s_x \le \theta_{C,i} \end{cases} \tag{4}$$

As with other models, the value of $\theta_{C,i}$ is important. If $\theta_{C,i}$ is extremely low, the model will call everything “Old,” producing lots of hits but also lots of false alarms. If θ is too high, the model will call everything “New.” If the model were to have access to the full expected distributions of strengths for target and lure items, it could choose an optimal value for $\theta_{C,i}$. But note that the mean matching strength is directly related to this attentional subset size, $\mu_{\text{target}} = n_{C,i}/n$. For non-presented items, μ_{lure}=0. Stretch and Wixted (1998) (for example) defined “optimal” criterion placement, as “the point that maximizes the proportion of correct responses,” which for symmetric reward conditions is halfway between the two means, thus $\theta_{C,i} = .5n_{C,i}/n$.

Let us pause to emphasize that we assume the model (participant) has direct access to the approximate number of features it has just processed of a given probe item. The unbiased threshold is then simply what one expects if half those attended features match memory. The participant does not need to remember anything else about the list, nor to keep track of criteria used during other trials. All that is relevant is current processing of the probe item. The number of features attended will in turn be influenced by the participant’s meta-knowledge of the task, so attention will be driven by characteristics of the list, but at the level of the list as a whole, not varying across items. On the other hand, the threshold derived for each item will be specific to the item, itself, since we have allowed for $n_{C,i}$ to differ as a function of both condition, C, and item, i. The threshold will change from one item to the next, but the meta-cognitive rule dictating that threshold is assumed to be relatively fixed over the course of a set of test trials.

Whereas it may seem implausible that participants can accurately enough estimate the full expected strength distributions (but there is support for the idea that participants have knowledge of the statistical properties of encoded stimuli; Dubé et al., 2019; Tong & Dubé, 2022a, 2022b; Tong et al., 2019), it is plausible that the participant has access to the approximate number of features attended on the probe item they are currently processing. However, this assumption remains to be tested in future research. Without knowing anything about any other probe items (nor even, at this stage, anything about their memory for the list, itself), the participant could plausibly select a criterion that is close to optimal. This heuristic only makes sense once we assume feature-subsetting. Without subsetting, there is no meaning to the idea of a particular number of features processed. This heuristic results in a criterion-shift, but unlike the prior criterion-shift models, the model needs absolutely no information about either the expected target or lure strength distributions. If the number of attended features varies across items ($n_C = n_{C,i}$), then so will the threshold, $\theta = \theta_{C,i}$, but this is due to immediate processing (attentional subsetting) of the current stimulus, not due to cumulative knowledge of the studied list or even of the probe set (Starns et al., 2010), consistent with a mirror effect emerging even on the first test trial (Koop et al., 2019).

Next we derive the hit and false alarm rates with a main focus on the model of stimulus duration, where the strong items include weak-item features plus additional features subsetted from a much larger-dimensional feature space.

Extension of attentional subsetting theory to hit rate and false alarm rate

Previous authors have proposed participants can make use of characteristics of the probe item to adjust their response criterion, for example, when considering high- versus low-frequency words (e.g., Gillund & Shiffrin, 1984; Stretch & Wixted, 1998). Here we make this process quite specific to the item. For a given item, we assume the participant is aware of how many features they readily extract from the item. For now, let us drop the index i and assume that this number is n_C and is constant for a set of stimuli. But note first, that we retain the index, C, because the number of attended features could vary as a function of condition, and second, the n_C specific features, themselves, will still tend to be different for each item. Moreover, we start with the simplest assumption that participants will tend to process items at test similarly to how they did so during study, so the same n_C is applied at test as at study for the case of pure lists (but we will amend this for mixed lists and in the nested model). For all models considered by Caplan (2023), $\mu_{\text{target}} = n_C/n$ and $\mu_{\text{lure}} = 0$. The criterion is simply

$$\theta_C = \frac{1}{2}\frac{n_C}{n}. \tag{5}$$

Next we let $C \in \{S, D\}$, where condition S represents something like a shallow level of processing and D represents something like a deep level of processing or short versus long stimulus durations.[7] In typical strength-based mirror effect experiments, hits and false alarms are compared between pure lists, so $\theta_S/\theta_D = n_S/n_D$. In other words, the criterion will be set higher when it can be — when the expected distribution of matching strengths is greater. This acts a bit against the increased hit rate for strong (D) versus weak (S) items, but not entirely. And in exchange, it reduces the false-alarm rate because the higher criterion will not be as often duped by high-strength lure items.

Given μ_{target}, μ_{lure}, σ_{target} and σ_{lure}, the hit rate is the proportion of strengths from the target distribution that will fall above the threshold, θ. The false-alarm rate is the same for the lure distribution. The use of d' implies a normal distribution, so we use the error function, erf(), to integrate strength from the threshold to infinity. Thus:

$$P(\text{hit}) = \int_{\theta_C}^{\infty} N(\mu_{\text{target}}, \sigma_{\text{target}}) = 1 - \left(0.5 + 0.5\ \text{erf}\left(\frac{\theta_C - \mu_{\text{target}}}{\sqrt{2}\sigma_{\text{target}}}\right)\right) \tag{6}$$

[7] Hintzman (1994) proposed that the participant tests the probe item for learnability or memorability and then it essentially rescales itself; that mini-test gives the participant the information needed to customize their threshold. We are proposing something similar but arguably asking even less of the participant because we are not suggesting participants do any learning test. Rather, they simply use meta-knowledge of the information they have just extracted about the stimulus.

$$P(\text{false alarm}) = \int_{\theta_C}^{\infty} N(\mu_{\text{lure}}, \sigma_{\text{lure}}) = 1 - \left(0.5 + 0.5\ \text{erf}\left(\frac{\theta_C - \mu_{\text{lure}}}{\sqrt{2}\sigma_{\text{lure}}}\right)\right) \quad (7)$$

$$\text{where } \text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Our primary focus in this manuscript is the nested model, applied to stimulus-duration, where the D condition includes the S condition feature subspace, and the latter is more compact than the additional deep subspace. However, in the Appendix we consider two different model variants considered by Caplan (2023), where the two conditions are assumed to dwell within the same feature subspace and both are sparse. This deepens our mathematical intuition, and those models may describe other experimental manipulations.

The full model of strength as duration

Caplan (2023) proposed that when an item is studied, the earliest features attended occupy a low-dimensional subspace, such as orthographic or phonological features. This builds on a suggestion by Tulving (1968) and Lewis (1979) that the order of retrieval of features from superficial to semantic (and see Criss & Malmberg, 2008; Malmberg & Nelson, 2003), although they did not suggest any differences in feature-space dimensionality. With more processing time (and perhaps also with repeated presentations of an item), the extracted features are sparsely subsetted from a much higher-dimensional subspace, such as semantic or imagery-based features.[8] Thus, in strength manipulations, it may often be the case that the S condition is in the non-sparse regime whereas the D condition includes both those S features plus additional features that are in the sparse regime. Consistent with the former assumption, Yonelinas et al. (1992) observed a large list-strength effect when duration was manipulated between 50 ms and 200 ms for weak and strong conditions, respectively; in our framework, this corresponds to a non-sparse regime. Even when RoR = 1, the so-called "null list-strength effect" may be a misnomer. The account proposed by Caplan (2023) implies that there in fact is an influence of other studied items, which can be seen when list-strength is manipulated around the strength levels of what is usually called the "weak" condition — due to non-sparse subsetting (as reported by Yonelinas et al., 1992). The null effect of typical list-strength manipulations is because "strong" items add sparsely subsetted additional features, and thus do not introduce any *more* non-negligible noise due to feature overlap with other items.

This arrangement can also produce a mirror effect. The mirror effect produced by the full-probe model in the Appendix (although arguably cognitively implausible) is lost with masked probes because the terms due to feature overlap became negligibly small ($O(n_C^2/n^3)$ rather than $O(n_C/n^2)$, where we use "big-O" notation to summarize terms of a particular order or higher, emphasizing the term that dominates in the limit). This is because with sparse subsetting, the chance of overlap of attended feature-subsets is quite small. But if the n_S subspace were low-dimensional, and not sparsely subsetted, the $O(n_C)$ terms are reintroduced. In this model, the false-alarm rate is reduced when the threshold is increased proportionally to n_C because similarity amongst weakly encoded items is substantial.

Disregarding. In a pure list, we can consider two cases, illustrated in Fig. 3a and b, respectively: First, we could assume the participant intuits that it is better to disregard the S features, since they produce similarity-based confusion between items, including between targets and lures, thus $\theta = (1/2)n_D/n$ and $\theta = (1/2)n_S/n$ for conditions D and S, respectively. Alternatively, the participant might be unable to ignore those early-attended features, in which case $\theta = (1/2)(n_S + n_D)$ for both conditions. Note that for simplicity, we are assuming the S and D feature spaces are strictly segregated, but if they are not, additional cross-terms would be added. Unlike the fully sparse regime considered in the masked probe model (Appendix), the lure distribution of strengths does not have a negligible variance, so reducing the threshold will substantially increase the false-alarm rate in condition S versus D.

In the first model version, where participants can successfully ignore the S features when judging a pure-D list, each cross-term contributes $V_{xy} = (\Omega_{CC}/n_C)/n^2$ (defined in the Appendix) but now the overlaps will differ. Overlap arises from choosing n_C features out of n_c, where the lowercase index refers to the size of the feature subspace specific to condition C, if $\binom{n}{k}$ denotes n choose k: $\Omega_{CC} = \binom{n_c}{n_C}^2/n_c$, which will be large for small n_c and small for large n_c. For the second model version, where participants cannot selectively ignore the S features, the pure-S lists are unchanged but for the pure-D lists, V_{xy} is the sum of V_{xy} for S and D conditions in the first version.

When the model can disregard the S features during pure-D lists, the hit rate increases but the false-alarm rate is constant as the "D" condition increases in strength, increasing n_D (Fig. 3a). When n_D is small, the hit rate for the so-called "deep" condition suffers, which makes sense, because when $n_D < n_S$, the "deep" items have fewer features encoded, and the advantage due to the sparseness of the deep feature space is insufficient to compensate for that. But the sparseness increasingly benefits the "D" hit rate as more features are attended and soon shows an advantage over the "S" condition. Meanwhile, the false-alarm rate is not just invariant to n_D, it is quite small due to the sparseness. Assuming a handful of "deep" features are attended in the strong condition (to the right of $n_D = 5$ for this parameter set), the D condition produces a mirror effect compared to the S condition.

The idea that a participant might be 100% successful in disregarding superficial features may be unrealistic — and likewise for the assumption that a participant might be absolutely unable to disregard any superficial features. A more realistic model might be in between the two models in Fig. 3: some superficial features might be successfully disregarded and some other portion not (or if the S and D feature spaces are not *strictly* segregated, the overlapping features may not be disregarded).

Discussion of the model

The heuristic for deriving the criterion item-by-item can produce strength-based mirror effects, without differentiation and without a good estimate of the expected distribution of null strengths. This stands as proof of principle, although we do not present any evidence ruling out the other accounts. Two arrangements produced a fairly symmetric mirror effect could be explained by immediate processing of the probe alone. The first is the full-probe model (Appendix). This model produced a substantial list-strength effect (Caplan, 2023), inconsistent with many list-strength findings, so it may provide insight into mirror effects in situations in which list-strength effects are sizeable (e.g., the production effect, Bodner et al., 2016; Hopkins & Edwards, 1972; MacLeod et al., 2010 and short presentation durations, Yonelinas et al., 1992). We chose the $(1/2)\mu_{\text{target}}/n$ value because it marks the midpoint between the two distribution means and is an optimal placement of the threshold. It is also robust; a threshold too high will produce 0% false alarms but 0% hits and a threshold too low will produce 100% hits but 100% false alarms. A participant who has a vastly miscalibrated threshold would thus not even be able to express the information they do have in memory. One half the expected target distribution keeps the participant close to the comfortable middle, reducing the risk of a drastically miscalibrated threshold. That said, it is clear that the expressions in Eqs. (A.5)–(A.6) would still produce a mirror effect if the 1/2 were replaced with some other coefficient, although the further one deviated from 1/2, the more asymmetric the mirror effect would become. The threshold does not need to be perfectly tuned to produce a mirror effect, speaking to its plausibility.

The second account works with a model appropriate for stimulus-duration that was found to produce small list-strength effects and

[8] This echoes Ratcliff and McKoon (1989) who showed that associative information is retrieved later than item or feature-matching information.

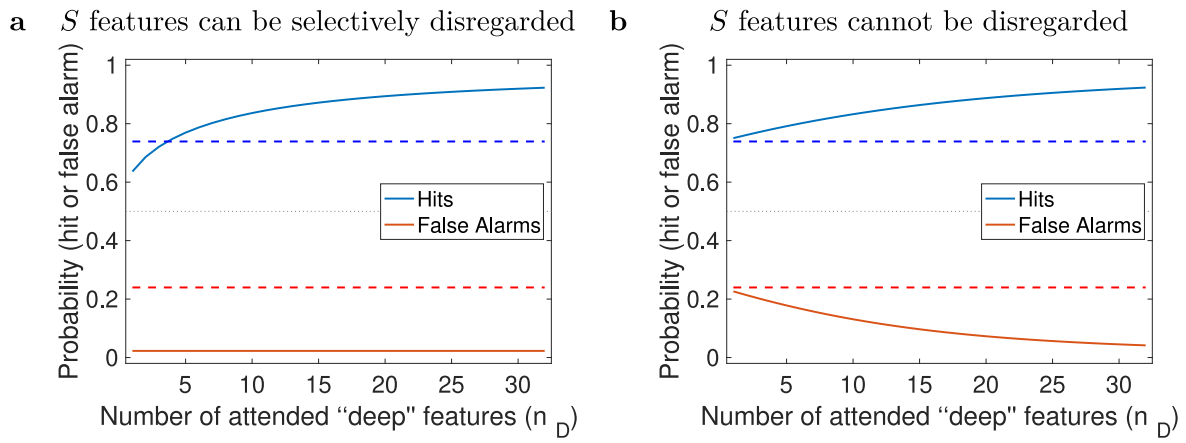

Fig. 3. The nested model of stimulus duration with the ability to selectively ignore S features when tested on a pure D list (a) or not (b). Keeping with Caplan (2023) and sticking close to the experimental design of Ratcliff et al. (1990), Experiment 1, list length is set to 32. $n_s = 64$ "superficial" features. $n_S = 16$ features subsetted per item. The dashed line plots the hit rate and false alarm rate for pure-S lists, which was not varied. n_d was fixed at 512. The solid lines plot hit and false-alarm rate for pure-D lists as a function of the number of (additional) D features encoded per item.

even inverted list-strength effects (Caplan, 2023). The important assumption was that the weak condition drew attention to features that occupy a low-dimensional subspace and are thus not sparse, leading to confusion due to feature-overlap. In both cases, n_C need not be estimated accurately. In fact, we suggest it is plausible that the participant processes the probe in much the same way as during the study phase (particularly for short study–test intervals). It then seems plausible that n_C should be immediately accessible. A more realistic model would include variability in n_C. This would produce largely the same effects, including a mirror effect, with a cost of adding variability across probes (easily confirmed with a simulation, replacing n_C with $n_{C,i} \sim N(n_C, \sigma_C)$). The implication would be that probe stimuli that draw attention to a larger number of features (Fig. 7) would be subject to a higher evidence criterion as probes (higher $\theta_{C,i}$), leading to fewer hits and fewer false alarms. However, this might be offset by factors that separately influence false alarms, such as the overlap of the shallow features (or even some deep features) with other studied items.

However, mirror effects are typically asymmetric, often with far less effect on false alarms than on hits. It is noteworthy that the other two model variants we explored in fact produce this type of result, where hit rate is influenced by strength but false-alarms not. Importantly, one of those conditions was found for the model of duration when features producing confusion due to feature-overlap can be disregarded in pure-strong but not in pure-weak lists, coinciding with conditions that produce an inverted list-strength effect.

Our theoretical account of stimulus duration can produce null list-strength effects and substantial mirror effects. But it specifically predicts true (not sampling error) inverted list-strength effects when shallow features can be at least partially disregarded in tests of pure long-duration lists. Next we report two new experiments aimed to test this result.

Experiment 1: A replication attempt of an inverted list-strength effect and asymmetric mirror effect

We conducted a pre-registered replication attempt of experiment 1 of Ratcliff et al. (1990), which produced a significantly inverted list-strength effect along with an asymmetric mirror effect (a large effect on hit rate but a small effect on false-alarm rate).

Rationale and goals. The theory can explain how the list-strength effect can sometimes invert, as in the first experiment of Ratcliff et al. (1990). This novel prediction distinguishes the theory from other accounts of the list-strength effect, which have been more focused on explaining the null (or near-null) effect. Some (accounts relying on orthogonal representations) have not suggested why it inverts. Others (accounts relying on differentiation) explain near-null list-strength effects by in fact assuming a cause of an inversion (differentiation) that nonetheless is often well offset by a source of upright list-strength effect. Models like REM, that function this way, have many ways in which this balance might be weighted towards a net inversion. But all accounts of list-strength effects would be justified in disregarding inversions if the scarce reports of inversion are not real, but perhaps a statistical fluke due to variability around a true null list-strength effect. If we can replicate the inverted list-strength effect, that would emphasize that the inversion, itself, needs to be explained. If we observe inverted list-strength effects under the kinds of conditions attentional subsetting theory implies they might be observed, that would reinforce our continuum account of list-strength effects. It would also provide data that could inform the conditions under which differentiation models might be expected to produce net inverted list-strength effects.

Second, we test for a mirror effect, where hits and false alarms both differ between strength levels. Also, we produced an inverted list-strength effect in the model that assumed that disregarding shallow features was possible in pure-strong lists; this assumption also produced a very pronounced asymmetric mirror effect, where hits vary considerably with strength but false alarms change very little (Fig. 3a). Granted, there are numerous ways a mirror effect can become asymmetric, but out of curiosity, we sought to test whether the same individuals who produced an inverted list-strength effect also show the asymmetry. Third, exploratory analyses tested how response times varied across conditions, potentially speaking to the timecourse of retrieval of shallow versus deep features, and whether the correct rejection rate exceeded the hit rate, as produced by the model. Fourth, we wondered whether the conclusions of Neath et al. (2021), that mirror effects are due to more than one separable underlying factor, is also seen at the item level, following Cortese et al. (2010, 2017). Within a condition, we sought to test whether individual items show a mirror effect: a word that is better identified as a target is also better ruled out as a lure. As evident in the theory (and the word-pool manipulations by Neath et al., 2021), hit rate and false alarm rate can be influenced by different, often independent factors, so this was far more exploratory and reported for both experiments together.

The stopping rule was to collect sets of 10 participants until the critical Bayes Factor for the Pure/Mixed×Strength interaction was conclusive (>3:1 or <1:3). However, after collecting 100 participants (after exclusions), although the p value was (just) under 0.05, the Bayes Factor was around 1.5, still quite inconclusive. Rather than spend more money collecting more data, we analysed the data. We had not

anticipated this prior to the pre-registration, but a considerable number of participants failed the basic manipulation-check — that is, if the 2 s duration produces "stronger" encoding than the 1 s condition, then performance should be better for strong than for weak items. We therefore follow up with unregistered exploratory analyses broken down into participants who passed that manipulation check (both for pure lists *and* for mixed lists) compared to those that do not. For the subset who passed both manipulation checks, the interaction indicating an inverted list-strength effect was clearly significant and clearly conclusive; thus, although one should not forget that this was post-hoc, for those "valid" participants, the stopping rule was already surpassed, despite being under our initial target sample size.

Data availability. Data, materials and scripts can be found at https://osf.io/39cz8.

Methods

This experiment was a pre-registered (pre-registration available at https://osf.io/rx9jb and data at https://osf.io/39cz8) replication of Experiment 1 of Ratcliff et al. (1990). Deviating from the pre-registration, we planned to check for a recency confound. Because plots of d' as a function of serial position and of test position produced no suggestion of such confounds, we did not pursue this further. Additional analyses that were not pre-registered are denoted as "exploratory". The procedures were approved by a University of Alberta ethical review board.

Participants. Participants were recruited via Prolific (prolific.co), (a) were native speakers of English, (b) were of British, American or Canadian nationality, (c) had normal or corrected-to-normal vision, (d) had no cognitive impairment or dementia, (e) had no language-related disorders, (f) were of ages between 18 and 30 years, and (g) had an approval rating of at least 90% on prior submissions at Prolific. Demographic information (Questions 1 through 6) from Prolific is self-reported by the participants and the approval rating is computed by Prolific. Participants were paid £7 for their participation in a session lasting around 30–45 minutes.

To keep the sample uniform, participants were excluded if they took more than a ten-minute break. Participants were also excluded if their overall d' (collapsed across list and item type, but excluding practice trials) was below 0 (chance), which would suggest they misunderstood the task or the response mapping or were not able to perform the task at the very basic level. On this basis, one participant was excluded for taking a break longer than 10 min and 4 because their overall $d' < 0$, leaving $N = 101$. We had planned to exclude any participant who responded with the same key (either "old" or "new") to more than 90% of the trials were to be entirely excluded, on suspicion of mindlessly pacing through the experiment, but there were no such participants.

Sample size and stopping rule. The original experiment (Ratcliff et al., 1990) had 5 participants with 6–9 sessions each for a total of 37 subject-sessions. This makes it tricky to estimate required power. Our first target sample size was 70, about double the number of subject-sessions, partly to take into account the fact that we expected having more subject-contributed variance. Due to the differences that should, in principle, be immaterial to the effect (different stimuli, one session/participant, randomly mixed lists; these are detailed below), it was conceivable that our sensitivity differs. As already mentioned, we deviated from our pre-registration, stopping at $N = 100$.

Materials. Stimuli were the 1000 words from the Toronto Word Pool (Friendly et al., 1982), displayed in 40 point size Times font in the centre of the screen. Each list was composed of 32 nouns for study, followed by 64 old/new recognition probes, half of which were just studied (targets) and the other half of which were never seen in the experiment (lures). Words were drawn at random, anew for each participant. Strong words were presented for 2 s and weak words for 1 s, with no inter-stimulus interval. Pure lists were composed of all strong items (pure-strong) or all weak items (pure-weak). Mixed lists were composed of half strong and half weak items, with strength order drawn at random. Following Ratcliff and colleagues, each counterbalance set of four lists included one pure-strong and one pure-weak list, but two mixed lists to equate data collection rates for all item types (Item Strength[strong, weak] × List Type[Mixed, Pure]). Condition-order was random within each counterbalance set of four lists.

Procedure. The online experimental session was controlled via PsyToolkit (Stoet, 2010, 2017). Each session started with one 10-word mixed practice list with interleaved instructions, excluded from analyses. The test phase was self-paced. Responses faster than 100 ms were trapped and a 5-s message displayed the message "Too Fast!" to prevent participants from speeding through the experiment.[9]

Data analyses. Single trials that were signalled "Too Fast!" (or under 100 ms) were excluded trial-wise. Participants were excluded entirely from any analysis for which they had missing data after trial-exclusions.

Our primary measure was d', with the log-linear correction favoured by Hautus (1995), adding 0.5 observation correction to hits, false alarms, misses and correct rejections,[10] computed for each participant and each of the four conditions separately. Because this correction can sometimes distort the results, in the pre-registration we planned to analyse hit and false-alarm rates separately, which we do. We also had planned to check the results with hits minus false alarms to check for complications to the interpretation of the results; this we have not done because the results were clear-cut in this regard and did not seem to warrant a separate analysis of hits — false alarms. The ratio-of-ratios was computed $[d'(\text{mixed strong})/d'(\text{mixed-weak})]/[d'(\text{pure-strong})/d'(\text{pure-weak})]$ and was log-transformed prior to statistical tests and correlation across participants. The pre-registration stated that an interaction, where hits increase and false alarms decrease in strong, compared to weak, lists would be considered support for the mirror effect. This seemed unnecessary; we report strength effects for hits and false-alarms individually and evaluate the difference of those differences, computing the index[11] $\omega = \left[\text{HR(pure strong)} - \text{HR(pure weak)}\right] - \left[\text{CR(pure strong)} - \text{CR(pure weak)}\right]$. If hits and false alarms move to the same degree in opposite directions, this index will be zero. If hits move more than false alarms, the index will be positive, indicating the predicted form of the asymmetry. The pre-registration erroneously stated that asymmetry of the mirror effect would be quantified with hit rate-(1–false alarm rate) for each participant. It is this, but the difference between strengths for hit rate and false alarm rate, respectively. We do report this within list-type analysis anyway, because while it may not strictly test the theory, at least our specific simulations lead to a clear prediction. In Experiment 2, we report this analysis as exploratory.

Our main analysis of interest for the list-strength effect was a repeated-measures ANOVA on d' with design Item Strength [Strong, Weak] × List Type[Mixed, Pure]. An interaction was considered evidence of deviation from the null list-strength effect.

[9] Due to a programming oversight, such trials were presented again to all participants except the last 10 participants and were sometimes still below 100 ms if the participant held the key down the throughout the too-fast message. There is presumably some contamination of the data from those participants, where some trials were immediate repeats of the prior probe. However for the final 10 participants, the implementation was fixed and the number of such trapped trials was low (9 experimental trials out of a total of 7680 trials across the 10 participants), so we think the effect on the overall results is minimal. This remained corrected for Experiment 2.

[10] We had this incorrect in our pre-registrations, where we had mistakenly stated the correction was to hits and false alarms only.

[11] This index is not perfect, but rather, a quick-and-dirty measure we can use for exploratory purposes. In evaluating data and model output, it is also important to examines hits and correct rejections individually to obtain a full picture, which we will do.

To test whether the inversion of the list-strength effect roughly co-occurs with the asymmetry of the mirror effect, we compared the mirror effect asymmetry index, ω, between participants with RoR < 1 versus RoR ≥ 1 (there were in fact no participants with RoR = 1). Statistical tests are reported with both Classical and Bayesian approaches. Significance is assessed with $\alpha = 0.05$ but p values near-threshold are interpreted with caution. Bayes Factors are considered to provide support for the null hypothesis if $BF_{10} < 1/3$ or for the hypothesis if $BF_{10} > 3/1$ (Kass & Raftery, 1995). Analyses of false alarm rate using the three-level factor have the Greenhouse–Geisser correction applied to correct for violations of sphericity and post-hoc pairwise comparisons are Holm-corrected t tests.

Deviations from the original study. Because this was a replication attempt, here we list all elements we pre-registered that deviated from the original study. These were deviations that we felt were superficial and should not reduce our expectation to replicate the original findings. Ratcliff and colleagues collected data from a small number of participants who each performed several sessions; we had more participants but only one session each. The original study was conducted in person whereas ours was conducted online, with no direct interaction between participant and researcher. The original study excluded trials shorter than 200 ms and greater than 2500 ms; we chose more inclusive criteria, excluding trials shorter than 100 ms and longer than 10,000 ms. The original study had 14 lists per session; we had 12. The original study had no practice list; we included a short 10-word mixed practice list. The original study blocked strength within the mixed lists; we constructed lists with a random shuffle of mixed/strong. Finally, we added the "Too Fast!" deterrent.

Results

Before manipulation checks

List-strength effect. A repeated-measures ANOVA on d' (Table 1 and Fig. 5) with design Mixed/Pure × Item Strength [Strong/Weak] revealed a significant main effect of Item Strength, $F(1, 100) = 45.15$, $MSE = 0.055$, $p < 0.001$, $\eta_p^2 = 0.31$, $BF_{\text{inclusion}} > 1000$. The main effect of Mixed/Pure was not significant, $F(1, 100) = 0.52$, $MSE = 0.08$, $p = 0.47$, $\eta_p^2 = .005$, $BF_{\text{inclusion}} = 0.18$. Our effect of interest, speaking to the nature of the list-strength effect, was the interaction. It was significant, $F(1, 100) = 4.40$, $MSE = 0.05$, $p = 0.038$, $\eta_p^2 = 0.04$, but the Bayesian ANOVA was not conclusive, $BF_{\text{inclusion}} = 1.54$. At face-value, this is consistent with our previous prediction: a reliable but small interaction. The Bayes Factor is biased against small-magnitude effects.

We next disentangled the effects on d' by analysing hit rates and false alarms separately. For Hit Rate, the main effect of Item Strength was significant, $F(1, 100) = 58.95$, $MSE = 0.003$, $p < 0.001$, $\eta_p^2 = 0.37$, $BF_{\text{inclusion}} > 1000$. The main effect of Mixed/Pure was not significant, $F(1, 100) = 2.85$, $MSE = 0.003$, $p = 0.095$, $\eta_p^2 = 0.03$, $BF_{\text{inclusion}} = 0.39$ (nearly a supported null, although just inside the "inconclusive" range). The interaction was not significant, $F(1, 100) = 1.29$, $MSE = 0.002$, $p = 0.26$, $\eta_p^2 = 0.013$, $BF_{\text{inclusion}} = 0.38$, nearly favouring a null interaction. For the false alarm rate, now a one-way ANOVA on a factor with three levels (to avoid duplicating false alarms for "strong" and "weak" items in mixed lists): Mixed/Pure-Strong/Pure-Weak produced a non-significant main effect, $F(1.8, 176) = 0.90$, $MSE = 0.003$, $p = 0.40$, $\eta_p^2 = 0.009$, $BF_{\text{inclusion}} = 0.079$.

The ratio of ratios (RoR), which is based on d', was slightly above 1 on average (mean) but the median was slightly below 1 (Table 1). A t test of the log(RoR) against zero was not significant, $t(98) = -0.51$, $p = 0.61$, $BF_{10} = 0.13$. Comparing the RoR to 0.88, the value reported by Ratcliff and colleagues, $BF_{10} = 0.12$, also favouring the null. This is, of course, a naïve application of Bayes Factors but tells us that we do not have the resolution to differentiate between RoR = 1 and RoR = 0.88.

Mirror effect. To evaluate the mirror effect, we turn to the pure lists only. The hit rate was greater in pure-strong than pure-weak lists, $t(100) = 6.36$, $p < 0.0001$, $BF_{10} > 1000$. If there were a perfect mirror effect, this would be parallelled by an equal effect in the opposite direction in the false-alarm rate but the difference for false alarms was not significant, $t(100) = -1.16$, $p = 0.25$, $BF_{10} = 0.21$. To be more direct, the change in the hit rate was significantly greater than the change in false-alarm rate, $t(100) = 3.53$, $p < 0.001$, $BF_{10} = 33.80$, resembling the model with feature-disregarding (Fig. 3a).

With manipulation checks

The following analyses were not anticipated in our pre-registration and should be read as exploratory. Fig. 4a plots the cumulative distribution functions of the item-strength effect, d'(Pure-Strong)−d'(Pure-Weak) and d'(Mixed-Strong)−d'(Mixed-Weak), respectively. If the manipulation of duration influences encoding strength as intended, both of these measures should be greater than zero. In fact, for each check, about one third of the participants failed; duration did not produce an overall increase in d'. Interestingly, the distribution of strength effects is broader for pure than for mixed lists, with both more big positive strength effects and more big negative strength effects than mixed lists. This may be due to the fact that for mixed lists, a single false-alarm rate went into the difference measure, whereas for pure lists, each list type had its own false-alarm measure, adding more measurement variance to the calculation.

If the additional duration is not helping those participants, the premise of measuring a list-*strength* effect is undermined. For those participants, one might even view the 1 s duration as the "strong" condition and the 2 s duration as the "weak" condition. This seems irrational; in fact, the duration conditions must be functioning differently than Ratcliff and colleagues had in mind, at least for some participants, but possibly for all participants. This echoes the theme of Caplan (2023), that the word "strength" has been overloaded, and may refer to a collection of different processes and effects, each of which should be understood in its own right. No participants had strict equivalence between pure-weak and pure-strong lists. We return to this in the General Discussion.

The implication of calling it the "null list-strength effect" is that list-strength has no effect. Taking this literally, the pure and mixed lists are essentially within-experiment replications of one another. The prediction is that the benefit of strong over weak items in pure lists (d') should covary with the benefit of strong over weak items in mixed lists, across participants. This is contradicted by the correlation we observed, $r(99) = 0.065$, $p = 0.52$, $BF_{10} = 0.10$.

The same outcome results if we correlate the difference in hit rate, $r(99) = 0.14$, $p = 0.15$, $BF_{10} = 0.22$. The near-equivalence of the strength effect in mixed and pure lists may not occur at the level of single subjects, but rather, differently, due to independent sources of variability in pure than in mixed lists. Alternatively, the data acquired with the 1 s versus 2 s comparison might be too subtle and swamped by noise.

Table 1b and 1c report results for the subset of participants who failed at least one, or passed both manipulation checks, respectively. The former performed worse than the latter. But closer inspections suggests the major difference, at least for the group as a whole, was that those who failed the manipulation check had higher false-alarm rates.

List-strength effect. Illustrated in Fig. 5a, for the 54 participants who failed at least one manipulation check, analysis of d' produced only non-significant effects, with $BF_{\text{inclusion}} < 0.32$, favouring null effects. For hit rate, only the main effect of Item Strength was significant ($p = 0.026$) but with an inconclusive $BF_{\text{inclusion}} = 0.93$. For false alarms, the main effect was non-significant ($p = 0.43$) and a favoured null ($BF_{\text{inclusion}} = 0.12$).

Table 1
Experiment 1: Hit rates, false alarm rates and d' as a function of List (Mixed, Pure) and Item type (Long: 2000 ms, Short: 1000 ms ≡ Strong, Weak), as well as the ratio-of-ratios. Note that false alarms for mixed lists are simply repeated under the mixed-strong and mixed-weak columns, as lure items are not identified with one or the other duration in mixed lists. In parentheses are the 95% confidence interval based on standard error of the mean. (a) Before the manipulation-checks. (b) Participants failing one or both of the manipulation-checks: d'(Pure-Long)> d'(Pure-Short) and d'(Mixed-Long)> d'(Mixed-Short). (c) Participants passing both manipulation-checks. (d) The original values reported by Ratcliff et al. (1990), Experiment 1.

a Before the manipulation-checks

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.66 (0.63, 0.69)	0.63 (0.60, 0.66)	0.66 (0.63, 0.69)	0.61 (0.58, 0.64)
False Alarm Rate	0.24 (0.21, 0.28)	0.24 (0.21, 0.28)	0.24 (0.21, 0.27)	0.25 (0.22, 0.28)
d'	1.25 (1.10, 1.40)	1.14 (1.00, 1.28)	1.28 (1.11, 1.44)	1.07 (0.95, 1.20)
			Mean Ratio of Ratios: 1.12 (0.87, 1.37)	Median: 0.97

b Participants who failed a manipulation-check

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.64 (0.59, 0.68)	0.62 (0.58, 0.66)	0.63 (0.59, 0.67)	0.61 (0.58, 0.65)
False Alarm Rate	0.29 (0.25, 0.34)	0.29 (0.25, 0.34)	0.30 (0.25, 0.35)	0.29 (0.25, 0.33)
d'	0.99 (0.82, 1.15)	0.95 (0.79, 1.10)	0.96 (0.79, 1.12)	0.95 (0.79, 1.11)
			Mean Ratio of Ratios: 1.22 (0.79, 1.65)	Median: 1.10

c Participants who pass both manipulation-checks

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.70 (0.65, 0.74)	0.63 (0.59, 0.68)	0.70 (0.65, 0.74)	0.61 (0.57, 0.65)
False Alarm Rate	0.18 (0.14, 0.22)	0.18 (0.14, 0.22)	0.16 (0.13, 0.19)	0.20 (0.16, 0.23)
d'	1.59 (1.37, 1.82)	1.39 (1.17, 1.61)	1.69 (1.43, 1.96)	1.24 (1.05, 1.43)
			Mean Ratio of Ratios: 0.99 (0.83, 1.15)	Median: 0.89

d Ratcliff et al. (1990) Experiment 1

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.705	0.646	0.740	0.646
False Alarm Rate	0.227	0.227	0.202	0.228
d'	1.30	1.12	1.48	1.12
			Ratio of Ratios (computed from the averaged d' values): 0.88	

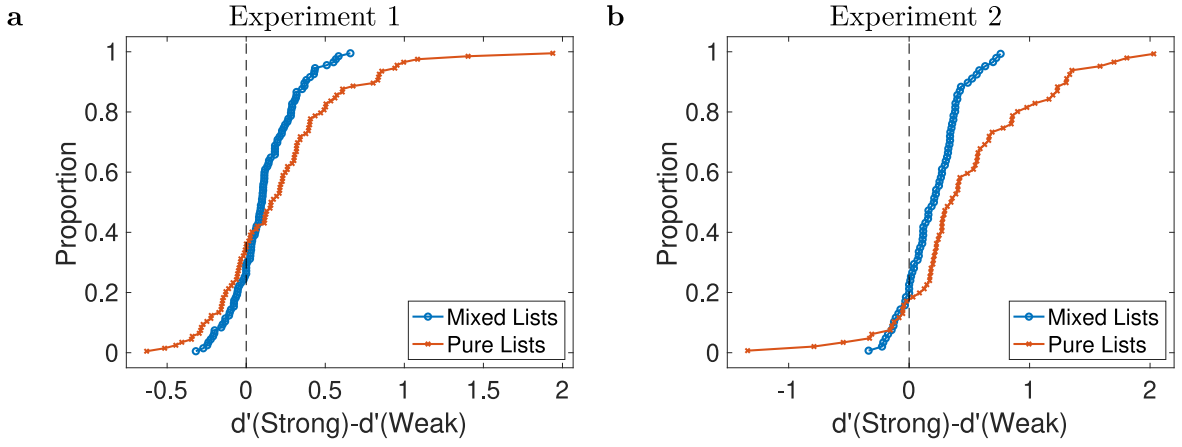

Fig. 4. Cumulative proportion distribution functions of the two measures that were used as strength manipulation checks, for mixed lists and pure lists, respectively. (a) Experiment 1. (b) Experiment 2. Each point represents one participant and the proportion is the cumulative proportion of participants.

More to the point, for the 44 participants who passed both manipulations checks and thus for whom the manipulation of duration was arguably influencing strength in the intended direction, a more robust picture emerged. For d', the main effect of Item Strength was significant, $F(1,43) = 94.61$, $MSE = 0.05$, $p < 0.001$, $\eta_p^2 = 0.69$, $BF_{\text{inclusion}} > 1000$. This is of course unsurprising, because the effect was robust in the full sample and this subsample was selected based on the individual strength effects. The Mixed/Pure main effect was not significant, $F(1,43) = 0.40$, $MSE = 0.09$, $p = 0.53$, $\eta_p^2 = 0.009$, $BF_{\text{inclusion}} = 0.26$.

The effect of interest, the interaction, was now quite robust, $F(1,43) = 17.03$, $MSE = 0.04$, $p < 0.001$, $\eta_p^2 = 0.28$, $BF_{\text{inclusion}} = 513$.

Breaking this down, for hit rates, the main effect of Item Strength was significant, $F(1,43) = 151.05$, $MSE = 0.003$, $p < 0.001$, $\eta_p^2 = 0.78$, $BF_{\text{inclusion}} > 1000$. The main effect of Mixed/Pure was not significant but also inconclusive, $F(1,43) = 2.01$, $MSE = 0.003$, $p = 0.16$, $\eta_p^2 = 0.04$, $BF_{\text{inclusion}} = 0.57$. Finally, the interaction was also not significant nor conclusive, $F(1,43) = 2.12$, $MSE = 0.002$, $p = 0.152$, $\eta_p^2 = 0.05$, $BF_{\text{inclusion}} = 0.99$.

For false alarms, the main effect of the three-level factor was significant, $F(1.8,79) = 8.63$, $MSE = 0.003$, $p < 0.001$, $\eta_p^2 = 0.17$, $BF_{\text{inclusion}} = 66$, suggesting the effects on d' are more driven by false alarms than by hits. Post-hoc pairwise comparisons with the Holm correction revealed significantly more false alarms for Pure-Weak than Pure-Strong lists

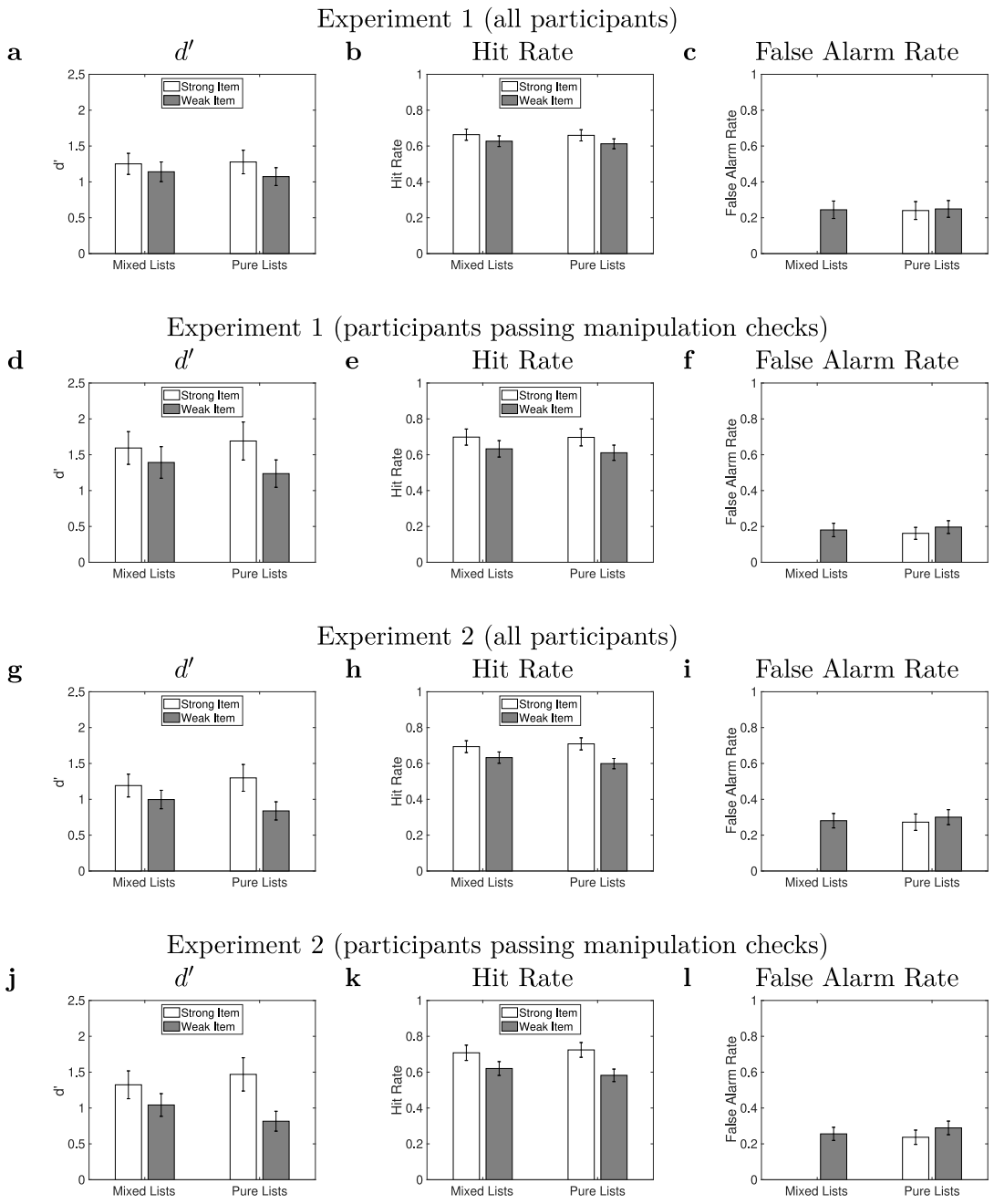

Fig. 5. Accuracy data for both experiments, plotting sensitivity (d'), hit rate and false alarm rate (note that for mixed lists, lures are not tied to a particular item-strength). Top two rows: Experiment 1. Bottom two rows: Experiment 2. First and third rows plot data for all participants; second and fourth rows plot data for participants who passed both manipulation checks: long duration > than short duration in mixed lists and in pure lists, respectively. Error bars plot 95% confidence intervals based on standard error of the mean.

($p < 0.001$, $BF_{10} > 1000$), but both other comparisons fell just short of significant (Mixed > Pure-Strong, $p = 0.056$, $BF_{10} = 2.93$; Mixed < Pure-Weak, $p = 0.059$, $BF_{10} = 0.90$.

The log of the ratio of ratios was nearly significantly different than zero (i.e., log(1)), based on a t test, $t(43) = -1.31$, $p = 0.11$ and significant based on a Wilcoxon test ($p = 0.011$) although with an inconclusive $BF_{10} = 0.57$. Comparing the raw RoR to 0.88 was non-significant based on a t test and a nearly favoured null $BF_{10} = 0.34$. $BF_{10} = 0.36$. In other words, in the subsample that passed the manipulation check, our findings are consistent with those of Ratcliff and colleagues: a robust interaction and a ratio of ratios clearly under 1.

As a final check, the benefit of strong over weak items in pure lists still did not covary with the benefit of strong over weak items in mixed lists, across participants for the subset who passed the manipulation checks. This was the case for d', $r(42) = 0.15$, $p = 0.33$, $BF_{10} = 0.19$ and for hit rate, $r(42) = -0.136$, $p = 0.38$, $BF_{10} = 0.17$.

Mirror effect. The hit rate was still greater in pure-strong than pure-weak lists, $t(45) = 7.77$, $p < 0.001$, $BF_{10} > 1000$ and the false-alarm rate exhibited the opposite change, now significant, $t(45) = -3.97$, $p < 0.001$, $BF_{10} = 100.83$, confirming a basic mirror effect. Next we asked if the asymmetry is still present, which it was. The change in the hit rate was significantly greater than the change in false-alarm rate, $t(45) = 3.18$, $p = 0.0027$, $BF_{10} = 12.16$.

Response times and deep processing

Our account of the manipulation of duration rests upon the assumption that deeper features take longer to process, resulting in more deep features encoded for long-duration (strong) items. We assume the participant attentionally subsets the probe similarly. This implies a speed–accuracy tradeoff at test. If deep features are available, they should afford more diagnostic evidence but the cost is that they take longer to process. This leads to the prediction that for pure-strong lists, there should be longer correct responses than for pure-weak lists. Hits and correct rejections are therefore expected to have longer mean response times in pure-strong than in pure-weak lists. That said, if participants have some metaknowledge, they might seek that speed–accuracy tradeoff in general in pure-strong lists and less so in pure-weak lists, so error responses may show the same response-time difference. For mixed lists, it is harder to derive predictions, but to the degree that participants seek a speed–accuracy tradeoff (due to the presence of strong items on the list), there should be some excess hits with long response-times (note that lures are not tied to an encoding strength in mixed lists). The following analyses were not pre-registered and should be read as exploratory.

Fig. 6 plots the response times (median for each participant in each trial type) for participants who passed the manipulation checks. Aligning with the predictions, response times on pure-strong lists were longer for all response types, significantly for all but correct rejections and supported by a Bayes factor for misses and false alarms (Hits: $t(43) = 2.48$, $p = 0.017$, $BF_{10} = 2.50$; Misses: $t(43) = 2.76$, $p = 0.009$, $BF_{10} = 4.51$; False alarms: $t(42) = 2.90$, $p = 0.006$, $BF_{10} = 6.25$; Correct rejections: $t(43) = 1.69$, $p = 0.098$, $BF_{10} = 0.61$). This was somewhat corroborated when all participants were analysed, apart from false alarm times which were inconclusive here (Hits: $t(100) = 3.01$, $p = 0.003$, $BF_{10} = 7.46$; Misses: $t(100) = 1.81$, $p = 0.073$, $BF_{10} = 0.53$; False alarms: $t(98) = 2.42$, $p = 0.018$, $BF_{10} = 1.76$; Correct rejections: $t(100) = 1.01$, $p = 0.32$, $BF_{10} = 0.18$).

Overlap between asymmetric mirror effects and inverted list-strength effects

With the subset of participants who passed both manipulation-checks, we examined the asymmetry of the mirror effect (ω, defined in the methods) in pure lists only, for participants depending on whether their RoR was above or below 1. Mean asymmetry was 0.061 (95% CI=[0.0230, 0.0995]) and 0.026 (95% CI=[–0.0238, 0.0767]) for RoR < 1 and RoR > 1, respectively. Although this is in the expected direction, the difference was not significant, $t(42) = 1.08$, $p = 0.29$, $BF_{10} = 0.50$. This may be due to the small (sub)sample size of the RoR > 1 group ($N = 13$) versus $N = 31$ with inverted RoRs). Individually, for the RoR < 1 group, the mirror effect index was significantly asymmetric, $t(30) = 3.14$, $p = 0.0038$, $BF_{10} = 10.16$ but not for the RoR > 1 group, $t(12) = 1.03$, $p = 0.32$, $BF_{10} = 0.44$.

Single-condition symmetry

While not a strong prediction, the simulations we presented produced approximately equal hit and correct-rejection rates as can be seen in Fig. 3b, when the shallow features could not be disregarded as well as in Fig. 3a, where the model can disregard the shallow features on pure-strong lists when there were a lot of sparse features attended. This symmetry breaks down when the model attends to only a small number of deep, sparsely subsetted features, which can be seen towards the left of Fig. 3a. In these figures, the "S" condition was simulated as though it comprised only attended shallow features and no deep features. Our 1 s and 2 s conditions are probably a mix of "S" features attended plus some number of "D" features. So the weak condition should be more like a "D" condition (solid lines) with parameters towards left of the figure and the strong condition relatively more towards the right. For both conditions, we expect the correct-rejection rate to be greater than the hit rate. The prediction is that for the strong condition, the hit rate will be closer to the correct-rejection rate than for the weak condition. As pre-registered for Experiment 1, the hit rate was less than the correct-rejection rate for all participants, pure-strong lists, $t(100) = -4.37$, $p < 0.0001$, $BF_{10} = 555.43$ and for pure-weak lists, $t(100) = -6.10$, $p < 0.0001$, $BF_{10} > 1000$ and the difference was significantly smaller for pure-strong than for pure-weak lists, $t(100) = -3.53$, $p = 0.0006$, $BF_{10} = 33.80$, consistent with the simulations. An exploratory follow-up with only participants passing both manipulation checks produced the same pattern; pure-strong, $t(43) = -5.52$, $p < 0.0001$, $BF_{10} = 9718.71$; pure-weak $t(43) = -7.02$, $p < 0.0001$, $BF_{10} > 1000$; and the difference, $t(43) = -3.24$, $p = 0.0023$, $BF_{10} = 14.07$.

Discussion of experiment 1

Comparing Table 1b and 1c to 1d, it is evident that despite some minor differences in the materials and procedures, our accuracy values come close to those of Ratcliff et al. (1990). Our data are clearly noisier, which may be due to the motivation of the participants. Our participants were recruited online and participated in a single session with no social pressure (i.e., no direct interaction with a human researcher). Ratcliff and colleagues collected data from a smaller number of participants, but in person, and notably, each participant did several sessions. These factors may have provided added incentive for the participants to engage earnestly in the task, and may also have selected for participants who were interested in performing well. The noisiness of our data is evident in the high number of participants with chance or below chance performance overall. More seriously: more than half of our participants showed nominally reverse effects of stimulus duration in either pure lists or mixed lists or both! The premise of a list-strength effect or strength-based mirror effect experiment is that one is manipulating encoding "strength". Although that may be arguable, at minimum the stronger condition should produce better memory than the weaker condition. The central question, whether there is an interaction between item-strength and list composition, is noised up if participants are included who did not respond to the intended strength manipulation or confounded by participants who showed a reverse effect. This may very well be entirely due to noise, which was the main motivation for running Experiment 2. This is indeed our view, having examined the results of Experiment 2, which follows.

Despite the noisiness of the data, Experiment 1 did produce a small-but-significant (also small through the lens of the inconclusive Bayes Factor) inversion of the list-strength effect. When restricted to participants passing both manipulation checks, the inverted list-strength effect became extremely robust. The inversion was driven more by false alarms than by hits, consistent with the simulations (Fig. 3). Finally, consistent with an assumption of our attentional subsetting theory of how stimulus duration works, there was some post-hoc evidence that response times were longer in pure-strong lists, which may be due to the deeper, more diagnostic features taking longer to process.

Experiment 2: A larger manipulation of strength

Experiment 1 replicated the inverted list-strength effect, which our attentional subsetting model of duration manipulations can explain. However, given the noisiness of the data, we conducted a second experiment with the aim of making the manipulation of strength more pronounced, to see if we could obtain an inverted list-strength effect under conditions that differed from those of Ratcliff and colleagues'. We looked to attentional subsetting theory as a guide. First, we note that the model figures already show that the inversion (RoR < 1) is a fickle result that is only produced in sweet-spots of the parameter space. If the weak condition were too much weaker, or the strong condition too much stronger, a near-null or even upright list-strength

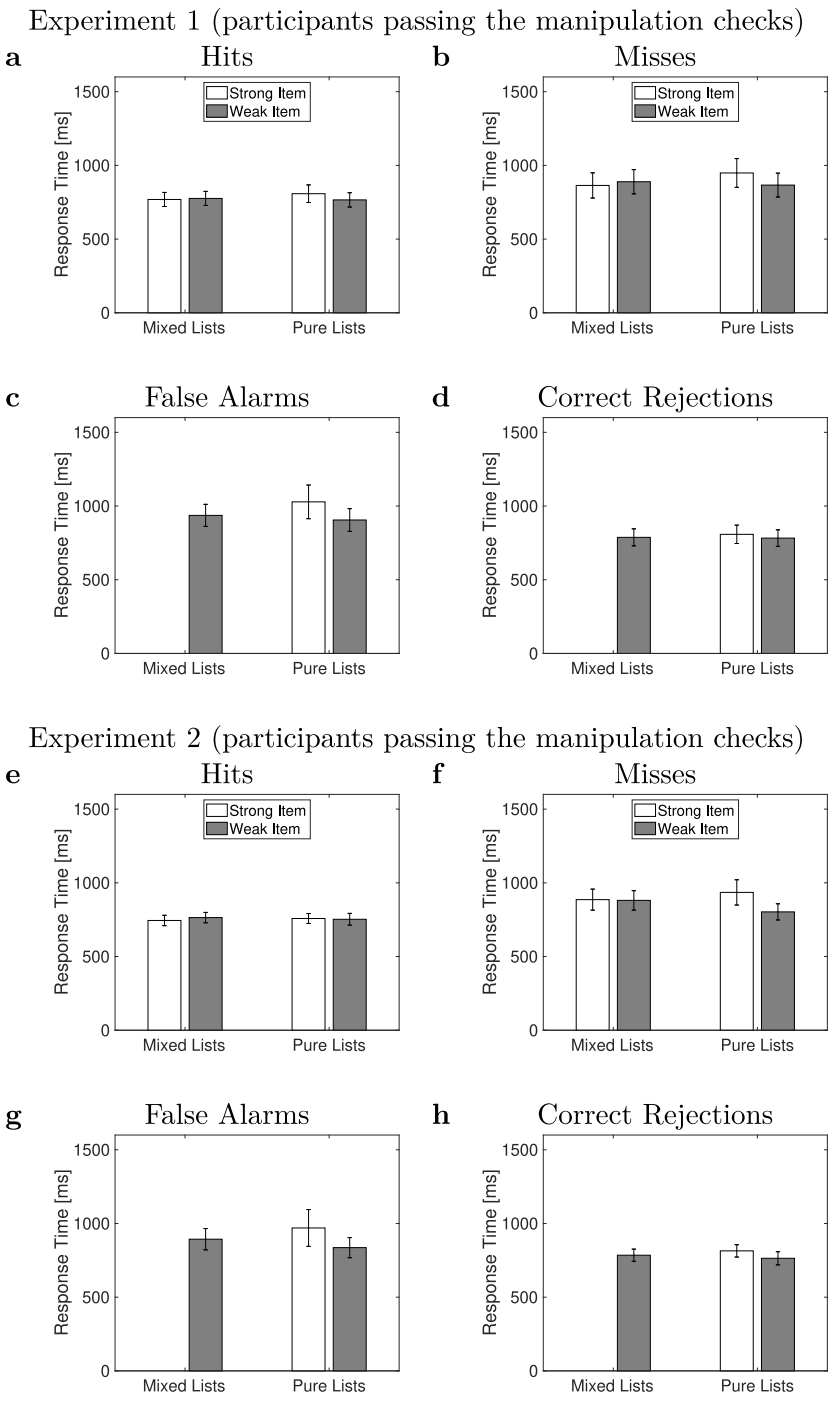

Fig. 6. Response times for both experiments for participants who passed the manipulation checks. The median was computed for each participant. Note that one participant was excluded from the false-alarm analyses for having no false alarms. Also note that for lure probes (false alarms and correct rejections), item-strength is only meaningful on pure lists. Error bars plot 95% confidence intervals based on standard error of the mean.

effect would be expected. In fact, Ratcliff et al. (1990) designed their second experiment with the same goal: to make the manipulation of strength more pronounced. But they made both the weak and strong condition longer; the short duration was 2000 ms and the long duration was 6000 ms. Our example model output (Fig. 2) makes it clear that this would likely produce an upright list-strength effect, which their experiment did (RoR = 1.10). To elaborate, if the 2000 ms duration already allows participants to process a number of deep, sparsely subsetted features, then both the short and long duration conditions are now squarely within that deep, sparse regime. There should be very little feature overlap. Although not zero, the false-alarm rates reduced from around 0.22 in their first experiment down to around 0.16 in their second experiment. If participants were able to disregard shallow features in their first experiment in 2000 ms pure lists, they would presumably be disregarding those features in all list types in the 2000 ms/6000 ms experiment. That experiment may have moved away from the regime that produces inverted list-strength effects.

Rather, if we keep the 2000 ms condition, which presumably provides a mix of shallow and deep features, but reduce the duration of the short condition to 500 ms, we thought that might reduce the number of deep/sparse features encoded during the short-duration items but leave relatively intact their shallow features. We thus predicted not only that

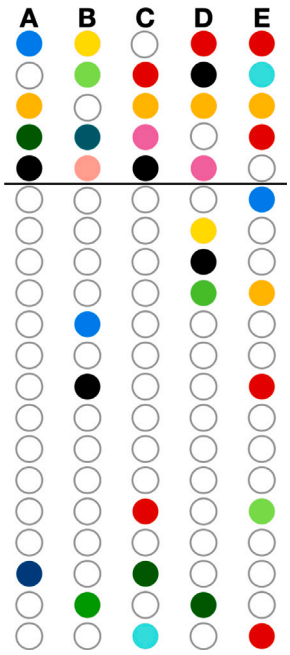

Fig. 7. Schematic depiction of how the number of attentional subsetted features might vary across items. Grey unfilled circles denote features that are not attended (and thus not encoded). The horizontal line separates the shallow feature-subspace (above) from the deep feature-subspace (below). For simplicity, items A through E have increasingly more attended deep features.

the strength effect should be larger in both pure and mixed lists, but the inversion of the list-strength effect should be more robust.

This experiment was pre-registered (pre-registration available at https://osf.io/y9dbj and data at https://osf.io/39cz8). Analyses that were not pre-registered are denoted as exploratory. All materials and procedures were identical to those of Experiment 1 apart from the short duration (500 ms in place of 1000 ms) and the "Too Fast!" bug was corrected.

Data availability. Data, materials and scripts can be found at https://osf.io/39cz8.

Methods

Methods were identical to Experiment 1 except the 1000 ms condition was replaced with 500 ms duration.

Participants. Our initial target sample size was $N = 70$, on the same basis as Experiment 1, but now with an upper limit of $N = 100$ to avoid high cost. We had planned to collect batches of 10 participants until the Bayes Factor for the critical interaction (Item Strength×List Type) moved outside the range (1:3, 3:1) but it already was after 73 included participants (a few extra participants were obtained due to some extra posted slots), so we stopped. To save money, we also implemented an early stopping rule to ensure that our manipulation of strength was successful. More exactly, if d'(pure-strong) – d'(pure-weak) > 0.4 (about twice as large as in Experiment 1) after 20 participants (no manipulation checks but after the basic exclusions based on chance performance, extra-long breaks and speed-through) data collection would be continued. This criterion was satisfied.

Results

Before manipulation checks

List-strength effect. A repeated-measures ANOVA on d' (Table 2 and Fig. 5) with design Mixed/Pure × Item Strength[Strong/Weak] revealed a significant main effect of Item Strength, $F(1,72) = 59.00$, $MSE = 0.13$, $p < 0.001$, $\eta_p^2 = 0.45$, $BF_{\text{inclusion}} > 1000$. $F(1,72) = 1.10$, $MSE = 0.044$, $p = 0.30$, $\eta_p^2 = .015$, although $BF_{\text{inclusion}} = 1134$. Speaking to the list-strength effect, the interaction was significant and quite robust, $F(1,72) = 18.79$, $MSE = 0.068$, $p < 0.001$, $\eta_p^2 = 0.21$, $BF_{\text{inclusion}} > 1000$.

Following up, for Hit Rate, the main effect of Item Strength was significant, $F(1,72) = 106.17$, $MSE = 0.005$, $p < 0.001$, $\eta_p^2 = 0.60$, $BF_{\text{inclusion}} > 1000$. The main effect of Mixed/Pure was not significant, $F(1,72) = 1.80$, $MSE = 0.003$, $p = 0.18$, $\eta_p^2 = 0.02$, although $BF_{\text{inclusion}} = 85.67$. The interaction was significant, $F(1,72) = 15.98$, $MSE = 0.003$, $p < 0.001$, $\eta_p^2 = 0.18$, $BF_{\text{inclusion}} = 391.38$. For the false alarm rate, now a one-way ANOVA on a factor with three levels (to avoid duplicating false alarms for "strong" and "weak" items in mixed lists): Mixed/Pure-Strong/Pure-Weak produced a significant main effect, $F(1.6,117) = 5.26$, $MSE = 0.003$, $p = 0.011$, $\eta_p^2 = 0.068$, $BF_{\text{inclusion}} = 4.42$. Holm-corrected post-hoc t tests found a significant advantage of Pure-Strong over Pure-Weak lists ($p = 0.006$, $BF_{10} = 2.86$), a nearly significant advantage of Mixed over Pure-Weak lists ($p = 0.055$, $BF_{10} = 2.42$) and no difference between Pure-Strong and Mixed lists ($p = 0.35$, $BF_{10} = 0.23$).

The ratio of ratios (RoR), which is based on d', was clearly below 1 for all participants except those who failed a manipulation check (Table 2). A t test of the log(RoR) against zero was significant, $t(69) = -4.69$, $p < 0.001$, $BF_{10} > 1000$.

Mirror effect. Turning again to the pure lists only, the hit rate was greater in pure-strong than pure-weak lists, $t(72) = 9.50$, $p < 0.0001$, $BF_{10} > 1000$. If there were a mirror effect, this would be parallelled by opposite effect for false alarms; this was significant but with an inconclusive Bayes Factor, $t(72) = -2.59$, $p = 0.012$, $BF_{10} = 2.86$. The change in the hit rate was significantly greater than the change in false-alarm rate, $t(72) = 6.19$, $p < 0.001$, $BF_{10} > 1000$.

With manipulation checks

Learning from Experiment 1, we had pre-registered the plan of conducting additional analyses with participants who passed the manipulation checks. Fig. 4b plots the cumulative distributions functions of the item-strength effect, d'(Pure-Strong)> d'(Pure-Weak) and d'(Mixed-Strong)> d'(Mixed-Weak), respectively. Compared to Experiment 1 (panel a), one can see that the minor change in duration of the weak condition resulted in far fewer participants who nominally failed the manipulation check (Strong<Weak). As in Experiment 1, the distribution of strength effects is broader for pure than for mixed lists, with both more big positive strength effects and more negative strength effects. Although the concern is not as pronounced in this experiment, due to greater signal-to-noise ratio with respect to the strength manipulation, as with Experiment 1, we report a second set of analyses restricted to participants who passed both manipulation-checks. First, however, unlike in Experiment 1, the two strength effects are now related to one another; the correlation of strength effects on d' was significant as well as large, $r(71) = 0.47$, $p < 0.001$, $BF_{10} = 582$, and likewise for hit rate, $r(72) = 0.31$, $p = 0.008$, $BF_{10} = 3.15$. The null correlation obtained in Experiment 1 may have been due to the small effect of the strength manipulation relative to noise.

Table 2 panels b and c report results for the subset of participants who failed at least one, or passed both manipulation checks, respectively. As in Experiment 1, for the group as a whole, those who failed the manipulation check had higher false-alarm rates.

List-strength effect. Illustrated in Fig. 5j, for the 51 participants for whom the manipulation of duration was arguably influencing strength in the intended direction, an even more robust picture emerged. For d', the main effect of Item Strength was significant, $F(1,50) = 115.80$, $MSE = 0.10$, $p < 0.001$, $\eta_p^2 = 0.70$, $BF_{\text{inclusion}} > 1000$. This is of course unsurprising, because the effect was robust in the full sample and this subsample was selected based on the individual strength effects. The

Table 2
Experiment 2: Hit rates, false alarm rates and d' as a function of List and Item type, as well as the ratio-of-ratios. Hit rates, false alarm rates and d' as a function of List (Mixed, Pure) and Item type (Long: 2000 ms, Short: 500 ms ≡ Strong, Weak), as well as the ratio-of-ratios. Note that false alarms for mixed lists are simply repeated under the mixed-long and mixed-short columns, as lure items are not identified with one or the other duration in mixed lists. In parentheses are the 95% confidence interval based on standard error of the mean. (a) Before the manipulation-checks. (b) Participants failing one or both of the manipulation-checks: d'(Pure-Long)> d'(Pure-Short) and d'(Mixed-Long)> d'(Mixed-Short). (c) Participants passing both manipulation-checks. Compare with Table 1.

a Before the manipulation-checks

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.69 (0.66, 0.73)	0.63 (0.60, 0.66)	0.71 (0.68, 0.74)	0.60 (0.57, 0.63)
False Alarm Rate	0.28 (0.25, 0.31)	0.28 (0.25, 0.31)	0.27 (0.23, 0.31)	0.30 (0.27, 0.34)
d'	1.19 (1.03, 1.35)	1.00 (0.87, 1.13)	1.30 (1.11, 1.49)	0.84 (0.71, 0.97)
			Mean Ratio of Ratios: 0.68 (0.44, 0.93)	Median: 0.71

b Participants who failed a manipulation-check

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.66 (0.61, 0.71)	0.66 (0.61, 0.71)	0.67 (0.62, 0.73)	0.64 (0.59, 0.69)
False Alarm Rate	0.34 (0.27, 0.41)	0.34 (0.27, 0.41)	0.36 (0.28, 0.43)	0.33 (0.25, 0.40)
d'	0.89 (0.65, 1.12)	0.89 (0.67, 1.11)	0.90 (0.66, 1.15)	0.89 (0.62, 1.17)
			Mean Ratio of Ratios: 0.97 (0.55, 1.39)	Median: 0.84

c Participants who pass both manipulation-checks

	Mixed-Long	Mixed-Short	Pure-Long	Pure-Short
Hit Rate	0.71 (0.67, 0.75)	0.62 (0.58, 0.66)	0.72 (0.68, 0.77)	0.58 (0.55, 0.62)
False Alarm Rate	0.26 (0.22, 0.29)	0.26 (0.22, 0.29)	0.24 (0.20, 0.28)	0.29 (0.25, 0.33)
d'	1.32 (1.13, 1.52)	1.04 (0.88, 1.20)	1.47 (1.24, 1.70)	0.82 (0.68, 0.95)
			Mean Ratio of Ratios: 0.56 (0.27, 0.85)	Median: 0.70

Mixed/Pure main effect was not significant, $F(1,50) = 2.13$, $MSE = 0.039$, $p = 0.15$, $\eta_p^2 = 0.041$, although $BF_{\text{inclusion}} > 1000$. The effect of interest, the interaction, was even more robust than with the full sample, $F(1,50) = 37.10$, $MSE = 0.05$, $p < 0.001$, $\eta_p^2 = 0.43$, $BF_{\text{inclusion}} > 1000$.

Breaking this down, for hit rates, the main effect of Item Strength was significant, $F(1,50) = 211.25$, $MSE = 0.003$, $p < 0.001$, $\eta_p^2 = 0.81$, $BF_{\text{inclusion}} > 1000$. The main effect of Mixed/Pure was not significant but also inconclusive, $F(1,50) = 2.42$, $MSE = 0.003$, $p = 0.13$, $\eta_p^2 = 0.05$, although $BF_{\text{inclusion}} = 826$. Finally, the interaction was very significant, $F(1,50) = 21.58$, $MSE = 0.002$, $p < 0.001$, $\eta_p^2 = 0.30$, $BF_{\text{inclusion}} > 1000$.

For false alarms, the main effect of the three-level factor was significant, $F(1.8,88) = 14.80$, $MSE = 0.002$, $p < 0.001$, $\eta_p^2 = 0.23$, $BF_{\text{inclusion}} > 1000$, suggesting the effects on d' are driven both by false alarms and hits. Post-hoc pairwise comparisons with the Holm correction revealed significantly more false alarms for Pure-Weak than Pure-Strong lists ($p < 0.001$, $BF_{10} = 720$), more false alarms for Mixed than Pure-Weak lists ($p = 0.002$, $BF_{10} = 36.77$) but the greater false-alarm rate for Mixed versus Pure-Strong lists fell just short of significance ($p = 0.055$, $BF_{10} = 1.61$).

The log of the ratio of ratios was significantly different than zero (i.e., log(1)), based on a t test, $t(48) = -6.96$, $p < 0.001$ and a Wilcoxon test ($p < 0.001$), $BF_{10} > 1000$.

For the subsample of participants passing both manipulation checks, the correlation of the strength effect (d') between mixed and pure lists remained significant, $r(49) = 0.51$, $p < 0.0001$, $BF_{10} = 176.90$.

Mirror effect. The hit rate was still greater in pure-strong than pure-weak lists, $t(50) = 12.24$, $p < 0.001$, $BF_{10} > 1000$ and the false-alarm rate exhibited the opposite change was significant, $t(50) = -4.61$, $p < 0.001$, $BF_{10} = 720.40$, indicating a robust mirror effect. The change in the hit rate was again significantly greater than the change in false-alarm rate, $t(50) = 5.62$, $p < 0.0001$, $BF_{10} > 1000$.

Response times and deep processing

Not pre-registered for Experiment 2, Fig. 6 plots the response times (median was computed for each participant in each trial type) for participants who passed the manipulation checks. Again aligning with the predictions, response times on pure-strong lists were longer for all response types but hits, significantly and with conclusive Bayes Factor (Hits: $t(50) = 0.39$, $p = 0.70$, $BF_{10} = 0.16$; Misses: $t(50) = 4.44$, $p < 0.001$, $BF_{10} = 439$; False alarms: $t(50) = 3.27$, $p = 0.0019$, $BF_{10} = 15.73$; Correct rejections: $t(50) = 3.40$, $p = 0.0013$, $BF_{10} = 22.21$). This was corroborated when all participants were analysed, although with less conclusive Bayes Factors (Hits: $t(72) = -0.75$, $p = 0.46$, $BF_{10} = 0.17$; Misses: $t(72) = 2.72$, $p = 0.0082$, $BF_{10} = 3.84$; False alarms: $t(72) = 2.17$, $p = 0.033$, $BF_{10} = 1.17$; Correct rejections: $t(72) = 2.52$, $p = 0.014$, $BF_{10} = 2.43$).

Overlap between asymmetric mirror effects and inverted list-strength effects

Of the 51 participants passing both manipulation checks, only 9 had RoR > 1. The comparison we did for the full sample and in Experiment 1 is thus underpowered. Given the predominance of RoR < 1 in this experiment, those nine participants may have produced RoR > 1 largely by chance. That said, the co-occurrence of a very robustly inverted list-strength effect and very pronounced mirror-effect asymmetry in this subsample aligns with the idea that the two features occur within a common task-space.

Single-condition symmetry

Although not pre-registered for Experiment 2, consistent with the simulations and Experiment 1, the hit rate was less than the correct-rejection rate for all participants, but a supported null for pure-strong lists, $t(72) = -0.75$, $p = 0.46$, $BF_{10} = 0.17$; pure-weak lists, $t(72) = -4.05$, $p = 0.0001$, $BF_{10} = 161.01$ and the difference was significantly smaller for pure-strong than for pure-weak lists, $t(72) = -6.19$, $p < 0.001$, $BF_{10} > 1000$. The subset of participants passing both manipulation checks produced the same pattern except that the Bayes Factor for pure-strong lists fell short of conclusive; pure-strong, $t(50) = -1.55$, $p = 0.13$, $BF_{10} = 0.47$; pure-weak $t(50) = -4.57$, $p < 0.0001$, $BF_{10} = 641$; and the difference, $t(50) = -5.62$, $p < 0.0001$, $BF_{10} > 1000$.

Item analyses (both experiments)

We report analyses of variability across words. Some of the following analyses were pre-registered in each experiment, but in retrospect,

we felt the full sequence of analyses is necessary to obtain a complete picture of the item-level effects. We advise the reader to view all these findings as fully exploratory.

A major assumption of our theoretical account of stimulus-duration effects was that the participant derives the criterion based on real-time processing of the probe item. Given the existence of effects of stimulus properties, including stimulus-based mirror effects (Glanzer & Adams, 1985, 1990; Neath et al., 2021), we know that items differ in their difficulty in recognition tasks. There must be item-variability in difficulty at least at the level of those stimulus properties such as word frequency, concreteness and contextual diversity. If participants tune their criterion based on characteristics of each individual stimulus, not only the pattern of hit rate across items should replicate, but also the pattern of false-alarm rates. We exploited the fact that our two experiments were quite similar in design, differing only in the stimulus-duration of the weak condition, to test these predictions. Including only participants who passed both manipulation checks in each experiment, all words had at least 7 trials as targets and at least 7 trials as lures, so we included all words. The Pearson correlation of hit rate across words, between the two experiments was significant and robust, $r(998) = 0.268$, $p < 0.0001$, $BF_{10} > 1000$. For false alarms, the correlation was even greater, $r(998) = 0.390$, $p < 0.0001$, $BF_{10} > 1000$, where $R^2 = 0.15$, indicating that 15% of the variance in false-alarm rate was attributable to item-difficulty effects; this would seem non-negligible, and suggests that a pre-condition to our account of criterion selection is met.

Next we can ask whether there is something like a stimulus-based mirror effect at the item level — rather than the stimulus-class level, as has been previously reported starting with Glanzer and Adams (1985) because Neath et al. (2021) found separate influences on hits than false alarms when stimulus-set properties were better controlled. Within each experiment, we correlated overall hit rate for a word with its overall false-alarm rate, still collapsing across list types to avoid too much missing data. The correlation was significant ($p < 0.05$), and even the expected sign (negative) in Experiment 1, but not significant in Experiment 2. Both correlations were quite small in magnitude and according to the Bayes Factors, inconclusive in Experiment 1 and a favoured null in Experiment 2;. HR-Experiment 1: $r(998) = -0.0909$, $p = 0.004$, $BF_{10} = 1.57$; Experiment 2: $r(998) = 0.0410$, $p = 0.195$, $BF_{10} = 0.058$. This seems to take the conclusions of Neath et al. (2021) further, suggesting that different factors influence hit rate versus false-alarm rate. This is unlikely to be due to lacking signal-to-noise ratio since the patterns of hit rate and false-alarm rate did replicate quite robustly between experiments (results in the previous paragraph). This is informative with respect to stimulus-based mirror effects, but it neither supports nor challenges attentional subsetting theory. The theory suggests different factors that could influence hits differently than false alarms, such as the number of features attended and the similarity of those features to other studied items, but the way in which those trade off is parameter-dependent.

These findings extend prior findings of item effects in old/new recognition. Analysing lists of 60 monosyllabic words (Cortese et al., 2010) or disyllabic words (Cortese et al., 2017) presented for 2000 ms each or self-paced, around 1/3 of the variance in hits — false-alarms was explained by stimulus factors such as word length, frequency, imageability and orthographic and phonological neighbourhood characteristics and similarity to other items in their stimulus pool (see Lau et al., 2018 for similar results). This exceeds the between-experiments variance explained in our analyses. They also tested for an item-level mirror effect and falsified it. In fact, for mono-syllabic words, Cortese et al. (2010) found a small but significant *positive* correlation between hits and false alarms across items ($r = .17$ and .15 for 2000 ms/word and self-paced, respectively). With disyllabic words, Cortese et al. (2017) found a non-significant item-level mirror effect, despite their highly powered datasets (combined: $r = -0.017$).

Cox et al. (2018) found item effects that explained similarities and differences between memory tasks, including one item-factor related to recognition bias and one related to propensity to be produced as a response in a recall task. These may be closely related to factors we find to separately influence hit and false-alarm rates. Between-sample consistencies have also been found in the visual domain, with high correlations across complex visual stimuli measured by hit rate for various categories of images (Isola et al., 2011) and both hit and false-alarm rate of complex images (Bainbridge, 2020; Bainbridge & Rissman, 2017) and faces (Bainbridge et al., 2013).

Finally, we asked if *strength* influences hit rate and false-alarm rate in tandem or independently. We correlated, across items, the difference in hit rate for pure-strong minus pure-weak lists with the difference in false-alarm rate for pure-strong minus pure-weak lists. For both experiments, this correlation was non-significant and a supported null effect, Experiment 1: $r(927) = -0.032$, $p = 0.330$, $BF_{10} = 0.042$; Experiment 2: $r(961) = 0.019$, $p = 0.562$, $BF_{10} = 0.030$. Although not a strong test of the theory, this is consistent with our assumption that strength (here, stimulus-duration) mainly influences attention to deeper, sparsely subsetted features and that the shallower features common to both conditions can be largely disregarded when participants are tested on pure-strong lists.

General discussion

We developed and extended an attentional subsetting theory of stimulus duration and tested it with two new experiments. The central assumptions of attentional subsetting theory (Caplan, 2023; Caplan et al., 2022) are:

Assumption (1) Most features of an item are *not* attended and thus not encoded; rather, only a small subset of features are attended and thus encoded.

Assumption (2) These subsets are stimulus-specific; thus, they tend to be different vector dimensions for different items.

Assumption (3) Given the same task-set (cf. Criss & Shiffrin, 2004), the features subsetted during study of a stimulus will largely (although not strictly) be the same as those subsetted when the same stimulus is presented as a recognition probe.

To separately model hit and false-alarm rates, we have introduced a way in which the response criterion could be derived from the probe item, itself, adding:

Assumption (4) The participant derives a criterion for each probe item as they process it.

Assumption (5) The criterion is a simple calculation related to the number of features processed in real-time (e.g., one half).

Critically, our specific assumptions about stimulus-duration are:

Assumption (6) The earliest features attended are "shallow" or "superficial" such as orthographic or phonological features (or potentially certain kinds of semantic or elaborative features). These early-extracted features are selected from a relatively small feature subspace, so they cannot be sparsely subsetted. This is the source of confusion due to feature similarity across items.

Assumption (7) Later features are "deeper" or more "semantic" and as such, are subsetted from a much larger feature subspace so that attended subsets are sparse vectors within that deeper subspace. This avoids most of the confusion due to feature overlap.

Assumption (8) In certain circumstances, participants may be able to disregard feature subspaces, such as orthographic or phonological features when deeper features are sufficient to support performance. We propose that participants may have a way to detect the presence of confusion due to superficial features, leading to this kind of meta-cognitive strategy.

We ran a replication study of the first experiment that was noted to violate the expected "upright" list-strength effect. The replication was not perfect, but came close to the original experiment, with some support that we obtained the original inverted list-strength effect. A second experiment with a larger manipulation of strength produced a more robust inverted list-strength effect. This was done by reducing the weak condition from 1000 ms to 500 ms, focusing the weak condition even more on superficial features that the theory presumes cannot be sparsely subsetted, while the strong condition, left at 2000 ms, gives participants ample time to process deeper features that are sparsely sampled from a high-dimensional space.

These replications mean that the inverted list-strength effect cannot be swept aside. Attentional subsetting theory anticipates inverted list-strength effects under certain conditions. Our empirical findings thus offer validation to the theory. That said, we cannot rule out other accounts of inverted list-strength effects. Yonelinas et al. (1992), for example, proposed that an upright list-strength effect might be partly due to "rehearsal-borrowing," where in a mixed list, strong items draw more rehearsal at the expense of the weak items. It is possible that we have the opposite kind of rehearsal-borrowing in the mixed lists, whereby weak items attract compensatory rehearsal, stealing rehearsal resources from the strong items. As described earlier, the phenomenon of differentiation produces an inverted list-strength effect. When incorporated into SAM (Shiffrin et al., 1990) or REM (Shiffrin & Steyvers, 1997), this can approximately offset the coexisting upright list-strength effect. In some conditions, it might more than offset the latter and result in a net inverted list-strength effect (as Ensor et al., 2021 produced; note that this simulation still produced an upright list-strength effect for hits, which mismatches the data; that said, REM has numerous moving parts which might counterbalance in different ways to match the fine structure of the data), which our data could test. Differentiation also produces a strength-based mirror effect, so it remains to be seen whether REM includes sufficient flexibility to produce a net inverted list-strength effect while simultaneously producing a much larger difference in hit rate between item-strengths than the difference in false alarm rate, as in our two experiments and Experiment 1 of Ratcliff et al. (1990).

We learned an important pragmatic lesson while analysing the data. While 1 s versus 2 s at first seems like a large experimental manipulation (twice as much time to study the "strong" items as the "weak" items), it produced very small effects on recognition in the hands of Ratcliff et al. (1990) as well as in our first experiment. The main characteristics of the original study replicated, so there is no reason to suspect the validity of the manipulation. However, the small magnitude of the manipulation of stimulus-duration on behaviour produces a dataset with more noise and less sensitivity than one would like. A specific problem arose because many of the participants did not even seem to pass a basic manipulation-check, that the longer duration should produce better memory than the shorter duration. At worst, it could be possible that some participants are immune to manipulations of stimulus-duration or have a paradoxical effect, where shorter duration leads to better memory than longer duration, which would clash with the basic premise that the experiment is a manipulation of encoded strength. Experiment 2 produced robust results and far fewer violations of the manipulation-checks, so our view is that it is unlikely that duration acts differently for some participants. However, the weak effect of the manipulation in Experiment 1 is a tiny signal amid a large amount of noise. In this kind of regime, null effects, even Bayes Factors favouring nulls, may be quite common.

At a conceptual level, however, we would like to draw attention to the overloaded nature of the term "strength" in so-called list-strength effect experiments and strength-based mirror effect experiments. Using a single term, "strength," for manipulations as different as duration, repetition and even qualitatively different processing tasks probably glosses over a number of very concrete and different mechanisms. For example, early use of the word "strength" referred to a scalar multiple of an encoded vector; longer vectors will be remembered better. This produces the long-expected "upright" list-strength effect, an advantage for strong items within mixed lists compared to pure lists, and the corresponding disadvantage for weak items within mixed lists (Caplan, 2023; Ratcliff et al., 1990). Distinct from that notion of strength, feature-level models have assumed that some forms of "strengthening" result in more features encoded and/or more features correctly rather than erroneously encoded (e.g., Caplan, 2023; Nairne, 1990; Shiffrin & Steyvers, 1997) as we have done in our formulation of stimulus-duration. Strengthening via repetition has been proposed to result in editing of existing local traces or the formation of a new trace (e.g., Criss, 2006; Ensor et al., 2021). And finally, some manipulations viewed as "strengthening" may result in the encoding of additional features potentially of a different type, such as the account of the production effect by Jamieson et al. (2016), or for different levels of processing, potentially completely non-overlapping feature subspaces (Caplan, 2023). Digging into these various specific mechanisms can add significantly more specificity and direct connections to model mechanisms than continuing to use "strength" as a catch-all term.

An attentional subsetting formulation of the effects of stimulus-duration. In REM, stimulus duration is modelled by increasing the probability that each feature is encoded (Shiffrin & Steyvers, 1997). We have incorporated that assumption into attentional subsetting theory (Caplan, 2023), but extended it as summarized in the previous section. Our formulation of stimulus-duration can produce a variety of list-strength effects, including near-null, upright/positive and inverted/negative list-strength effects. It can also now produce a robust strength-based mirror effect but also leads to large changes in hit rates with very little change in false-alarm rates when superficial features can be disregarded (or in other paradigms, if subsetting is entirely sparse or the two feature subspaces are non-overlapping).

Published data, including the data reported here, offer constraints on the putative timecourse of processing of shallow versus deeper features. The presence of an upright list-strength effect and pronounced strength-based mirror effect (Yonelinas et al., 1992) when stimulus-duration was very short (50–200 ms) is consistent with the idea that at these timescales, superficial features dominate, and very few deep, sparsely subsetted features are encoded (but the effect may be fragile; an inverted and null list-strength effect was found by Ratcliff et al. (1994) with 50 ms versus 200 ms and 100 ms versus 400 ms, respectively, with strength blocked within mixed lists). This book-ends the continuum alongside the second of experiment of Ratcliff et al. (1990), who varied stimulus-duration between 2000 ms and 6000 ms. Their finding of an upright list-strength effect, although small in magnitude, along with strength predominantly influencing hit rate with very little effect on false-alarm rate, is consistent with the idea that well before 2000 ms, participants are already able to process a number of deep, sparsely subsetted features that they can rely primarily on those features to make recognition judgements. The first experiment of Ratcliff et al. (1990) and our Experiment 1 compared 1 s to 2 s stimulus durations. This seems to be close to the transition point. 1000 ms gives participants some, but not much spare time to encode deep features. 2000 ms offers more time to encode deep features, but shallow features are still somewhat useful in making recognition judgements. Consequently, the list-strength effect is close to null but slightly inverted. Finally, our Experiment 2 produced clearer results presumably because we designed it to straddle that transition point.

Reducing the short duration to 500 ms may have further reduced the availability of deep features during encoding.

This may explain why clear, significant and robust inversions of the list-strength effect in recognition have been so elusive. Fig. 2 shows how, according to our model, the inverted list-strength effect is a fragile finding, that is highly parameter-dependent. Comparing 500 ms to 2000 ms durations of visually presented words appears to be a sweet spot for further understanding inverted list-strength effects.

The assumption that shallower features are processed earlier than deeper features has some support, and with similar timescales, during the test phase of recognition experiments using the response-deadline procedure. In a response-deadline experiment, participants are trained to make a decision by a particular time following stimulus-onset. For example, Gardiner et al. (1999) trained participants on response deadlines of 500 ms and 1500 ms. These deadlines are close to the range of durations we investigated (500, 1000 and 2000 ms). When studied with a shallow, "phonemic" processing task (rate how easy it is to find a rhyme to the word), the hit rate increased from 0.48 to 0.59 from the short to the long response deadline. For lists studied with a deeper, "semantic" processing task (rate how easy it is to find a semantic associate of the word), performance was overall better, but the increase with longer response deadline was also bigger, with hit rate increasing from 0.56 to 0.77. Mulligan and Hirshman (1995), sampling more response deadlines, found evidence for levels of processing influencing primarily the asymptotic accuracy (d') level reached, with very little influence on evidence-accumulation rate.

Brockdorff and Lamberts (2000) have an interesting take on response-deadline data. They assumed features are sampled with some probability, but different features had different sampling probabilities per unit time, to explain different timecourses for different forms of information. Their first target finding was an experiment by Hintzman and Curran (1994), testing recognition of target words, dissimilar lures and similar lures that varied only in whether they were singular or plural (frog versus frogs). The results showed a non-monotonicity in the false-alarm rate to those similar lures: at early response deadlines, false-alarms to similar lures increased, then decreased at later deadlines. Hintzman and Curran (1994) reasoned that this non-monotonicity was evidence of two distinct processes used to drive recognition, familiarity, which accumulates early, and recollection, which is possible only later, that supports a recall-to-reject strategy. Brockdorff and Lamberts (2000) showed that two processes are not needed to explain the data, and instead proposed different rates of feature sampling. They adapted the Generalized Context Model (Nosofsky, 1986) that had been developed for categorization behaviour. They modelled the word stimuli with binary vectors of six features, where the sixth feature stood in for plurality. The model produced the non-monotonic function of false alarms to similar lures by fitting the sampling probability of the sixth feature to be much lower (about 1/10 smaller) than the remaining features. Their account has a lot in common with our theory of stimulus duration. It includes more detailed temporal dynamics which we have omitted, but which would presumably be compatible with our account. What our theory adds, however, are two things. First, Brockdorff and Lamberts (2000) had very low-dimensional representations of stimuli. Such a model would quickly break down if the list length were substantially greater than 6 (the list length they modelled was 12). They assumed that all features were encoded, and that probabilistic sampling only occurred at test. To make this more realistic and add probabilistic sampling, increasing the vector dimensionality would have the same problem as Glanzer et al. (1993), for example, that the chance of randomly sampling the same features at test that were encoded becomes quite small. In our framework, a subset of features is stored, but they will tend to be similar upon repeated presentations of an item, including between a study and a test trial. This reiteration is what supports high performance levels even while item dimensionality increases. Second, we add the idea that later-sampled features will tend to be sparse, derived from a high-dimensional feature-space.

Our theory suggests an addendum to the Brockdorff and Lamberts (2000) account of the Hintzman and Curran (1994) data, which in a sense, harmonizes the two accounts. We assume that later-attended features tend to be more sparse. Two stimuli differing only in one letter, frog and frogs, are highly similar within the orthographic feature-space, since they have almost identical spelling. What is retrieved later is not the letter 's' or its omission, but semantic or imagery information as the participant more deeply contemplates a frog or many frogs. A visual image involving a single frog may be quite dissimilar to a visual image of many frogs. This is a concrete way we can understand how semantic of deeper feature spaces can offer distinctiveness through sparseness that is generally not available in the more superficial, but earlier processed feature space.

Feature depth or dimensionality of the subspace. In expressing the effect of feature-space dimensionality, we have used as shorthand the idea that perceptual features are densely subsetted from a low-dimensional feature space whereas semantic or imagery-related features are sparsely subsetted from a high-dimensional space. However, there may be features we would like to think of as "deep" that are nonetheless within a low-dimensional subspace, such as, perhaps, attributes like animacy, pleasantness, function (furniture, tool, etc.). The early availability of semantic features in response-deadline data offers some support for the idea that semantic features are available as rapidly as perceptual features (Mulligan & Hirshman, 1995) and an early interaction of different levels of features was expressed by Gibson (1971). Conversely, it is possible that in some conditions, some perceptual features are sparsely subsetted from within a very large feature space. Along these lines, Johnson (1975) provided evidence that the whole word can be identified before all the composite features, such as letters, have been processed. This raises the interesting possibility that semantic or elaborative information about an item may even feed back to prioritize attention to particular low-level features of the stimulus. The deeper logic we present is that as time unfolds, to a large degree, the earlier features will be extracted from lower-dimensional feature spaces than those that are extracted later. The denseness versus sparseness of those features determines the form of the list-strength effect.

How duration may differ from repetition. Although both have been described as manipulations of "strength," manipulations of spaced repetition may function differently than manipulations of stimulus-duration. For example, Caplan (2023) suggested that repetition forces more attention to the superficial features, because the additional study time is also accompanied by a new stimulus-onset. Before noticing the repetition, the participant must surely need to process its superficial features anew, which may result in stronger encoding of superficial features with, say, two presentations at 1 s/word than one presentation at 2 s/word. Moreover, holding constant total duration, the need to re-process the superficial features upon repetition would displace a few hundred ms of study time that would otherwise be used to process deeper features. Reduced encoding of deeper features combined with additional obligatory encoding of superficial features may reduce the potential benefit of disregarding superficial features in pure-strong lists. This may explain why inversions of the list-strength effects are rare or non-significant when strength is manipulated via repeated presentation, and remains to be tested.

The "one-shot" contextual encoding hypothesis. The "One-Shot" hypothesis, as integrated within the REM framework by Malmberg and Shiffrin (2005) for free recall, proposes that a fixed amount of context is stored during an initial brief exposure to an item, typically for at least 1 or 2 s, and this storage is deemed sufficient for supporting necessary context information for later retrieval. This was particularly effective in explaining list-strength effect in free recall. According to their hypothesis, additional study duration or deeper levels of processing do

not substantially increase the amount of context information stored beyond this initial "shot"; instead, they enhance content knowledge, such as meanings and associations. According to Malmberg and Shiffrin (2005), context information should continues to accumulate within this timeframe, which suggests that our short, 500-ms duration in experiments might encode less context than a 2000-ms duration. However, the one-shot hypothesis posits that initial brief context capture should suffice for later retrieval, so it may not be able to explain the inverted list-strength effect observed in our experiments. When differentiation based on item-features becomes dominant, as proposed, it could lead to inverted list-strength effects, especially under conditions where context features are either down-weighted or not used, which might be the case in our experiments. This mismatch between their hypothesis and observed data suggests a need to further explore whether variations in context encoding between short and long durations might underpin the mirror effects and inverted list-strength effects observed, calling for a nuanced application or reconsideration of the one-shot hypothesis in these contexts.

The list-strength effect. Despite the label, "null list-strength effect," there is in fact a range of published list-strength effects, including inversions (Ratcliff et al., 1990, 1994; Sahakyan, 2019). Our modelling and findings suggest these cannot be dismissed as due to noise, but must be taken seriously and addressed by models.

Local-trace differentiation models, including REM, are able to produce upright list-strength effects. Specifically, Criss (2006) noted that for spaced repetitions, the null list-strength effect occurs because of the assumption (Shiffrin & Steyvers, 1997) that upon repetition, the participant notices the repetition and this allows the participant/model to edit the earlier trace rather than forming a new one. Additional features can be stored (and in some versions of REM, features that were copied erroneously can be corrected). As Criss (2006), Ensor et al. (2021) demonstrated, if the participant does not notice the repetition, they presumably form a new local trace for the same item instead of editing. This reduces the benefit of the differentiation process, so a list-strength effect emerges. Inspiration for this came from Sahakyan and Malmberg (2018) and Sahakyan (2019) who found pronounced list-strength effects in recognition under divided attention. We do not rule out the trace-editing account, so it may very well be valid. But it may be straight-forward to explain effects of divided attention without local traces or notions of participants "noticing" repetitions. A plausible account of divided attention is that it reduces the participant's ability to attend to deeper features, or at least sparsely subsetted features. This would place the divided-attention data within the regime of very short presentation duration (Yonelinas et al., 1992) or our account of the production effect (MacLeod et al., 2010), which do produce pronounced upright list-strength effects (Caplan, 2023; Caplan & Guitard, 2024).

That said, although the trace-editing account has an air of plausibility for spaced repetitions, it does not immediately seem amenable to manipulations of stimulus-duration. It does not seem likely that participants would fail to notice an extended duration and thereby encode two traces rather than one. This leaves it unclear how local-trace differentiation models might explain upright list-strength effects due to strength manipulations such as reported by Ratcliff et al. (1990), Experiment 2 with 2 s versus 6 s per item, or by Yonelinas et al. (1992), with 50 ms versus 200 ms per item. Our attentional subsetting account expects an upright list-strength effect in the former conditions because the two strength levels are both within the sparse regime, and in the latter because the two strength levels are both within the non-sparse regime. Likewise, for the production effect, it is not obvious why participants might occasionally store two traces for a single presentation read aloud versus read silently, whereas our attentional subsetting perspective would anticipate an upright list-strength effect, as is the case (MacLeod et al., 2010), because of the assumption that production acts primarily on orthographic or phonological features, which are not sparsely subsetted.

Moreover, trace-editing does not explain inverted list-strength effects, first reported as significant by Ratcliff et al. (1990) and now here. But the differentiation mechanism in REM produces an inverted list-strength effect, which was proposed to approximately cancel out an upright list-strength effect produced by ambiguity in the context cue. As in most models, a strong item produces more evidence than a weak item in favour of a hit. But differentiation means that the strong traces also produce more evidence than weak traces *against* the likelihood that a lure item was on the list. The false-alarm rate, therefore, is positively related to the number of weak items in the list and negatively related to the number of strong items in the list. This produces an intermediate rate of false alarms in mixed lists, between pure-strong and pure-weak lists. This in fact does resemble the pattern of false alarms we observed in our two experiments. The idea that two opposite list-strength effects coexist leaves open the idea that in cases such as Ratcliff and colleagues (and our) manipulations of duration, there is a net dominance of the pattern produced by differentiation. Additional experiments could test the two accounts directly, bearing in mind that they may both coexist.

Finally, one reason list-strength effects have been so challenging to models is that the near-null effect in recognition is found despite robust upright list-strength effects in free recall and cued recall; this contrasting pattern was already shown by Ratcliff et al. (1990). Briefly, although we have not yet developed attentional subsetting theory for recall tasks, Caplan (2023) pointed out that clear that positive list-strength effects would generally be predicted. To summarize: the reason sparse attentional subsetting produces near-null list-strength effects is because in item recognition, the probe is the item, itself. Given the item, the model produces the item-specific attentional mask (in many conditions, the same mask as was produced when first studying the item, in the case of a target). Sparseness makes overlapping features quite rare, sidestepping most opportunities for other list items to introduce noise into the judgement. In recall tasks, the probe is an item (cued recall) or an instruction to recall (free recall; usually it is implied that the participant self-cues with some sort of representation of context) but the participant's goal is to find an item to produce as a response. Without a specific item in hand, there is no item-specific mask. Thus, cueing with an attentional mask determined by general task-context or plausibly, something like the union of all attended features during the study phase, there will inevitably be a fan effect. What makes the task a recall task is precisely what prevents the cue from being item-specific. Supporting this reasoning, Caplan (2023) indeed found sizeable positive list-strength effects even in the "sparse" regime when the full vector was used as the recognition probe. Although Caplan rejected this (in favour of the idea that probes are also masked) as a plausible model of recognition, it demonstrates the mathematical effect that would presumably be operating in recall. This account also implies a near-null list-strength effect for associative recognition; although it is associative like cued recall, because both items are presented at test, both items could be fully masked. Indeed, associative recognition produces negligible list-strength effects (Osth & Dennis, 2014, 2015).

Criterion tuned based on the probe, itself. Prior accounts of the strength-based mirror effect have differed in whether they assume differentiation-based local traces or criterion shifts. Criterion-shift accounts assume participants use their meta-knowledge of the expected distribution of target and lure strengths to set an approximately optimal criterion between them. Differentiation accounts assume that a stronger trace provides both more evidence for the corresponding target probe having been studied, and more evidence for lure probes having not been studied, thus moving hits and false alarms in opposite directions without invoking any metacognitive strategy. Numerous articles have gone back and forth over whether participants in fact are able to adjust their criterion for different types or strengths of items, and whether or not they can make use of knowledge about expected strength distributions. Attentional subsetting could be entirely compatible with both accounts, without the need for a new mechanism. But in extending the

formulation of the theory to produce separate estimates of hit rates and false-alarm rates, we noticed an opportunity, provided by attentional subsetting, to consider a third mechanism. Namely, we suggest that the participant customizes the criterion for each item, based upon real-time processing of the item, itself. This retains some desirable characteristics of each of the two positions. The criterion changes from one item to the next, but the *principle* by which the criterion changes may be fixed across an entire list (or block of test trials; see below).

That said, the idea that the criterion changes from one item to the next may seem at odds with numerous findings that have been viewed as evidence that participants do not change their criterion, at least within a single list (e.g., Starns et al., 2010; Verde & Rotello, 2007) although they may have this ability because Verde and Rotello (2007) found that accuracy feedback did induce a criterion shift, as evidenced by a change in false-alarm rates over the course of multiple test trials. Stretch and Wixted (1998) (continued by Morrell et al., 2002) tried over several experiments to neutralize or reverse word-frequency effects on recognition by strengthening (with spaced repetitions) high-frequency items, and then cueing participants as to the strength/frequency during the recognition probes. Participants were apparently unable or unwilling to adjust their criterion, given those cues. However, in separate lists (e.g., different in strength), there was support for a change in criterion. Corroborating findings were reported by Singer and Wixted (2006). In lists composed of multiple categories, where some categories were studied in one list and the other categories in a list presented after a delay, making it more recent. Recognition probes were intermixed from the two lists. The authors thought that the category structure would enable participants to adjust their criterion based on meta-knowledge of the recency of the category, but findings were inconsistent with participants making use of that knowledge when the delay was 20 or 40 minutes. When the delay between the two lists was two days, finally there was evidence of participants using a more lenient criterion for the less recent categories, producing more false alarms. This result may be evidence of participants scaling their criterion to account for forgetting, as we speculate about in the future-directions section. However, the lack of adjustment of the criterion for shorter but still substantial delays (20 and 40 minutes) suggests that such scaling may often not vary across a set of recognition probes of a single list. Hicks and Starns (2014) also found little effect of strength-cueing or even performance feedback, but blocking test trials by strength did seem to induce criterion shifts (see also Verde & Rotello, 2007). This suggests some adaptation of the participant to their experience with test probes, although apparently not in response to explicit information about strength or performance. In contrast, Koop et al. (2019) found a reliable strength-based mirror effect only after a few trials, and argued that this is not long enough to expect participants to adjust their criterion.

In our account, the criterion depends on the number of features of the current probe that are attended. This is presumably readily accessible information. If the criterion depends primarily on the probe, itself, and how the participant typically processes it, that would explain why different stimulus classes can have systematically different false-alarm rates, but also why within a list, across a set of probes, the criterion appears relatively invariant (with the exception of the blocked strengths of Hicks and Starns 2014). However, from one list to another, assumptions about the task or the contents of memory could change. We have already proposed that in a pure list of strong (long-duration) items, participants figure out that they have the luxury of being able to disregard superficial features such as orthography or phonology, which would otherwise introduce a lot of feature overlap producing similarity-based confusion. Selective attention in a pure-strong list thus benefits from this meta-knowledge, and results in fewer features being evaluated — although those features are more diagnostic because they dwell within a high-dimensional feature space and are thus sparse. The criterion follows from that, and the false-alarm rate reduces a lot because the shallow features that produce confusion due to feature overlap are disregarded. So in our view, the way in which the criterion is calculated may not change, which might explain the apparent invariance of "the" criterion (although we presume a different criterion for each probe item). Rather, selective attention influences the set of features attended on a probe item, which can then produce a downstream effect on the criterion. We think this is broadly consistent with the findings we just summarized. In other words, the criterion adapts or varies from one probe to the next, but this is based on selective attention; factors that influence selective attention, which should usually be invariant over the course of a list, may influence the set of attended features, and by that route, potentially influence the final criterion used.

Finally, consider that for lists of mixed strengths, there is a distinction between a strong and a weak target, but lures are not distinguished by strength, because strength is determined by processing of the stimulus during the study phase. For a given probe item, the participant must select the criterion without knowledge of the possible strength. For the nested model, where the D feature subspace includes the S feature subspace, Caplan (2023) assumed that when testing mixed lists, the probe item would be based on the D condition, the greater of the two. This leads to a prediction of no difference in hit rate for strong items in mixed versus pure lists, but a reduction in hit rate for weak items. The false-alarm rate should be the same as for pure-strong lists. If some S features can be disregarded, the hit rate should still change but far less. This does, in fact, resemble our findings (Tables 1 and 2 and Fig. 5).

Response times and access speed for shallow versus deep features. A critical assumption of our model implementation of stimulus-duration was that the shallower, more fully subsetted features are accessed earlier than the deeper, more sparsely subsetted features. Paired with our assumption that probe stimuli are processed largely the same as study items, this led us to expect a speed–accuracy tradeoff, where more processing time, and thus longer response time, should result in greater accuracy when judgements are based more often on those deeper features. Ratcliff and Murdock (1976) reported this kind of effect in a between-subjects manipulation of stimulus duration. In our data, response times were longer for many probe types following pure-strong than pure-weak lists (note that for a manipulation of strength via spaced repetitions, Criss, 2010 found response times were generally faster for pure-strong than pure-weak lists, so repetition may function differently, perhaps if shallow features cannot be disregarded, as we suggested above). The main exception was response times for hits in Experiment 2, which was equivalent for pure-strong and pure-weak lists, which suggests some additional nuance. Adding to the model a formal process to produce response times, such as the diffusion model (like Cox, 2024; Criss, 2010; Osth et al., 2017) may shed further light on this. And of course, the response-time effects may be well explained in numerous other ways, so although we view them as supportive of our theoretical account of stimulus-duration, they might also be unrelated to attentional subsetting.

Mirror effects based on stimulus class and at the item level. In the introduction, we cited Neath et al. (2021) to justify a focus on the strength-based mirror effect, setting aside the older findings of mirror effects when stimulus properties such as word-frequency were manipulated. This was because Neath et al. (2021) found that when they put more effort than previous researchers into controlling stimulus characteristics that were not of interest, manipulations of single item-properties affected predominantly the hit rate or predominantly the false alarm rate but not both. The main lesson from their findings is that in manipulations that compare two sets of stimuli, the mirror effect is often mimicked by two separable effects that happened to both differentiate the two stimulus sets. This was reinforced by our finding of a null mirror effect at the item-level, corroborating similar reports by Cortese et al. (2010, 2017).

The Neath et al. (2021) experiments would suggest that there could be two separable factors influencing primarily hit rate or primarily false alarm rate, respectively. In our formulation, hit rate is primarily influenced by $n_{C,i}$ (Fig. 7). Because $\theta_{C,i}$ is proportional to $n_{C,i}$, items with a greater number of attended features in the test phase will have greater strengths, which will be partly (but not entirely) offset by the threshold being greater, increasing the hit rate for the item.

False alarms are produced by accidental matches of attended features of the probe item to features stored in memory — i.e., attended because of the presence of other items. If attentional subsetting were strictly sparse, this would never happen — there would be no false alarms at all, thus no mirror effect. So if items are processed primarily within a large, sparsely subsetted feature space, we expect to see some items with a higher hit rate than others, but with no associated reduction in false-alarms — because the false-alarm rate is virtually zero. In most recognition experiments (presumably tuned based on experimenter's intuition, precedent, and the desire to calibrate the task to achieve sensitivity), participants do make false alarms. If the attentional subsetting framework is valid, then the fact that participants produce false alarms at any reasonable rate suggests that features common to the lure items do get encoded. Encoding of features attended in a lure item will happen when there are common features across the stimuli *and* those features are attended (subsetted). This will occur frequently when the attended feature space is relatively low-dimensional, so that subsetting cannot be sparse, such as with shallow encoding conditions or short stimulus duration during study. Matching features to lure items will also be more prevalent for stimuli that have more features in common with other words in the stimulus pool, which is how word frequency has been modelled (e.g., Criss & Shiffrin, 2004; Malmberg et al., 2002; Shiffrin & Steyvers, 1997). If other factors are controlled, high-frequency items would be expected to produce more false alarms, with little effect on hit rate, which is the pattern reported by Neath et al. (2021) in their third experiment. When uncontrolled for other stimulus properties, their fourth experiment replicating a classic manipulation of word-frequency produced a mirror effect, which we would presume is because an uncontrolled factor resulted in greater $n_{C,i}$ for items within the "high-frequency" set compared to the "low-frequency" set.

When we treated our two experiments roughly as replications of each other, we found that there was indeed quite a lot of replication of both hit rate and false-alarm rate across items, between experiments, replicating verbal recognition studies, resonating with what has been found in both hit rate and false-alarm rate of continuous recognition of faces and complex visual images (Bainbridge, 2020; Bainbridge et al., 2013; Bainbridge & Rissman, 2017; Isola et al., 2011). That suggests that in tasks similar to these, item-difficulty effects are strong relative to subject-variability. Especially the reproducibility of the pattern of false-alarms across words aligns well with our assumption that participants tune their criterion based on the current probe word, so that the criterion effectively changes from one word to the next. The remaining item-level analyses weakened the argument that a single factor influences both hit and false-alarm rates in tandem, and that strength manipulations also influence hits differently than false-alarms across items.

Future direction: forgetting and scaling the criterion by an estimate of the dimensionality of the memory. In deriving the mirror effect, we avoided making use of any meta-knowledge of the memory, itself, but in practice, some knowledge of the characteristics of the memory would be necessary to adjust the criterion effectively. There is evidence that participants have some knowledge of ensemble properties of stimuli (e.g., Dubé et al., 2019; Tong & Dubé, 2022a, 2022b; Tong et al., 2019). For simplicity, we have omitted forgetting from the model. Forgetting would have the tendency to either reduce the amplitude of each encoding term in the memory (multiplication by a scalar) or potentially the degradation of individual features, so the number of encoded features of any given item would be effectively reduced. Without adjusting the criterion to account for forgetting, the current heuristic would eventually place the criterion too high, leading to a high (or 100%) rate of misses. Movement in this direction was seen by Singer and Wixted (2006) in probe sets mixed from one list just studied and a second list studied 20 or 40 minutes prior, where lists were composed of different categories. For any non-negligible delay, the heuristic would need to be adjusted downward to reflect this, but this might be achieved simply by a scalar factor, not demanding any detailed knowledge of the forms of the expected strength distributions. In their last two experiments, Singer and Wixted (2006) used a delay of 48 h and found evidence for a criterion adjustment depending on list-recency. So for delays up to tens of minutes, participants may not substantially adjust their criterion, but for long enough delays, they evidently do. In the current formalism, there are two ways in which forgetting might perturb the memory of a list.

First, suppose that forgetting results simply in a gain factor, scaling down the overall length of the memory vector,

$$\mathbf{m}_T = \rho(T)\mathbf{m}, \tag{8}$$

where $\rho(T) < 1$ and is a monotonic function decreasing with increasing study–test time, T. Scaling the criterion the same way,

$$\theta_T = \rho(T)\theta = \rho(T)\frac{1}{2}\frac{n_{C,i}}{n} \tag{9}$$

will compensate for this, placing the criterion midway between the expected mean strength for targets and lures, taking into account forgetting. Moreover, $\rho(T)$ can be derived with a simple calculation from $\mathbf{m}_T$. Without further assumptions, the full dimensionality of the memory is $n_1 = n_C$ for one item (where now $n_C = \mathrm{E}\left[n_{C,i}\right]$, the average across items) and about $n_2 = 2n_C - n_C^2/n$ for 2 items. This iterates, such that $n_m = mn_C - n_{m-1}n_C/n$. In the sparse limit (small n_C/n), $n_L = Ln_C$, linear in n_C. But the more we deviate from sparseness, the more overlap there will be in attentional masks across items, and the lower n_m will be. Thus, n_L is a sublinear function of L and of n_C. If the participant has a reasonable estimate of L, and of n_C (using the $n_{C,i}$ for the current probe item), then $\rho_T \sim (n_L/L)/n_{C,i}$, where n_L is the number of non-zero (attended) features in the memory.

Second, suppose that forgetting results not in a scaling down of $\mathbf{m}$, but of zeroing out of features with some probability, Π_T, dependent on study–test time, T. In this scenario, $n_L \simeq (1 - \Pi_T)Ln_C$. Thus, if the participant has access to an estimate of the number of attended features in the memory, that becomes the scale factor.

Conclusion. In sum, the attentional subsetting account of manipulations of duration anticipated (retroactively) an inverted list-strength effect. This was in fact a feature of the first experiment in the long line of results viewed as "null" list-strength effects (Experiment 1 of Ratcliff et al., 1990). We reproduced the inverted list-strength effect in a replication of that experiment and a follow-up experiment designed to increase the effectiveness of the duration manipulation, suggesting it is not a statistical accident but needs explaining. This provides support for our model account. We extended attentional subsetting theory to model hits and false alarms separately, introducing the idea that participants customize their response criterion based on the number of features extracted from a probe stimulus. This avoids past criticisms of criterion-shift account of mirror effects, because the participant does not need to have unrealistic levels of knowledge of the expected distributions of matching strengths. And yet, the item-wise customization of the criterion can explain why it has appeared that participants must flexibly adapt their criterion (or local-trace differentiation mechanisms are at play). Finally, our theory of duration draws direct attention to how the dimensionality of the attended feature space may unfold over the course of processing of a stimulus from dense to sparse subsetting. This account is compatible with a range of models that assume a vector representation of items: The attentional mask can be applied by elementwise multiplication, to most vector representations with similar effect. Then the subsequent computations applied to the vector (e.g., echo

strength for MINERVA 2 or likelihood ratio for REM) can be carried out as usual. The threshold would need to be adapted to align with the nonlinearities of nonlinear matching functions. It will therefore be interesting, in the future, to investigate how attentional subsetting might be productively combined with current well developed and more complete local trace and global matching models.

CRediT authorship contribution statement

Jeremy B. Caplan: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Dominic Guitard:** Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Partly supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Appendix. Preliminary models

To increase the coverage of the theory, we derive hit and false alarm rates for two variants of the model as formulated by Caplan (2023) with potential application to experimental manipulations other than stimulus duration, such as two different processing tasks. In the first variant, the probe consists of the full lexicon vector (length n). In the second, attentional subsetting is applied to the recognition probe.

Preliminary model 1: no attentional subsetting of the probe

Using solutions from Caplan (2023), we start with the simplest case, where we (unrealistically) assume the full vector representation of the probe is matched against memory. Each item stored in memory contributes a term to the variances. We denote the variance due to the dot produce a target item with itself V_{xx}, and the variance due to the dot product of the probe item (target or lure) with each other list item V_{xy}. We find

$$V_{xx} = 2n_C/n^2 \tag{A.1}$$

$$V_{xy} = n_C/n^2,\ y \neq x. \tag{A.2}$$

Targets will be subject to one V_{xx} term (itself) and $L-1$ V_{xy} terms (all other studied items). Lures will be subject to L V_{xy} terms since no studied item is an exact match. The expressions for the variances are thus

$$\sigma^2_{\text{target}} = V_{xx} + (L-1)V_{xy} \tag{A.3}$$

$$\sigma^2_{\text{lure}} = LV_{xy}. \tag{A.4}$$

Thus, $\sigma^2_{\text{target}} = 2n_C/n^2 + (L-1)n_C/n^2 = (L+1)n_C/n2$ and $\sigma^2_{\text{lure}} = Ln_C/n^2$. For large L, the variances approach equality. The numerator within the erf() function for hits is $n_C/2n - n_C/n = -n_C/2n$ and for lures is simply $n_c/2n$, and note that n in the numerator and denominator of the fraction cancel.

$$P(\text{hit}) = 1 - \left(0.5 + 0.5\ \text{erf}\left(-\frac{1}{2\sqrt{2}}\sqrt{\frac{n_C}{L+1}}\right)\right) \tag{A.5}$$

$$P(\text{false alarm}) = 1 - \left(0.5 + 0.5\ \text{erf}\left(\frac{1}{2\sqrt{2}}\sqrt{\frac{n_C}{L}}\right)\right) \tag{A.6}$$

The absence of n in these expressions offers some realism to the model; the full set of knowledge (full vector representation of an item) is immaterial to hit versus false-alarm rates. What matters is the dimensionality of the attentional masks, as well as list length.

Importantly, without differentiation and without knowledge of the expected distributions of strengths, this produces a mirror effect. These closed-form expressions show with θ chosen as the midpoint between the expected matching strength for lures (0) and that for the probe (which we here are assuming is determined by the participant's own immediate meta-knowledge of the approximate number of attended features of the probe), that the chief difference between the hit rate and the false alarm rate is the sign of the expression within the erf(). Given that $\text{erf}(x) = -\text{erf}(x)$, as L becomes arbitrarily large, as we vary n_C, P(false alarm) will move nearly symmetrically in the opposite direction to P(hit), a nearly symmetric mirror effect (Fig. A.1a). For small L, the $\sqrt{L+1}$ versus $\sqrt{L}$ in the denominator means that a given change in n_C will produce a larger shift in the P(false alarm) than in P(hit), a lopsided mirror effect. If we add realism by assuming participants either under-estimate or over-estimate the number of features they attend, the mirror effect can easily become more asymmetric, resembling published data.

Preliminary model 2: masked probe

The second model considered by Caplan (2023) adds realism by assuming that the probe is attentionally subsetted in the identical manner as it would have been had it been an item presented for study. Bearing in mind that we are considering pure lists only (for the purposes of the mirror effect), it is plausible to presume the participant processes the probe stimulus in the same manner as they had been doing generally during the study phase. This also suggests how the participant might have more or less direct access to n_C. Having processed a probe item, the participant might (a) have conscious access to the number of features extracted from the stimulus (with expectation equal to n_C) or (b) have the ability to compute $\nu = \|\mathbf{w}_{C,x} \otimes \mathbf{f}_x\|$, where $\mathbf{f}_x$ is the probe item, $\mathbf{w}_{C,x}$ is the item-specific mask in condition C, and $\|\cdot\|$ is the norm (vector-length). Because $\mathrm{E}[\nu] = n_C/n$, if $\theta = \nu/2$, overall this will result in θ being set midway between the expected target and lure distributions. If participants derive θ in real-time for each probe, this would optimize the criterion at the item-level, and might influence accuracy to the extent that n_C varies across items.

The masked model produced a (near-)null list-strength effect when n_C was low enough relative to n to produce sparse subsets. This was because the masked probe was unlikely to overlap with other list items. As n_C increased, mask overlap was more probable, and list-strength effects became large. The proposed criterion, midway between μ_{target} and μ_{lure}, is still $\theta_C = n_C/2n$, because the mean matching strengths are identical for the masked probe as for the full probe. Only the variances need to be adjusted for the variance due to cross-terms. These cross-terms are non-zero only where there is chance overlap between the mask of the probe with the mask of other items, $\Omega_{CC} = n_C^2/n$ features. Each non-target list item contributes $V_{xy} = n_C^2/n^3$ to the variance. Targets are subject to $L-1$ of these (plus variance due to the encoded target item, itself) and lures are subject to the full L of these. Thus:

$$\sigma^2_{\text{target}} = \frac{2n_C}{n^2} + (L-1)\frac{n_C^2}{n^3} \tag{A.7}$$

$$\sigma^2_{\text{lure}} = L\frac{n_C^2}{n^3}. \tag{A.8}$$

Substituting these expressions for the means and variances into Eqs. (6) and (7):

$$P(\text{hit}) = 1 - \left(0.5 + 0.5\ \text{erf}\left(\frac{-\sqrt{n_C}}{2\sqrt{2(2+(L-1)n_C/n)}}\right)\right) \tag{A.9}$$

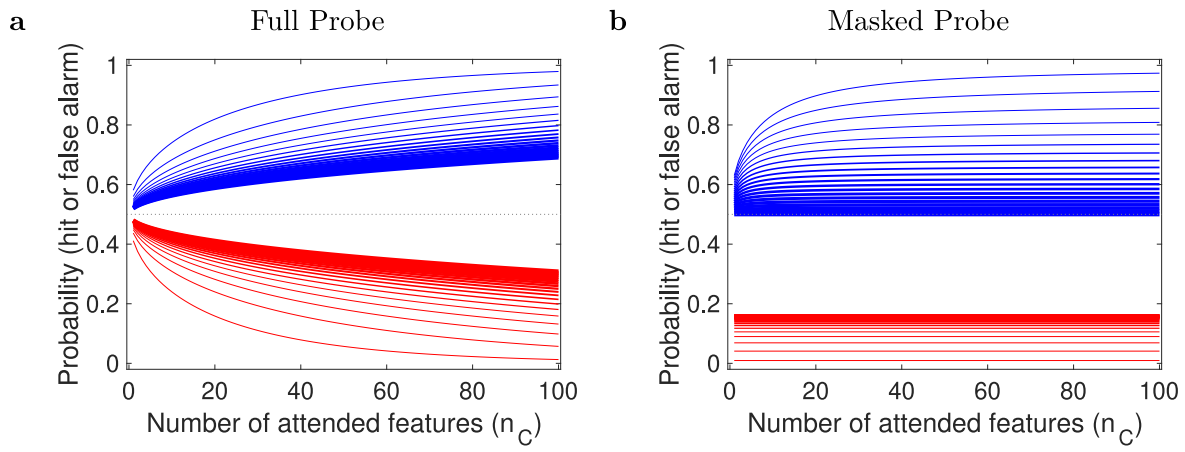

Fig. A.1. Hit rate (blue) and false alarm rate (red) as a function of the number of item features subsetted by selective attention (n_C), for list lengths varying from $L = 5$ items (thinnest line) to $L = 100$ items (thickest line), in steps of 5 items. (a) Full-probe model. (b) Masked-probe model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$P(\text{false alarm}) = 1 - \left(0.5 + 0.5\ \text{erf}\left(\frac{\sqrt{n_C}}{2\sqrt{2}\sqrt{Ln_C/n}}\right)\right)$$
$$= 1 - \left(0.5 + 0.5\ \text{erf}\left(\frac{1}{2\sqrt{2}\sqrt{L/n}}\right)\right). \quad \text{(A.10)}$$

The $\sqrt{n_C}$ in the numerator and denominator cancel in the expression for false alarms. In other words, although hit rate increases with increasing n_C, the false alarm rate is invariant in n_C (although it increases with L and decreases with n), so we have lost the mirror effect.

Interestingly, consider the effect of list length. The false-alarm rate is influenced by $\sqrt{L}$ in the denominator of the expression inside the erf(). For the hit rate, $(L - 1)$ multiplies n_C/n but is then added to the number 2. If we vary L but hold n_C fixed, recalling that our assumption is that in general, $n_C \ll n$, $(L - 1)n_C/n \ll 2$ so varying L will have negligible effect. The list-length effect is thus expected to have a substantial influence on false-alarms with very little effect on hits. This is what Ratcliff et al. (1990) found in their paradigm with lists composed of many categories, where "list-length" was effectively manipulated by varying the number of items within a given category. A strength manipulation (via spaced repetition) produced a large shift in hit rate with little effect on false-alarm rate, also resembling this model variant (the dependence on n_C in Fig. A.1b).

References

Anderson, J. A. (1970). Two models for memory organization using interacting traces. *Mathematical Biosciences*, *8*, 137–160.

Bainbridge, W. A. (2020). The resiliency of image memorability: a predictor of memory separate from attention and priming. *Neuropsychologia*, *141*(107408).

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323–1334.

Bainbridge, W. A., & Rissman, J. (2017). Dissociating neural markers of stimulus memorability and subjective recognition during episodic retrieval. *Scientific Reports*, *8*(8679).

Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D.-L., & Bernstein, D. M. (2016). The production effect in recognition memory: weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology*, *70*(2), 93–98.

Brockdorff, N., & Lamberts, K. (2000). A feature-sampling account of the time course of old-new recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(17), 77–102.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: a network model of the phonological loop and its timing. *Psychological Review*, *106*(3), 551–581.

Caplan, J. B. (2023). Sparse attentional subsetting of item features and list-composition effects on recognition memory. *Journal of Mathematical Psychology*, *116*(102802).

Caplan, J. B., Chakravarty, S., & Dittmann, L. (2022). Associative recognition without hippocampal associations. *Psychological Review*, (6), 1249–1280.

Caplan, J. B., & Guitard, D. (2024). A feature-space theory of the production effect in recognition. *Experimental Psychology*, *71*(1), 64–82.

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

Chappell, M., & Humphreys, M. S. (1994). An auto-associative neural network for sparse representations: analysis and application to models of recognition and cued recall. *Psychological Review*, *101*(1), 103–128.

Cortese, J., Khanna, M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory*, *18*(6), 595–609.

Cortese, J., McCarty, D. P., & Schock, J. (2017). A mega recognition memory study of 2897 disyllabic words. *Quarterly Journal of Experimental Psychology*, *68*(8), 1489–1501.

Cox, G. E. (2024). An integrated dynamic approach to item and associative recognition: unraveling the roles of capacity, attention, and decision. PsyArXiv, (submitted for publication).

Cox, G. E., Hemmer, P., Aue, W. R., & Criss, A. H. (2018). Information and processes underlying semantic and episodic memory across tasks, items, and individuals. *Journal of Experimental Psychology: General*, *147*(4), 545–590.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*, 671–684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Criss, A. H. (2006). The consequences of differentiation in episodic memory: similarity and the strength based mirror effect. *Journal of Memory and Language*, *55*(4), 461–478.

Criss, A. H. (2009). The distribution of subjective memory strength: list strength and response bias. *Cognitive Psychology*, *59*(4), 297–319.

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

Criss, A. H., & Malmberg, K. J. (2008). Evidence in favor of the early-phase elevated-attention hypothesis: the effects of letter frequency and object frequency. *Journal of Memory and Language*, *59*(3), 331–345.

Criss, A. H., & Shiffrin, R. M. (2004). Interactions between study task, study time, and the low-frequency hit rate advantage in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 778–786.

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 18–33.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, *54*(3), 304–313.

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

Dubé, C., Tong, K., Westfall, H., & Bauer, E. (2019). Ensemble coding of memory strength in recognition tests. *Memory & Cognition*, *47*(5), 936–953.

Ensor, T. M., Bancroft, T. D., Guitard, D., Bireta, T. J., Hockley, W. E., & Surprenant, A. M. (2020). Testing a strategy-disruption account of the list-strength effect are sampling bias and output interference responsible? *Experimental Psychology*, *67*(4), 255–275.

Ensor, T. M., Surprenant, A. M., & Neath, I. (2021). Modeling list-strength and spacing effects using version 3 of the retrieving effectively from memory (REM.3) model and its superimposition-of-similar-images assumption. *Behavior Research Methods*, *53*(1), 4–21.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080words. *Behavior Research Methods and Instrumentation*, *14*, 375–399.

Gardiner, J. M., Ramponi, C., & Richardson-Klavehn, A. (1999). Response deadline and subjective awareness in recognition memory. *Consciousness and Cognition*, *8*(4), 484–496.
Gibson, E. J. (1971). Perceptual learning and the theory of word perception. *Cognitive Psychology*, *2*(4), 351–368.
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 & Cognition*, *13*(1), 8–20.
Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 5–16.
Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*(3), 546–567.
Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51.
Hicks, J. L., & Starns, J. J. (2014). Strength cues and blocking at test promote reliable within-list criterion shifts in recognition memory. *Memory & Cognition*, *42*(5), 742–754.
Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551.
Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 201–205.
Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, *33*(1), 1–18.
Hirshman, E. (1995). Decision processes in recognition memory: criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 302–313.
Hockley, W. E., & Niewiadomski, M. W. (2007). Strength-based mirror effects in item and associative recognition: evidence for within-list criterion changes. *Memory & Cognition*, *35*(4), 679–688.
Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*, 534–537.
Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *24th IEEE conference on computer vision and pattern recognition* (pp. 145–152). IEEE.
Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: still playing twenty questions with nature. *Canadian Journal of Experimental Psychology*, *70*(2), 154–164.
Johnson, N. F. (1975). On the function of letters in word identification: some data and a preliminary model. *Journal of Verbal Learning and Verbal Behavior*, *14*(1), 17–29.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Society*, *90*(430), 773–795.
Kiliç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology*, *92*, 65–86.
Kim, K., & Glanzer, M. (1993). Speed versus accuracy instructions, study time, and the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 638–652.
Koop, G. J., Criss, A. H., & Pardini, A. M. (2019). A strength-based mirror effect persists even when criterion shifts are unlikely. *Memory & Cognition*, *47*(4), 842–854.
Lau, M. C., Goh, W. D., & Yap, M. J. (2018). An item-level analysis of lexical-semantic effects in free recall and recognition memory using the megastudy approach. *Quarterly Journal of Experimental Psychology*, *71*(10), 2207–2222.
Lewis, D. J. (1979). Psychobiology of active and inactive memory. *Psychological Bulletin*, *86*(5), 1054–1083.
MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685.
MacMillan, M. B., Ensor, T. M., Surprenant, A. M., & Neath, I. (2022). Stimulus-based mirror effects in associative recognition revisited. *Canadian Journal of Experimental Psychology*.
Malmberg, K. J., & Nelson, T. O. (2003). The word frequency effect for recognition memory and the elevated-attention hypothesis. *Memory & Cognition*, *31*(1).
Malmberg, K. J., & Shiffrin, R. M. (2005). The one-shot hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2).
Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, *30*(4), 607–613.
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.
Medin, D. L., & Shoben, E. J. (1988). Context and structure in conceptual combination. *Cognitive Psychology*, *20*(2), 158–190.
Morrell, H. E. R., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(6), 1095–1110.
Mulligan, N., & Hirshman, E. (1995). Speed–accuracy trade-offs and the dual process model of recognition memory. *Journal of Memory and Language*, *34*(1), 1–18.
Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(3), 689–697.
Nairne, J. S. (1990). A feature model of immediate memory. *Memory & Cognition*, *18*(3), 251–269.
Neath, I., Hockley, W. E., & Ensor, T. M. (2021). Stimulus-based mirror effects revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.
Osth, A. F., & Dennis, S. (2014). Associative recognition and the list strength paradigm. *Memory & Cognition*, *42*(4), 583–594.
Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260–311.
Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101–126.
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. (1989). Similarity information versus relational information: differences in the time course of retrieval. *Cognitive Psychology*, *21*, 139–155.
Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 763–785.
Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190–214.
Sahakyan, L. (2019). List-strength effects in older adults in recognition and free recall. *Memory & Cognition*, *47*(4), 764–778.
Sahakyan, L., & Malmberg, K. J. (2018). Divided attention during encoding causes separate memory traces to be encoded for repeated events. *Journal of Memory and Language*, *101*, 153–161.
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.
Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 179–195.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.
Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: evidence for a criterion shift. *Memory & Cognition*, *34*(1), 125–137.
Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language*, *63*(1), 18–34.
Starns, J., White, N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: the evidence for differentiation can be produced without differentiation. *Memory & Cognition*, *40*(8), 1189–1199.
Stoet, G. (2010). A software package for programming psychological experiments using linux. *Behavior Research Methods*, *42*(4), 1096–1104.
Stoet, G. (2017). A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, *44*(1), 24–31.
Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1379–1396.
Tong, K., & Dubé, C. (2022a). Modeling mean estimation tasks in within-trial and across-trial contexts. *Attention, Perception, & Psychophysics*, *84*(7), 2384–2407.
Tong, K., & Dubé, C. (2022b). A tale of two literatures: a fidelity-based integration account of central tendency bias and serial dependency. *Computational Brain & Behavior*, *5*, 103–123.
Tong, K., Dubé, C., & Sekuler, R. (2019). What makes a prototype a prototype? averaging visual features in a sequence. *Attention, Perception, & Psychophysics*, *81*(6), 1962–1978.
Tulving, E. (1968). Theoretical issues in free recall. In T. R. Dixon, & D. L. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 2–36). Prentice-Hall, Inc..
Verde, M. F., & Rotello, M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition*, *35*(2), 254–262.
Wu, L.-l., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: evidence from property generation. *Acta Psychologica*, *132*(2), 173–189.
Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(2), 345–355.