Spaced repetition and list-strength in recognition memory

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

Intuition and classic mathematical models suggest that "strong" list items should see a competitive recognition-memory advantage when mixed with "weak" list items within a single list. This holds for manipulations of reading aloud but not other manipulations. Study time per item (e.g., 500 vs. 2000 ms), can even invert the effect, with a larger strength effect between pure-strength lists than within mixed lists. This inversion remained when the longer study time comprised three massed repetitions. An attentional subsetting theory account explained the inverted list-strength effect by assuming weak items had primarily shallow features encoded, and participants disregarded those on pure-strong lists, when the better, deeper features were available. Here we asked whether spaced repetition is an exception, rendering disregarding ineffective, or like massed repetition, produces an inverted list-strength effect if presentation time is short. In three experiments, spaced repetition produced a null or upright list-strength effect. Under equated conditions, massed repetition and long duration produced inverted or null effects. Thus, despite being within an experimental regime that has robustly produced inverted list-strength effects, spaced repetition appears resilient. These boundary conditions constrain models, including attentional subsetting theory and Retrieving Effectively from Memory, and we discuss how various theories may need to adjust to explain the full range of list-strength data.

Keywords: List-strength effect, recognition memory, spaced repetition, selective attention, differentiation

In press at *Memory & Cognition*, December 19, 2025.
Corresponding author: Jeremy B. Caplan, jcaplan@ualberta.ca, Department of Psychology, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada. The authors thank Aditi Simran Roy and Devon Yanitski for helpful feedback on the manuscript. Partly supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

Introduction

In episodic old/new recognition memory tasks, participants study a list of items, such as 32 words, and then decide if each subsequent probe item is a target (present on the list), responding “old,” or a lure (not on the list but nonetheless part of the same stimulus pool), responding “new.” Many experimental factors can increase memory overall, such as study time per item, levels of processing and spaced repetitions. These are conventionally referred to as influence item “strength,” although as we shall argue, the precise memory consequences of various manipulations of “strength” are quite particular. In a “list-strength” design, lists are composed entirely of strong items (pure-strong lists) or entirely of weak items (pure-weak lists) or of both, typically in equal proportions (mixed lists). By definition, strong items are remembered better than weak items, so performance on pure-strong lists should exceed that on pure-weak lists (although the magnitude of those effects may be far smaller than one expects). But it is also intuitive to many that this strength effect should be even more pronounced in tests of mixed lists. That is because in a strong list, a strong item faces competition from L items (where L is the list length) but from only $L/2-1$ strong items in a mixed list; the remaining $L/2$ items are weak and thus should present less of a challenge. For the same reason, weak items should suffer more within mixed lists than pure lists. Until 1990, mathematical models of recognition embodied this concept and produced pronounced list-strength effects. However, Ratcliff et al. (1990) reported strikingly null, or at least non-upright list-strength effects.[1] This was despite finding the expected upright list-strength effect in free recall. We follow Ratcliff et al. (1990), who quantified the list-strength effect in recognition with the ratio-of-ratios, the ratio of the strength effect in mixed lists to that in pure lists:

$$\mathrm{RoR} = \frac{d'(\text{Mixed-Strong})/d'(\text{Mixed-Weak})}{d'(\text{Pure-Strong})/d'(\text{Pure-Weak})}. \tag{1}$$

Importantly, in list-strength effect studies, strength is typically manipulated with *spaced repetitions* (e.g., Criss and Koop, 2015; Ensor et al., 2020; Murnane and Shiffrin, 1991a, 1991b; Ratcliff et al., 1990; and see our summaries of published recognition list-strength effects in Tables 1, A1 and A2), which is the reason we investigate spaced repetition here. However, with other manipulations of strength, deviations from null list-strength effects have been seen. Production (reading words aloud, compared to reading silently) produces an upright (RoR>1) list-strength effect (Bodner et al., 2014; MacLeod et al., 2010). Stimulus study duration, at least if the weak condition lasts around 500–1000 ms and the strong condition lasts about 2000 ms or more, produces an inverted (RoR<1) list-strength effect (Caplan & Guitard, 2024b, 2025; Ratcliff et al., 1990, 1994). The older published inverted list-strength effects have been largely viewed as false positives or presumed special-cases and have largely been unaddressed by models. However, the findings of near-null list-strength effects in recognition have had a profound influence on theory, motivating the development of major influential modern mathematical memory models, the most influential of which is Retrieving Effectively from Memory (REM; Shiffrin and Steyvers, 1997).

[1] Ratcliff and colleagues have used “positive” for the expected list strength effect and “negative” for the reverse effect. Because positive and negative are also used to refer to significance, to avoid this confusion we have opted to use “upright” for the expected list-strength effect and “inverted” for its opposite.

These models have been designed in large part with the goal of explaining why manipulations of item strength could fail to produce (upright) list-strength effects. We describe two such classes of models next, as well as our own recent class of model, attentional subsetting theory, which takes a more continuum view and brought us to the hypotheses about spaced repetition that we test here (although we will consider other theoretical interpretations of the findings).

Account 1: Orthogonal representations. Arguably the most natural answer to the approximately null list-strength effect is that item representations are orthogonal, or approximately so. Some models assume orthogonal representations and explain the presence of list-length effects through associations between items and list context rather than direct interference between item representations, themselves (e.g., Chappell & Humphreys, 1994; Dennis & Humphreys, 2001). Deviations from orthogonality can re-introduce the expected upright list-strength effect, as in one model-account of the production effect (Jamieson et al., 2016). However, this account has no obvious way, on its own, to produce an inverted list-strength effect, as have now been shown to be robust (Caplan & Guitard, 2024b, 2025; Ratcliff et al., 1990, 1994).

Account 2: Differentiation models. First proposed by Ratcliff et al. (1990) and Shiffrin et al. (1990), differentiation is a property of some models wherein a stronger trace produces more evidence in favour of a target item having been studied but also, more evidence of a mismatch between a lure item and the strong studied item. In REM (Shiffrin & Steyvers, 1997), each item is stored in a separate so-called "local" trace. If strength entails having more features of the item vector accurately stored, then a stronger trace provides better evidence that a target item was indeed studied. But REM computes a likelihood ratio for each trace that uses mismatching features as evidence *against* a probe item having been studied. Consequently, a stronger trace will also produce more non-matches, reducing the false-alarm rate to lures. Differentiation introduces a force toward inverted list-strength effects because having L strong traces in a pure-strong list produces greater evidence to rule out lures, offering an extra boost to strong items within pure lists. The opposite occurs for pure-weak lists. Thus, an upright list-strength effect due to strength-based competition between traces is typically largely offset by that source of an inverted list-strength effect. This account thus, in principle, has ways of producing null, upright and inverted list-strength effects. It has ways, in principle, to produce upright list-strength effects, for example, when context effects dominate (e.g., Wilson & Criss, 2017) or when instead of refining an existing memory trace, the model produces a new trace for each repetition (Ensor et al., 2021). However, the most widely used and validated differentiation model, REM, has a lot of complexity and it is not clear what principles determine how these forces toward upright versus inverted list-strength effects trade off. For inverted list-strength effects in particular, the mechanism that inverts the list-strength effect (without adding assumptions to the model) makes the wrong predictions about false-alarm rates, for example (Caplan & Guitard, 2025). We investigate this in simulations of REM below and in the Appendix.

In our reading of REM papers, we found that modellers have not come out confidently on the parameter-dependence of the form of the recognition list-strength effect. and have even used careful wording such as "null/negative list-strength effect," perhaps symptomatic of the uncharted flexibility of the model. This is the case even in the original Shiffrin and Steyvers (1997) article (they argue the terminology should be the "... *list strength noneffect,* or even *list strength reverse effect,*" page 152), where their Figure 5 data (from Ratcliff et al., 1994) and model both show a hint of an inverted effect. The text reads that the REM behaviour "... appear quite close to the observed pattern." (page 152). The plots are quite small but if one looks closely, both model and data exhibit an inverted list-strength effect. Likely due to the large breadth of the paper, the authors are concise and do not draw attention to the fact that the list-strength effect is in fact inverted, nor do they weigh in on whether or not the model should be predicting inverted or null effects.

We were curious whether REM always produces that inversion or is highly parameter-dependent. To this point, Ensor et al. (2021) conducted parameter-sensitivity checks of their implementation of REM and their results suggest that REM in fact robustly produces slightly inverted list-strength effects (RoR$\simeq$0.95), not null list-strength effects. This RoR is inverted but less so than numerous published inverted list-strength effects (Table 1). Next they incorporated the idea (from the original Shiffrin and Steyvers, 1997 paper) that with spaced repetition, not all repetitions would edit the existing trace but rather, encode a new trace. Ensor et al. (2021) argued and showed that producing additional traces can be beneficial to recognition performance, and could explain why spaced repetitions are often remembered better than massed repetitions. In this way, the model could produce RoR values above 1. But RoR values above 1.2 were nearly (though not entirely) out of reach within the parameter ranges they investigated. This leaves REM perhaps over-predicting (slightly) inverted list-strength effects and predicting slightly positive but near-null list-strength effects with spaced repetition. It is not obvious how REM could produce very large upright list-strength effects such as are found for very novel stimuli (Osth et al., 2014) and the production effect (Bodner et al., 2014; Caplan & Guitard, 2024a; Jamieson et al., 2016; MacLeod et al., 2010). Production (reading aloud versus reading silently) seems a hard case to argue for participants encoding multiple memory traces for a single word. If this were, in fact, the mechanism, the modelling outcome of Ensor et al. (2021) suggests the model may still limit how large that list-strength effect could become. This model-exploration inspired us to conduct our own parametric explorations of REM, which we present below.

Account 3: Attentional subsetting theory. Developed recently and applied to list-strength effects (Caplan, 2023; Caplan & Guitard, 2024a, 2024b, 2025), this framework assumes that of all the features in an item's representation (vector; Figure 1a), only a small handful of features are attended and thus encoded (Figure 1b–d). For a given item in similar circumstances, largely the same feature subset will be attended. For example, the word HONEY may evoke features like sweetness, stickiness and golden-yellow colour when studied. Upon a later exposure, such as a recognition test trial, these same features are likely to reiterate. The feature subset is also proposed to be largely item-specific. Thus, the word GASOLINE may evoke features such as expensive, liquid and energy, which are not only different feature values, but (in this example) entirely different features, themselves. This offers a potentially large amount of inter-item distinctiveness (which can approximate account 1, orthogonal representations). If feature-subsets are sparse, a small number of features that are subsetted from a very high-dimensional feature subspace, that offers such high discriminability while maintaining a limberness and computational simplicity of having only a few non-zero features to work with within a typical episodic memory task.

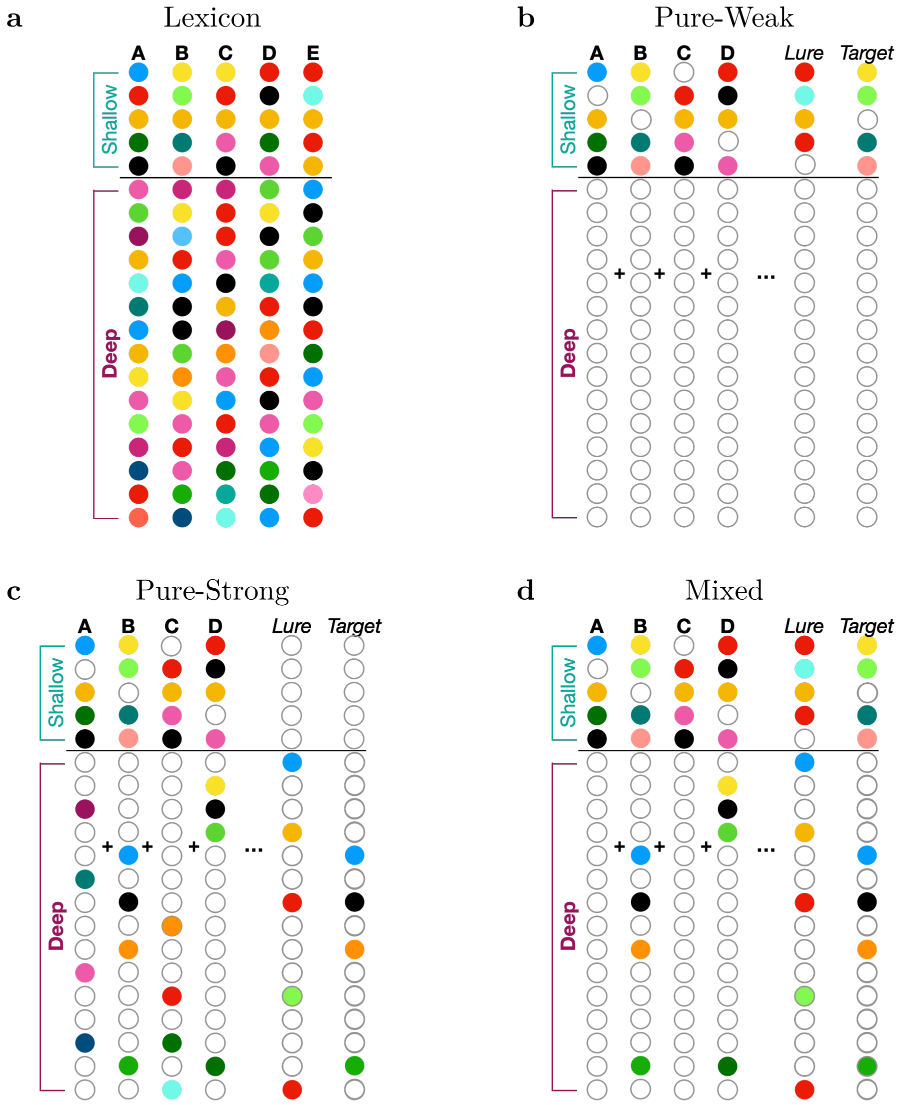


Figure 1

Attentional subsetting theory schematic. (a) Example items (A, B, etc.), n-dimensional column vectors. Each circle stands for a vector dimension and colour (arbitrarily chosen) denotes its value. (b) In a pure-weak list, only shallow features are stored and processed at test. Grey unfilled circles denote features that are not attended. (c) In a pure-strong list, shallow and deep features are stored. At test, because deeper features are more diagnostic of items due to their sparseness, we assume participants disregard *shallow features in their decision process, because the shallow features offer far less evidence of whether or not a word was studied. (d) In a mixed list, weak and strong items are encoded as in pure lists but disregarding is not possible, so shallow as well as deep features are evaluated at test. In this example, A and C are weak and B and D are strong. If disregarding is not possible (as will appear to be the case for spaced repetition and for massed repetition when spaced repetitions are also present), (c) would have shallow features processed just as in mixed lists (the study phase as in panel c and the test probes as in panel d), with the net effect being a near-null list-strength effect. For illustration purposes only, the example lure is the same in all cases and the target is always B. This figure is modified from Caplan and Guitard (2024b).*

This account readily explains upright list-strength effects similar to the traditional way, due to similarity-based confusion when the strong condition adds features that are *not* sparsely subsetted. This was our account of the production effect, where reading aloud presumably enhances encoding of phonological features (depicted as "shallow" in Figure 1), presumably subsetted from a relatively small phonological feature subspace (Caplan & Guitard, 2024a).

This account also readily explains near-null list-strength effects if the additional features subsetted in the stronger condition are extracted from a large feature subspace (such as those depicted as "deep" in (Figure 1) and are therefore sparse (Caplan, 2023). Sparseness leads to very little overlap of those features across items, introducing negligible levels of inter-item similarity-based confusion, so list-composition matters very little. An important nuance of this account is that it holds even if the features encoded in the weak condition are not sparse. Those non-sparse features (such as orthographic or phonological features) will contribute inter-item similarity-based confusion, but they will be equally present in the strong condition as in the weak condition.

Disregarding can explain inverted list-strength effects. With one additional element, this account can produce inverted list-strength effects as well as showing why they require special experimental conditions. When, for example, stimuli are presented for only a short amount of time, like 500 ms, the theory assumes that mainly shallow features have been attended, such as orthographic features (Figure 1b). Whatever those early-accessed features are, the assumption is that there are relatively few of them; in other words, the size of the total feature subspace is small. When the feature subspace is small, the subsets of features attended for each item can no longer be sparse. There is then considerable overlap of attended features across items, exposing the model to a large amount of similarity-based confusion, which produces an upright list-strength effect. However, an additional element of the theory is that participants calibrate their behaviour based on what features are available. When sufficient numbers of deep, distinctive features suffice (for example, imagery-related or semantic features, presumably comprising a very large total feature subspace), such as in tests of pure-strong lists, participants will *disregard* those confusing shallow features to achieve greater performance (Figure 1c, compared to b and d). The pure-strong condition gets an extra boost, increasing the strength effect between pure lists. The combination of largely shallow features attended for weak items, with the inclusion of disregarding in pure-strong lists, produces an inverted list-strength effect (Caplan, 2023; Caplan & Guitard, 2024b, 2025).

The attentional subsetting account is intrinsically a continuum account of list-strength effects. It places primary emphasis on the dimensionalities of feature-subspaces that are encoded and then used to draw recognition inferences. Attentional subsetting theory was the first theory to be explicitly used to suggest that inverted list-strength effects are not only latent within behavioural data but within certain regimes, could dominate, producing an observable net inverted list-strength effect (Caplan, 2023). Caplan and Guitard (2024b, 2025) validated and replicated inverted list-strength effects numerous times: 1) comparing display times of 1000 ms to 2000 ms (Caplan and Guitard, 2024b, with the caveat that the inverted list-strength effect was only robust for the subset of participants who passed screening criteria, where longer duration led to better memory[2]), replicating

[2] It was surprising to us that such a large manipulation of study time per item would not reliably produce

Experiment 1 of Ratcliff et al. (1990), 2) comparing 500 ms to 2000 ms (Caplan and Guitard, 2024b, and replicated in Caplan and Guitard, 2025), 3) comparing 500 ms to 500 ms display + 1500 ms blank inter-stimulus interval, 4) replacing the 1500 ms blank screen with a 1500 ms visual mask (replicated within Caplan and Guitard, 2025), 5) replacing the 1500 ms mask with two alternating masks (Caplan and Guitard, 2025 and 6) comparing 500 ms to three massed (immediate) repetitions of 500 ms separated by 250 ms of mask (Caplan & Guitard, 2025).

Caplan (2023) had speculated that for both spaced and massed repetitions, each time a word is displayed, the participant presumably needs to re-process the visual and orthographic features of the word simply to identify that it is indeed the same word. This could have the consequence of increasing encoding of shallow features, while cutting into the time available to process and store deeper features. This would mean that disregarding shallow features may be less effective for pure lists with massed repetition, leading us to predict a less-inverted (possibly null or even upright) list-strength effect. The fact that Caplan and Guitard (2025) still found an inverted list-strength effect with massed repetition contradicts that. However, consider that in that experimental session, participants only ever saw repeat presentations occur one after the other, never spaced. Perhaps participants caught on to the task design and learned to quickly identify massed repetitions as such, resulting in very little additional shallow-feature processing and consequently, also very little cost to deep-feature processing.

Either way, the idea also remains possible for spaced repetition. In a list-design with spaced repetitions, the participant never knows whether the upcoming stimulus onset is a repeat of a previous item or not. This may be more effective in compelling participants to re-process (and thus further encode) shallow features, at the expense of total study-time available to attend and encode deep features, rendering it ineffective (or at least less so) to disregard shallow features on tests of pure-strong lists. This was our first hypothesis regarding spaced repetition; namely, spaced repetitions oblige participants to process shallow features more than massed (or longer-duration) repetitions, rendering disregarding less effective, explaining the prevalence of published approximate-null list-strength effects with spaced repetition.

Published list-strength effects

To check what was initially our casual reading of previous data, we take a closer look at the pattern emerging from published recognition list-strength effects. Table 1 lists the most directly relevant studies that have reported list-strength effects for recognition tasks. We also look to Table A1 for studies that asked participants to study items in pairs. Ratcliff et al. (1990) had good reason to shift to ask participants to study words in pairs. They were concerned that participants might be shifting study processes or time from one item to another (also why they usually blocked strength within a list). However, theirs and subsequent studies showed that such effects are small or absent (see also Caplan and Guitard, 2025). One concern is that studying in pairs potentially changes the nature

a superiority of long- over short-duration words at the subject level. But our hit and false-alarm rate averages were not far from the original Ratcliff et al., 1990 values, so we presume they would have had comparable numbers of violations. Reassuringly, when we increased the manipulation in Experiment 2, 500 ms versus 2000 ms made a larger difference on average, and we found fewer reverse-strength participants.

of the task, for example, introducing explicitly associative processing which may function differently in models. We find those studies relevant but they should be interpreted with caution, which is why we summarize them in a separate table. We also report a few additional potentially relevant studies that are further afield in Table A2. Here we check these findings against the attentional subsetting theory framework but below we re-evaluate REM as well, in light of these published findings.

First, attentional subsetting theory predicts an upright list-strength effect (RoR>1) when strength acts on shallow features, such as the production effect. Table 1 aligns with that; the production effect studies have significantly upright list-strength effects, as does the experiment using novel stimuli (fractals) reported by Osth et al. (2014) in Table A2, for which we assume participants have insufficient expertise to produce sparse attentional subsets.

Second, near-null list-strength effects (RoR$\simeq$1) are expected, aside from the just-described conditions, whenever participants have longer than about 1 s study time per item. All studies meeting this criterion in Table 1 have non-significant RoRs. This includes manipulations of spaced repetition that use long study times (with some exceptions such as when study order or test order is deliberately confounded with strength; Ensor et al., 2020).

Third, inverted list-strength effects (RoR$<$1) are expected to require particular conditions (Caplan, 2023) but those should include cases where the weak condition is studied very briefly, such as 500 ms/item. Inverted list-strength effects are confirmed when the weak condition is around 500 ms (Caplan & Guitard, 2024b, 2025), including experiments that asked participants to study in pairs (Table A1). When the weak condition is a bit longer, such as 1.25 s split over two items studied together, the RoR values hover closer to 1. At 1 s for the weaker items, Ratcliff et al. (1990) found a significant inverted list-strength effect but Caplan and Guitard (2024b) found that inversion to be robust only after screening for participants who first showed strong>weak in both pure and mixed lists. Thus, around 1 s or even a bit shorter, the inversion becomes less robust.

Massed repetition, to our knowledge, has been investigated only twice. The first was by Ratcliff et al. (1990), with pairs at study for 1.25 s/pair, and with a mild manipulation (2P versus 4P), who found RoR<1 although not significantly so (Table A1). The second, Caplan and Guitard (2025), found a significantly and conclusively inverted list-strength effect, with 500 ms/pair. Thus far, massed repetition appears to function similarly to manipulations of duration, where inverted list-strength effects can arise when the presentation time per item is quite short.

Although spaced repetition has been extensively studied (see all tables), there is still some mystery. With presentation time per item on the long side, list-strength effects are typically non-significant (exceptions are in Ensor et al., 2020 when they manipulated study and test order). At 750 ms/item, Ratcliff et al. (1992) found a nominally upright list-strength effect but reported no statistical test of it. This could be an example of a non-inverted list-strength effect when the duration is short, although the 750 ms used in that experiment might not have been quite short enough. Especially without statistical tests, this study is hard to conclude from. Ratcliff and colleagues', as well as Ensor and colleagues' experiments where participants studied in pairs at $\sim$1.25 s/pair produced null or non-significance list-strength effects. This might be evidence that spaced repetition

Article	Exp.#	Strength	Duration(s)	RoR	Notes
Duration					
Ratcliff et al. (1990)	1	Dur	1 s & 2 s	0.88*	Strength blocked
Kahana et al. (2005)	1	Dur	2 s & 8 s	1.12	Not sig.
Caplan & Guitard (2024b)[a]	1	Dur	1 s & 2 s	*0.97*	Median
Caplan & Guitard (2024b)	2	Dur	0.5 s & 2 s	**0.71***	Median
Caplan & Guitard (2025)	1 group 1	Dur	0.5 s & 2 s	**0.80***	Median
Caplan & Guitard (2025)[b]	1 group 2	Dur	0.5 s & 2 s	**0.69***	Median
Caplan & Guitard (2025)[c]	1 group 3	Dur	0.5 s & 2 s	**0.66***	Median
Caplan & Guitard (2025)[c]	2 group 1	Dur	0.5 s & 2 s	**0.61***	Median
Caplan & Guitard (2025)[d]	2 group 2	Dur	0.5 s & 2 s	**0.69***	Median
Massed Repetition					
Caplan & Guitard (2025)	2 group 3	1P/3P	0.5 s & 2 s	**0.82***	Median
Spaced Repetition					
Ratcliff et al. (1992)	2	1P/5P	0.75 s	1.13	Mean. No stats
Diana & Reder (2005)	2	1P/11P	1.5 s	–	Non-sig. interaction
Kahana et al. (2005)	2	1P/4P	2 s	0.88	Not sig.
Ensor et al. (2020)	4	1P/2P	2 s	***1.63***	random study and test order
Ensor et al. (2020)	5	1P/2P	2 s	**1.14**/**0.942**	strong/weak tested first
Levels of Processing					
Kiliç & Öztekin (2014)	4	LoP	SP	–	
Production (read aloud versus silently)					
MacLeod et al. (2010)[e]	1&2	Prod	SP/2.5 s	1.53	No stats.
Bodner et al. (2014)		Prod	2.5 s	1.18	No stats.

Table 1

Summary of the characteristics and list-strength effects (ratio-of-ratios, RoR) for the most directly relevant published studies. Where available, we report authors' RoR and statistical test. We give priority to median RoR across participants, then mean RoR, then RoR calculated from the aggregate d' values (computed ourselves when necessary). *significant *($p < 0.05$)*. Conclusive Bayes Factor: RoR is set in* ***boldface****. When reliably inverted, RoR is underlined. Supported nulls are* italicized. *Exp.#: which experiment (annotated with a group condition where relevant). Strength: how item strength was manipulated (Dur = Duration, LoP = levels of processing, Prod = Production). For massed or spaced repetitions, 1P denotes once-presented items, etc. Duration(s) = time available to study each item. SP = self-paced. Notes: "strength blocked:" strength was blocked within a list. "No stats:" we could not find reported statistical tests of the LSE.* [a] *LSE was significant (and inverted) and supported by the Bayes Factor, after restricting to participants who showed better memory for strong than weak items: RoR=****0.89*****.* [b] *Strong=500 ms + 1500 ms blank.* [c] *Strong=500 ms + 1500 ms mask.* [d] *Strong=500 ms + alternating masks.* [e] *Mixed/Pure between subjects. Non-sig. interaction.* [f] *Computed from hit and false-alarm rates.*

eliminates the possibility of inverted list-strength effects. However, because participants studied in pairs, and because 0.625 s/item is still longer than the conditions that very robustly produce inverted list-strength effects (500 ms), these findings still leave open the possibility that the list-strength effect might invert with shorter presentation times and spaced repetition.

Interim conclusion and motivation for the experiments. If the weak condition has 1000 ms or more study time, there should be sufficient deep features to support recognition judgements even for weak items (Caplan & Guitard, 2024b) and in such conditions, disregarding of shallow features applies to all list conditions, removing the cause of the inversion of the list-strength effect (Caplan, 2023). Our second (alternative) hypothesis was that if presentations are short enough (e.g., lasting 500 ms), that would put the task into the regime for which selective disregarding on pure-strong lists is relatively effective, resulting in an inverted list-strength effect. If found, this would be the first report, to our knowledge, of a statistically robust inverted list-strength effect with spaced repetition.

Revisiting REM and the form of the list-strength effect

After some friendly nudging from anonymous reviewers we realized that from reading alone, we really did not know how REM would behave, and especially, which characteristics (e.g., list-strength effect directions and magnitudes) were intrinsic to the model and which were malleable. We were also specifically curious about the idea that a breakdown of trace-editing might offset an inverted list-strength effect. We therefore conducted parametric explorations of REM to get a better handle on what the model predicts, and under which conditions inverted, upright and near-null list-strength effects are actually expected. We report the full details in the Appendix and in Figures A1–A4.

In brief, the original 1997 choice of parameters appears to rather reliably produce near-null, and slightly or even substantially inverted list-strength effects, with no trace of upright list-strength effects. The failure of trace-editing, leading to multiple (here, two) traces per strong item in many cases, can, as promised, produce *slightly* upright list-strength effects but only under conditions in which editing results in only slightly more stored features than a single trace. Even then, RoR> 1 only occurs in conjunction with a second parameter-value adjustment that makes similarity of items more pronounced. There are several ways to achieve the latter; here we have demonstrated upright list-strength effects with trace-editing failure in conjunction with reduced diversity of feature values, reduced number of features per item, and increased frequency of feature copy-errors. When we look at this the other way, 100% trace-editing always (in our investigations, at least) trumped the other factor we introduced to increase noise or reduce performance of the model; with 100% trace-editing, the list-strength effect could be inverted or very close to null, but not upright. Trace-editing thus appears to protect the model against upright list-strength effects, which was an explicit goal in designing the original model (Shiffrin & Steyvers, 1997). What was unclear, but appears to be the case in our explorations, is that with 100% trace-editing, RoR> 1 in recognition is very hard, if not impossible, for REM to achieve.

Where REM stands with respect to published list-strength effects. Re-examining Tables 1, A1 and A2, at least within the parameter space we have explored, it appears that the designers of REM may have buried the lead: this model very robustly produces inverted list-strength effects. These are not just slightly inverted as in the simulations by Ensor et al. (2021), but can be very far below RoR=1, potentially sufficient to explain the accumulating published inverted list-strength effects.

The failure of trace-editing was promised to add an upright list-strength effect that could potentially cancel the inverted list-strength effect that is due to differentiation. This offsetting does appear to occur, and can be amplified when combined with adaptations of the model that increase item-similarity. However, in our hands, the model never produced upright list strength effects that are large enough to plausibly explain those found with the production effect or unfamiliar stimuli (fractals). To the extent that list-strength effects were slightly upright, that required a failure of trace-editing, which we do not see how to justify as an account of either production or unfamiliar stimuli. In other words, to produce sizeable upright list-strength effects, REM would appear to need to assume that when read aloud, participants form more than one memory trace for the item, and likewise when they view unfamiliar stimuli. There might be a way to reinterpret this effect, but even if so, the list-strength effect never seems to venture far above RoR=1.

REM's weakness seems to be that sizeable upright list-strength effects are beyond its reach, but on the other hand, it is a significant strength that inverted and near-null list-strength effects may be well accommodated.

With respect to spaced repetition (although the model is highly parameter-sensitive), if trace-editing succeeds at a high rate, REM would appear to predict inverted list-strength effects. If trace-editing fails at a high rate, if the model is in the same regime as an experiment with high trace-editing success (such as massed repetition or duration manipulations; compared Figure A1, panel c to panel a), the list-strength effect should probably remain inverted.

All that said, REM is a flexible model, so it seems desirable to give it the opportunity to fit empirical data quantitatively. We revisit REM in the General Discussion, with reference to quantitative fits of the model to our data, reported in the Appendix.

As noted earlier in the Introduction, other models produce upright or possibly null list-strength effects but REM (in many parameter regimes) and attentional subsetting theory (with particular conditions including the feasibility of disregarding) are the only current models that can produce inverted list-strength effects. This of course does not rule out undiscovered mechanisms or amendments to existing models. However, it does suggest that inverted list-strength effects are powerful in terms of challenging or constraining current models. The flexibility noted for REM is also paralleled in flexibility in attentional subsetting theory which, indeed, is not a model but a theory consisting of principles that can be integrated into numerous models, REM included. That said, attentional subsetting theory, in our previous implementations within the matched filter model, is more rigid with respect to inverted list-strength effects.

Goals and design of the three experiments

Our main objective was to test whether or not an inverted list-strength effect can be observed with spaced repetition, when study time per item is sufficiently brief. Specifically, all our tasks were designed to be within an experimental regime where inverted list-strength effects have been robust before, with 500 ms study time for the weak condition and a total of 2000 ms for the strong condition. Both attentional subsetting theory (Caplan & Guitard, 2024b, 2025) and REM (see Appendix) have no *a priori* reason to predict anything different about spaced versus massed repetitions (or duration) that would predict a different form of list-strength effect for spaced repetition. Thus, the fallback prediction is inverted list-strength effects in all cases. Any exception to this, particularly for spaced repetition, will demand a modification of both these two models to accommodate.

We manipulated strength via spaced repetition, where the weak (1P) condition had words displayed one time for 400 ms plus 100 ms blank inter-stimulus interval, and the strong (4P-spaced) condition had four such presentations scattered across the list. We included a matched massed-repetition condition where the strong (4P-massed) condition had the four repetitions (400 ms + 100 ms inter-stimulus interval) in immediate succession. The massed condition also served as a replication check of the inverted list-strength effect reported by (Caplan & Guitard, 2025), with a change in display time (400 ms versus 500 ms), inter-stimulus interval (100 ms blank screen versus 250 ms mask) and number of repetitions (4 versus 3). In Experiment 1, one group of participants only saw massed but not spaced repetitions and the other group saw spaced but not massed repetitions. Experiment 2 followed this with a fully within-subjects design, and where all mixed lists (including the practice list) contained all three item types: once-presented, massed-repeated and spaced-repeated words. Experiment 3 paired long study duration with spaced repetition within subjects.

One additional subtlety needs to be addressed. Researchers have noted that spaced repetition introduces a recency confound. Spaced 4P items are repeated, so in a design like ours, all 4P items were presented within the last quarter of the list. Some spaced-repetition designs have presented 1P items only intermixed with 4P items in that last list-quarter. The idea is that then, 1P and 4P items are equated for the recency of their *most recent* presentation. The cost, however, is that 1P items are never seen earlier in the list, adding a treatment-difference between 1P and 4P words. We preferred to intermix one quarter of the 1P words with 4P words during each quarter of the list, so that recency of any or all presentations cannot be used to determine weak/strong status. The cost of our chosen design is that, again, the recency of the *most recent* presentation of each 4P item is greater than that of 3/4 of the 1P items. However, our design allows us to check for a possible recency confound by conducting an additional set of analyses where we only include mixed-list weak items that were studied within the last quarter of the list. To foreshadow, the form of the list-strength effect for spaced repetition did not change, ruling out a recency confound.

Three new experiments

All three experiments include spaced repetitions, with short presentation times (400 ms plus 100 ms blank ISI), compared to once-presented items as the weak condition, to check if their list-strength effect can be inverted. To anticipate, all three experiments produced non-inverted list-strength effects: supported nulls in Experiments 1 and 3, and a significantly upright list-strength effect in Experiment 2.

Experiment 1 manipulates whether the strong condition uses massed repetitions or spaced repetitions, between-subjects. Experiment 2 follows up on a peculiarity of the results of Experiment 1, where it appeared that the inclusion of spaced repetitions reduced performance on weak (once-presented) items. Experiment 2 manipulates spaced and massed repetitions within-subjects as well as mixing them together with weak items within the mixed lists. Because this experiment unexpectedly neutralized the inverted list-strength effect for massed repetitions, we conducted Experiment 3 to check whether the effect of including spaced repetitions reduces all inverted list-strength effects or just for the case of massed repetition. Experiment 3 thus replaces the massed repetition condition of Experiment 2 with a long study duration condition (400 ms display time followed by 1600 ms blank screen) to ask whether prior findings of inverted list-strength effects would also be neutralized by the inclusion of spaced repetitions.

Experiment 1: Massed and spaced between subjects

The first experiment was a list-strength manipulation with short presentation times (500 ms total per presentation). Weak words had one such presentation and strong words had four repetitions, either massed for one group of participants or spaced for the other group. We tested two hypotheses:

H1) The inverted list-strength effect is a consequence of total study time, with no difference between whether repetitions are massed or spaced. Because similar task parameters produced an inverted list-strength effect with massed repetition, the prediction is a significant interaction List Type×Item Strength in the spaced repetition group. H1a) a strong version of H1: when both groups are analyzed together, List Type×Item Strength will be significant and the three-way interaction, adding group, will be a supported null.

H2) Spaced repetition functions differently than massed repetition or extended display duration, for example, because of multiple onsets following different preceding items. Prediction: When analyzed together, the three-way interaction will be significant, with a form such that the spaced repetition group produces a less inverted, possibly even null or upright, list-strength effect. H2a) Stronger version of H2: Spaced repetition, even with a single-presentation duration of 500 ms, will be functionally equivalent to prior list-strength effect studies that have typically used longer presentation durations with spaced repetitions. These tend to produce near-null list-strength effects. Prediction: supported null List Type×Item Strength interaction for the spaced repetition group.

Methods

Experiment 1 was pre-registered (pre-registration and data available at https://osf.io/pn2bh; comparison data from Caplan and Guitard, 2025, group 3 (massed repetition with three presentations), are posted along with that experiment's pre-registration at https://osf.io/h5u9g). These were designed to be close replications and extensions of experiments reported previously (Caplan & Guitard, 2024b, 2025), which in turn, were replication/extensions of Experiment 1 of Ratcliff et al. (1990). The procedures were approved by a University of Alberta ethics review board.

Participants. Participants were recruited through an introductory undergraduate Psychology course pool at the University of Alberta, in exchange for partial course credit for participation. Screening text required participants to 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 ($N = 1$ and 2, in the massed and spaced conditions, respectively) or if their overall $d' \leq 0$ ($d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$ ($N = 4$ and 0); if d' collapsed across list and item type (excluding practice) was below 0 (chance), that would suggest they misunderstood the task or the response mapping or were not able to perform the task at the very basic level. 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. Of the 188 total, $N = 91$ and $N = 90$ were the final analyzed samples for massed and spaced conditions, respectively.

Sample sizes and stopping rules. Our first target sample size was 50/group (Total $N = 100$), based on previous related experiments (Caplan & Guitard, 2024b, 2025). After achieving the initial target sample size, we planned to run 20 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. Our upper limit ($N = 100$/group) was not reached.

Materials. Following Caplan and Guitard (2024b, 2025) we used the 1000 words from the Toronto Word Pool (Friendly et al., 1982), displayed in 40 point Times font in the centre of the screen. Each list had 32 words for study, followed by 64 old/new recognition probes, half of which were in the latest study list (targets) and the other half of which were not previously seen in the experiment (lures). Words were drawn at random, anew for each participant.

Pure-strong lists were composed of all strong items and pure-weak lists, all weak items. Mixed lists were composed of half strong and half weak items, with strength order drawn at random. Each of three sets of four lists comprised a counterbalance set that 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]), following Ratcliff et al. (1990). Condition-order was random within each counterbalance-set.

During the study phase, each word was always presented for 400 ms with a 100-ms blank inter-stimulus interval. Strong words were presented four times each. In the massed condition, a strong word was presented four times in immediate succession (400 ms displayed, plus the 100 ms inter-stimulus interval). In the spaced condition, each presentation had to be in a different one of the four quarters of the list. Thus, average spacing = list length = 32. This means there are accidental massed repetitions when a strong word is at the end of one list-quarter and also at the very start of the next list-quarter. Because these are rare, we have ignored them (but the reader has the ability to investigate this in the data posted at OSF).

Mixed lists were also designed by dividing the list into quarters. For the massed condition, order of strong/weak was selected at random. For the spaced condition, each quarter included: a) a new 1/4 subset of the weak words (4 per quarter) plus b) all the strong words (16 per quarter) in a random order.

Procedure. The experiments were run online via PsyToolkit (Stoet, 2010, 2017). After confirming consent, participants did one 10-word mixed practice list with interleaved instructions, excluded from analyses. The test phase was self-paced. Each probe word (targets and an equal number of lures) was displayed as in the study phase, and participants were instructed to press the 'Z' key if they thought the word was old (on the list they just studied) and 'M' if they thought it was new. Responses faster than 100 ms were trapped and a 5-s message displayed the message "Too Fast!" to prevent participants speeding through. Participants were able to take short breaks between lists.

Data analyses. Trials with responses under 100 ms (signalled "Too Fast!") were excluded and participants were excluded entirely from any analysis for which they had 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 defined in Equatoin 1 and was log-transformed prior to analyses.

The list-strength effect was evaluated for each repetition group 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 whether the list-strength effect was, in turn, different between the repetition groups, we tested for a three-way interaction in a repeated-measures ANOVA with design Repetition Group[2] × Item Strength [Strong, Weak] × List Type[Mixed, Pure]. Repetition Group×Item Strength tells us whether massed versus spaced repetition influenced the magnitude of the effect of strength aside from list-composition.

Statistical tests are reported with both Classical and Bayesian approaches. Our significance $\alpha = 0.05$ but we are cautious with any p value near 0.05. Bayes Factors measure a ratio of evidence for an effect explaining variability in the data to the removal of the effect producing a better fit. We consider an effect favoured if $BF_{10} > 3/1$ and the null favoured $BF_{10} < 1/3$ (Kass & Raftery, 1995). For ease of reading, we always report Bayes Factors with the effect in the numerator and the null in the denominator so that large BF values indicate support for the effect and low values indicate support for the null. ANOVAs with multi-level factors have the Greenhouse-Geisser correction applied to correct for violations of sphericity. Post-hoc pairwise comparisons are Holm-corrected t tests. Any effects that are not mentioned have both $p > 0.1$ and BF < 0.3.

Results and discussion

List-strength effect

Our main pre-registered analyses concerned d' (Figure 2). The three-way interaction was significant, $F(1, 179) = 10.64$, $MSE = 0.055$, $p = 0.0013$, $\eta_p^2 = 0.056$, $BF_{\text{inclusion}} = 87$, suggesting the shape of list-strength effect differed between the two repetition groups. The two-way interaction, Repetition Group×Item Strength was also significant, $F(1, 179) = 34.34$, $MSE = 0.094$, $p < 0.0001$, $\eta_p^2 = 0.16$, $BF_{\text{inclusion}} > 1000$. Post-hoc Holm-corrected pairwise t tests were all significant ($p < 0.01$) aside from one, with the rank-ordering Spaced Strong > Massed Strong > Spaced Weak = Massed Weak. This replicates the common finding that spaced repetitions are remembered better than massed repetitions.

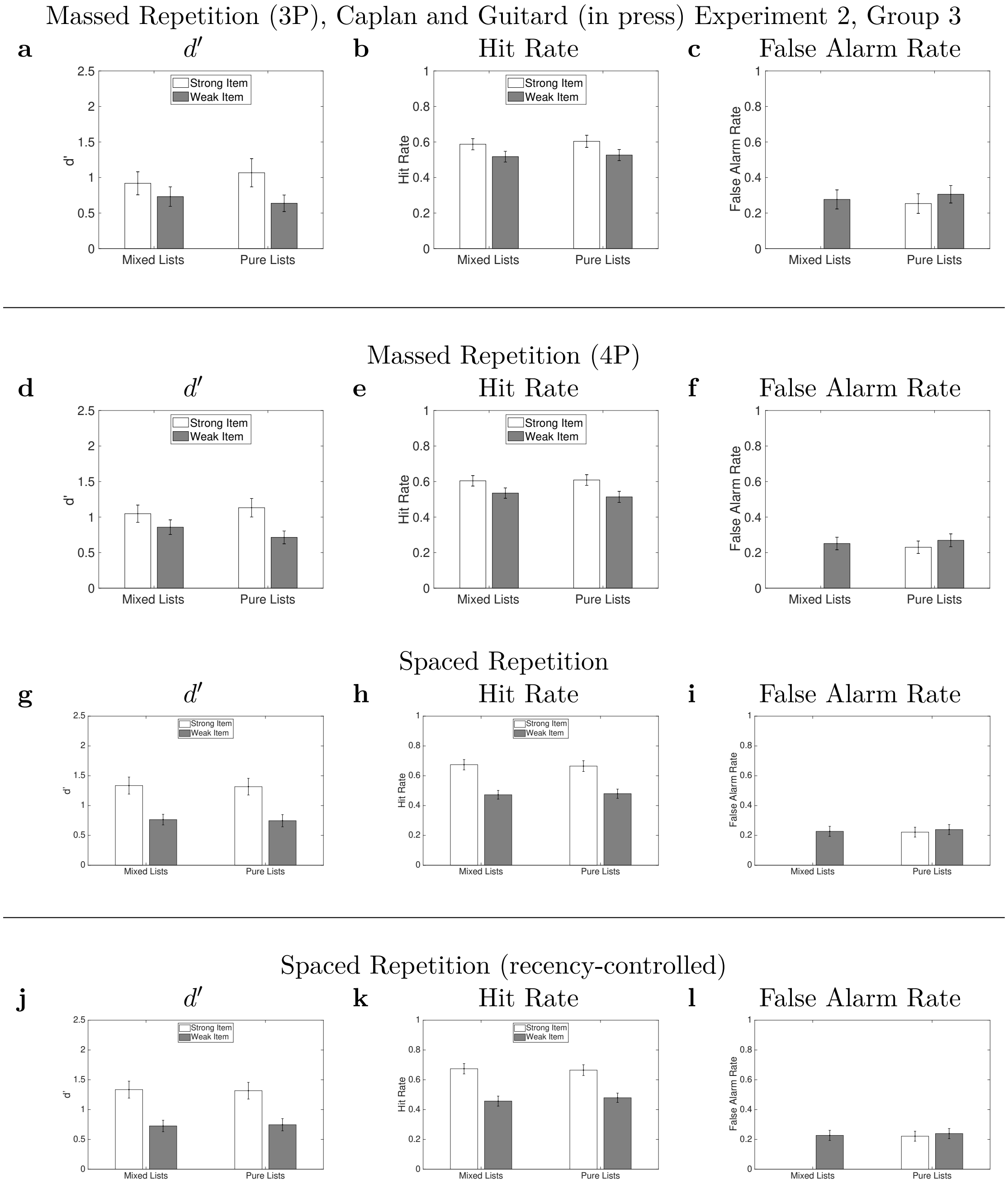
Item Strength was a significant main effect, $F(1, 179) = 367.93$, $MSE = 0.094$, $p < 0.0001$, $\eta_p^2 = 0.67$, $BF_{\text{inclusion}} > 1000$, as was List Type×Item Strength, confirming an overall inverted list-strength effect, $F(1, 179) = 10.67$, $MSE = 0.055$, $p = 0.0013$, $\eta_p^2 = 0.056$, $BF_{\text{inclusion}} = 97$. All other effects were non-significant ($p > 0.1$) although the Bayes Factors were large (but presumably because they are nested under the supported three-way interaction).

Analyzing each repetition group alone, the list-strength effect, List Type×Item Strength, was still significant for the Massed Repetition group, $F(1, 90) = 20.04$, $MSE = 0.059$, $p < 0.0001$, $\eta_p^2 = 0.18$, $BF_{\text{inclusion}} > 1000$. Holm-corrected post-hoc t tests confirmed that all pairwise comparisons were significant, thus Pure-Strong > Mixed-Strong > Mixed-Weak > Pure-Weak, exhibiting an inverted list-strength effect, and replicating the Massed condition of Caplan and Guitard (2025). This replication also satisfies our pre-registered pre-condition (HA; compare Figure 2d to a), allowing us to proceed as planned with the main hypothesis-tests. For the Spaced Repetition group, a null list-strength effect was found, $F(1, 89) = 0.000009$, $MSE = 0.051$, $p = 0.998$, $\eta_p^2 < 0.0001$, $BF_{\text{inclusion}} = 0.11$. The difference in form (inverted versus null) of the list-strength effects between the two repetition groups explains the three-way interaction.

In sum, H2 (even the stronger, H2a) was supported: equating the repetition groups as well as we could, massed repetition replicated an inverted list-strength effect whereas spaced repetition produced a non-inverted list-strength effect, with Bayes factor suggesting a supported null effect.

Checking for a recency confound. As pre-registered, we re-ran the main analyses with only the last-quarter mixed-weak items included, so that all analyzed items would have been presented at the same (most recent) ranges of recencies. As can be seen in Figure 2d (compare to c), the d' and hit rates of mixed-weak items change almost imperceptibly and consequently, the ANOVA on d' is nearly identical, with slightly less statistical robustness, as expected due to having less data overall. The three-way interaction was significant, $F(1, 179) = 13.33$, $MSE = 0.059$, $p = 0.00034$, $\eta_p^2 = 0.069$, $BF_{\text{inclusion}} = 129$. Repetition Group×Item Strength was also significant, $F(1, 179) = 32.79$, $MSE = 0.11$, $p < 0.0001$, $\eta_p^2 = 0.16$, $BF_{\text{inclusion}} > 1000$. Item Strength was a significant main effect, $F(1, 179) = 324.58$, $MSE = 0.11$, $p < 0.0001$, $\eta_p^2 = 0.65$, $BF_{\text{inclusion}} > 1000$, as was List Type×Item Strength, $F(1, 179) = 6.56$, $MSE = 0.06$, $p = 0.011$, $\eta_p^2 = 0.035$, $BF_{\text{inclusion}} = 495$. In sum, equating recency does not appear to materially change the results.

Summary of the most important results. The null list-strength effect for spaced repetition replicates the typical pattern reported for spaced repetition (Criss & Koop, 2015; Ensor et al., 2020; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990) and deviates from the pattern we have reported numerous times with short (∼500 ms) versus long (2000 ms total) study time (Caplan & Guitard, 2024b, 2025). That co-occurs with our replication and extension of an inverted list-strength effect with massed repetition (compare Figure 2d–f to a–c). At this stage, the results are more consistent with our second hypothesis, that spaced repetitions remove the benefit to be gained from disregarding shallow features— even when the strength manipulation is such that the weak condition presumably has mainly shallow features, and very few deep features stored. In fact, the stronger version of the second hypothesis, H2a, was supported, in that not only is there a



Massed Repetition (3P), Caplan and Guitard (in press) Experiment 2, Group 3

a d' **b** Hit Rate **c** False Alarm Rate

Massed Repetition (4P)

d d' **e** Hit Rate **f** False Alarm Rate

Spaced Repetition

g d' **h** Hit Rate **i** False Alarm Rate

Spaced Repetition (recency-controlled)

j d' **k** Hit Rate **l** False Alarm Rate

Figure 2

Experiment 1 accuracy data, plotting sensitivity (d′), hit rate and false alarm rate (note that for mixed lists, lures are not tied to a particular item-strength so there is only one false-alarm rate "for" all strengths within mixed lists). Each row is a different group of participants (the top row is previously published data, reprinted here for comparison). The bottom row plots the Spaced Repetition group but with a recency-check, retaining only weak items that were studied in the last quarter of the list. Error bars plot 95% confidence intervals based on standard error of the mean.

difference in the list-strength effect between repetition groups, the spaced group produced a null effect favoured by the Bayes Factor. Finally, the recency confound was ruled out as a concern.

Exploratory analysis: ratio-of-ratios to quantify the list-strength effect

Although not pre-registered for this experiment, we did pre-register this for Experiment 2 and report it here for comparison. We conducted a t test (independent-samples) between the log of the ratio of ratios, which quantifies the list-strength effect (Ratcliff et al., 1990), the effect of strength between items within in mixed lists versus between pure lists. They quantified the list-strength effect with the RoR (Equation 1). The log(RoR) is thus 0 if the list-strength effect is null (same effect of strength in mixed as in pure lists), positive for an upright list-strength effect and negative for an inverted list-strength effect. Our main interest, however, was whether the list-strength effects were the same or different for massed versus spaced repetition. Note that log() is undefined for values 0 and below, so 13 participants with RoR$\leq$ 0 were excluded from this test.[3] The result was a slightly favoured null effect, $t(166) = 1.55$, $p = 0.12$, $BF_{10} = 0.28$ (M (SEM)=–0.26 (0.10) and –0.06 (0.09) for massed and spaced, respectively). In retrospect, however, the log transform was meant to correct for the RoR being ratio, but because it is a ratio of ratios, a single log seems insufficient. We followed this with a Mann-Whitney U test, with the same participant exclusions, which produced a result that aligns more closely with the ANOVA $U = 4471$, $z = 2.99$, $p = 0.0028$, where the medians were 0.77 (inverted) and 1.08 (close to null) for massed and spaced, respectively.

Exploratory analyses: hits, false alarms, response times and bias

So far we have only looked at d', which combines hit and false-alarm rates. To dig deeper into the form of the inverted list-strength effects, we analyzed hit rate and false-alarm rate individually (Figure 2).

Hit Rate. Item Strength was a significant main effect, $F(1, 179) = 443.19$, $MSE = 0.0078$, $p < 0.0001$, $\eta_p^2 = 0.71$, $BF_{\text{inclusion}} > 1000$, as was Repetition Group×Item Strength, $F(1, 179) = 72.44$, $MSE = 0.078$, $p < 0.0001$, $\eta_p^2 = 0.29$, $BF_{\text{inclusion}} > 1000$. Holm-corrected pairwise t tests were all significant ($p < 0.05$), with the rank-order Spaced Strong > Massed Strong > Massed Weak > Spaced Weak (the latter was the weakest, with $p = 0.020$). We found it surprising and noteworthy that the weak (once-presented) items fared worse for the Spaced group than for the Massed group, with no trace of that interacting with pure versus mixed lists. We revisit this below, where we motivate Experiment 2 to settle the interpretational ambiguity that this raises. The main effect of Repetition Group was not significant, nor were any other effects apart from the three-way interaction, although just barely significant, with a very small effect size and with the Bayes Factor favouring the null, $F(1, 179) = 4.41$, $MSE = 0.0046$, $p = 0.037$, $\eta_p^2 = 0.024$, $BF_{\text{inclusion}} = 0.019$.

False Alarm Rate. For false alarms, an ANOVA with design Repetition Group[2]×Item Type[3] produced a significant main effect of Item Type, $F(1.6, 292) = 13.22$, $MSE = 0.0033$, $p < 0.0001$, $\eta_p^2 = 0.069$, $BF_{\text{inclusion}} > 1000$ while the other effects

[3] As noted above, weak>strong is seen for some participants, in part due to noise and in part presumably due to item effects overpowering the strength manipulation.

were non-significant ($p > 0.1$), although the Bayes factor for Repetition Group was inconclusive (0.44) and likewise for the interaction (0.36), leaning toward null effects. Post-hoc pairwise comparisons found significant ($p_{\text{holm}} < 0.05$) differences between all pairs of Item Types, with order Pure-Weak > Mixed > Pure-Strong.

Taken together, the three-way interaction that was robust for d' was fairly subtle for hit and false-alarm rates alone, suggesting the full effect is due to a combination of hit and false-alarm values.

Response times. Caplan and Guitard (2024b) found evidence for response times to longer-duration words to be longer than those for shorter-duration words, which is loosely consistent with the idea that longer duration leads to more "deep" (and sparsely encoded) features stored and thus available at time of test. In other words, if features are available, that take longer to access, participants may adopt a speed-accuracy tradeoff to take advantage of those when possible.

Figure 3 plots the means of the response times (median for each participant in each condition). Comparing pure lists only, for the Massed Repetition group, response times were longer for the strong (massed) than the weak (once-presented) trials in all cases aside from hits, which was a supported null (Hits: $t(90) = 1.29$, $p = 0.20$, $BF_{10} = 0.26$; Misses: $t(90) = 5.13$, $p < 0.0001$, $BF_{10} > 1000$; False alarms: $t(88) = 3.09$, $p = 0.0027$, $BF_{10} = 9.61$; Correct rejections: $t(90) = 2.73$, $p = 0.0077$, $BF_{10} = 3.72$).

For the Spaced Repetition group, response times for hits were in fact shorter for spaced than for once-presented items ($t(89) = -2.77$, $p = 0.0069$, $BF_{10} = 4.14$). Misses were longer on spaced pure lists than on once-presented pure lists ($t(89) = 2.95$, $p = 0.0041$, $BF_{10} = 6.55$). False alarms did not differ ($t(87) = 0.060$, $p = 0.95$, $BF_{10} = 0.13$) and Correct rejections, although trending toward longer on pure spaced than pure once-presented lists, were inconclusive ($t(89) = 2.49$, $p = 0.015$, $BF_{10} = 2.16$).

These findings should be interpreted with caution, because there are many possible explanations for response times varying across conditions. That in mind, the Massed Repetition group showed some evidence of a speed-accuracy tradeoff indicating participants taking longer to access those putative deeper features, similar to Caplan and Guitard (2024b). The Spaced Repetition group showed less difference in response time between spaced and once-presented items, even with shorter times for spaced hits than once-presented hits. This raises the possibility that the Spaced Repetition participants judge probes differently than the Massed Repetition participants.

Bias, *c*. Finally, complementing d', one can also compute the position of the putative response criterion, or bias, $c = .5(z(HR) + z(FAR)$. We did so, again using the log-linear correction, and analyzed each Repetition group with an ANOVA with design List Type×Item Strength. For the Massed Repetition group, there was a significant main effect of Item Strength, $F(1, 90) = 27.58$, $MSE = 0.019$, $p < 0.0001$, $\eta_p^2 = 0.24$, $BF_{\text{inclusion}} > 1000$. The main effect of List Type ($F(1, 90) = 0.40$, $MSE = 0.021$, $p = 0.53$, $\eta_p^2 = 0.004$, $BF_{\text{inclusion}} = 0.14$) and the interaction ($F(1, 90) = 1.55$, $MSE = 0.022$, $p = 0.22$, $\eta_p^2 = 0.017$, $BF_{\text{inclusion}} = 0.21$) were not significant. The corresponding means and standard errors were Mixed Massed: 0.242 (0.039), Mixed Weak: 0.337 (0.042), Pure Massed: 0.271 (0.039), Pure Weak: 0.327 (0.043).

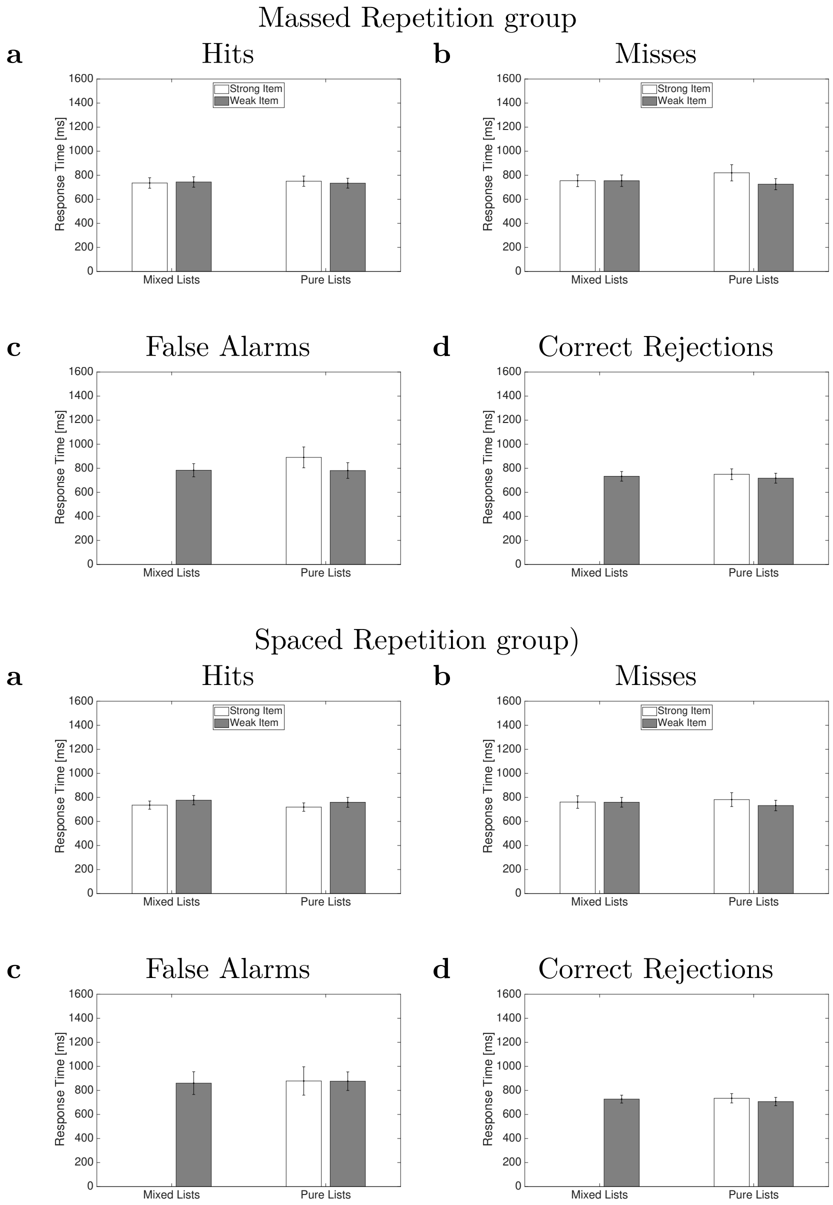


Figure 3

Response times for both Repetition groups of Experiment 1. Median response time was computed for each participant in each condition. Note that for lure probes (false alarms and correct rejections), item-strength is only meaningful on pure lists. Error bars plot 95% confidence intervals based on standard error of the mean.

For the Spaced Repetition group, the same general pattern was found. The main effect of Item Strength was significant, $F(1, 89) = 254.33$, $MSE = 0.024$, $p < 0.0001$, $\eta_p^2 = 0.74$, $BF_{\text{inclusion}} > 1000$. The main effect of List Type was not significant, $F(1, 89) = 0.13$, $MSE = 0.017$, $p = 0.72$, $\eta_p^2 = 0.001$, $BF_{\text{inclusion}} = 0.23$. The interaction was not significant but inconclusive, $F(1, 89) = 3.83$, $MSE = 0.014$, $p = 0.053$, $\eta_p^2 = 0.041$, $BF_{\text{inclusion}} = 0.62$. The corresponding means and standard errors were Mixed Spaced: 0.185 (0.043), Mixed Weak: 0.470 (0.047), Pure Spaced: 0.205 (0.043), Pure Weak: 0.449 (0.047).

However, it makes little sense for the criterion to depend on item strength but not interact with whether the list was mixed or pure. For example, in a mixed list, the lures are the same items that go into the false-alarm calculation for strong as for weak items. A criterion difference between strong and weak items within mixed lists would seem to imply that participants adjust their criterion, but only on target trials! But that is circular, because how would the participant know that a probe is a target without first making the recognition judgement, itself? We view these findings as a case in which the equal-variance assumptions of this simplified version of signal-detection theory are violated, leading to strange results. If we could independently measure the difference in putative variances of strong and weak strength distributions, we could correct for this, but we have no simple way of doing so. Instead, the false-alarm data speak to the question of bias without this confound.

Summary of Experiment 1

We replicated and extended the inverted list-strength effect for massed repetition. We found no inversion, but an approximately null list-strength effect for spaced repetition, despite the weak 1P condition having a study time similar to conditions that have so far only produced inverted list-strength effects. This is at least weakly incompatible with the hypothesis that inverted list-strength effects are always found when the weak and strong condition straddle about 1 s/item (such as 500 ms versus a total of 2000 ms across all repetitions). However, unexpectedly, the spaced repetition group had strikingly low hit rates for weak (1P) items in both pure and mixed lists. Experiment 2 switched to a fully within-subjects design to address this, as follows.

Experiment 2: Massed and spaced within subjects

We could not find evidence to explain why weak items, even in pure-weak lists, produced much lower performance for the spaced-repetition group than the massed-repetition group. However, inspection of the list-by-list effects suggested that this difference appeared even upon the very first list of the session. That "first" list, however, was really the second list because we had included a practice list as well. Our reasoning was that the participants should see both 1P and 4P items in the practice list, practice lists were always mixed. This leads to the conclusion that the effect of repetition group on pure-weak lists must be induced by experience with a mixed list, the nature of that mixed list differing between repetition groups in Experiment 1.

To address this, we designed Experiment 2 fully within-subjects. This would ensure that the pure-weak condition was identical and thus matched between massed and spaced repetition manipulations. But more importantly, participants would have full knowledge of and experience with once-presented, massed-repeated and spaced-repeated items throughout the entire session. Lists were pure-weak, pure-massed, pure-spaced or mixed. Mixed lists always included 50% weak items (once-presented), 25% massed-repeated items and 25% spaced-repeated items. To ensure participants approached the experimental lists with full knowledge of the three item types, the practice list was mixed, and apart from coming first and with extra instructions, was identical to the experimental lists (also with a random selection of words).

The hypotheses from Experiment 1 apply equally to Experiment 2, but with knowledge of the outcome of Experiment 1, we refine them further. As pre-registered:

H1) Spaced repetition produces null (or non-inverted) list-strength effects in recognition memory, even when the display time for each presentation is short. Predictions: When considering the strong condition to be the massed condition, there will be an inverted list-strength effect. When considering the strong condition to be the spaced condition, there will be a less inverted list-strength effect (perhaps even near-null).

H2) Experience with spaced repetitions (e.g., even in the practice list) causes participants to study weak (1P) items and massed 4P items differently than if participants only see 1P or massed repetitions. Predictions: The spaced list-strength effect will be near-null, as in the previous experiment, but the massed-repetition list-strength effect will also be near-null.

H3) Experience with massed repetitions (e.g., even in the practice list) causes participants to study weak (1P) items and spaced 4P items differently than if participants only see 1P or massed repetitions. Predictions: The massed list-strength effect will be inverted, as in the previous experiment, but the spaced-repetition list-strength effect will also be inverted.

Methods

Experiment 2 was pre-registered (pre-registration and data are available at https://osf.io/4hrpw). Due to project-permission issues on the OSF platform and subsequent auto-save errors, some inconsistencies remained between the registered protocol and the final implementation. Although we believed these discrepancies had been corrected in earlier drafts, we only became aware of them after data collection was well underway. These inconsistencies are documented here for transparency. The procedures were approved by a University of Alberta ethics review board.

The materials were identical to those used in Experiment 1, as were the procedures, with the exception of the following differences. While pure lists were unchanged from Experiment 1, mixed lists now included massed and spaced repetitions as well as 1P items. Each mixed list was divided into four quarters, with each quarter including one presentation of all eight spaced words, four weak words, and two strong massed words.

In total, participants completed twelve experimental lists. Each half of the experiment (i.e., the first and second set of six lists) included one pure strong spaced list, one pure strong massed list, two pure weak lists, and two mixed lists that were presented in a randomized order for each participant. Due to a programming error, we deviated from the pre-registered plan, which had specified 16 experimental lists and eight mixed lists in total. However, we believe this deviation is not materially different than our initial plan.

Each session began with a practice list which was a 32-word mixed list constructed identically to the experimental lists (rather than a shorter list as used in Experiment 1).

Participants. Participants were recruited as in Experiment 1. To keep the sample uniform, participants were excluded if they took more than a ten-minute break ($N = 1$), or if d' collapsed across list and item type (excluding practice) was below 0 (chance; $N = 15$). Of the $N = 203$ recruited participants, this left $N = 187$ analyzed.

Due to a bug in our data-analysis code, we thought we had inconclusive Bayes Factors so we ended up collecting up to our pre-registered upper-limit sample size ($N = 200$). The study may be overpowered, in some sense, so we advise the reader to rely on Bayes Factors in addition to p values to interpret the findings.

Results and discussion

List-strength effect

Our first pre-registered analyses concerned d' (Figure 4), analyzed with strong=massed and then strong=spaced repetitions separately. That is, in each analysis, we analyzed the data as though massed versus spaced repetition were manipulated between subjects (as it was in Experiment 1).

For massed repetitions, the two-way interaction was not significant and a favoured null, $F(1, 186) = 1.97$, $MSE = 0.062$, $p = 0.16$, $\eta_p^2 = 0.011$, $BF_{\text{inclusion}} = 0.29$, supporting an non-inverted list-strength effect. The mean log(RoR) was slightly negative, at -0.045, but not significantly different than zero, and a supported null, $t(168) = -0.74$, $p = 0.46$, $BF_{10} = 0.11$. In other words, the inversion of the list-strength effect found for massed repetition in Experiment 1 and in Caplan and Guitard (2025) was not replicated and became a very near-null list-strength effect in this experiment. Item Strength was a significant main effect, $F(1, 186) = 216.54$, $MSE = 0.11$, $p < 0.0001$, $\eta_p^2 = 0.545$, $BF_{\text{inclusion}} > 1000$ and List Type was not significant, $F(1, 186) = 2.67$, $MSE = 0.081$, $p = 0.10$, $\eta_p^2 = 0.014$, $BF_{\text{inclusion}} = 0.28$.

For spaced repetitions, the pattern was different. The two-way interaction was significant, $F(1, 186) = 11.69$, $MSE = 0.062$, $p = 0.0007$, $\eta_p^2 = 0.059$, $BF_{\text{inclusion}} = 75$, supporting not an inverted but an upright list-strength effect. The main effect of Item Strength was significant, $F(1, 186) = 429.83$, $MSE = 0.12$, $p < 0.0001$, $\eta_p^2 = 0.70$, $BF_{\text{inclusion}} > 1000$ and the main effect of List Type was also significant, $F(1, 186) = 6.02$, $MSE = 0.089$, $p = 0.015$, $\eta_p^2 = 0.031$, $BF_{\text{inclusion}} = 34$. The mean log(RoR) was 0.14, significantly above zero although the Bayes factor was still in the inconclusive range, $t(168) = 2.66$, $p = 0.0085$, $BF_{10} = 2.61$. This resembles the spaced repetition group from Experiment 1 and previous published findings, with no trace of an inverted list-strength effect and even evidence of an upright effect.

The other main pre-registered analysis was a t test[4] between the log of the ratio of ratios, as we reported for Experiment 1. In this experiment, the t test was paired-samples. Again, 18 participants with RoR$\leq$ 0 were excluded. The result was a clear, significant

[4] An error was left in the submitted pre-registration, presumably due to auto-save problems, describing this both as an ANOVA and a t test.

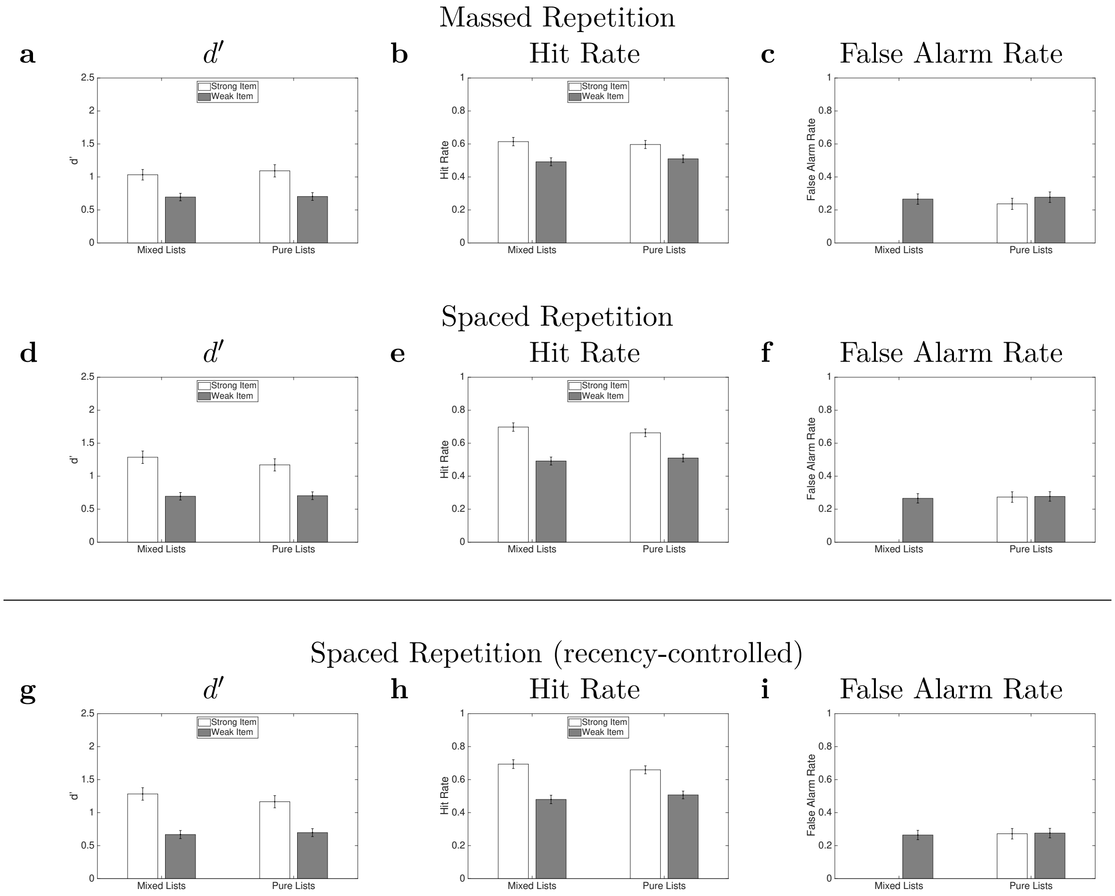


Figure 4

Experiment 2 accuracy data, plotting sensitivity (d′), hit rate and false alarm rate (note that for mixed lists, lures are not tied to a particular item-strength so there is only one false-alarm rate "for" all strengths within mixed lists). Top row: massed repetition, middle row: spaced repetition. Because this is a fully within-subjects design, note that the weak-item values are the same values entering into both the massed-analysis and the spaced-analysis. The bottom row plots the Spaced Repetition data but with a recency-check, retaining only weak items that were studied in the last quarter of the list. Error bars plot 95% confidence intervals based on standard error of the mean.

difference, $t(168) = 3.18$, $p = 0.0018$, $BF_{10} = 10.67$ (M (SEM)=–0.0447 (0.0602) and 0.1383 (0.0519) for massed and spaced, respectively). Following Experiment 1 (although not pre-registered), a Wilcoxon signed rank test on the difference in RoR was also significant, $W = 13973$, $z = 10.66$, $p < 0.0001$, where the median RoR values were 0.96 and 1.12 for massed and spaced, respectively.

Checking for a recency confound. As pre-registered and in Experiment 1, we re-ran the main analyses with only the last-quarter mixed-weak items included. Again, the d' and hit rates of mixed-weak items change almost imperceptibly (Figure 4h; compare to e) and consequently, the ANOVA on d' is similar, with slightly less statistical robustness, as expected due to having less data overall. The two-way interaction was more significant, $F(1, 186) = 13.46$, $MSE = 0.067$, $p < 0.0001$, $\eta_p^2 = 0.067$, $BF_{\text{inclusion}} = 217.8$. Item Strength was a significant main effect, $F(1, 186) = 426.53$, $MSE = 0.13$, $p < 0.0001$, $\eta_p^2 = 0.70$, $BF_{\text{inclusion}} > 1000$ as was List Type, $F(1, 186) = 7.23$, $MSE = 0.096$, $p = 0.008$, $\eta_p^2 = 0.037$, $BF_{\text{inclusion}} = 128.2$. Again, the emerging list-strength was in the upright direction. In sum, equating recency does not appear to materially change the results.

Exploratory analyses of hit and false-alarm rates. Not pre-registered, again we analyzed all the hit rates together and then all the false-alarm rates together. For hit rate, an ANOVA with design List Type[Mixed, Pure] × Item Strength[Weak, Strong-Massed, Strong-Spaced] produced a significant main effect of List Type, $F(1, 186) = 4.43$, $MSE = 0.008$, $p = 0.037$, $\eta_p^2 = 0.023$, $BF_{\text{inclusion}} = 545$, a significant main effect of Item Strength, $F(1.95, 362.9) = 335.41$, $MSE = 0.009$, $p < 0.0001$, $\eta_p^2 = 0.64$, $BF_{\text{inclusion}} > 1000$. Holm-corrected post-hoc t tests confirmed significant differences across strength levels with the rank-order Spaced > Massed > Weak. The interaction was significant, $F(2.00, 371.5) = 11.57$, $MSE = 0.006$, $p < 0.0001$, $\eta_p^2 = 0.059$, $BF_{\text{inclusion}} > 1000$. Holm-corrected pairwise t tests were relatively weak for comparisons of pure versus mixed of a given strength (spaced: $t = 3.82$, $p = 0.0005$, massed: $t = 1.89$, $p = 0.06$, weak: $t = -2.68$, $p = 0.02$). The rest of the pairwise comparisons were all significant ($t > 5$, $p < 10^{-6}$).

For false alarms, an ANOVA with just List Type[Mixed, Pure Weak, Pure Massed, Pure Spaced] was significant, $F(2.57, 477.5) = 13.89$, $MSE = 0.005$, $p < 0.0001$, $\eta_p^2 = 0.069$, $BF_{\text{inclusion}} > 1000$, which Holm-corrected pairwise t tests found Pure-Massed had the fewest false alarms, significantly different than the other three conditions. The remaining three conditions were not significantly different from each other.

Summary of Experiment 2

We replicated the lack of inverted list-strength effect for spaced repetition found in Experiment 1, despite the presence of massed repetitions within the same experiment. In fact, the list-strength effect for spaced repetition was slightly upright. Unexpectedly, massed repetition no longer exhibited an inverted list-strength effect. If the inverted list-strength effect relies on some sort of special-case strategy, perhaps the presence of spaced repetitions renders that ineffective or unavailable. Experiment 3 swapped out massed repetitions for long durations within this same within-subjects design to check whether spaced repetitions always neutralize the cause of inverted list-strength effects or potentially only in combination with massed repetitions.

Experiment 3: Long duration and spaced repetition within subjects

The unexpected finding that the inverted list-strength effect was neutralized for massed repetitions in Experiment 2 led us to think that the inclusion of spaced repetitions undermines the special-casing that massed repetitions can benefit from, which in turn, inverts the list-strength effect. In the attentional subsetting framework, the obvious candidate is disregarding of shallow features during pure-strong lists. In the REM/differentiation framework, the obvious candidate is trace-editing. This led us to wonder: Does the inclusion of spaced repetitions undermine that special-casing across the board or only for massed repetitions? Consider that the presentation rate is fast. A weak, massed or spaced item will be displayed for 400 ms followed by 100 ms of blank screen. So in order to identify that an item is being massed-repeated, the participant has to wait at least for the second (of four) successive repetition and note that it is the same item. This may make it hard work and particularly hard to keep up with at such a fast presentation rate, leading participants to treat massed repetitions more like weak and spaced-repeated items. On the other hand, when no spaced repetitions are expected (as in Experiment 1, Massed group), participants may be able to better specially handle massed repetitions.

For this reason, we thought that if we swapped massed repetitions for long duration single presentations (400 ms display time followed by 1600 ms blank screen), participants might find it easier and quicker to identify those strong items and specially handle them as in an experiment with no spaced repetitions, where we have consistently seen inverted list-strength effects (Caplan & Guitard, 2024b, 2025). Experiment 3 was thus identical to Experiment 2, with the massed repetitions exchanged for long duration as just described.

Experiment 3 was pre-registered (pre-registration and data are available at https://osf.io/6hyma). The procedures were approved by a University of Alberta ethics review board.

Participants. Participants were recruited as in Experiments 1 and 2. Our first target sample size was $N = 50$, which we exceeded due to more sign-ups than we anticipated, for a total of $N = 72$. Again, participants were excluded if they took more than a ten-minute break ($N = 4$), or if their overall d' collapsed across list and item type (excluding practice) was below 0 (chance). This left $N = 63$ analyzed. Because our initial stopping rule was met (conclusive Bayes Factors for the two-way interaction for both the long-duration and spaced-repetition list-strength effects), we terminated data-collection.

Results and discussion

List-strength effect

Our first pre-registered analyses concerned d' (Figure 5), analyzed with strong=long and then strong=spaced repetitions separately similar to Experiment 2.

For long duration, the two-way interaction was significant, $F(1, 62) = 31.89$, $MSE = 0.043$, $p < 0.0001$, $\eta_p^2 = 0.34$, $BF_{\text{inclusion}} > 1000$, supporting an inverted list-strength effect. The mean log(RoR) was substantially negative, at -0.4565, significantly different than zero, $t(53) = -3.69$, $p = 0.0005$, $BF_{10} = 49$. Unlike massed repetition, the inverted list-strength effect remains for long versus short duration, despite the presence of spaced repetitions. Item Strength was a significant main effect, $F(1, 62) = 45.53$,

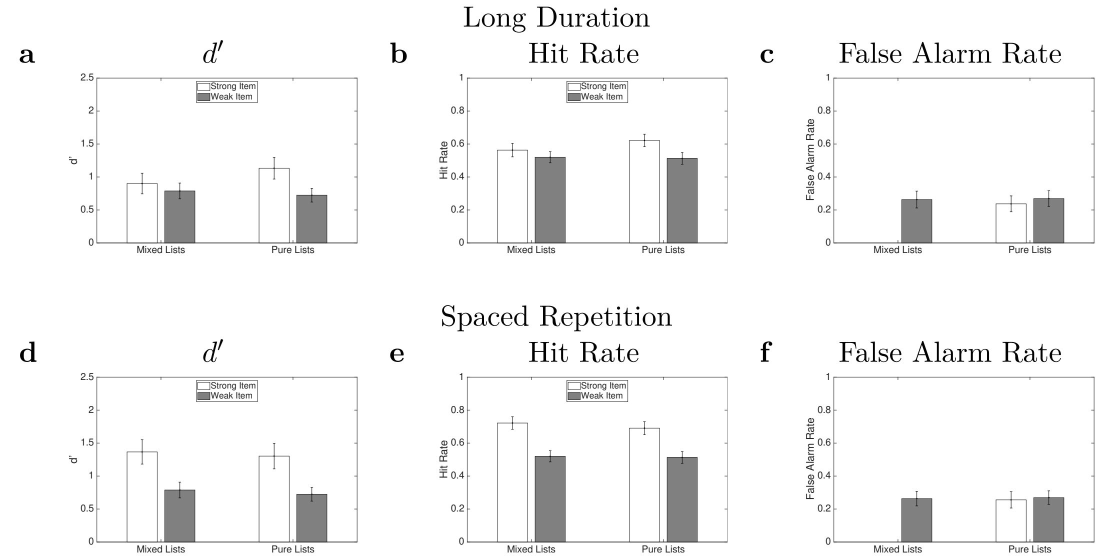


Figure 5

Experiment 3 accuracy data, plotting sensitivity (d′), hit rate and false alarm rate (note that for mixed lists, lures are not tied to a particular item-strength so there is only one false-alarm rate "for" all strengths within mixed lists). Top row: long duration, bottom row: spaced repetition. Because this is a fully within-subjects design, note that the weak-item values are the same values entering into both the long-duration-analysis and the spaced-analysis. Error bars plot 95% confidence intervals based on standard error of the mean.

$MSE = 0.094$, $p < 0.0001$, $\eta_p^2 = 0.42$, $BF_{\text{inclusion}} > 1000$ and List Type was not significant, $F(1, 62) = 3.56$, $MSE = 0.13$, $p = 0.06$, $\eta_p^2 = 0.054$, $BF_{\text{inclusion}} > 1000$.

For spaced repetitions, the pattern was different. The two-way interaction was not significant, $F(1, 62) = 1.1 \times 10^{-4}$, $MSE = 0.062$, $p = 0.99$, $\eta_p^2 < 0.0001$, $BF_{\text{inclusion}} = 0.26$, supporting a null list-strength effect. The main effect of Item Strength was significant, $F(1, 62) = 148.39$, $MSE = 0.14$, $p < 0.0001$, $\eta_p^2 = 0.71$, $BF_{\text{inclusion}} > 1000$ and the main effect of List Type was not significant, $F(1, 62) = 2.32$, $MSE = 0.11$, $p = 0.13$, $\eta_p^2 = 0.036$, $BF_{\text{inclusion}} = 0.41$. The mean log(RoR) was -0.1017, not significantly non-zero, $t(53) = -1.17$, $p = 0.25$, $BF_{10} = 0.28$. This resembles the spaced repetition group from Experiment 1.

Comparing the log(RoR) between long and spaced list-strength effects (excluding 9 participants with negative RoR values), the result was a clear, significant difference, $t(53) = 3.86$, $p = 0.0003$, $BF_{10} = 81$ (M (SEM)=–0.4565 (0.1239) and –0.1017 (0.0867) for long and spaced, respectively). Not pre-registered, a Wilcoxon signed rank test on the difference of RoR was also significant, $W = 1450$, $z = 6.09$, $p < 0.0001$, where the median RoR values were 0.71 and 1.03 for long and spaced, respectively.

Summary of Experiment 3

In sum, Experiment 3 is a third replication of spaced repetition failing to produce an inverted list-strength effect, despite a short presentation duration. In contrast to massed repetitions (Experiment 2), the inverted list-strength effect for long versus short duration was not neutralized in the presence of spaced repetitions. This also establishes that the within-subjects experimental design we used in Experiment 2 is, at least in other conditions, able to simultaneously produce two different forms of list-strength effects.

General Discussion

Our main goal was to ask whether strengthening through spaced repetitions has produced null or slightly upright list-strength effects due to some characteristic of spaced repetition, or could produce an inverted list-strength effect if total study time straddles the $\sim$ 1 s regime, as has been previously found with manipulations of study time per presentation (Caplan & Guitard, 2024b, 2025; Ratcliff et al., 1990, 1994) and with massed repetition (Caplan & Guitard, 2025). Our attentional subsetting theory account of inverted list-strength effects depended on the idea that following pure-strong lists, participants may figure out to disregard the less diagnostic, shallow item features, given that deep features would be plentiful. We had hypothesized (Caplan, 2023; Caplan & Guitard, 2024b) that repeated stimulus onsets may undermine this adaptive disregarding strategy by inducing participants to re-process the shallow features of the word, such that the stronger condition is in part stronger due to those shallow features, similar to our account of the production effect (Caplan & Guitard, 2024a). Across all three experiments, spaced repetition produced null (Experiments 1 and 3) or upright (Experiment 2) list-strength effects, similar to other spaced repetition list-composition manipulations that have used longer presentation times (Diana & Reder, 2005; Ensor et al., 2020; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990). These findings suggest that spaced repetition is indeed not conducive to whatever causes inverted list-strength effects (perhaps disregarding shallow features), and intriguingly, the expectation that there could be spaced repetitions (Experiment 2) appears to undermine such a process even for massed repetition. Experiment 3 suggests this may be peculiar to massed repetitions because the long versus short duration manipulation still showed a very robust inverted list-strength effect in the presence of spaced repetitions.

These non-inverted list-strength effects with spaced repetition occurred despite in the same experiment: 1) massed repetition in Experiment 1 produced an inverted list-strength effect, replicating one prior such instance with different, although close methods (the massed-repetition group in the second experiment of Caplan and Guitard, 2025); 2) massed repetition produced a null list-strength effect in Experiment 2 when combined in a session with spaced repetitions; and 3) long study duration produced an inverted list-strength effect when combined in a session with spaced repetitions, replicating several prior findings in the absence of spaced repetitions (Caplan & Guitard, 2024b, 2025; Ratcliff et al., 1990, 1994).

Finally, in Experiment 2, with massed and spaced repetitions present for all participants, even in the very first, practice list, the massed repetition list-strength effect was no longer inverted, and the massed and spaced repetition pattern was quite similar. Thus, the boundary conditions for inverted list-strength effects are clearer: So far, inversions do not occur when: presentation times are longer than about 1 s, repetitions are spaced, repetitions are massed repetition but are in the presence of spaced repetitions. Experiment 3 found that this final boundary condition for inverted list-strength effects does not apply to a manipulation of study duration.

There may be numerous plausible accounts of the whole pattern of findings, we first reconsider our *a priori* hypotheses and then revisit REM.

If a necessary condition for inverted list-strength effects is the (selective) disregarding of shallow features, then knowing that repetitions *can* be spaced may render disregarding ineffective or unavailable to participants. This could be because repeated onsets now need to be more fully visually processed, to check whether, indeed, the word is the same as the just-prior presentation or potentially a much earlier presentation. This may then increase the encoding of shallow features, whereas when repetitions are always massed, participants may in fact disregard those multiple onsets during the study phase, itself. Experiment 3 shows that this challenge may be avoided with long duration stimuli, avoiding multiple onsets.

A peculiar feature of Experiment 1 was that weak items, even in pure lists, suffered in the Spaced Repetition group compared to the Massed Repetition group. This suggests that when participants had been exposed to spaced repetitions but not massed repetitions, they for some reason threw the weak (once-presented) words under the bus. Comparing Figures 2 and 4, Experiment 2 produced weak-item hit rates in between the two groups of Experiment 1 but closer to the spaced repetition group. Hit rates for spaced and massed repetitions were also both elevated. Although comparing between experiments should be done with caution, this suggests that the presence of spaced repetitions influences the way that weak items are handled, even within pure-weak lists.

However, the important characteristic of the findings across both experiments is that despite our suspicions that experiment composition modulated participants' strategies, spaced repetition in both experiments produced a non-inverted, near-null list-strength effect, suggesting this characteristic is robust for spaced repetition, even with very short display times. On the other hand, massed repetition produced an inverted list-strength effect that was neutralized when embedded with spaced repetitions, suggesting the inversion relies on a customized strategic choice such as selectively disregarding shallow features on tests of pure-strong lists only.[5]

Although the three experiments were inspired by our thinking about attentional subsetting theory, including our pre-registered hypotheses, that does not imply that attentional subsetting is the only framework that could explain the full pattern of list-strength effects. Next we turn to other theoretical accounts of list-strength effects.

Consider Account 1, described in the Introduction, that items are orthogonal (or nearly so). If this were the case, that could explain the spaced-repetition data alone. If orthogonality is not perfect, that can produce the long-expected slightly upright list-strength effect such as we found for the spaced repetition data of Experiment 2. However, this would be hard to reconcile with the inverted list-strength effects found in some other very closely matched conditions, because this account does not have an obvious way to produce inverted list-strength effects. Equally, the orthogonality demands an account for the presence of large, upright list-strength effects in some conditions, such as the production

[5] As we have argued previously (e.g., Caplan, 2023), we assume that participants always disregard as many features as they can. For example, readily identifying that all stimuli are words, participants do not continually process the wordness, the font, the colour, etc., because those features cannot help the participant distinguish items from one another. Disregarding shallow features when deep features are plentiful is one case of that.

effect (Bodner et al., 2014; Caplan & Guitard, 2024a; Jamieson et al., 2016; MacLeod et al., 2010).

Next, consider Account 2, differentiation. First, our own investigations of the dominant differentiation model, REM, caused us to change our assumptions about how REM produces list-strength effects. Inverted list-strength effects are readily produced by REM in its traditional formulation and parameter values, presumably due to differentiation. However, upright list-strength effects are more out of reach of the model.

Working within the REM framework, Ensor et al. (2021) argued that massed repetitions are quite likely to be noted by participants as repetitions and thus stored within a single trace. Spaced repetitions, when noted as repetitions, could presumably be stored within a single trace, producing differentiation. But as also suggested by Shiffrin and Steyvers (1997), they thought spaced repetitions may often go unnoticed as repetitions, inducing participants to create new traces for each repetition. Because differentiation is the force toward inverted list-strength effects, this account aligns to some degree with Experiment 1, where the massed-repetition group produced an inverted list-strength effect but the spaced repetition produced a near-null (less inverted) list-strength effect. However, as can be seen in the parameter optimization explorations in the Appendix, in our hands, REM did not seem to be able to exploit separate traces to sufficiently offset differentiation. Part of the challenge may relate to the particular regime we are in, when performance is relatively low, as is the case for very short presentation durations.

While REM could fit the inverted list-strength effect found for the massed repetition group (Figure A5), its best fits to the spaced data were consistently extremely inverted. The biggest deviation was that the model seems to produce a very high false-alarm rate for pure-weak lists. The best-fitting models also were parameter sets that assume trace-editing happens most of the time. In other words, the traditional version of REM and our small variants all found it difficult to fit the full pattern of data using the idea that participants form multiple separate traces of a strong (spaced repeated) item. Because of this, the model retained its differentiation functionality. One symptom of differentiation is the mirror effect, where the stronger condition (pure spaced) produces more hits but fewer false alarms than the weaker condition (pure weak). This can be seen in the model but not in the data, where the false-alarm rate changes only very slightly (Figure A6).

Setting aside REM, it is not a given that the logic of the differentiation/trace-editing account works for Experiment 2, where the inversion for massed repetition was neutralized and both spaced and massed repetitions produced a near-null (slightly upright) list-strength effect. Even if the separate-trace account were to be achieved with some modified version of REM, that account would have to presume that when participants know to expect some spaced repetitions, they form new traces to *massed* repetitions. If one agrees that participants are storing local traces, does a participant really form multiple memory traces for an item that repeats in immediate succession? Aligning with this, the list-strength effect was more upright for spaced than for massed repetitions in Experiment 2, suggesting at least that the separate traces are formed more often in the spaced than massed repetition conditions.

In sum, the current formulation of the trace-editing hypothesis in REM, based on our own explorations of the model in the Appendix, seems insufficient to explain our global pattern of results. On the other hand, if the separate-trace strategy can be applied to massed repetitions, that could harmonize with a similar possible account of upright list-strength effects in the case of production (reading aloud). Finally, if differentiation is what holds REM back from fitting some findings, it may be promising to adapt REM by softening the likelihood calculation. For example, non-matching features might be processed less precisely than matching features. That might be a way to add missing flexibility to REM, to partially decouple hits (the effectiveness of an item's trace as evidence supporting the probe) from false alarms (its use as evidence that the probe was not on the list).

Aiming to explain list-strength effects in free recall, Malmberg and Shiffrin (2005) proposed their "one-shot" hypothesis, that with 1–2 s of study time, context is stored but then asymptotes. Upon a spaced repetition, more context can be stored, producing an upright list-strength effect in free recall with spaced repetition. Our study times are shorter than this, so the one-shot account would seem to predict more context stored already for 2-s durations and our massed repetitions compared to the weak condition that has only 500 ms per item. This account has in its favour the starting conditions, that REM already produces a net inverted list-strength in the simpler manipulations of duration and massed repetition. This could then be partly offset by cumulative contextual encoding during the longer conditions and resulting in even more contextual storage with spaced repetitions, even of short individual duration. Such an account is not ruled out but remains to be developed and tested. That said, such context-related features in REM are described as changing only very gradually over the course of a list. They would presumably function to generally increase item-similarity, similarly to our model manipulations of g, n and c. Those attempts to increase similarity were still stubbornly against producing upright list-strength effects and could not explain a near-null list-strength effect such as we observed in the spaced repetition group of Experiment 1. This makes it hard to see how contextual features could sufficiently neutralize the strong force of differentiation to convert an inverted list-strength effect into a null or upright list-strength effect.

Some limitations are worth noting. Because the presentation rate is fast, performance is low (although in the ballpark of similar prior studies, starting with Ratcliff et al., 1990). We needed these short presentation times because we needed the initial condition of an inverted list-strength effect and this is the only way we currently know that will reliably do so. However, if inverted list-strength effects can be produced along with higher performance, it would be interesting to do similar comparisons with spaced repetitions in that regime. The materials and list lengths are also particular, and may possibly be pertinent to the generality of these effects.

Conclusion. In sum, despite our team having observed numerous inverted list-strength effects with manipulations of stimulus duration, reinforcing inverted list-strength effects dating to the early 1990s, spaced repetition appears to be inaccessible to some ingredient that is necessary for inversions. Spaced repetitions, when present in the experimental session, also undermine this ingredient with respect to massed repetitions. We have suggested that this ingredient is the combination of weakly studied items having primarily shallow features encoded, strongly studied items having plenty of deep (sparsely subsetted) features stored, and the opportunity to disregard the shallow features only in tests of pure-strong lists. This of course does not rule out other theoretical accounts, such as a modification of REM with a softened implementation of differentiation.

These findings also help explain why near-null list-strength effects have appeared to be dominant in recognition experiments, because spaced repetition appears to be the favoured means of manipulating item-strength. In contrast, manipulating duration over short ranges can produce inverted list-strength effects (even with massed repetitions as the "long" duration) and strengthening via production such as reading aloud produces upright list-strength effects, suggesting that a more complete view of list-composition manipulations is that a broad range of effects can be observed.

Open Practices Statement

The data and program code for Experiment 1, which was pre-registered, are available at https://osf.io/pn2bh/. The data, program code, and pre-registration for Experiment 2 are available at https://osf.io/4hrpw/. The data, program code, and pre-registration for Experiment 3 are available at https://osf.io/6hyma/.

Declarations

Funding: This research was partly supported by a grant from the Natural Sciences and Engineering Research Council of Canada awarded to Jeremy B. Caplan.

Conflicts of interest/Competing interests: The authors declare no conflicts of interest.

Ethics approval: This research (Structure of Human Memory: Pro00105383) was approved by a University of Alberta Research Ethics Board.

Consent to participate: All participants provided electronic informed consent prior to participation.

Consent for publication: All participants provided consent for publication.

Availability of data and materials: The data are available on the OSF pages associated with this manuscript.

Code availability: The code is available on the OSF pages associated with this manuscript.

References

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin & Review*, *21*(1), 149–154.

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).

Caplan, J. B., & Guitard, D. (2025). Inverted list-strength effects in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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.

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

Criss, A. H., & Koop, G. J. (2015). Differentiation in episodic memory. In J. Raaijmakers, A. H. Criss, R. Goldstone, R. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin* (pp. 112–125). Psychology Press.

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

Diana, R. A., & Reder, L. M. (2005). The list strength effect: A contextual competition account. *Memory & Cognition*, *33*(7), 1289–1302.

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

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

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

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.

Jamieson, R. K., Nevzorova, U., Lee, G., & Mewhort, D. J. K. (2016). Information theory and artificial grammar learning: Inferring grammaticality from redundancy. *Psychological Research.*

Kahana, M. J., Rizzuto, D. S., & Schneider, A. R. (2005). Theoretical correlations and measured correlations: Variability and output encoding in four distributed memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 5.

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

Kiliç, A., & Öztekin, I. (2014). Retrieval dynamics of the strength based mirror effect in recognition memory. *Journal of Memory and Language*, *76*, 158–173.

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685.

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 322–336.

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.

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *Quarterly Journal of Experimental Psychology*, *67*(9), 1826–1841.

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.

Rose, R. J., & Sutton, L. T. (1996). Encoding conditions and the list-strength effect. *Canadian Journal of Experimental Psychology*, *50*(3), 261–268.

Sahakyan, L., & Malmberg, K. J. (2018). Divided attention during encoding causes separate memory traces to be encoded for repeated events. *Journal of Memory and Language*, *101*, 153–161.

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.

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.

Wilson, J. H., & Criss, A. H. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88.

Appendix

Further related list-strength effect studies

Table A1 summarizes studies that speak to our questions about list-strength effects in recognition but had participants study items in pairs (but tested them individually). Presentation in pairs was introduced by Ratcliff et al. (1990) in the hope of minimizing the degree to which participants might be using presentation time of one item to study another item. However, a consequence is that this may change the task materially. Without a thorough dive into these issues, these findings should be viewed with caution. For example, paired-presentation may induce participants to study words interactively or even form interactive images between items. Second, stimulus duration is hard to think about because we don't know how participants allocated, for example, 1.25 s of study time between the two items. If they split study time equally, the time per item would be 0.625 s. But if participants, for example, tended to prioritize the left-hand item, they might have devoted 1 s to that item, leaving just 0.25 s for the other item. That would put one item effectively longer than the short-duration regime and the other item shorter. In other words, study-time-sharing between a pair of words could be modulating the effective study time per word in ways we cannot verify.

Table A2 lists additional studies of interest that, however, have some complicating factor for our present purposes (see notes in that table).

Parameteric explorations of REM

We briefly describe our implementation of REM (model code posted along with Experiment 1, at https://osf.io/pn2bh) and then report our parameteric explorations of the model with respect to the list-strength effect. We start with the model and main parameter values Shiffrin and Steyvers (1997) used in their list-strength effect demonstration, although we have changed the notation slightly to harmonize with our attentional subsetting theory notation. An item (word), $\mathbf{f}_i$, is a n-dimensional vector and again, vectors are set in boldface and i denotes unique items. Features, $\mathbf{f}_i(k)$ (indexed by k in parentheses) are geometrically distributed, such that the probability of integer value j is

$$P[f_i(k) = j](1-g)^{j-1}g, \quad j = 1, \ldots, \infty. \tag{A1}$$

Studying a word involves creating a new "trace" (initialized to all zeroes, standing for the absence of a value for each feature) for the item and copying feature values. In each unit of time, of which there are t, a value has u^* probability of getting copied, given that the feature has not yet been stored. With c probability, the copied value is the correct value for the studied item but with $1 - c$ probability, the stored value will be drawn fresh from the geometric distribution (Equation A1), representing a storage error.

When a probe item, $\mathbf{f}_x$, is presented for an old/new judgement, all n features enter into a likelihood calculation. For each stored trace, i, the matching features and mismatching features are identified. We denote the matching features as the set, M_i, and the mismatching features as the set, Q_i. For each trace, the likelihood that trace i was stored if it were the same item as the probe is denoted λ_i, where

Article	Exp.#	Strength	Duration(s)	RoR	Notes
		Duration			
Ratcliff et al. (1990)	2	Dur	1 s & 3 s	1.1	Strength blocked
Ratcliff et al. (1990)	3	Dur	1 s & 3 s	0.93	Strength blocked
Ratcliff et al. (1990)	4 PI	Dur	0.5 s & 2.5 s	0.8	Non-sig. Strength blocked
Ratcliff et al. (1990)	4 no-PI	Dur	0.5 s & 2.5 s	0.77*	Non-sig. Strength blocked
Ratcliff et al. (1992)	3	Dur	0.5 s & 2.5 s	0.99	Mean. No stats. Strength blocked
		Massed Repetition			
Ratcliff et al. (1990)	5 massed	Massed 2P/4P	s & 0.625 s	0.89	Non-sig. but p close. Strength blocked
		Spaced Repetition			
Ratcliff et al. (1990)	5 spaced	Spaced 2P/4P	0.625 s	1.03	Non-sig. Strength blocked
Ratcliff et al. (1990)	6 strong<weak	Spaced 1P/4P	0.625 s	1.02	No stats.
Ratcliff et al. (1990)	6 weak<strong	Spaced 1P/4P	0.625 s	1.08	No stats.
Ratcliff et al. (1990)	6 random	Spaced 1P/4P	0.625 s	0.97	No stats.
Wilson & Criss (2017)	4	Spaced 1P/4P	0.630 s	0.8254	Non-sig. Computed from their d' values.
Wilson & Criss (2017)	5	Spaced 1P/4P	0.630 s	***0.9383***	Non-sig., BF favours the null. Computed from their d' values.

Table A1

Summary of the characteristics and list-strength effects (ratio-of-ratios, RoR) for the most relevant published studies that had participants study in pairs but tested items singly. *Where available, we report authors' RoR and statistical test. We give priority to median RoR across participants, then mean RoR, then RoR calculated from the aggregate d' values (computed ourselves when necessary). * significant ($p < 0.05$). Conclusive Bayes Factor: RoR is set in* ***boldface****. When reliably inverted, RoR is underlined. Supported nulls are* italicized. *Exp.#: which experiment (annotated with a group condition where relevant). Strength: how item strength was manipulated (Dur = Duration, massed = massed repetition, spaced = spaced repetition). For repetitions, 1P denotes once-presented items, etc. Duration(s) = time available to study each item, divided in half to get the average time per word. Notes: "strength blocked:" strength was blocked within a list. "No stats:" we could not find reported statistical tests of the LSE.*

Article	Exp.#	Strength	Duration(s)	RoR	Notes
		Spaced Repetition			
Murnane & Shiffrin (1991a)[a]	1	1P/3P	7 s/sentence	1.032	Non-significant interaction
Murnane & Shiffrin (1991b)[a,b]	1	1P/3P	8 s/sentence	1.4417*	Significant (sum of differences)
Murnane & Shiffrin (1991b)[a]	2	1P/3P	8 s/sentence	1.0782	Non-significant (sum of differences)
Rose & Sutton (1996)[c]	1	3P/3P	SP	0.95	No stats.
Rose & Sutton (1996)[d]	1	1P/3P	SP	1.04, 1.00	No stats.
Osth et al. (2014)	1	1P/3P	3.25 s	LSE*	Missing pure-weak. Fractals
Osth et al. (2014)	2	1P/3P	3.25 s	Non-sig.	Missing pure-weak. Faces
Osth et al. (2014)	3	1P/3P	3.25 s	Non-sig.	Missing pure-weak. Photos
Ensor et al. (2020)	1	1P/2P	2 s	***1.092***	strong tested before weak
Ensor et al. (2020)	2	1P/2P	2 s	**1.983***	strong tested before weak
Ensor et al. (2020)	3	1P/2P	2 s	0.665*	weak tested before strong
Ensor et al. (2020)	1	1P/2P	2 s	**1.983***	strong tested before weak
Sahakyan & Malmberg (2018)	3 FA	2P	2 s	0.8786	Non-sig. Full attention group. Weak=massed, Strong=spaced
Sahakyan & Malmberg (2018)	3 DA	2P	2 s	1.41*	Non-sig. Divided attention group. Weak=massed, Strong=spaced
		Levels of Processing			
Kiliç & Öztekin (2014)[e]	1 and 2	LoP	SP	inverted	No stats.

Table A2

*Summary of the characteristics and list-strength effects (ratio-of-ratios, RoR) for relevant published studies that had some complicating factor. Where available, we report authors' RoR and statistical test. We give priority to median RoR across participants, then mean, then RoR calculated from aggregate d′ values (computed ourselves when necessary). * significant (*$p < 0.05$*). Bayes Factor supporting RoR<>* 1*: RoR is set in* ***boldface****. When reliably inverted, RoR is underlined. Supported nulls are* italicized*. Exp.#: which experiment (annotated with a group condition where relevant). Strength: how item strength was manipulated (LoP = levels of processing). For repetitions, 1P denotes once-presented items, etc. Duration(s) = time available to study each item. SP = self-paced. Notes: "No stats:" we could not find reported statistical tests of the LSE.*
[a] Words were embedded in sentences. [b]For repetitions, three different sentences. [c]Recognition inferred from frequency judgements. Spaced repetitions, weak=same orienting question, strong= 3 different questions. [d]Weak=1P, Strong=3P, same phrase or different phrase contexts. Recognition inferred from frequency judgements. Spaced repetitions, weak=same orienting question, strong= 3 different questions. RoR based on #presentations or #contexts. [e]RoR seen in SAT plots. Between-experiments.

$$\lambda_i = \prod_{k \in M_i} \frac{c + (1-c)g(1-g)^{f_i(k)-1}}{g(1-g)^{f_i(k)-1}} \prod_{k \in Q_i} 1 - c \tag{A2}$$

For a given probe, the overall odds ratio for the probe,

$$\Phi_x = \frac{1}{L} \sum_{i=1}^{L} \lambda_i, \tag{A3}$$

where L is the list length. The model makes an "old" response if $\Phi_i > 1$ and a "new" response otherwise. Ensor et al. (2021) found little effect on the form of the list-strength effect when they varied that threshold (values 1, 2 and 3), so we save ourselves the trouble and fix the threshold to the traditional value of 1 throughout. We fix the list length to 32 items throughout our simulations for direct applicability to our experiments and the first experiment of Ratcliff et al. (1990) which inspired them. We start with the Shiffrin and Steyvers (1997) parameter values: $n = 20$, $g = 0.4$, $u^* = 0.04$ and $c = 0.7$.

Because Shiffrin and Steyvers (1997) and subsequent REM papers have suggested that the inverted list-strength effect due to differentiation can be approximately offset by assuming this is fallible, and the model has some probability of forming a new trace (especially upon a spaced repetition), we first asked what happens to the form of the list-strength effect when the probability of editing a trace upon item repetition varies from $p_e = 0.1$ to 1, in steps of 0.1, where $p_e = 1$ is the original model (100% trace-editing). We split t into two parameters, t_{weak} and t_{strong}. We assumed only a single repetition, so that the first presentation is stored with t_{weak} time steps, the second again with t_{weak} time steps but if trace-edited, the single trace will have t_{strong} total time steps. We varied t_{weak} from 1 to 19 and t_{strong} from $t_{\text{weak}} + 1$ to 20 to check how these generalize across different putative manipulations of strength. Each parameter set was simulated for 1000 lists.

Ensor et al. (2021) had a much richer implementation of trace-editing, where a trace was selected to be edited given that it was retrieved, with additional parameters, which we do not adopt. We assumed only a 2 total presentations to keep the model both parameterically simpler and also to rein in run-times.

"Traditional" REM. Figure A1 shows that in the original model, with $p_e = 1$ so that the repetition always results in a single, cumulative trace (panel c), there is no trace of an upright list-strength effect. The effect approaches a null effect where the two strength levels are quite similar (close to the diagonal). Quite pronounced inverted list-strength effects are produced especially when the weak condition is very weak (low values of t_{weak}). Whereas Ensor et al. (2021) consistently found only slight inversions, RoR$\simeq 0.95$, our expansion of their parameter search shows considerably more flexibility in the model to produce much lower RoR values. As the model increasingly avoids trace-editing and instead, forms new traces for the strong condition (moving through panel b to panel a), inverted list-strength effects are still readily produced when t_{weak} is relatively low. Upright list-strength effects are to be found, but with RoR values only slightly above 1, and only very close to the diagonal, when the trace-edited strength encodes hardly more features than the weak condition, but in contrast, when the trace is not edited, a strong item will have twice as many traces of the same level of quality as a weak item. In other words, the model so far can produce slightly upright list-strength effects, but in a regime where

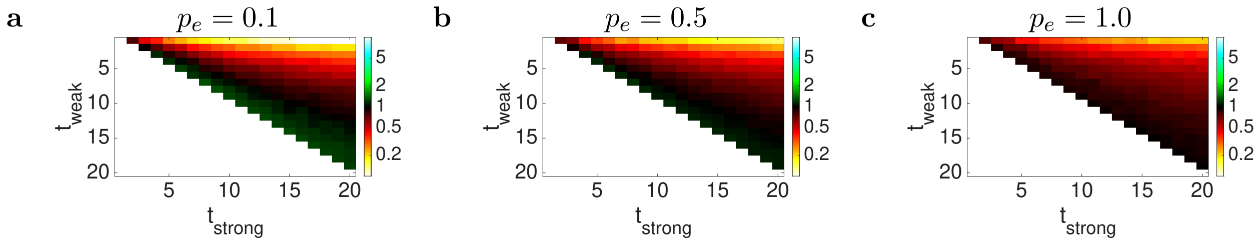


Figure A1

"Traditional" REM. REM simulations with the typical parameter values and list-length, $L = 32$: $n = 20$, $g = 0.4$, $u^* = 0.04$ and $c = 0.7$. t_{weak} and t_{strong} are jointly varied. Panels a–c plot three levels of p_e (0.1, 0.5 and 1.0, respectively), the probability that a repeated item results in an edited trace versus a new trace encoded for the item. $p_e = 1.0$ is the original REM model, where the trace is always edited (c). The colour scale, plotting the ratio-of-ratios (RoR) for each parameter set, is log-transformed to reveal the detail in the plot, but labelled in the original RoR scale.

most 2P items are doubly encoded whereas edited traces are nearly unchanged from a single presentation.

From this starting point, taking cues from Criss (2006) who showed that item similarity can undermine differentiation in REM (and matching her empirical findings on the strength-based mirror effect), we investigated ways in which item similarity might increase, reducing differentiation and thereby potentially revealing an underlying upright list-strength effect.

Varying g, feature frequencies. First, increasing g from 0.4 to 0.9 (analogous to Shiffrin and Steyvers, 1997 modelling higher-frequency words), which generates items and probes with more feature values of low values and fewer rare features, should increase similarity across items in general. When we do this, Figures A2a–c show that this magnifies some of the upright list-strength effects where the trace-edited strong condition is close to the weak condition (close to the diagonal) and substantial numbers of items have additional traces stored (panels a and b). Changing g in the opposite direction, $g = 0.1$ (similar to low-frequency words), appears to further entrench the predominance of inverted and null list-strength effects (d–f).

Varying n, item vector length. Next, we increased inter-item similarity by reducing the vector dimensionality to $n = 10$ features per item and compared that to the flipside, doubling the dimensionality to $n = 40$. Figure A3 shows that reduced dimensionality still produces inverted list-strength effects at similar combinations of the two t parameters, and produces upright list-strength effects when trace-editing frequently fails. Again, increasing n to 40, presumably increasing the distinctiveness of items, further entrenches the predominance of inverted and null list-strength effects.

Varying c, storage accuracy. Finally, we varied the fidelity of stored features, comparing low copying accuracy, $c = 0.3$, to high copying accuracy, $c = 0.9$. Figure A4 again produces near-null and slightly inverted list-strength effects but with frequent copy errors, in conjunction with frequent trace-editing failures, upright list-strength effects can be produced. Just as with g and n, changing c in the opposite direction, to 0.9, further

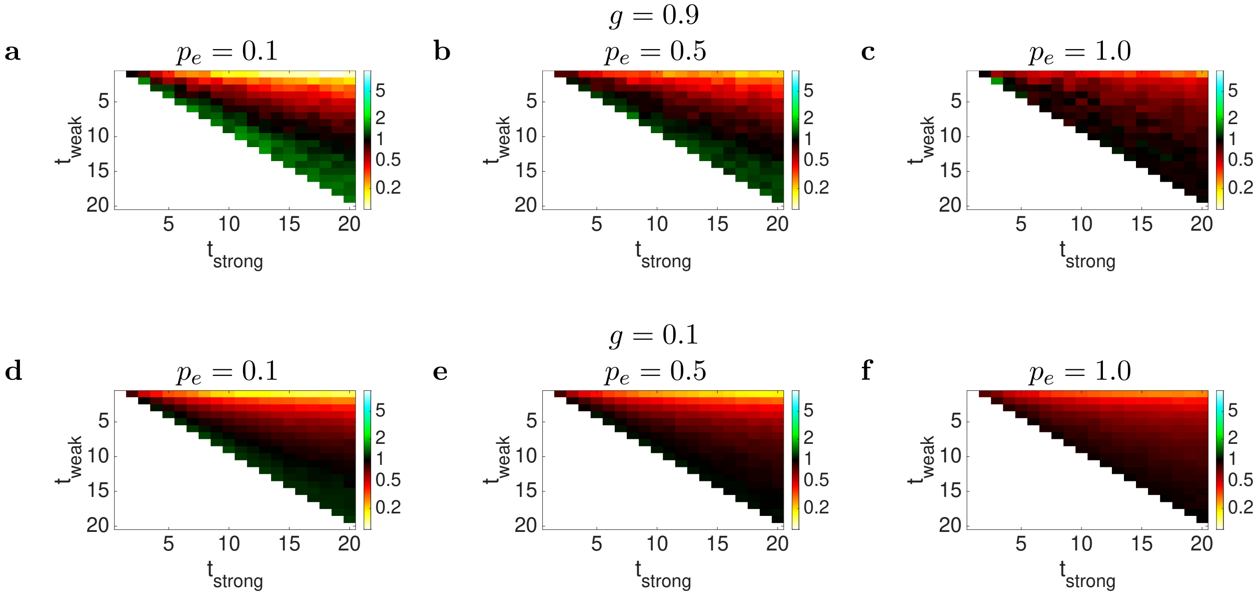


Figure A2

Varying g, feature frequencies. REM simulations with a steep ($g = 0.9$ instead of the typical 0.4, a–c) or shallow ($g = 0.2$, d–f) geometric distribution of features . List-length, $L = 32$ and $n = 20$, $u^* = 0.04$ and $c = 0.7$. t_{weak} and t_{strong} are jointly varied. Panels a–c plot three levels of p_e (0.1, 0.5 and 1.0, respectively), the probability that a repeated item results in an edited trace versus a new trace encoded for the item. $p_e = 1.0$ is the original REM model, where the trace is always edited (c). The colour scale, plotting the ratio-of-ratios (RoR) for each parameter set, is log-transformed to reveal the detail in the plot, but labelled in the original RoR scale.

entrenches the predominance of inverted and null list-strength effects.

Application of REM to our massed and spaced repetition data

Thus far we have been looking only at the RoR, to evaluate the form of the list-strength effect. The RoR could be a good match to the data with or without the d', hit rate and false-alarm rate patterns fitting well. Next we asked if REM could fit our data quantitatively. Because REM, especially with our added variability, has a large number of parameters, the full parameter space is massive and unwieldy to search, particularly because time-consuming simulations need to be done for each parameter set. The following are not exhaustive parameter-space search but rather, we simply found the minimum root-mean-squared error (RMSE) for the 7 parameter explorations we did (including all ten values of p_e although we have embedded only three values of p_e in this appendix). RMSE was calculated for the 4 hit rate and 3 false-alarm rate values produced by a list-strength design. Reports of this exercise are shared for the massed group of Experiment 1 in Table A3 and for the spaced group in Table A4.

Experiment 1, massed repetition group. The massed data are extremely well reproduced by REM (Figure A5), with a moderate amount of trace-editing. Although the

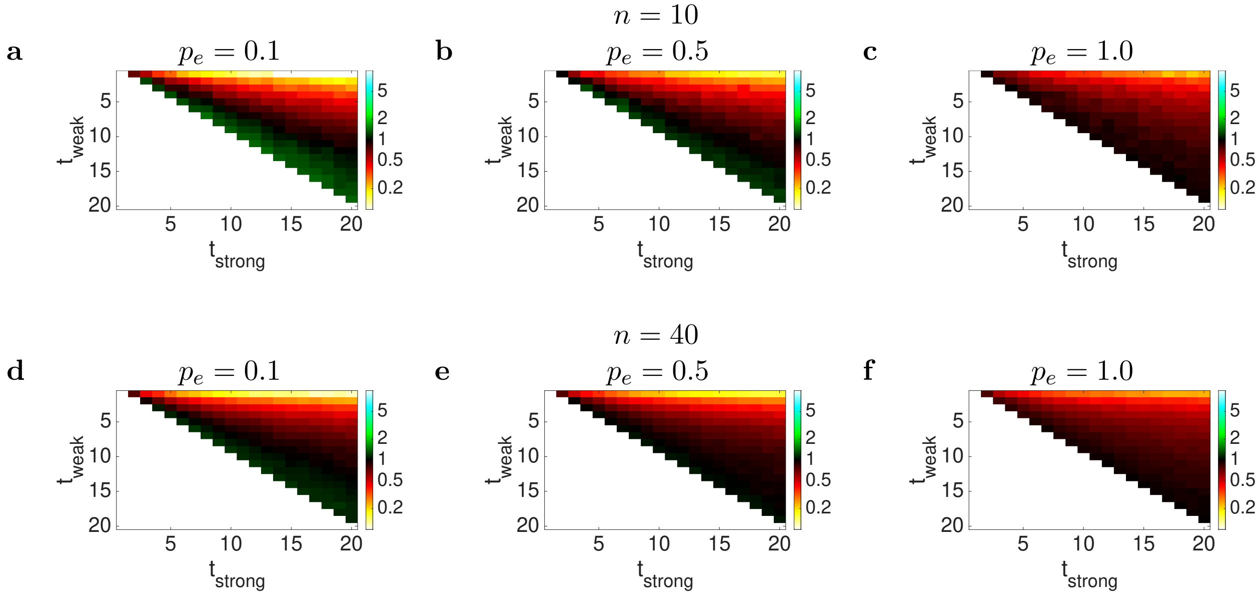


Figure A3

Varying n, item vector length. REM simulations with half ($n = 10$, a–c) or double ($n = 40$, d–f) the number of features per item . List-length, $L = 32$, $g = 0.4$, $u^* = 0.04$ and $c = 0.7$. t_{weak} and t_{strong} are jointly varied. Panels a–c plot three levels of p_e (0.1, 0.5 and 1.0, respectively), the probability that a repeated item results in an edited trace versus a new trace encoded for the item. $p_e = 1.0$ is the original REM model, where the trace is always edited (c). The colour scale, plotting the ratio-of-ratios (RoR) for each parameter set, is log-transformed to reveal the detail in the plot, but labelled in the original RoR scale.

numbers are not identical, recall that this is not a comprehensive parameter-search. The d', hit rate and false-alarm rate values are all quite close to the data, and more importantly, all the rank-ordering of the various conditions match those in the data. In order to fit the data, presumably the low performance levels due to the very short study times per word, the model needed $t_{\text{weak}} = 1$ and $t_{\text{strong}} = 2$, substantially lower than the parameters used, for example, in the demonstrations of Shiffrin and Steyvers (1997). In fact, however, this fit is quite robust; as can be seen in the first row of Table A3, even the “Traditional” variant (with $g = 0.4$; not shown) also fit extremely well and matched the rank-orderings of the conditions, in that case fitting the probability of trace-editing much higher (0.8) and greater t values, closer to the original parameters. REM in its original form, with plenty of trace-editing in the strong condition, quite easily and robustly fits our observed inverted list-strength effect, even though our RoR is more extreme (lower) than the RoR inverted-list-strength effect RoR values that were known prior to 2024. In this regard, REM is better than has been overtly appreciated. Because authors were uncertain whether inverted list-strength effects were to be taken as robust inversions versus measurement error (see discussions in Caplan and Guitard, 2024b, 2025), it may have not been clear whether the stability of REM producing inverted list-strength effects was a strength or a weakness. Given our numerous significant, conclusive (by Bayes Factors) replications of inverted list-

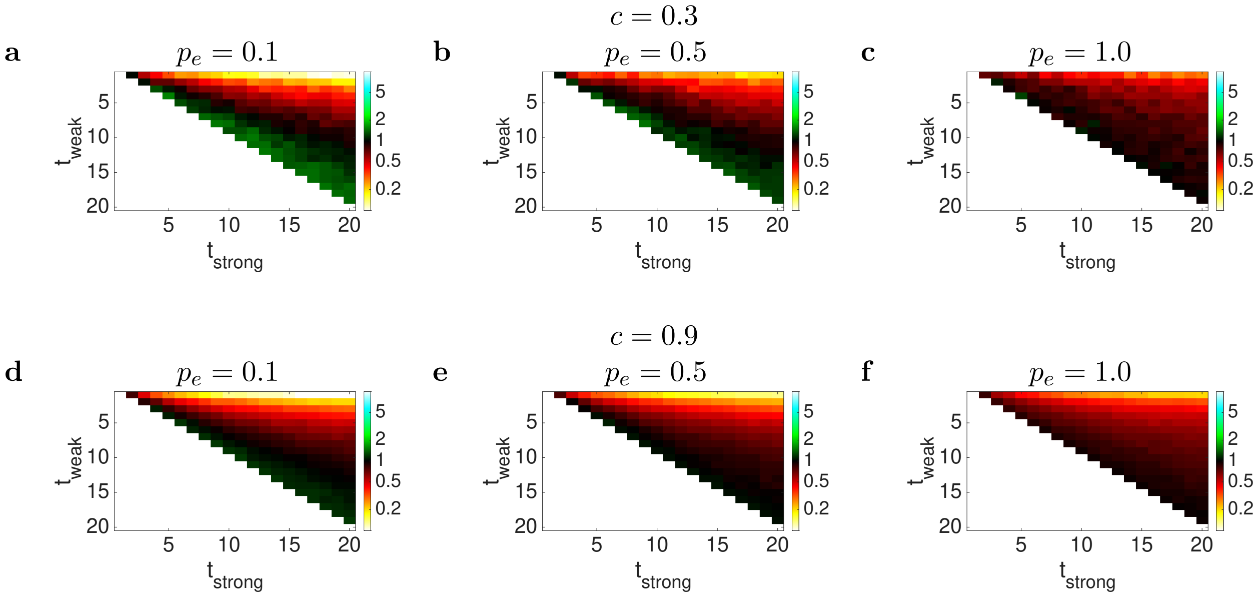


Figure A4

Varying c, storage accuracy. REM simulations with low ($c = 0.3$, a–c) or high ($c = 0.9$, d–f) correct feature-storage probability. List-length, $L = 32$, $n = 20$, $g = 0.4$ and $u^* = 0.04$. t_{weak} and t_{strong} are jointly varied. Panels a–c plot three levels of p_e (0.1, 0.5 and 1.0, respectively), the probability that a repeated item results in an edited trace versus a new trace encoded for the item. $p_e = 1.0$ is the original REM model, where the trace is always edited (c). The colour scale, plotting the ratio-of-ratios (RoR) for each parameter set, is log-transformed to reveal the detail in the plot, but labelled in the original RoR scale.

strength effects (Caplan & Guitard, 2024b, 2025), plus the massed repetition data in the current Experiment 1, we suggest this was an uncelebrated strength of REM (which we, ourselves, had also been confused about).

Experiment 1, spaced repetition group. The spaced data are not well fit within any of the parameter ranges we investigated. Also, the best-fits within each exercise fit to high rates of trace-editing; in other words, the relaxation of trace-editing does not seem to help REM fit the spaced repetition data. This does not mean REM cannot explain the data, only that we have not found parameter sets that do so. It may likewise be the case that these kinds of parameter explorations would cover parameter sets that might fit other published list-strength effect data using spaced repetition. Figure A6 plots the best-fitting parameter set from our parameter explorations. Although most of the numbers come close to the data, the model seems to be drawn to producing an inverted list-strength effect, in this case producing a very high false-alarm rate in pure-weak lists. Restricting the parameter searches to $p_e \leq 0.5$ (not shown) did not help even with the rank-ordering of conditions. One can see in Figures A1–A4 that wherever t_{weak} is low, the list-strength effect tends to be inverted. So we suspect that the low performance regime of the data, due to the very short study time per item, makes it difficult for the forces toward upright list-strength effects (storing multiple traces, adding noise and item-similarity) to be effective, leaving dominant the differentiation

Variant	RoR	RMSE	p_e	t_{weak}	t_{strong}
Traditional	0.6875	0.0145	0.9	2	5
$g = 0.9$	0.6551	0.0560	1.0	5	19
$g = 0.1$	0.7540	0.0134	0.6	1	2
$n = 10$	0.7291	0.0162	1.0	5	10
$n = 40$	0.7567	**0.0125**	0.8	1	2
$c = 0.3$	0.9389	0.0279	0.8	14	20
$c = 0.9$	0.5802	0.0220	0.8	1	3

Table A3

Best fit for each parameter-exploration (see Figures A1–A4) for the massed group of Experiment 1. For the data, RoR=0.7707. The lowest RMSE is set in boldface. "Traditional" refers to $g = 0.4$, $u^* = 0.04$, $c = 0.7$, $n = 20$ and $L = 32$. The other variants are described based on the single parameter that was varied from "Traditional."

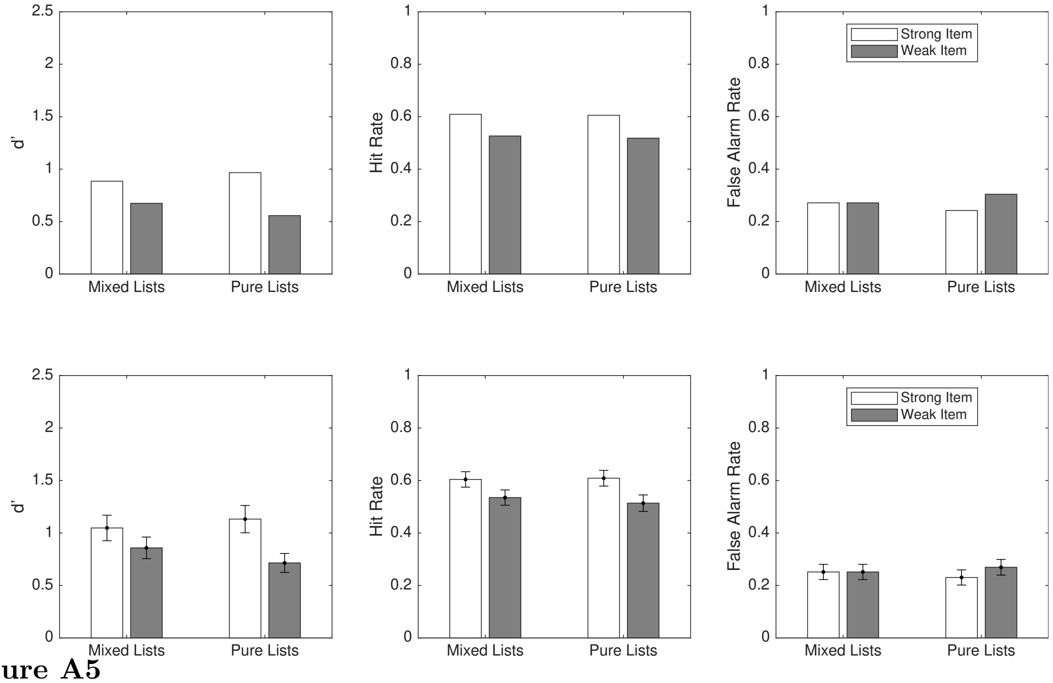


Figure A5

Model (top) output based on the best-fitting parameters from the parameter explorations (Model variant "n = 40" in Table A3 to the massed-repetition data from Experiment 1 (bottom). Parameter values are $g = 0.4$, $p_e = 0.8$, $t_{weak} = 1$, $t_{strong} = 2$, $u^* = 0.04$, $c = 0.7$, $n = 40$ and $L = 32$.

Variant	RoR	RMSE	p_e	t_{weak}	t_{strong}
Traditional	0.4055	0.0171	0.9	1	6
$g = 0.9$	0.4980	0.0711	1.0	2	20
$g = 0.1$	0.6417	0.0365	1.0	1	3
$n = 10$	0.4369	**0.0160**	1.0	2	13
$n = 40$	0.6108	0.0318	0.9	1	3
$c = 0.3$	0.6355	0.0553	1.0	4	20
$c = 0.9$	0.5326	0.0284	1.0	1	4

Table A4

Best fit for each parameter-exploration (see Figures A1–A4) for the spaced group of Experiment 1. For the data, RoR=0.9860. The lowest RMSE is set in boldface. “Traditional” refers to $g = 0.4$, $u^* = 0.04$, $c = 0.7$, $n = 20$ and $L = 32$. The other variants are described based on the single parameter that was varied from “Traditional.”

mechanism that produces inverted effects. One interesting observation is that the best-fitting model set t_{strong} to 6 times t_{weak} (and many of the other parameter explorations did something similar). Because we are only modelling one possible additional trace per strong item, this means that a trace-edited item has $6t$ encoding time steps compared to a multiple-trace item which has $2t$ split across two traces. This presumably gives trace-edited items a massive advantage over multiply-encoded items. But the argument for relaxing the trace-editing assumption was that massed repetitions would have more edited traces while spaced repetitions would have more separate traces. But with this parameterization, the former would not only produce more differentiation, but also better performance implying that massed repetition should outperform spaced repetition. In our data (and typically), data show the reverse pattern.

This, of course, is not grounds for rejecting REM. It shows first of all the accuracy (to data) of the traditional model with high frequency of trace-editing, and second, a limitation of REM and two proposals for how an upright list-strength effect might approximately offset the inverted list-strength effect due to differentiation— namely, storing multiple traces and increased similarity across items. Obvious directions for developing REM include modelling up to 4 encoded traces, to better match the experimental design and following the lead of Ensor et al. (2021) to develop and validate criteria that determine precisely when a trace is edited and which trace (including potentially wrong traces) get edited. We also wonder if trace-editing introduces an overall cost to encoding. If the model (and participant) needs to probe their memory while each item is presented, to check whether they can edit an existing trace, that presumably takes some time away from the time allotted to the item (in our case, this was only 500 ms), which presumably reduces the time available to encode the current item or add features to said existing trace. That would harm not only repeated items but also once-presented items, which would need to be checked for an existing trace and then often fail, requiring a new trace anyway. If participants have any insight into such an encoding-time cost, they might not do that kind of checking on each item, e.g., not checking for item 1 and then checking increasingly more often on each successive item. For this kind of reason, perhaps trace-editing is rare when presentation rates are fast, and more

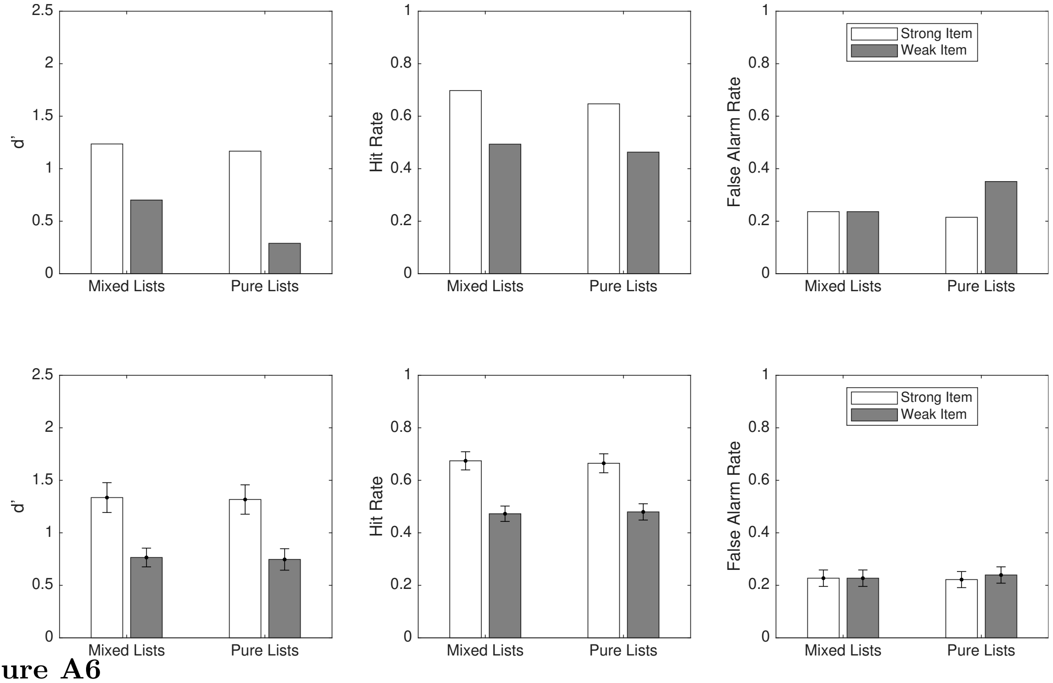


Figure A6

Model (top) output based on the best-fitting parameters from the parameter explorations (Model variant "n=10" in Table A4 to the spaced-repetition data from Experiment 1 (bottom). Parameter values are $g = 0.4$, $p_e = 1.0$, $t_{weak} = 2$, $t_{strong} = 13$, $u^* = 0.04$, $c = 0.7$, $n = 10$ and $L = 32$.

common when there is more time available during the study phase. If so, our attempts to fit REM do not seem to align with this; the probability of trace-editing was quite high for massed repetition, which seems sensible, but also quite high for spaced repetition when the checking costs would seem like a lot, given the 500 ms/item study time, suggesting that the version of REM we have investigated requires something more.