Inverted list-strength effects in recognition

Jeremy B. Caplan[1] and Dominic Guitard[2]
[1]Department of Psychology and Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Alberta, Canada
[2]School of Psychology, Cardiff University, Cardiff, United Kingdom

Abstract

If some list items are studied strongly and others weakly, many memory models predict the effect of strength on memory will be larger when strengths are mixed within a list than between pure lists of a single strength: a list-strength effect. In explaining why list-strength effects were elusive in old/new recognition, Shiffrin et al. (1990) introduced differentiation. This gave the model a way to produce an inverted list-strength effect which they thought was usually offset by the co-existing expected "upright" list-strength effect. Alternatively, attentional subsetting theory (Caplan, 2023; Caplan & Guitard, 2024b) predicted inverted list-strength effects in some circumstances by considering how the dimensionalities of attended feature spaces might differ for strong and weak items. Inversions were indeed found in manipulations of stimulus duration (e.g., 500 ms versus 2000 ms study time/word). Here we replicated the pattern when display time was equated (Experiment 1) and with massed-repetition (Experiment 2), ruling out the relevance of vision-locked features and number of stimulus onsets. Both theoretical accounts of inverted list-strength effects, however, miss the fine structure of the data, namely, reduced hit rates for weak items in pure than mixed lists and the reverse effect (albeit less robust) for strong items. Model fits suggested the critical factor is that list composition parameterically influences the number of deep features processed at test combined with participants response bias adapting to list composition. In sum, inverted list-strength effects are robustly found in manipulations of item study time and point to differential processing of probe features depending on list composition, compatible with most models.

Keywords: List-strength effect, recognition memory, selective attention, matched filter model

Jeremy B. Caplan https://orcid.org/0000-0002-8542-9900. Dominic Guitard https://orcid.org/0000-0002-4658-3585. Partly supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Corresponding author: Jeremy Caplan, jcaplan@ualberta.ca, Department of Psychology, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada.

Introduction

Whenever an empirical finding cannot be explained by any existing memory model, that can inspire research that can lead to major advances in how we understand memory. One of the best cases of this is the so-called “null list-strength effect” in old/new recognition. In old/new recognition, participants study a list of items, usually words, and typically a large number of them, such as 30 or 50. Then words are presented one at a time and participants judge whether the word was on that list or not. In a list-strength design, list items processed one way, such as with a short presentation time like 1 s, are labelled “weak” and items processed another way, such as with a longer presentation time like 2 s, are labelled “strong.” Strong items are recognized better than weak items; this is the starting point of the list-strength design. Each list is composed of either all one strength (pure-weak or pure-strong lists) or mixed, usually 50:50 weak and strong items. Ratcliff et al. (1990) reasoned that *all existing models predicted* that strong items should have a competitive advantage when embedded in a mixed lists. That is because strong items face less competition from the other items: in a mixed list, there are $L/2-1$ other strong items and $L/2$ weak items ($L =$ list length), whereas in a pure list there are more $(L-1)$ strong other items. Conversely, a weak item should be disadvantaged in a mixed list, facing more competition than in a pure-weak list.

Following the history of this line of thought, inour recent publicatinos we have written “**upright list-strength effect**” when the strength effect is greater in mixed than in pure lists (the traditionally expected effect) and “**inverted list-strength effect**” when the strength effect is greater in pure than in mixed lists (opposite what was expected). Thus, that effect is “upright” or as expected, and the opposite is an inversion of that effect. Ratcliff, Shiffrin, Criss and others have used the terms “positive” and “negative.” Because “negative” can be confused with a null effect (non-significance), we find “upright” and “inverted” to be less confusing, although arguably more theoretically loaded. We note that there may be no perfect terminology. The important thing is to be able to align our exposition with others’, so the mapping is upright=positive and inverted=negative.

To the contrary, Ratcliff and colleagues found little evidence for such a list-strength effect. So at least at the time, this meant that strictly speaking, *all models were wrong.* A more nuanced view, of course, is that the models may not have been entirely wrong but at least demanded amendments or changes in assumptions to accommodate this unexpected result. This led to a very fruitful chapter in which models were developed with null list-strength effects in mind. To summarize, two types of model accounts of near-null list-strength effects were advanced: 1) Models that assumed items are orthogonal, hence no influence of other items (Chappell & Humphreys, 1994; Dennis & Humphreys, 2001). In such models, which the authors also called “context-noise” models, the critical factor driving old/new recognition of an item was the item’s association to the target-list (aka “context”), whereas judgements of an item were nearly immune to variability in the encoded strengths of other list items due to their orthogonality. 2) Differentiation models, typically implemented in local trace models (each item is represented in its own separate memory record) such as McClelland and Chappell (1998) and Retrieving Effectively from Memory, (REM; Shiffrin and Steyvers, 1997). Differentiation refers to a strong item producing a higher likelihood of having been studied, but that same strong trace producing more evidence against lure items.

A strong item may have a similar hit rate in mixed as in pure lists, but the false alarms will be reduced in pure-strong lists, producing an inverted list-strength effect. To produce a near-null list-strength effect, Shiffrin et al. (1990) proposed that this is typically offset by an upright list-strength effect due to cueing with context features, which are common to all items. That is because a single cue associated with a number of items produces what has been called a fan effect, or cue-overload, and this ambiguity is supposed to produce competition (when item representations are not forced to be orthogonal), very much like what Ratcliff et al. (1990) saw the older models predicting. So the standard differentiation model produces a near-null list-strength effect due to the presence of both an upright and an inverted list-strength effect that roughly cancel out.

Inverted list-strength effects. But actually, Ratcliff et al. (1990) reported an *inverted* list-strength effect comparing 1000 ms (weak) to 2000 ms (strong) durations; strength had a larger effect when comparing pure lists than when comparing items within mixed lists. Another significant inverted list-strength effect was reported by Ratcliff et al. (1994) but like most authors on the topic, they treated the inversions as due to suspected measurement error; they were more struck by the fact that the average list-strength effect across all their experiments was close to null. However, those inverted list-strength effects may have been legitimate and not just null effects that appeared to invert due to chance because Caplan and Guitard (2024b) replicated their inverted list-strength effect manipulating strength between 1000 ms and 2000 ms duration. They also found an even more robust inverted list-strength effect comparing 500 ms to 2000 ms durations. This suggests that inverted list-strength effects may not false positives, but legitimate results that need to be explained. In this manuscript, we seek further evidence of replication of the inverted list-strength effect and test some ideas about its potential boundary conditions.

Regarding the two classes of theoretical accounts of list-strength effects in recognition, orthogonal representation accounts do not have an obvious way to produce inverted list-strength effects, so if inverted list-strength effects are not rare exceptions but quite common, this would pose an increasing challenge to such models. Differentiation models, as just described, produce near-null list-strength effects by assuming the presence of both an upright effect due to the use of context as a cue and an inverted effect due to differentiation at the level of representations of items. If recognition were driven more by item features than by context features, the inverted list-strength effect could predominate. Additional information about conditions that produce inverted list-strength effects could thus test differentiation models, especially the relative reliance on item versus context features.

However, our interest in inverted list-strength effects arose from a third theoretical account we recently developed, attentional subsetting theory, which seemed to anticipate inverted list-strength effects under certain conditions (i.e., model-parameter values). First proposed by Caplan (2023), this theory starts with the assumption that in an episodic memory task, participants do not process all known features of a stimulus (Figure 1a), but rather, just a few features (Figure 1b), generally producing sparse functional representations (mostly zeroes). Next, it is assumed that upon each encounter with a stimulus such as a word, the participant often processes the same subset of features (Figure 1b–d), unless conditions change substantially. For example, each time reading the word HUMMINGBIRD, the participant is likely to think about the iridescent feathers, hovering and long beak (and perhaps not think about the item's food qualities or predator/prey status). Near-null list-

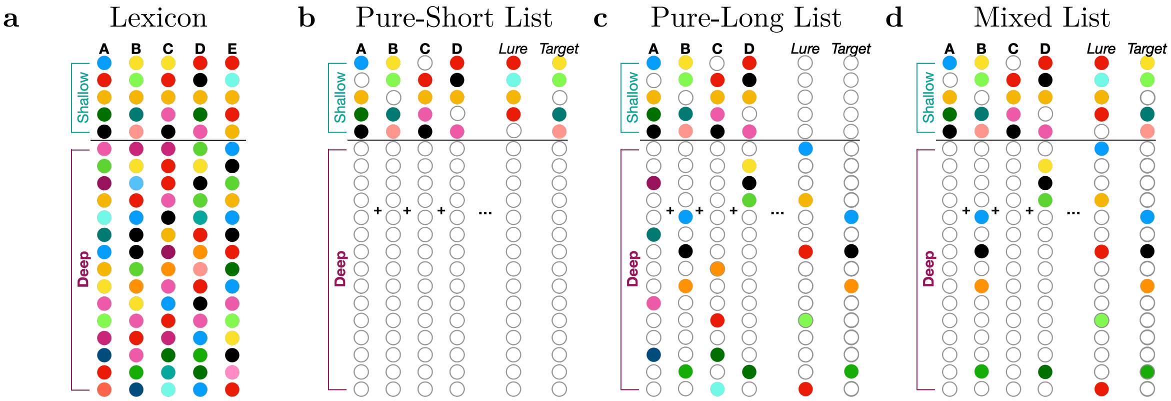

Figure 1

Illustration of the attentional subsetting theory account of inverted list-strength effects in some manipulations of stimulus duration. (a) Each item (A, B, etc.) is an n-dimensional column vector, with some of them depicted here. Each circle stands for a vector dimension and its colours (arbitrarily chosen) denotes its value. (b) In a pure-short list, only shallow features are stored and at test, correspondingly, participants only process shallow features of the probe. Grey unfilled circles denote features that are not attended (and thus not encoded during the study phase). (c) In a pure-long list, both shallow and deep features are stored, but at test, because of the greater diagnostic ability of the deeper features, we assume participants disregard (and thus do not take into account in their recognition judgements) shallow features. (d) In a mixed list, short and long items are encoded as in the pure lists and both shallow and deep features are processed at test. In this mixed-list example, A and C are short items and B and D are long. For illustration purposes, the example lure is the same item in all cases and the target item is item B.

strength effects arise because under many circumstances, the attended features of one item will have little overlap with the attended features of another item. For example, the word PENGUIN may evoke features like their affinity to the cold, inability to fly, and penchant for fish (Figure 1, “deep” features). Not only are the feature values different, the types of features, themselves, will often be different. This produces minimal (but not zero) confusion across items, and thus readily produces list-strength effects that are not strictly null, but rather small, in line with the bulk of the recognition-list-strength effect data (Caplan, 2023).

Caplan and Guitard (2024b) elaborated attentional subsetting theory for manipulations of stimulus duration. At short durations, around 500 ms or so, there is sufficient time to process shallow features (Figure 1b), such as visual word-form, orthographic and phonological features, but not deeper features such as semantic or imagery-related features. Longer duration results in more stored deep features (Figure 1c). The assumption is that the shallow features are drawn from a low-dimensional feature space, so even if only a handful of such features are stored per item, those will introduce a lot of similarity-based confusion across list items and between studied items and lure probes. The additional deep features available with longer durations will be drawn from a much higher-dimensional feature space, so they will be sparse. Because the subset is assumed to be item-specific, those sparse vectors will introduce very little similarity-based confusion. Finally, we assume that when sufficient deep features are available to make recognition judgements, participants will be able to disregard those confusing shallow features— namely, when tested on pure-strong lists. As a result, strong (long-duration) items have a bit of an advantage in pure lists because the confusing shallow items can be disregarded (Figure 1c), exaggerating the effect of the manipulation of strength (duration).

Attentional subsetting theory provided a good account of the data. However, what led us to the current two experiments was that it is not clear what the nature of those putative shallow features is. They might be low-level visual features such as points, line- and curve-segments, etc. Or the shallow functional features might be related to the visual forms of the letters comprising the word. Alternatively, the features might be unrelated to immediate processing of the visual stimulus: they might be at the conceptual level of orthographic (or other) features. If we remove the stimulus from the screen after 500 ms in the strong as well as the weak condition, we can ask what happens when the immediate sensory processing of the stimulus no longer differs by item-strength. Experiment 1 therefore tests the hypothesis that the inverted list-strength effect depends on the strong items having more real-time visual processing of the stimuli.

Secondly, to our knowledge, spaced repetitions of single items have not yet been found to produce an inverted list-strength effect. We speculated that this could be a consequence of repeated onsets of the stimuli (Caplan, 2023; Caplan & Guitard, 2024b). Suppose the onset of a word evokes obligatory processing of the shallow features of the word. That means that words studied over spaced repetitions would have repeated processing of those shallow features and thus have more shallow features stored. The additional obligatory shallow processing might also displace some of the time available to encode deeper features. The net result might be that the "strong" condition, with spaced repetitions, have more encoding of shallow features, so that the shallow features can no longer be disregarded. Without disregarding, the pure-strong lists no longer have an advantage and the list-strength effect cannot invert. Here we do not study spaced repetitions, but we use massed (immediate) repetitions to evaluate the hypothesis that an onset of a word induces obligatory re-processing of the shallow (visual or orthographic) features of the word. A massed repetition has those repeated onsets but sticks closer to our other experimental conditions. Thus, our second hypothesis was that repeated onsets, by adding shallow features, prevent participants from potentially benefitting from disregarding shallow features, thus eliminating the inverted list-strength effect.

Our first general goal was thus to test these two hypotheses, regarding whether various factors might weaken or reverse the inverted list-strength effect. To foreshadow, rather than find limits, our findings expanded the generality of the inverted list-strength effect, replicating it in six experimental groups comprising five different task conditions.

Upon close inspection of the precise patterns of hit rates and false-alarm rates, we noted that current formulations of both the differentiation account and the attentional subsetting account make some wrong predictions about the fine structure of the data. We fitted the data with six variants of the attentional subsetting model and found just one variant that captures all the qualitative (rank-order) effects on hit and false-alarm rates. Following the reports of the experiments, we report these model fits and conclude with a new insight into what may make inverted list-strength effects a robust phenomenon, which could be incorporated into most memory models.

The favoured model variant was designed to incorporate the idea that encoding time is redistributed across items. Before describing the experiments, we take a brief detour to consider early debates about redistribution; the lack of clear resolution of this issue sets us up for additional, unplanned analyses we conducted to test the new theoretical account.

Redistribution of study time. Just because the experimenter allots an item a particular amount of time does not mean that participants necessarily study the item for exactly that amount of time and no more. Rehearsal, for example, redistributes study processes across presentation of the list (e.g., Rundus, 1971; Tan & Ward, 2000). Early investigations of the near-null list-strength effect in recognition, there was concern about the possibility of so-called "rehearsal borrowing." Although in such a fast-paced task, there might not be very much rehearsal *per se*, it seems plausible that participants continue to process or extract features of a word as study-time progresses. We remain agnostic as to whether the redistribution is of rehearsals or of item-processing. Suppose there is an underlying upright list-strength effect. In a mixed list, perhaps short-presented items could "borrow" some study time from a subsequent long-presented item. That would redistribute encoding processes from the strong to the weak items and might approximately offset that putative list-strength effect. This is analogous to a diffusion process or a Robin-Hood strategy: stealing from the rich to give to the poor. This process can be considered when list-strength effects are close to null. Interestingly, Yonelinas et al. (1992) considered a form of study-time redistribution but in the opposite way, which they called "reverse rehearsal borrowing." They were worried that they found an upright list-strength effect which seemed to contradict the near-null list-strength effect demonstrated by Ratcliff et al. (1990). They speculated that a long-duration item might capture the participant's interest more than a short-duration item and thus steal rehearsal time from the weak to the strong items— or as Ratcliff et al. (1994) later put it, that participants simply give up on the short-presented items. This could produce the expected but generally rare upright-list-strength effect in some conditions, a rich-get-richer concept.

Worried about redistribution, Ratcliff and colleagues typically blocked item strengths to minimize the opportunity for redistribution to a large degree (Ratcliff et al., 1990, 1992, 1994). Other early researchers conducted analyses looking for sequential dependencies that might signify a weak item stealing encoding resources from a subsequent strong item or a strong item stealing encoding resources from a neighbouring weak item. Using a more complicated procedure, with several words embedded within sentences but tested with single-word recognition probes and strength manipulated by spaced repetitions rather than stimulus duration, Murnane and Shiffrin (1991a, 1991b) found no empirical support for accuracy depending on the subsequent or preceding item strength. Checking for reverse-redistribution, Yonelinas et al. (1992) found sequential-dependencies all non-significant, although they were nominally in the expected direction to potentially explain the presence of an upright list-strength effect. If a weak item preceded the current item, the current item was recognized worse, in nearly all cases, albeit never significantly so. The authors felt they could not entirely rule out redistribution. One more intriguing finding may be relevant: Murnane and Shiffrin (1991a) found no list-strength effect in a final recognition test. That is, all lists were tested initially, but half the items were held out, untested. Those held-out items were tested in final recognition, so they could be considered uncontaminated by re-test effects. The authors reasoned that if list composition influences how items are actually encoded, those effects should remain even in the final recognition test. That is, initial-list composition should still matter (in the initial-test data, there were effects of list composition). Because it did not (nominally, at least), that could be evidence that the entire list-strength effect needs to be explained through processes that occur at test and not during the study phase, ruling out redistribution of encoding resources. Yonelinas et al. (1992) noted that Murdock had an argument against this logic but only cited a conference paper and did not enlighten us on that reasoning. Still, the particularities of Murnane and Shiffrin's methods, differing from ours, leave open either possibility: that redistribution occurs during the study phase, or that it does not.

In sum, redistribution of encoding processes has been considered, with very little empirical support, but also without overwhelming evidence to the contrary. Because our winning model was designed to implement redistribution, we were motivated to conduct post-hoc, unplanned and not pre-registered sequential-dependency analyses. After reporting the model fits, we present those findings and then conclude on a different theoretical interpretation of a model that mathematically mimics the favoured model, drawing only upon processes at test that adapt to list composition.

Summary of experiments. We present two experiments, each with three groups of participants differing in specific ways (detailed below). Each participant studied lists of 32 words, where lists were pure or mixed strength, in a 2×2 design.

First, we note that in previous manipulations of stimulus duration, two factors were actually varied together: display time and total time available to study a word. In the second experiment of Caplan and Guitard (2024b), the short condition had 500 ms during which the word was displayed and could be processed visually, and 500 ms during which the participant could process the word and encode it. The long condition had 2000 ms of display time, again, comprising time that could be used for visual processing but also for other processing and encoding of the stimulus. In both experiments here, the short condition was always the same: the word was displayed for 500 ms, followed by no inter-stimulus interval. The long condition always provided a total of 2000 ms time for study. The timing of the six groups is visualized in Figure 2.

Experiment 1. In Experiment 1, aside from the replication condition (Group 1), we fixed the display time at 500 ms for all words. We allowed only the total study time to vary by adding blank screen time (Group 2) or a mask (Group 3) after words in the "strong" condition. This way we could test whether the general strength effect remained, compared to a replication group, and also whether the inverted list-strength effect changed when conditions were equated for display time.

Group 1 was a replication of the second experiment of Caplan and Guitard (2024b): weak items were presented for 500 ms with no inter-stimulus interval and strong items were presented for 2000 ms with no inter-stimulus interval.

Group 2 was identical except that the strong condition displayed the word for 500 ms followed by a 1500 ms blank inter-stimulus interval— totalling 2000 ms available to study each strong word.

Group 3 was identical to Group 2, except that the 1500-ms inter-stimulus interval was replaced with a 1500-ms visual mask, again providing a total of 2000 ms of study

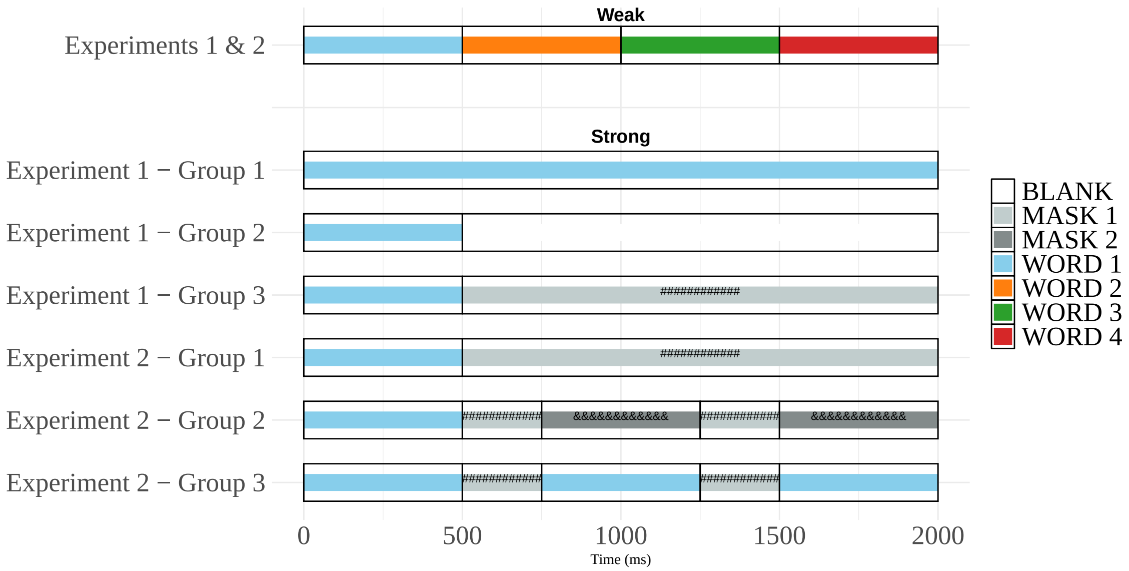

Figure 2

Depiction of the timelines of each condition across the six groups in Experiment 1 and 2. In all groups, weak items were presented for 500 ms with no inter-stimulus interval. In all groups, the total study time for each strong word was 2000 ms; only what was displayed on the screen over the course of those 2000 ms varied. Note: Colours are exclusively to illustrate the design and highlight commonalities across the conditions; all stimuli were presented the same and no colour was actually used.

time per “strong” word. The mask (see Figure 2), consisting of a series of hash symbols (############), was displayed immediately after the stimulus at the center of the computer screen, superimposed at the location where the stimulus had just been presented. Its purpose was to limit further processing of early visual features, a standard technique in memory research (see Enns and Di Lollo, 2000 for a review).

If the equated display time conditions of the experimental groups differed from the longer display time condition of the replication group, the interaction Group[3]×Item Strength[2] (general effect of strength) or Group[3]×List Type[2]×Item Strength[2] (influence on the list-strength effect) should be significant. Non-significant interactions (which, to foreshadow, were indeed favoured null effects) would suggest that total study time per word, rather than display time, is the driving factor.

Experiment 2. Experiment 2 tested the hypothesis that multiple onsets render disregarding ineffective, combined with the hypothesis that disregarding is necessary for the list-strength effect to invert. Again, the weak condition always displayed the word for 500 ms with no inter-stimulus interval.

Group 1 was now a replication of the masked group of Experiment 1, where strong items were displayed for 500 ms followed by a mask for 1500 ms.

Group 2 controlled for the multiple onsets in Group 3 (the main condition of interest). The 500 ms display of the word was followed by 250 ms of mask 1

(############), 500 ms of mask 2 (&&&&&&&&&&&&), 250 ms of mask 1 (############) and 500 ms of mask 2 (&&&&&&&&&&&&), ensuring there was still a total of 2000 ms study time allotted to each strong word.

Group 3 was a massed-repetition condition; strong items were displayed three times for 500 ms each, within a mask displayed for 250 ms between: the word was displayed for 500 ms followed by 250 ms of mask 1 (############), 500 ms of the word again, 250 ms of mask 1 (############) and 500 ms of the word again. Thus, Group 2 was like Group 3 except that instead of repeating the word, mask 1 (############) stood in for the word.

The primary hypothesis that onsets induce obligatory additional processing of the word, resulting in over-storage of shallow features, thus reducing the benefit of disregarding shallow features would be supported by a significant Group[3] × List Type[Pure, Mixed]×Item Duration[Short, Long], where List Type×Item Duration is greater (more inverted) for Group 1 than Group 3. The alternative hypothesis was that participants readily learn to ignore the superfluous additional stimulus onsets, with the driving factor being total time available for study.

Group 2 in Experiment 2 mainly served to follow up on a potential three-way interaction, to test whether the rapidly changing stimulus could explain away the interaction, but because the three-way interaction was not supported, this proved unnecessary.

Because the three-way interactions were all supported null effects, we report the methods and results of both experiments together.

Methods

Experiment 1 and Experiment 2 were pre-registered (pre-registration and data available at https://osf.io/swvpt and https://osf.io/h5u9g, respectively). They were designed to be close replications and extensions of Experiment 2 of Caplan and Guitard (2024b), which in turn, was replication/extension of Experiment 1 of Ratcliff et al. (1990). The procedures for these experiments were approved by a University of Alberta ethics review board.

Participants. Participants were recruited through an introductory undergraduate Psychology course pool from the University of Alberta, receiving partial course credit for participation. If we needed more data, we would have supplemented with participants recruited via Prolific (paid participants). For eligibility through the undergraduate course pool, participants must 1) have learned to speak English before the age of 6, 2) have normal or corrected to normal vision, and 3) use a desktop or laptop computer. A session lasted around 30–45 minutes.

To keep the sample uniform, participants were excluded if they took more than a ten-minute break or if their overall d' ($d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$; for these exclusions, 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 (reported next). 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 sizes and stopping rules. Our first target sample size was 50/group (Total $N = 150$), based on the precedent experiment (Experiment 2 of Caplan and Guitard, 2024b), which produced conclusive results with $N = 73$ (one group). We included an early stopping rule in Experiment 1 only: If the new conditions are much more difficult and performance is close to chance, the data will presumably be too noisy to draw meaningful conclusions. After running 10 subjects in each group, we checked if overall $d' > 0.5$ in more than half the participants in each group; this was satisfied so we proceeded with data-collection. After achieving the initial target sample size, we planned to run 30 participants ($N = 10$/group) until Bayes Factors were conclusive for the three-way interaction and individually for the two-way interaction, List Type×Item Strength for each group. We imposed an upper-limit on sample size to avoid endlessly collecting data, particularly in case the p value were to be significant while the Bayes Factor were inconclusive. That said, our attempt to make use of the counterbalancing feature of PsyToolkit was unsuccessful in numerous ways, leading to disparate sample sizes (although we always posted participation slots for all three groups in an experiment in each block of postings). The groups should be viewed as approximately, not strictly, counterbalanced. For Experiment 1, after excluding 6 participants with overall chance performance ($d' \leq 0$) and 2 for taking a ten-minute break or more, there were 89 participants in group 1; excluding 3 chance participants (one of whom took a ten-minute break), left 95 participants in Group 2; and excluding 5 chance and 4 ten-minute break participants (one participant had both exclusions) left 93 participants in Group 3. For Experiment 2, excluding no participants left 47 participants in Group 1; excluding 5 chance participants left 39 participants in Group 2; and excluding 5 chance participants left 54 participants in Group 3.

Materials. Following Caplan and Guitard (2024b) 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 previously seen in the experiment (lures). Words were drawn at random, anew for each participant.

Weak words were always presented for 500 ms with no inter-stimulus interval. The only difference across the six groups across the two experiments was the strong condition, as follows (illustrated in Figure 2):

- Experiment 1, Group 1, Replication Group: displayed for 2000 ms with no inter-stimulus interval.
- Experiment 1, Group 2, Fixed-Display Group: displayed for 500 ms followed by 1500 ms blank screen
- Experiment 1, Group 3, Fixed-Display Masked Group: displayed for 500 ms followed by 1500 ms mask (############).
- Experiment 2, Group 1, Replication of Fixed-Display Masked Group: Same as Experiment 1, Group 3.
- Experiment 2, Group 2, Onset Control Group: displayed for 500 ms followed by 250 ms of mask 1 (############), 500 ms of mask 2 (&&&&&&&&&&&&),

250 ms of mask 1 and 500 ms of mask 2.

- Experiment 2, Group 3, Massed Repetition Group: displayed for 500 ms followed by 250 ms of mask 1 (############), 500 ms of the word again, 250 ms of mask 1 (############) and 500 ms of the word again.

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 experiments were run online 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, , responding with 'Z' for old and 'M' for new. Responses faster than 100 ms were trapped and a 5-s message displayed the message "Too Fast!" to prevent participants speeding through.

Data analyses. Trials signalled "Too Fast!" (under 100 ms) were excluded. Participants were excluded entirely from any analysis for which they have missing data.

Our primary measure was d', with the log-linear correction favoured by Hautus (1995): +.5 observation always added to hits, misses, false alarms and correct rejections, computed for each participant in each List Type×Item Strength combination. The ratio-of-ratios was computed, following Ratcliff et al. (1990), as $[d'(\text{mixed strong})/d'(\text{mixed-weak})]/[d'(\text{pure-strong})/d'(\text{pure-weak})]$ and was log-transformed prior to analyses. Note that this d' calculations assumes equal variances (following Ratcliff and colleagues), which is not realistic. However, in the Appendix we confirm that the pattern does not substantially change when we use d_a, assuming unequal variances.

For each group, we evaluated the list-strength effect with 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. To test for the effects of the group treatments, we conducted a repeated-measures ANOVA with design Group[3] × Item Strength [Strong, Weak] × List Type[Mixed, Pure]. The three-way interaction tells us whether the inverted list-strength effect differs across the groups. The interaction Group×Item Strength tells us whether the various groups' variants of the strong condition influenced the magnitude of the effect of strength aside from list-composition.

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 (BF), a ratio of evidence for an effect present in the data versus the effect absent (similar to the null hypothesis in null-hypothesis-testing except; but here, a better way to think about the alternative is that if the alternative is supported, the data are better understood by assuming the effect is absent) 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). Importantly, for ease of reading, we always report Bayes Factors with the effect in the numerator and the null in the denominator (i.e., always BF_{10} and never BF_{01}). The reader can thus always view large BF values as indicating support for the effect (or for its inclusion in the ANOVA model) and very low values as favouring the null (or omission of the effect from the ANOVA model). 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. Effects that are not mentioned have both $p > 0.1$ and $BF_{\text{inclusion}} < 0.3$; near-significant and near-conclusive effects are described and carefully considered.

Results

Accuracy, expressed as d', hit rate and false-alarm rate, are plotted for each group in Experiment 1 in Figure 3 and for Experiment 2 in Figure 4.

List-strength effect

Our main pre-registered analyses concerned d'. For Experiment 1, the three-way interaction was not significant, $F(2, 274) = 1.65$, $MSE = 0.065$, $p = 0.19$, $\eta_p^2 = 0.012$, $BF_{\text{inclusion}} = 0.002$. The two-way interaction, Group×Item Strength was also not significant, $F(1, 274) = 1.85$, $MSE = 0.092$, $p = 0.16$, $\eta_p^2 = 0.013$, $BF_{\text{inclusion}} = 0.018$. Item Strength was a significant main effect, $F(1, 274) = 188$, $MSE = 0.092$, $p < 0.0001$, $\eta_p^2 = 0.41$, $BF_{\text{inclusion}} > 1000$, as was List Type×Item Strength, confirming an overall inverted list-strength effect, $F(1, 274) = 52.98$, $MSE = 0.065$, $p < 0.0001$, $\eta_p^2 = 0.16$, $BF_{\text{inclusion}} > 1000$. All other effects were non-significant ($p > 0.1$) and favoured null effects ($BF_{\text{inclusion}} < 0.3$). Analyzing each group alone, the list-strength effect, List Type×Item Strength, was consistently significant (Group 1: $F(1, 88) = 9.45$, $MSE = 0.076$, $p = 0.003$, $\eta_p^2 = 0.097$, $BF_{\text{inclusion}} = 13.03$; Group 2: $F(1, 94) = 18.64$, $MSE = 0.045$, $p < 0.0001$, $\eta_p^2 = 0.17$, $BF_{\text{inclusion}} = 491$; Group 3: $F(1, 92) = 28.25$, $MSE = 0.075$, $p < 0.0001$, $\eta_p^2 = 0.23$, $BF_{\text{inclusion}} > 1000$).

For Experiment 2, the outcome was much the same. The three-way interaction was not significant, $F(2, 137) = 0.019$, $MSE = 0.076$, $p = 0.98$, $\eta_p^2 = 0.00027$, $BF_{\text{inclusion}} = 0.030$. The two-way interaction, Group×Item Strength was also not significant, although marginal and with a strictly inconclusive Bayes Factor, $F(2, 137) = 2.93$, $MSE = 0.11$, $p = 0.057$, $\eta_p^2 = 0.041$, $BF_{\text{inclusion}} = 0.42$, so there may have been a small-magnitude difference across groups in the magnitude of the strength manipulation. Item Strength was a significant main effect, ($F(1, 137) = 72.01$, $MSE = 0.11$, $p < 0.0001$, $\eta_p^2 = 0.35$, $BF_{\text{inclusion}} > 1000$, as was List Type×Item Strength, confirming an overall inverted list-strength effect, $F(1, 137) = 30.62$, $MSE = 0.076$, $p < 0.0001$, $\eta_p^2 = 0.18$, $BF_{\text{inclusion}} > 1000$. All other effects were non-significant ($p > 0.1$) and favoured null effects ($BF_{\text{inclusion}} < 0.3$).[1] Analyzing each group alone, the list-strength effect, List Type×Item Strength, was consistently significant (Group 1: $F(1, 46) = 8.40$, $MSE = 0.086$, $p = 0.006$, $\eta_p^2 = 0.15$, $BF_{\text{inclusion}} = 20.75$; Group 2: $F(1, 38) = 13.15$, $MSE = 0.053$, $p = 0.0008$, $\eta_p^2 = 0.26$, $BF_{\text{inclusion}} = 109$; Group 3: $F(1, 53) = 11.32$, $MSE = 0.084$, $p = 0.0014$, $\eta_p^2 = 0.18$, $BF_{\text{inclusion}} = 28.2$).

Not pre-registered, we characterized the form of the list-strength effect interaction with two uncorrected t tests for each group. Strong items had a greater d' in pure than

[1] The exception is the main effect of List, which had a very large $BF_{\text{inclusion}}$ but with a very non-significant p-value, which we understand can happen when a higher-order interaction is supported.

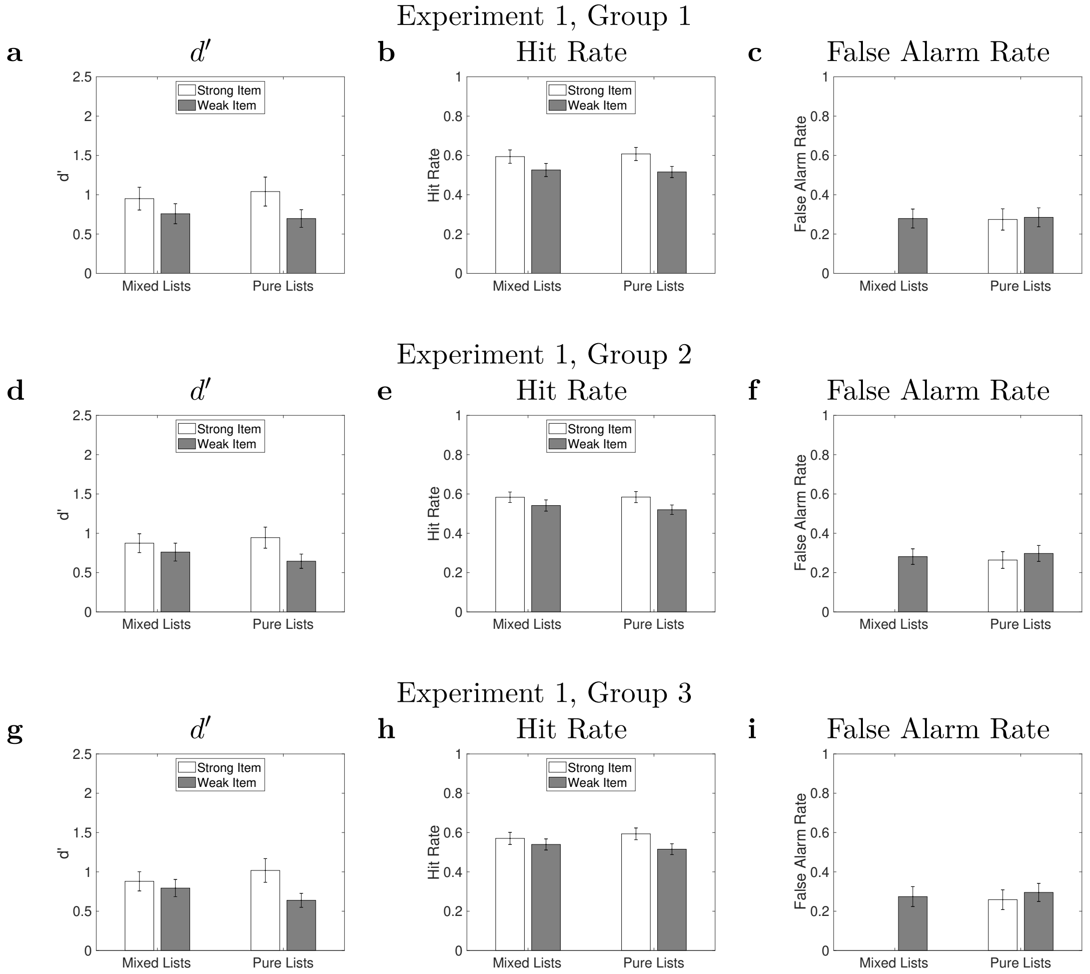

Figure 3

Accuracy data for Experiment 1, plotting sensitivity (d′), hit rate and false alarm rate (note that for mixed lists, lures are not tied to a particular item-strength). Each row is a different group. Error bars plot 95% confidence intervals based on standard error of the mean.

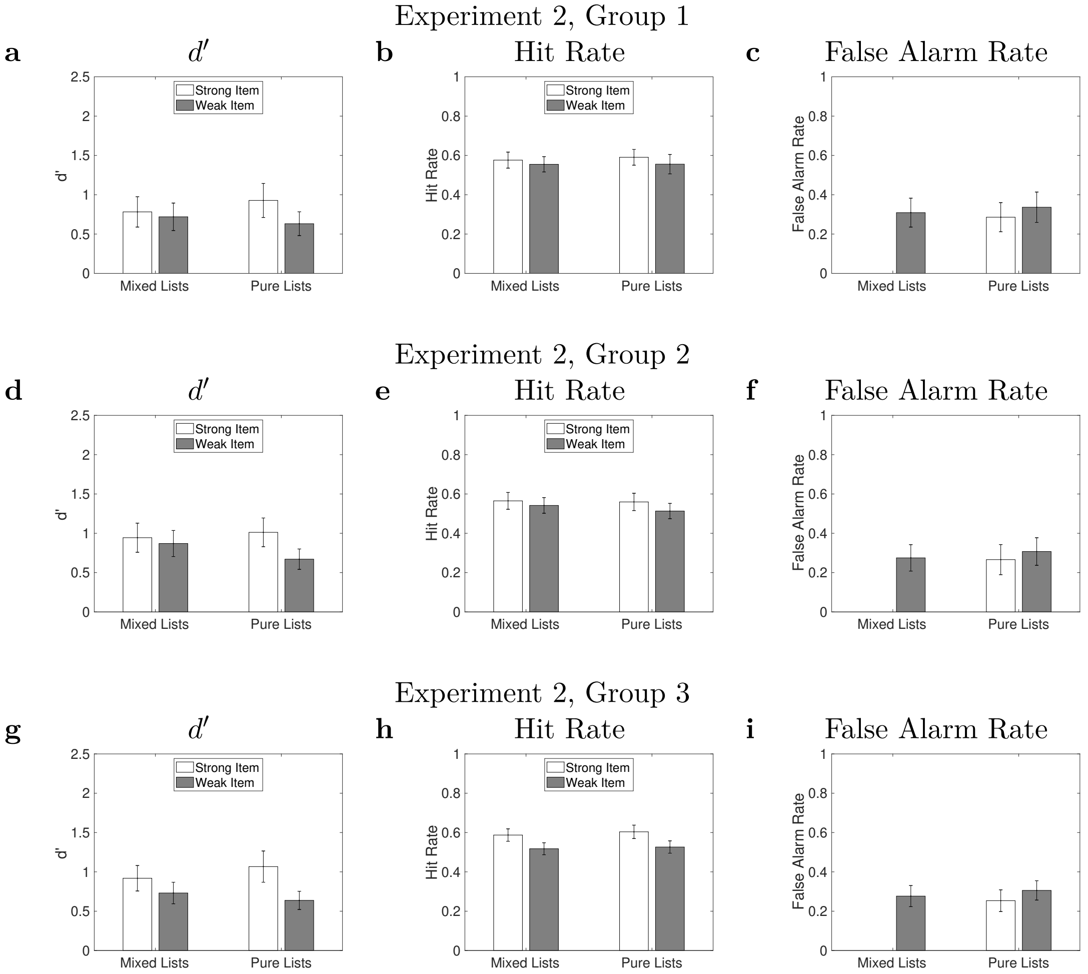

Figure 4

Accuracy data for Experiment 2, plotting sensitivity (d′), hit rate and false alarm rate (note that for mixed lists, lures are not tied to a particular item-strength). Each row is a different group. Error bars plot 95% confidence intervals based on standard error of the mean.

mixed lists (Experiment 1: $t(276) = -4.26$, $p < 0.0001$, $BF_{10} = 409$; Experiment 2: $t(139) = -3.44$, $p = 0.0008$, $BF_{10} = 25$) and weak items had lower d' in pure than mixed lists (Experiment 1:$t(276) = 5.54$, $p < 0.0001$, $BF_{10} > 1000$; Experiment 2: $t(139) = 4.03$, $p < 0.0001$, $BF_{10} = 268$).[2]

Summary of the most important results. In sum, the list-strength effect was inverted in all groups (List Type×Item Strength) but that effect, in turn, was not substantially different across the experimental groups (null three-way interaction), suggesting that equating display time and additional stimulus onsets did not sabotage the effect. To the contrary, the inversion of the list-strength effect appears robust to display time (holding total study time constant between experimental groups) and to additional stimulus onsets (holding total time constant), indicating more generality of the effect that we previously thought.

Exploratory analyses: hits, false alarms and response times

To dig deeper into the form of the inverted list-strength effects, we analyzed hit rate, false-alarm rate and response time individually.

Hit Rate. Breaking down the d' results, hit rate in Experiment 1, Item Strength, $F(1, 293) = 205.00$, $MSE = 0.005$, $p < 0.0001$, $\eta_p^2 = 0.41$, $BF_{\text{inclusion}} > 1000$, and Item Strength×List Type, $F(1, 293) = 19.03$, $MSE = 0.004$, $p < 0.0001$, $\eta_p^2 = 0.06$, $BF_{\text{inclusion}} = 251$, were significant, the latter confirming a reliable (inverted) list-strength effect. The three-way interaction was a favoured null, $F(2, 293) = 1.40$, $MSE = 0.004$, $p = 0.25$, $\eta_p^2 = 0.009$, $BF_{\text{inclusion}} = 0.023$, implying similar-magnitude list-strength effects in all groups. The interaction Group×Item Strength was, however, significant, $F(2, 293) = 4.75$, $MSE = 0.005$, $p = 0.009$, $\eta_p^2 = 0.031$, although the Bayes Factor favoured the null, $BF_{\text{inclusion}} = 0.053$, suggesting, along with the small effect size, that this reflects a significant but small-magnitude difference. All other effects were non-significant and favoured null effects ($p > 0.1$, $BF_{\text{inclusion}} < 0.3$) except the main effect of List Type, which was not significant but had $BF_{\text{inclusion}} = 45.9$ presumably because of its participation in the List Type×Item Strength interaction.

Experiment 2 differed somewhat. The effect of Item Strength was quite reliable, $F(1, 137) = 54.88$, $MSE = 0.0064$, $p = 0.0074$, $\eta_p^2 = 0.29$, $BF_{\text{inclusion}} > 1000$. But the interaction, List Type×Item Strength fell short of significance, $F(1, 137) = 3.59$, $MSE = 0.0043$, $p = 0.060$, $\eta_p^2 = 0.26$, $BF_{\text{inclusion}} = 0.16$, suggesting the list-strength effect in d' was at least not primarily driven by hit rates. The interaction, Group×Item Strength was significant, $F(2, 137) = 5.09$, $MSE = 0.0064$, $p = 0.0074$, $\eta_p^2 = 0.069$, $BF_{\text{inclusion}} = 2.33$. Although not clearly supported by the Bayes Factor, we conducted Holm-corrected post-hoc t tests but found that hit rate for weak items did not differ amongst the groups, nor did hit rate for strong items.

To characterize the form of the list-strength effects, we conducted uncorrected t tests. The hit rate was greater for strong items in pure than mixed lists, although with an inconclusive Bayes Factor in Experiment 1 and not even significant in Experiment 2 (Experiment 1: $t(295) = -2.50$, $p = 0.013$, $BF_{10} = 1.39$; Experiment 2: $t(139) = -1.82$,

[2] When broken down by group, the means had the same relationships, although not all individual tests were significant or supported by Bayes Factors.

$p = 0.071$, $BF_{10} = 0.47$). The hit rate was lower for weak items in pure than mixed lists although not significant in Experiment 2 (Experiment 1: $t(295) = 3.28$, $p = 0.0012$, $BF_{10} = 12$; Experiment 2: $t(139) = -0.89$, $p = 0.37$, $BF_{10} = 0.14$).

False Alarm Rate. For false alarms in Experiment 1, an ANOVA with design Group[3]×Item Strength[2] produced a significant main effect of Item Strength, $F(1.86, 509) = 25.24$, $MSE = 0.0029$, $p < 0.0001$, $\eta_p^2 = 0.084$, $BF_{\text{inclusion}} > 1000$ while the other effects were non-significant, favoured nulls. Similarly, for Experiment 2, only Item Strength was significant, $F(1.69, 231) = 22.06$, $MSE = 0.0045$, $p < 0.0001$, $\eta_p^2 = 0.14$, $BF_{\text{inclusion}} > 1000$, although the main effect of Group was in the inconclusive range, $BF_{\text{inclusion}} = 0.45$.

To characterize the form of the list-strength effect, we conducted uncorrected t tests. The false-alarm rate was significantly lower for pure-strong than mixed lists (Experiment 1: $t(276) = 3.25$, $p = 0.001$, $BF_{10} = 11$; Experiment 2: $t(139)$, $p = 0.002$, $BF_{10} = 9.66$) and was significantly greater for pure-weak than mixed lists (Experiment 1: $t(276) = -4.47$, $p < 0.0001$, $BF_{10} = 939$; Experiment 2: $t(139) = -4.50$, $BF_{10} = 995$).

Mirror effect. Important to the attentional subsetting account was that the inverted list-strength effect was expected to be accompanied by a very asymmetric mirror effect (Caplan & Guitard, 2024b). This was confirmed in each of the six groups: Strong items produced significantly ($p < 0.05$ and $BF_{10} > 3$) more hits in all but Experiment 2, Group 1 (consistent effect but weaker: $t(46) = 2.15$, $p = 0.037$, $BF_{10} = 1.289$) and fewer false alarms. The strong–weak difference was significantly greater for hits than for false alarms in Experiment 1, Groups 1 and 3. For Experiment 1, Group 2, the effect was nominally in the same direction but not conclusive ($t(94) = 2.27$, $p = 0.025$, $BF_{10} = 1.31$) and for Experiment 2, Groups 2 and 3, although nominally in the expected direction, neither significant nor supported by the Bayes Factor ($t(38) = 0.33$, $p = 0.75$, $BF_{10} = 0.18$ and $t(53) = 1.75$, $p = 0.086$, $BF_{10} = 0.62$, respectively). In Experiment 2, Group 2, the effect was nominally reversed, although not significant and a favoured null ($t(46) = -0.23$, $p = 0.82$, $BF_{10} = 0.16$), fairly symmetric mirror effect. This provides some weak support for the inverted list-strength effect being accompanied by a very asymmetric mirror effect and suggests there are boundary conditions to these features of the data co-occurring.

Model fits

The inverted list-strength effect replicated six times, speaking to its robustness. However, the fine-grained analyses of hit rate and false-alarm rate suggest that the nature of the inversion is different than what was predicted by both differentiation and attentional subsetting theories. Specifically, the hit rate for strong items is nearly equal and perhaps slightly greater in pure than in mixed lists. At the same time, the hit rate for weak items is nearly equal but slightly lower in pure than in mixed lists.

The differentiation mechanism in REM, for example, works as follows (Shiffrin & Steyvers, 1997). The old/new decision is based on a likelihood ratio, λ, which is the average of the likelihoods of all L (list length) stored traces when compared to the probe item. The likelihood ratio is greater when more features match, and lower when more features present in the trace mismatch the probe. The typical assumption is that a strong item has more features encoded in its trace, thus more features that could match features of a target probe, but also more features that could mismatch a lure probe. The likelihood ratio contribution for a probe item from its own stored trace is not dependent on list composition. Call this m (weak item) and M (strong item), where $m < M$ (strength effect). Likewise, a mismatch to a lure, or from a stored trace to a different-item probe will be smaller if the stored trace is strong, N than if the stored trace is weak, n, thus $N < n$. The final λ for a strong target item in a pure list will be the sum over the L local traces $= M + (L-1)N$. But in a mixed list, λ will be $M + (L/2-1)N + (L/2)n$. $\lambda_{\text{pure strong}} - \lambda_{\text{mixed strong}} = L/2(N-n) < 0$ because $n > N$. This is in the opposite direction than what is observed. For weak items, $\lambda_{\text{pure weak}} = m + (L-1)n$ and $\lambda_{\text{mixed weak}} = m + (L/2-1)n + (L/2)N$ for a difference of $L/2(n-N) > 0$ because $n > N$ but again, the data are in the opposite direction.

Turning to our previous attentional subsetting theory account, the inversion was due to participants disregarding the more confusing shallow features when tested on a list of pure strong items. This reduces the effective vector-length of the item, but in exchange, it reduces the false-alarm rate on pure-strong lists. However, without additional assumptions, the model does not produce a *greater* hit rate for strong items in pure than in mixed lists. Likewise, because the response criterion (threshold) is by default assumed to be set to half the mean matching strength, the hit rate for weak items is, if anything, greater in pure than in mixed lists because the threshold can be placed lower in the absence of strong items.

The fine structure of the inverted list-strength effect clearly demands a new theoretical account. To find such an account, we fit a set of variants of the attentional subsetting model to the data collapsed across all participants in both experiments apart from the massed-repetition group in Experiment 2. As we shall see, only one of the model-variants produced a good quantitative and qualitative fit. The implication is not that this model is correct, nor that attentional subsetting theory is a better account than REM. In fact, the mechanism that achieved a good fit could be easily implemented in REM. We present all variants we considered. This provides a more transparent view of the theoretical assumptions we felt were well motivated but which were inconsistent with the full pattern.

The basic attentional subsetting model

We start with the matched filter model proposed by Anderson (1970), which is the item-memory term Murdock (1982) used in Theory of Distributed Associative Memory (TODAM). We are not endorsing the matched filter model, which is overly simplistic and lacking mechanisms that current models of recognition have. Rather, we exploit the simplicity of the model to seek proofs of principle and develop an intuition for how the model explains the data. The mathematical simplicity of the model also means we can compute exact solutions and avoid more computationally costly Monte Carlo simulations.

Items are n-dimensional column vectors, which we denote in boldface, $\mathbf{f}_i$ and we use subscripts to denote different items. The features are assumed to be drawn at random from a Gaussian distribution, $\mathcal{N}(0, 1/\sqrt{n})$ which for large n, makes them nearly (but not perfectly) normalized, $||\mathbf{f}_i|| \simeq 1$. A memory is a simple vector sum, but we deviate from the matched filter model by assuming not all features are encoded, setting the remaining features to 0, as Murdock has typically done. Deviating from Murdock, we assume the subset of features is specific to the item, itself, potentially modulated by task conditions. We write this as a mask, $\mathbf{w}_{i,c}$, that multiplies the item vector features elementwise (denoted $\otimes$) before being added to the memory vector, $\mathbf{m}$. Note that the mask is indexed by i, meaning that it is item-specific, and by c, which can stand in for task conditions. Thus, a list of L items is the sum over those masked vectors:

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

At test, each recognition probe, $\mathbf{f}_x$, is also masked and then a matching strength s_x, is computed as the dot product of that masked probe vector with $\mathbf{m}$. If the conditions at test are the same as during study, we can make the simplification that the mask is the same, when the probe is a target. (We will later deviate from this to some degree in the case of mixed lists).

$$s_x = (\mathbf{w}_{x,c} \otimes \mathbf{f}_x) \cdot \mathbf{m}, \tag{2}$$

which is the sum of the dot product of the masked probe with each (masked) encoded item. The model responds "old" if $s_x \geq \theta$ and "new" otherwise, where θ is a response criterion we specify later.

Next we add assumptions about how different types of features are handled differently. In our model of stimulus duration (Caplan & Guitard, 2024b), we assume that when processed for a short duration, such as 500 ms, only shallow features are processed, such as the orthographic or phonological features of a word. With more study time, additional features are processed, but these will tend to be deeper features, such as related to the meaning or imagery of a word. We denote the shallow features "S" and the deep features "D." The important point is that we assume the total number of available "S" features is a lot smaller than the total number of available "D" features, $n_s \ll n_d$, where for simplicity, we assume $n = n_s + n_d$, thus no features are unused. We then assume that upon processing an item, the participant processes ν_s of those n_s items, and in the longer duration condition, an additional ν_d of the n_d features. Because the shallow-feature subspace is compact, there is a high chance the multiple items, including lure items, will have features in common, introducing similarity-based confusion. The deep features are sparsely subsetted because n_d is so large, making the likelihood of two items drawing attention to the same "D" features much more rare. ν_s and ν_d should be variable across items, but because of the linear properties of the model, we treat them as constant with little loss of generality.

As elaborated in Caplan (2023) and Caplan and Guitard (2024b), we can compute the means, μ, and variances, σ^2, of the matching strengths for targets and for lures by partitioning the terms into the contributions from the shallow and deep features, respectively and then use those to compute d':

$$d' = \frac{\mu_{\text{target}} - \mu_{\text{lure}}}{\sqrt{.5\left(\sigma^2_{\text{target}} + \sigma^2_{\text{lure}}\right)}}. \tag{3}$$

In all models we consider, $\mu_{\text{lure}} = 0$. $\mu_{\text{target}} = \nu_c/n$, where ν_c stands for the total number of features stored, for example, ν_s for a short-duration item. The lure variance is the sum over all studied items of the number of chance-overlap features between two different items, within the relevant subspace. Each of these multiplies $1/n^2$. For a short-duration item, the average number of common shallow features between two features is $\nu_s\nu_s/n_s$. The target variance has $L - 1$ such terms, reflecting cross-talk, chance-similarity between the probe and other list items. The term contributed by the probe item's encoded representation, itself, includes all attended (masked) features, multiplied by 2, thus for a short-duration item, this is $2\nu_s/n^2$. Note that the size of the feature subspace, n_s and n_d, does not affect the term contributed by the match to the item, itself. But the larger the feature subspace, the weaker cross-item interference will be, reducing the matching strength both of targets and of lures.

The previously published attentional subsetting models (Caplan, 2023; Caplan & Guitard, 2024a, 2024b) have assumed that following a pure-strong list, when possible, participants will disregard shallow features during the recognition phase, because they can rely on the high-resolution of the deeper features. This simply means that for the pure-strong condition, all terms involving shallow features are removed.

The hit rate is computed by integrating the presumed Gaussian distribution function with mean μ_{target} and variance σ^2_{target} from the response criterion or threshold, θ, up to infinity. The false-alarm rate is the integral of the distribution with μ_{lure} and variance σ^2_{lure} from θ to infinity. So finally, we need to determine θ. Because we have assumed (Caplan & Guitard, 2024b) the probe is processed similarly as a study item, and only a subset of probe features are attended (this deviates from other models like REM and TODAM), we assume the model (and participant) has direct access to the number of features just attended and can use this to tune their threshold. We have previously assumed the participant places the criterion at the optimal unbiased location, $\theta_{i,c} = \mu_{i,c}/2$ for a given item, i and condition, c. However, if one looks at the empirical hit and false-alarm rates (Figures 3 and 4), it looks like participants deviate considerably from the midpoint of the distributions. As we have previously noted (Caplan & Guitard, 2024b), the correct-rejection rate (1–false alarm rate) is much greater than the hit rate. Also, the hit rates are quite close to the balanced-criterion chance rate (0.5) whereas the false-alarm rates are considerably lower. This explains why our early attempts at fitting the model to the current data failed dramatically. We include a bias parameter, α, such that $\theta = \alpha\mu/2$ as a free parameter that (apart from the first variant) can vary across the three list types (pure strong, pure weak, and mixed).

The model variants

The model and model-fitting code (written for MATLAB/Octave) is available at https://osf.io/swvpt, along with the subject-level data files that were fit to seven data points (hit rate and false-alarm rate as a function of item and list composition). Models were searched by maximizing likelihood computed from root-mean-squared deviation (rmsd). Bayesian Information Criterion (BIC) was computed for model selection. By convention, $\Delta\text{BIC} > 2$ is considered some evidence for the lower-BIC model over the higher-BIC model. Six model variants were considered. These are described as follows. Each model variant has several free parameters.

Rationale and description of the sequential, direct-search fitting strategy. When the full parameter space (all combination of parameter values) is large, and/or computing the model fit for a given parameter set is time-consuming, modellers resort to algorithms that cleverly subsample the full parameter space such as SIMPLEX (Nelder & Mead, 1965). But when a model takes little time to compute and the number of parameters is low, it is possible to compute model-fit for all combinations of parameter values to some reasonable resolution. We figured out that not all parameters are relevant to all data points.

A naïvely applied subsampled search (e.g., with SIMPLEX) would be performing a lot of unnecessary computations of parameter sets that are equivalent to each other. So rather than use non-deterministic parameter optimization on the entire set of free parameters, we figured out how to break down the model-fits into sequential steps, each of which only required a search of a small parameter space, with at most, three free parameters fit in one step. This enabled us to do direct search, meaning that we computed the model for all combinations of 1, 2 or 3 parameters at a time, sampled with appropriate resolution over a meaningful range of parameter values. Log-likelihood, computed from root-mean-squared deviation (rmsd), was maximized. Parameters from earlier stages are fixed and applied without further modification to the later stages. A total of seven data points were fit: hit rates for strong and weak items, in mixed and pure lists, as well as one false-alarm rate for each of the three list types.

Step 1 of all model-fits. In all models, the pure-weak list was fit first. Target data points were the hit rate and false-alarm rate for pure-weak lists. The model had to fit those two values with three free parameters, ν_s, n_s and the bias for the pure-weak lists, α_s. One might think this is an overdetermined problem, using three parameters to fit two data points. However, the constraints of the model led to an unambiguous best fit. As already noted, previous attempts to fit the hit rate and false-alarm rate but fixing $\alpha_s = \mu_s/2$ failed. This stage of the fit, however, should be viewed as descriptive, not explanatory. Next we describe the subsequent fitting procedures for each variant in turn.

Variant 1: A single α. After step 1, fitting the pure-weak lists, step 2 fit the pure-strong condition the same as the pure-weak condition, fitting the hit rate and false-alarm rate for pure-strong lists, with three free parameters: ν_d, n_d and α_d. Note that the pure-strong condition does not depend on ν_s or n_s because we assume complete disregarding of shallow features. In this variant, we sought the single $\alpha_s = \alpha_d$ that would optimize the fit to both list types together. This was done by finding the single α value that maximized the maximum (within each condition) log-likelihood, summed together, over the searched ranges. In step 3, to fit the mixed lists, ν_s, ν_d, n_s and n_d as well as α were not allowed to vary further, so they become fixed parameters. The remaining three data points, the hit rate for strong and weak items on mixed lists, plus the false-alarm rate for mixed lists, were not searched further, but log-likelihood was simply computed directly after solving the model for those remaining points.

Variant 2: One α for each of the three list types. In this variant, step 1, fitting the pure-weak lists yielded a bias that we considered unique to pure-weak lists, now calling it α_s. Then in step 2, pure-strong lists were likewise optimized completely separately with their own α_d as well as ν_d and n_d. Finally, in step 3, ν_s, ν_d, n_s and n_d from fits to the pure lists were fixed while the mixed lists were fit with only one free parameter, α_{mixed}.

Variant 3: Excess deep features on weak items in mixed lists. We thought a short item embedded within a mixed list might draw more deep processing than the same item on a pure-weak list. This variant is like Variant 2, including the three α parameters. In step 3, as well as fitting α_{mixed}, it has one additional free parameter, ξ, denoting the average number of additional features, within the deep feature subspace, processed on short items within mixed lists.

Variant 4: Encoding-time sharing. Inspired by classic speculations about "rehearsal-borrowing" (Yonelinas et al., 1992), we thought weak items might be processed during some of the time allotted to a strong item. Thus, in step 3, the excess features, ξ, from Variant 3, are added to weak items but also now (with no further free parameters) subtracted from the strong items. Consider that such redistribution of encoding time might happen all the time. In a pure-strong list, one strong item might benefit from ξ extra features, at the expense of another strong item. But because those are both strong items, there is no net change in the number of features stored, and hence no net change in the mean matching strengths, etc. In a mixed list, short-duration items possess less encoding time to be stolen. Thus, the direction of diffusion of stolen study time may result in a net loss of study time for the strong items and a net gain for the weak items. Note that we provide an alternative interpretation of the mathematical formulation of this model in the next section (see "One final retrofit model").

Variant 5: Encoding-time sharing but no disregarding. We wondered if encoding-time sharing might remove the need for the disregarding assumption in explaining the data. This variant is the same as Variant 4 but the disregarding assumption is dropped, so that in step 2, for pure-strong lists, both shallow and deep features drive the recognition judgement. As always, pure-weak lists were fit first. But then those ν_s and n_s parameters were fixed and are used without further modification to fit the pure-strong data where they are added to the deep features. For example, $\mu_{\text{pure strong}} = (\nu_s + \nu_d)/n$ and cross-item interference occurs in both the shallow and deep feature subspaces, summating in the variance terms. The mixed lists were fit next in step 3, and as in Variant 4, only α_{mixed} and ξ were allowed to vary.

Variant 6: Incomplete processing of mixed probes. This variant drops the excess feature processing, time-sharing assumption. Instead, adding to Variant 2, in step 3 it assumes that in mixed lists, fewer deep features are processed than in pure-strong lists. This adds a parameter, $\nu_{dm} < \nu_d$. Because the mask multiplies all masked-out features by zero, ν_{dm} replaces ν_d where the probe acts in the mixed-list equations.

The following is a list of the searched parameter values were (where the parameter was relevant), along with brief notes about the cognitive and experimental meaning of each parameter.

ν_s**:** 1..32, integer steps

Number of shallow features of each word processed. This is presumably variable but we use only an average here. We have proposed (Caplan & Guitard, 2024b) that this number increases at short stimulus durations and asymptotes between 500 ms and 1000 ms. It may also be influenced by stimulus quality, stimulus-noise, etc.

n_s**:** 1..128, integer steps

Number of available shallow features. This should be relatively fixed for each participant but may vary for different types of stimuli, and could differ between different types of what we are calling "shallow" features, such as orthographic, acoustic and articulatory (Caplan, 2023; Caplan & Guitard, 2024a).

ν_d**:** 1..32, integer steps

Number of deep features of each word processed. We have proposed that this number will increase as stimulus duration increases (Caplan & Guitard, 2024b), particularly above 500 ms

n_d: 1..128, integer steps

Number of available deep features. This may also vary with the task-relevant deep features (such as imagery versus meaning-based features)

$\alpha_{\{s,d,\mathbf{mixed}\}}$: 0.1..3.0, steps of 0.1

Response biases for pure-weak, pure-strong and mixed list types, respectively. These could be manipulated in established ways such as monetary compensation/penalty for hits relative to false alarms.

ξ: 0..ν_d (where ν_d was fit from a previous step), steps of 0.1

Depending on the model variant: Excess deep features attended for weak stimuli in mixed lists, or number redistributed from strong to weak items

ν_{dm}: 0..($\nu_d - 0.1$) (where ν_d was fit from a previous step), steps of 0.1

Number of deep features attended on strong words when they are embedded in mixed lists. This is a hypothetical process and we do not know what would influence it, if present

Threshold method. Finally, each model variant was fit three times, with each of three rules for selecting θ_{mixed}: based on μ_{sm} only, based on μ_{dm} only or based on the average of both.

Model fitting results

First, threshold method made very little difference (ΔBIC< 1) for all model variants except for variant 4. For that variant, BIC was by far lowest when the threshold was based on the expected mean weak-item strength alone (BIC=-173.3, -156.4 and -154.2 for weak-item only, both, and strong-item only, respectively). We focus on that threshold method alone, although it should be noted that variant 4 produced the lowest BIC by far using any threshold method (ΔBIC to all other model variants > 10).

Table 1 summarizes the best fits of all six variants (model-output for all six model variants is plotted in Figure A2), and includes Ratcliff and colleagues' Ratio-of-Ratios (RoR) value, the ratio of strong:weak d' in pure lists divided by that ratio in mixed lists. Only one model, the one with by far the lowest BIC, produced all the target qualitative features found in the data: an inverted list-strength effect (RoR<1), hit rate of strong items greater in pure than mixed lists, hit rate of weak items lower in pure than mixed lists, and false-alarm rates in rank-order pure strong < mixed < pure weak. That model, variant 4, is plotted in Figure 5 above the data collapsed across the five groups of participants.

The functional feature subspaces fit to rather low dimensionality values, 12 and 28 features for shallow and deep feature spaces, respectively. The average number of attended/subsetted features also fit to small values, 7 and 4, respectively. These are not far off from the best-fit parameter values found for production-effect data (Caplan & Guitard,

Variant	BIC	ν_s	n_s	ν_d	n_d	α_s	α_d	α_{mixed}	ξ	n_{dm}	RoR	S	W	FAR
1	–70.4	1	19	4	28		1.4				3.14			
2	–133.3	7	12	4	28	1.8	1.4	2.0			1.09			√
3	–142.9	7	12	4	28	1.8	1.4	1.7	2.0		1.08		√	√
4	**–173.3**	7	12	4	28	1.8	1.4	1.7	1.3		**0.82**	√	√	√
5	–127.5	7	12	32	18	1.8	1.5	1.7	14.3		**0.79**		√	
6	–139.6	7	12	4	28	1.8	1.4	1.8		1.9	**0.89**	√		

Table 1

BIC values and parameter values for all the best fits of all the model variants considered. The winning BIC is bolded. Variants producing an inverted list-strength effect will have a ratio-of-ratios, RoR< 1; these RoRs are bolded. The final three columns report rank-order relationships present in the data. √ denotes that the condition was met for S=strong items' hit rates greater in pure than in mixed lists; W=weak items' hit rates lower in pure than in mixed lists; FAR=false alarm rates pure strong < mixed < pure-weak.

2024a). That said, it is plausible that participants completely lapse in attention during some proportion of study and test trials; for such "throwaway" trials, performance would be at chance. The functional subspaces and number of attended features are probably underestimated (for the proportion of trials with good participant engagement) for this reason. The deep subspace did, in fact, fit to a higher dimensionality than the shallow subspace, fitting with our notion that the sparseness versus denseness of the attended subset may be a driving force in the behavioural pattern and in specifying exactly how list-strength effects look. Incorporating the idea of encoding-time sharing did seem to be necessary, since the only model that hit all the empirical target characteristics had this process. That said, the ξ value was 1.3; only a bit over one feature was "stolen" on average from a strong item to a weak item. While not tiny, the modest value of ξ speaks to its realism. Finally, the assumption of disregarding shallow features when tested on pure-strong lists appears necessary, because Variant 5, dropping the disregarding assumption, fit far worse, both quantitatively and in terms of the rank-order target characteristics. Consistent with Caplan (2023), the best-fitting parameter set of Variant 5 did produce an inverted list-strength effect (RoR$= 0.89 < 1$) but the way in which this inversion came about was qualitatively different than the pattern of observed hit and false-alarm rates.

Tests of the redistribution account

As summarized in the introduction, for previous datasets, redistribution was not well supported. Either our datasets are different enough that redistribution *can* explain the net inverted list-strength effects we observed or it is not, in which case, redistribution might still be ruled out. We therefore conducted unplanned, not pre-registered sequential-dependency analyses. Because we are trying to explain an inverted list-strength effect, the kind of redistribution we are looking for is where a weak item steals time from a subsequent strong item. The prediction is that on a mixed list, a weak item will have a greater hit rate when followed by a strong item than a weak item, and a strong item will have a lower hit rate when preceded by a weak item than a strong item (Figure 6a). Mean (95% con-

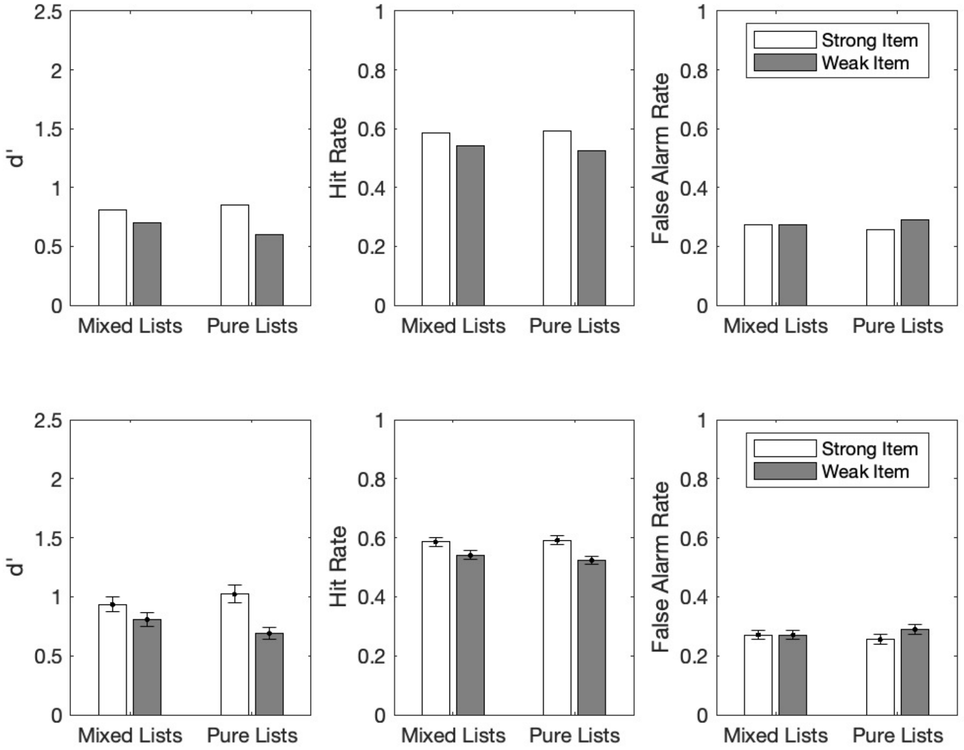

Figure 5

Model output in the top row for the best-fitting parameter set of model variant 4 (see text for details and Table 1 for parameter values). The target data, collapsed across all participants apart from those in the massed-repetition condition of Experiment 2, are plotted with 95% confidence intervals in the bottom row for comparison. Note that d' is plotted but was not fit to; hit rate and false-alarm rate were fit.

fidence interval based on standard error of the mean) hit rate, first for weak items, when followed by a strong or weak item, respectively, nominally (and significantly, where noted with an asterisk, $^*p < 0.05$, $^\dagger p < 0.1$) matched the prediction in all groups except Experiment 2, Group 2: Experiment 1, Group 1*: 0.5470 (0.0376) vs. 0.51181 (0.0359); Group 2†: 0.5522 (0.0299) vs. 0.53194 (0.0312); Group 3: 0.5472 (0.0292) vs. 0.53372 (0.0331); Experiment 2, Group 1: 0.5478 (0.0395) vs. 0.56640 (0.0436); Group 2: 0.5459 (0.0474) vs. 0.55335 (0.0402); Group 3*: 0.5348 (0.0362) vs. 0.50069 (0.0366). For strong items, hit rate when *preceded* by a weak versus strong item (Figure 6b) also nominally fit the expected pattern in all groups apart from Experiment 2, Group 3: Experiment 1, Group 1: 0.5915 (0.0381) vs. 0.60462 (0.0374); Group 2: 0.5725 (0.0275) vs. 0.58777 (0.0303); Group 3†: 0.5636 (0.0310) vs. 0.58205 (0.0343); Experiment 2, Group 1: 0.5625 (0.0418) vs. 0.58235 (0.0450); Group 2*: 0.5511 (0.0485) vs. 0.59912 (0.0497); Group 3: 0.6007 (0.0379) vs. 0.57892 (0.0365). At first glance, there does appear to be evidence of participants process-

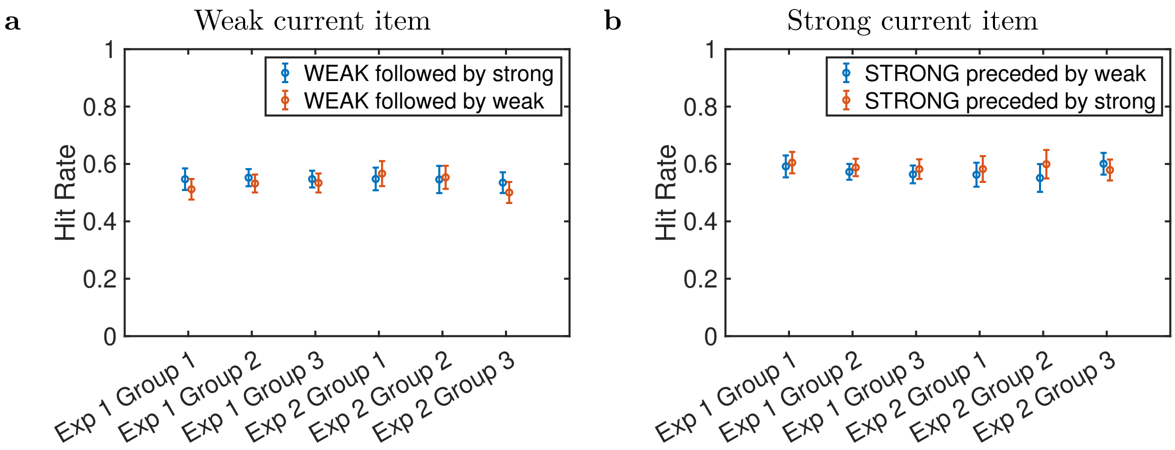

Figure 6

Tests of evidence of an encoding redistribution-like process, shifting encoding resources from a subsequent strong item to a preceding weak item. (a) Hit rate of weak *items conditional on whether the* subsequent *item was strong (blue) or weak (red). The hypothesized signature of redistribution is more hits when the subsequent item is strong than weak. (b) Hit rate of* strong *items conditional on whether the* preceding *item was weak (blue) or strong (red). The hypothesized signature of redistribution is more hits when the preceding item is strong than weak.*

ing a weak item more during a subsequent strong item. Although not statistically robust in all cases, the magnitudes of these effects are of a similar size as the list-composition effects.

We next asked if redistribution might play a causal role in the inversion of the list-strength effect. If so, then the greater the difference in weak followed by weak minus weak followed by strong, the *smaller* the ratio-of-ratios (i.e., more inverted) should be. Spearman correlations contradicted this; for Experiment 1, Groups 1–3 and Experiment 2, Groups 1–3, these correlations were all non-significant and some were even nominally negative: $\rho(87) = 0.11$, $p = 0.31$; $\rho(93) = 0.060$, $p = 0.56$; $\rho(91) = -0.11$, $p = 0.29$; $\rho(45) = -0.17$, $p = 0.24$; $\rho(37) = 0.10$, $p = 0.53$; and $\rho(52) = 0.068$, $p = 0.62$, respectively. Similarly, the greater the difference in strong preceded by strong minus strong preceded by weak, the smaller the ratio-of-ratios should be. These were also all non-significant and varied in sign: $\rho(87) = 0.0015$, $p = 0.99$; $\rho(93) = -0.13$, $p = 0.23$; $\rho(91) = 0.098$, $p = 0.35$; $\rho(45) = 0.077$, $p = 0.60$; $\rho(37) = -0.04$, $p = 0.80$; $\rho(52) = 0.089$, $p = 0.52$, respectively.

Reminiscent of similar analyses reported by Yonelinas et al. (1992), this leaves us with little support for the idea that redistribution plays the major role in producing inverted list-strength effects.

One final retrofit model

Model Variant 4 did not withstand additional tests. In addition, prior tests of redistribution effects were similarly not very supportive of their presence nor causal role in determining the form of list-strength effects. No other model variant we tried fit as well quantitatively, and Variant 4 was the only model variant that captured all the rank-order

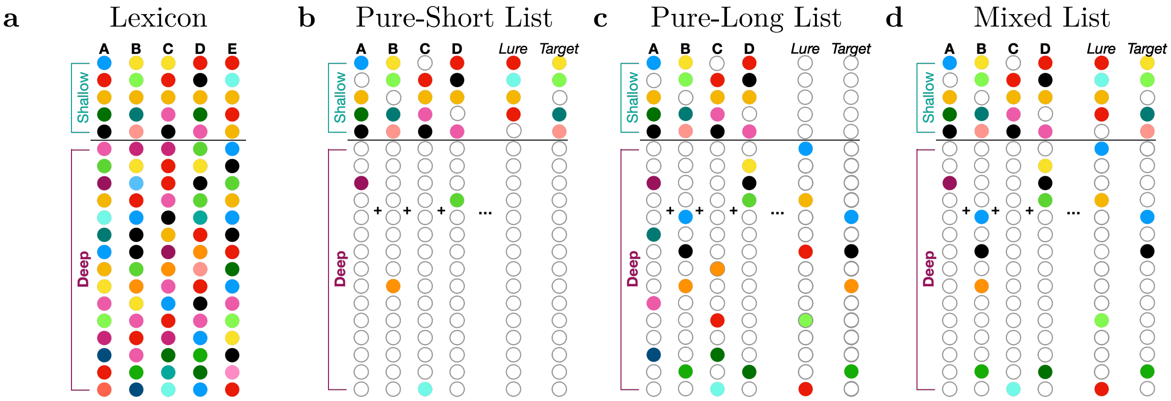

Figure 7

Illustration of our new attentional subsetting theory account of inverted list-strength effects in some manipulations of stimulus duration. (a) As in Figure 1, the full vectors are depicted with coloured circles standing feature values. (b) In a pure-short list, shallow features plus some deep features (here, $\xi = 1$ feature for illustration) are stored and at test. We still assume participants only process shallow features of the probe so those excess deep features exert no influence on the recognition judgement. Grey unfilled circles denote features that are not attended. For illustration purposes, the example lure is the same item in all cases and the target item is item B. (c) In a pure-long list, both shallow and deep features are stored as before, including disregarding of shallow features. (d) In a mixed list, short and long items are encoded as in the pure lists and both shallow and deep features are processed at test, but fewer (by ξ, 1 feature in this demonstration) than in tests of pure-long lists. In this mixed-list example, A and C are short items and B and D are long. Note that there are therefore 1 deep feature encoded for each short item, 4 deep features encoded for each long item as in the pure-long lists, but the probe includes 1 less (only 3) deep features.

characteristics of the hit rate and false-alarm rate data. The model is at a higher level; we did not implement the kinds of sequential effects we tested for in the data. This suggests that the model, itself, might be “correct” but our interpretation of the parameters might be wrong. We thought about how we might “retrofit” a theory that could mathematically mimic Variant 4 but re-conceptualize it. We propose the following idea (illustrated in (Figure 7; compare with Figure 1), where nothing changes with list composition during encoding, but everything depending on list composition takes place at test:

First, we assume that short items are always stored with ν_s features from the n_s-dimensional shallow feature-space, *plus* an additional ξ features from the n_d-dimensional deep feature-space (Figure 7b, study phase). Long items are always stored with ν_s shallow features and ν_d features, where $\nu_d > \xi (Figure\ 7c)$.

On a pure-weak list, participants only attend to shallow features (Figure 7b). Because unattended features have zeroes in the attentional mask, those zeroes multiply through and thus *disregard* deep features when participants are tested on a pure-weak list. We allow the bias, α_s, to fit freely. This mimics Variant 4 on pure-weak lists, where we had instead assumed no deep features were even encoded.

On a pure-strong lists, identical to Variant 4, participants disregard shallow features (Figure 7c); thus, only the more sparse, deep features weigh into the recognition decision, which produces a high hit rate but also a lower false-alarm rate because of the disregarded cross-item similarity from the shallow features. Again, the bias, α_d, can vary freely and this is mathematically identical to Variant 4 for pure-strong lists.

Finally, on mixed lists, we assume that the presence of both long and short items leads the participant to process some, but not all, deep features of each probe item (Figure 7d). Although we could introduce another parameter, we can also dual-purpose ξ such that each probe has ν_s shallow features processed and $\nu_d - \xi$ deep features processed. If $(\nu_d - \xi) > \xi$, there will still be an advantage for strong over weak items within a mixed list. That is because $\nu_d - \xi$ deep features will be available to match a strong encoded item, but weak items only have ξ deep features encoded in the first place, so only ξ of those can match a weak target probe. We again leave the bias as a free parameter. Again, this remains mathematically identical to Variant 4, but with a major reinterpretation of the ξ parameter and a different explanation of why there are ξ more deep features that can match to a weak-mixed item: not because those features are "stolen" from a strong item, but because they were encoded all along. The participant simply makes a bit more of an effort to check those small number of additional potentially matching deep features.

Discussion

The inverted list-strength effect replicated in all six experimental groups, representing five different (albeit slightly) task designs. This adds to the inversion reported in one manipulation of duration reported by Ratcliff et al. (1990), another by Ratcliff et al. (1994), a replication of that former finding in the first experiment of Caplan and Guitard (2024b) and an extension (here, replicated in Experiment 1, Group 1) in their second experiment. This total of ten findings of inverted list-strength effects with manipulations of study time per item makes the phenomenon hard to ignore and deserving of a thoughtful theoretical explanation.

In both experiments, the three-way interactions were non-significant, favoured null effects, finding no non-negligible influence of display time, masked display time or multiple onsets on the magnitude of the inverted list-strength effect. This challenges our two main hypotheses, as we expand upon below. But the other way to view the results is that we replicated the inverted list-strength effect six times, including four new variants of the study procedure. Until recently (Caplan & Guitard, 2024b), inverted list-strength effects in item recognition were acknowledged but not given much attention (except Ensor et al., 2021; Ratcliff et al., 1990; Shiffrin et al., 1990). One class of models, assuming approximately orthogonal representations, appears to have no way to produce such inversions (e.g., Chappell & Humphreys, 1994; Dennis & Humphreys, 2001). The other account of list-strength effects, differentiation in local-trace models, can produce inverted list-strength effects but have not been extensively tested on data that show clear net inverted list-strength effects (e.g., Shiffrin & Steyvers, 1997; Shiffrin et al., 1990). The scant prior reports of inverted list-strength effects had made it easy to overlook those results, perhaps even viewing them as symptomatic of false-positives of an underlying null list-strength effect. Meanwhile, attentional subsetting theory, while explaining near-null and upright list-strength effects, appeared to predict inverted list-strength effects, albeit with considerable parameter-sensitivity (Caplan, 2023; Caplan & Guitard, 2024b). It is thus quite non-trivial that the inverted list-strength effect appears so robustly across our six participant groups. These replications at face-value provide additional support for the attentional subsetting account, although as we summarize next, the fine structure of the data required new assumptions to explain and thus do not currently provide evidence to select between attentional subsetting and differentiation accounts. Although we propose an account embedded within a model, the clearer value of the data is that they provide boundary conditions (in fact, broad boundary conditions) and specific empirical targets for future theoretical accounts of how strength via duration is implemented in models.

Experimental parameters that produce inverted list-strength effects

As noted in the introduction, previous manipulations of study time per item have confounded stimulus duration, the visual display time, with the amount of time following stimulus-onset during which participants might be able to process the stimulus. Experiment 1 was designed to test the hypothesis that the longer-duration condition benefitted from prolonged immediate extraction of visual features or study time regardless of the stimulus continuing to be visible. Group 1 was a replication of the second experiment of Caplan and Guitard (2024b), which included that confound. The other groups displayed the word for 500 ms regardless of conditions. Group 2 had a blank screen for the remaining 1500 ms and replicated the pattern of Group 2, suggesting, contrary to our hypothesis, that immediate visual processing of the stimulus does not drive the strength effects. Group 3 replaced the blank screen with a backward mask in an effort to more severely terminate further visual processing and it still produced the same pattern of results. This rules out the hypothesis that the relevant shallow features are primarily related to superficial visual properties of the word stimuli.

Experiment 2 targeted a different question, testing the hypothesis that the onset of a word induces something like obligatory (re-)processing of the shallow features of the word. If so, we predicted that Group 3, with three massed repetitions, should exhibit a less inverted list-strength effect than Group 2 or Group 1. Again, the supported-null three-way interaction argues against this. We think the hypothesis might still apply to strengthening via spaced repetition, which could be tested in the future. That is because in Group 3, the mask 100% reliably preceded exact repetitions of the prior word. Participants may have been able to disregard those additional onsets, due to the regularity of the procedure, whereas that might not be possible in a typical spaced-repetition experiment. Our findings at least argue against obligatory re-processing of shallow features being entirely out of the participant's control, especially when they expect an immediate repetition.

Bottom line. The one consistent characteristic of the task design in all six groups was that the strong condition gave participants more time to think about the stimulus, irrespective of what was displayed on the screen. The centrality of the amount of allotted study time is compatible with the interpretation that is favoured by the model fits, as we discuss next.

Implications for theories

Theories that have explained near-null list-strength effects in recognition by assuming item representations are nearly orthogonal (e.g., Chappell & Humphreys, 1994; Dennis & Humphreys, 2001) do not anticipate inverted list-strength effects. When orthogonality is relaxed, incorporating some similarity across item representations, an upright list-strength effect is produced (e.g., Caplan, 2023), converging with the older models.

Differentiation theories like REM (e.g., Shiffrin & Steyvers, 1997) can produce inverted list-strength effects when context is downweighted and the old/new decision is based primarily on the encoded item-representation portions of the local traces. This is because strong items produce more evidence against a lure probe than weak items. Mixed lists have fewer strong traces than pure-strong lists, producing more false alarms, but mixed lists have more strong traces than pure-weak lists, producing fewer false alarms in that comparison. However, as shown earlier, without additional mechanisms, the predictions that the differentiation mechanism makes about hit rates are inconsistent with the observed pattern.

Attentional subsetting theory produced inverted list-strength effects in hand-tuned demonstration models (Caplan, 2023), as illustrated in Figure 1. This was due to participants disregarding shallow features when tested on pure-strong lists. Although the one successful model did appear to need to retain the disregarding assumption (compare model variant 4 with model variant 5), disregarding alone was insufficient to explain the full pattern of rank-ordering of hit rates and false-alarm rates.

All models we investigated missed at least one desired characteristic of the hit and false-alarm rate pattern (Table 1) apart from Variant 4, which included disregarding, variable response biases across the three list types and the encoding-time sharing assumption. Variant 4 was initially designed to model redistribution of encoding processes. However, follow-up analyses failed to support that mechanism, consistent with early tests that were largely inconclusive on this matter.

Heeding lessons from prior list-strength effect research suggesting inversions must be explained by processes at test rather than study, we finally proposed a model that mathematically mimics Variant 4 but can be located entirely at the test phase (Figure 7). First, we assume short items have a small number of encoding deep features but those are disregarded in tests of pure-short lists (at least within the experimental parameters considered here, comparing 500 ms to 2000 ms available time during study for each item). Pure-long lists are as before, with shallow features disregarded at test. In mixed lists, we assume that some undetermined metacognitive process leads participants to process fewer deep features of probes than they would on pure-long lists, partially acting against what should be an advantage for long over short items in mixed lists. But that additional processing of deep features offers a benefit to short-studied items that they would not receive in tests of pure-short lists.

This account is well within the spirit of attentional subsetting theory. However, recall that our implementation of attentional subsetting within the matched-filter model was only for tractability. In fact, this attentional subsetting account is quite compatible with models that assume vector representations of items, including those that have incorporated differentiation. Those that assume strict orthogonality would be hard to reconcile with the attentional subsetting account unless orthogonality could be softened, a possibility that could be interesting to explore. It may also be possible for a model, perhaps REM, to produce the full pattern of hits and false alarms by letting the response criterion vary across list types. In our simple vector-sum model, we also appeared to need a process for disregarding features; with freely varying response criteria but no disregarding, the model was insufficient. However, it could be that models such as REM have another way to accomplish a similar effect.

As another aside, item-specific attentional subsetting does arguably achieve some of the desirable effects that differentiation accomplishes. Is attentional subsetting in fact a mechanism for differentiation? On the one hand, greater subsets of features will lead to greater matching strengths, which would also be true in a local-trace model. However, the key to differentiation is that the mismatches to lure probes is, at the same time, supposed to shift downward, to lower matching strengths (producing a so-called "mirror effect," which we explore in Caplan and Guitard, 2024b). Attentional subsetting theory does not, itself, offer a means to do this. In fact, like other classic models, encoding more features will generally increase both hits and false alarms, shifting them in the same direction. The way Caplan and Guitard (2024b) produced mirror effects, to speak to some (but not all) experimental findings, was to introduce the heuristic we use here, whereby the criterion is adjusted based on current processing of the item, itself. Inspired by variable-threshold models (e.g., Cary and Reder, 2003; Glanzer and Adams, 1985; Starns et al., 2012; Stretch and Wixted, 1998), without differentiation, mirror effects can be produced by a combination of attentional subsetting and criteria tuned to each probe item, also dependent on knowledge of list composition. That said, disregarding does something similar to the criterion adjustment. With disregarding, the weaker items may become even weaker, functionally, because those shallow features are left out of the retrieval-strength calculation. So it may be fair to say that disregarding achieves some of what differentiation is meant to do, but it does so quite differently, and with respect to strong versus weak items (in some circumstances) but not with respect to the lure distribution. In all our models, $\mu_{\text{lure}} = 0$ and never shifts leftward, even when the variance changes.

Our favoured account is consistent with some known boundary conditions. For example, inverted list-strength effects have not been robustly reported when the range of stimulus durations are all quite short under 500 ms, or all quite long, 1000 ms or more (and that characterizes most manipulations of numbers of repetitions as well). Our account depends critically on the durations spanning the very short regime, under about 500 ms, where we presume very few deep features are processed, and the very long regime, above around 1000 ms, where we presume enough deep features are processed that recognition judgements can be effectively driven by those deep features alone (as elaborated in Caplan, 2023; Caplan and Guitard, 2024b). If "strong" as well as weak durations are both under 500 ms, the recognition judgement must rely on those shallow features, leading to upright or potentially close to null list-strength effects. But if "weak" as well as strong durations are both above about 1000 ms, even the short duration provides enough deep features, so shallow features may be safely disregarded during tests of all lists. The prediction in this regime is therefore a null or slightly upright list-strength effect. Finally, when strong items are read aloud versus weak items read silently, the "production effect," as we have argued before (Caplan, 2023; Caplan & Guitard, 2024a, 2024b), strength is increased through the storage of more shallow features. This makes it counterproductive to disregard shallow features. Due to the time it takes to produce words, durations are typically longer than 1000 ms as well, placing those experiments within the regime where near-null list-strength effects are expected due to deep features and upright list-strength effects are expected due to the shallow features that cannot be disregarded. Consistent with this, only upright or near-null list-strength effects have been reported when strength is manipulated via production (Caplan & Guitard, 2024a; Taikh & Bodner, 2016).

Clearly the repertoire of models tested here is not comprehensive. There may be other ways in which various models might explain inverted list-strength effects in manipulations of study duration.

The apparent necessity of letting the response bias vary across list types will also need further investigation. There might be a way to derive the response bias from characteristics of the memory or, as we have proposed (Caplan & Guitard, 2024b), of properties of the currently processed probe item.

Returning to the three classes of theories we considered, orthogonal representations, differentiation and attentional subsetting, the adapting of probe-processing depending on list composition seems quite compatible with all models. This insight does not select amongst the theories, although it might ultimately mean that the other proposed mechanisms to explain null and inverted list-strength effects are not necessary. The other important characteristics, variable response bias and disregarding of shallow features on pure-strong lists, could also be incorporated into any model that assumes a vector representation of items.

Limitations. We hope the reader appreciates (and enjoys) the fact that the theoretical account we land upon is extremely post-hoc. It certainly demands future testing and creative researchers may come up with even more theoretical accounts of inverted list-strength effects. We clearly do not wish to oversell this final account. Rather, the value of our data and model fits derives more from the extensive light we feel these shed upon current theories. First, the inverted list-strength effect is not hard to obtain and has broader boundary conditions than we previously knew. Second, one way to potentially explain away inverted list-strength effects, that was thoroughly considered in the early 1990s, that encoding time is not restricted to presentation time, was not able to explain away early list-strength effect data. Although our redistribution model produced excellent quantitative and qualitative fits of our data, it was challenged by additional sequential-dependency analyses, which resonates with those early unsuccessful attempts to look to redistribution as a way to trivialize the effect. Third, strictly orthogonal representations are inadequate to explain inverted list-strength effects. Fourth, differentiation models inherently produce inverted list-strength effects but modellers have mostly been aiming such models to roughly cancel out, to produce near-null effects as typifies most of the published recognition list-strength-effect findings. Attentional subsetting theory anticipated inverted list-strength effects but (fifth) both differentiation models and attentional subsetting models in existing forms missed the fine structure of the data.

Effects of study time are quite small

We pause to appreciate that the effects of stimulus duration or study time per item are remarkably small. We noted this in our replication of the 1000 ms versus 2000 ms manipulation of duration that Ratcliff et al. (1990) used, which led us to follow up with

500 ms versus 2000 ms in Caplan and Guitard (2024b) and in all experiments here. Despite the fact that the pattern is quite robust, replicating numerous times with large sample sizes, even quadrupling the time available for study of an item, the differences in hit rate and false-alarm rate are quite small in magnitude (Table 1). Thinking of study time or presentation rate as a manipulation of encoding strength might be missing the big picture. Using a response-deadline procedure and a manipulation of levels of processing, Mulligan and Hirshman (1995) found that d' asymptoted shortly after 1000 ms of study time, with a similar rate of accumulation for deeply and shallowly studied items, but the deeply studied items reaching a higher asymptote. Now note that most list-strength effect studies use study times of 1000 ms or more. This suggests that additional study time, per se, does not enhance encoding quality or strength. Rather, processing of the stimulus that unfolds within the first second or so determines the quality of the item in memory. This resonates with our previous finding of large item effects on hit-rate and false-alarm rate (Caplan & Guitard, 2024b), where the hit and false-alarm rate for a word were themselves not substantially correlated (consistent with Bainbridge, 2020; Bainbridge & Rissman, 2017; Bainbridge et al., 2013; Cortese et al., 2010, 2017; Cox et al., 2018; Isola et al., 2011; Lau et al., 2018). It may be more fruitful to follow item effects through recognition tasks or to consider how processing tasks such as phonological and orthographic processing, imagery, semantic judgements, etc., may interact with item characteristics.

Conclusion

Inverted list-strength effects are robust in old/new recognition of words, with manipulations of study time on the order of 500 ms versus 2000 ms per word. The critical factor appears to the amount of time allotted to study of each item, not the duration of the display, nor the number of stimulus onsets (e.g., with massed repetition). Current theories, however, are insufficient to explain why not only the false-alarm rate is lower in pure-strong than mixed lists but greater in pure-weak than mixed lists, but also why the hit rate for strong items is equal or greater in pure than mixed lists and lower for weak items in pure than mixed lists. The idea that processing or encoding is redistributed across items was tentatively supported but failed follow-up tests. Avoiding effects during the study phase, we suggest list composition influences the extent to which participants process items, as well as adjust their criterion (in terms of a bias at the whole-list level), and that they disregard shallow features when feasible. Although it calls for further testing, this stands as a simple, model-agnostic way to explain why inverted list-strength effects can be readily observed with manipulations of duration.

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).

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

Caplan, J. B., & Guitard, D. (2024b). Stimulus duration and recognition memory: An attentional subsetting account. *Journal of Memory and Language*, *139*(104556).

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*, 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, M. J., Khanna, M. M., & Hacker, S. (2010). Recognition memory for 2,578 monosyllabic words. *Memory*, *18*(6), 595–609.

Cortese, M. 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., 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.

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

Enns, J. T., & Di Lollo, V. (2000). What's new in visual masking? *Trends in Cognitive Sciences*, *4*(9), 345–352.

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 1,080 words. *Behavior Research Methods and Instrumentation*, *14*, 375–399.

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*(1), 8–20.

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.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2022). *Detection theory: A user's guide* (3rd ed.). Routledge.

Isola, P., Xiao, J., Torralba, A., & Oliva, A. (2011). What makes an image memorable? In *24th IEEE conference on computer vision and pattern recognition (cvpr)* (pp. 145–152). IEEE.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Society*, *90*(430), 773–795.

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.

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.

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

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

Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory & Cognition*, *19*(2), 119–130.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, *7*, 308–313.

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., & 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., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global models using ROC curves. *Psychological Review*, *99*(3), 518–522.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, *89*(1), 63–77.

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.

Starns, J. J., White, C. 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.

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology*, *70*(2), 186–194.

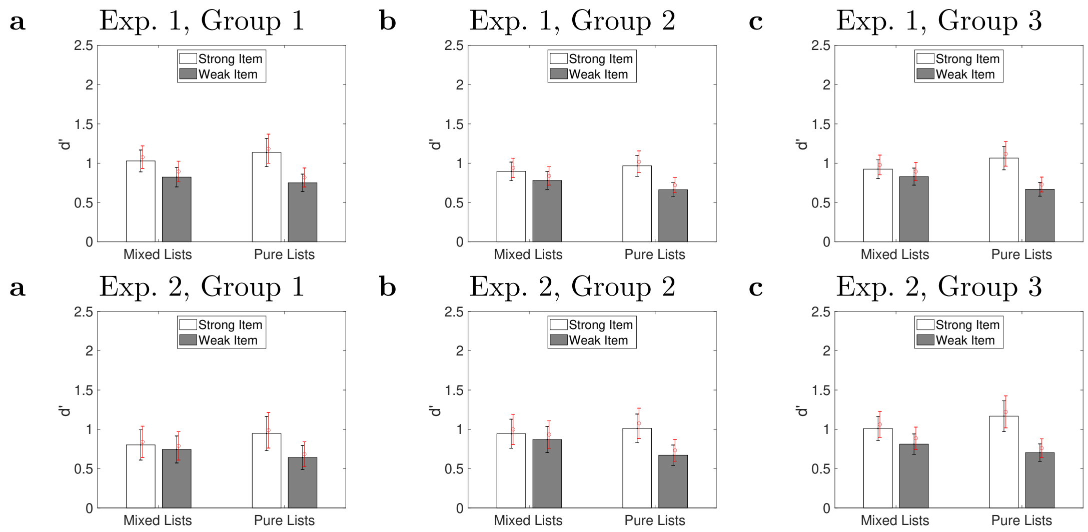

Figure A1

d' for all groups across both experiments. In red are plotted d_a assuming a ratio of target:lure variance of 1.25.

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1589–1625.

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.

Appendix

Check the unequal variance assumption for d'. We have focused mainly on hit and false-alarm rates. d' incorporates both these measures but we used a calculation that assumes equal variances. In Figure A1 we compare with d_a (Hautus et al., 2022) where we have arbitrarily assumed a ratio of 1.25 between the target and lure variances. The correction makes a small adjustment upward to the d' values but to a similar extent in all conditions. Importantly, the rank-ordering of the conditions does not change. This suggests that the assumption of equal variance is not misleading as to the general pattern of results for d'.

Output for all model variants. Figure A2 plots model output over the data for all model variants, assuming the threshold, θ, is based on μ_{sm} only.

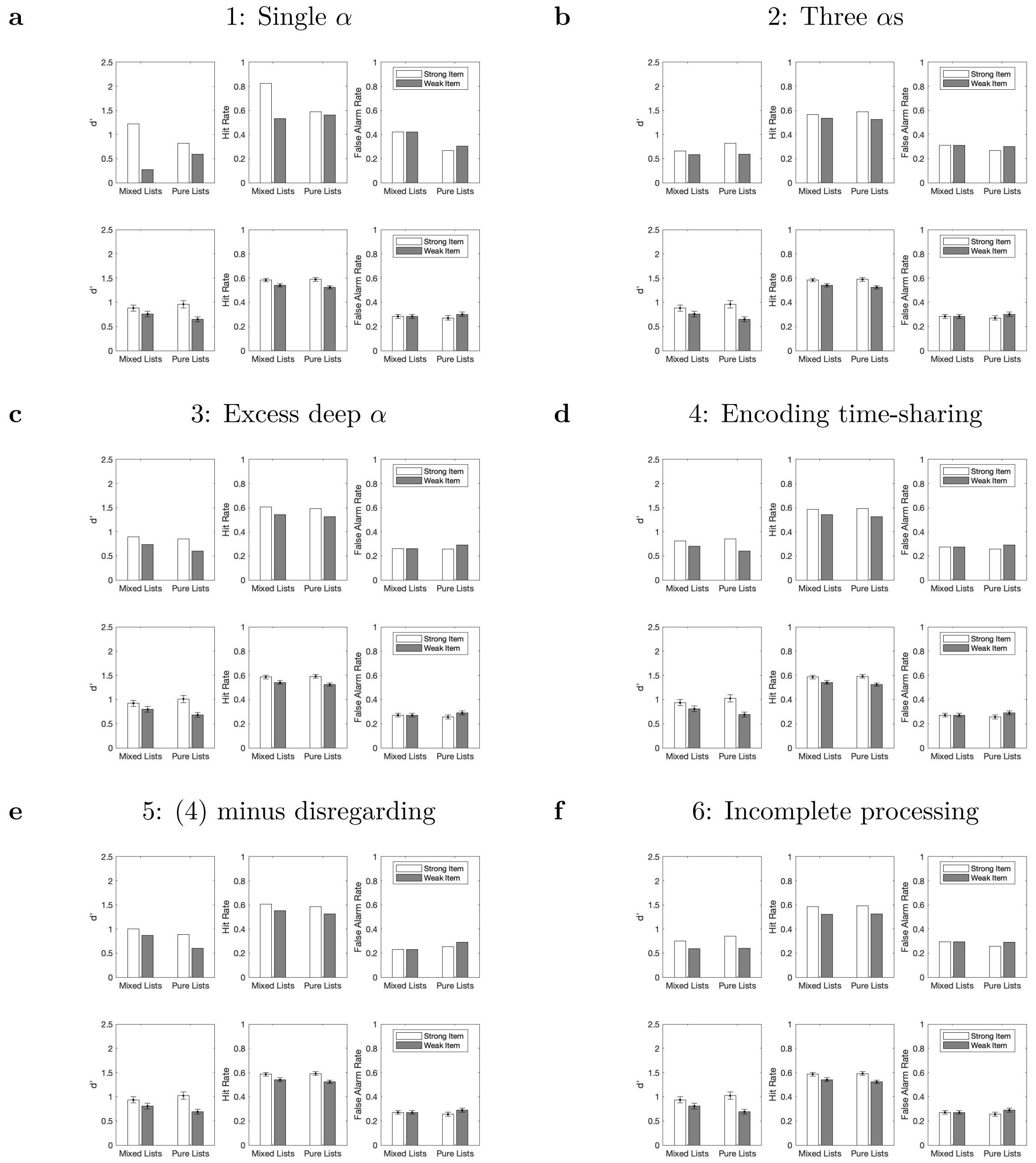

Figure A2

Model output for the winning (4) and non-winning model variants (compare with Figure 5) in each top row (see text for details and Table 1 for parameter values). They all assume θ is based on μ_{sm} only. The target data, collapsed across all participants apart from those in the massed-repetition condition of Experiment 2, are plotted with 95% confidence intervals in each bottom row to enable direct comparisons.