Behavior Research Methods
https://doi.org/10.3758/s13428-019-01324-z

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

Tyler M. Ensor[1,2] · Aimée M. Surprenant[2] · Ian Neath[2]

Abstract
Shiffrin and Steyvers (1997) introduced a model of recognition memory called retrieving effectively from memory (REM) and successfully applied it to a number of basic memory phenomena. REM incorporates differentiation, wherein item repetitions are accumulated in a single mnemonic trace rather than separate traces. This allows REM to account for several benchmark findings, including the null list-strength effect in recognition (Ratcliff, Clark, & Shiffrin, 1990). The original REM treated massed and spaced repetitions identically, which prevents it from predicting a mnemonic advantage for spaced over massed repetitions (i.e., the spacing effect). However, Shiffrin and Steyvers discussed the possibility that repetitions might be represented in a single trace only if the subject identifies that the repeated item was previously studied. It is quite plausible that subjects would notice repetitions more for massed than for spaced items. Here we show that incorporating this idea allows REM to predict three important findings in the recognition memory literature: (1) the spacing effect, (2) the finding of slightly positive list-strength effects with spaced repetitions, as opposed to massed repetitions or increased study time, and (3) list-strength effects that have been observed using very large strong-to-weak ratios (see Norman, 2002).

Keywords Retrieving effectively from memory · Differentiation · Recognition · Item strengthening · Memory models

Retrieving effectively from memory (REM; Shiffrin & Steyvers, 1997) is a well-known model of human memory that has successfully accounted for a number of memory phenomena, including the word-frequency effect (Malmberg & Murnane, 2002), the strength-based mirror effect (Criss, 2006), output interference (Criss, Malmberg, & Shiffrin, 2011), the list-strength effect (Malmberg & Shiffrin, 2005), intentional forgetting (Lehman & Malmberg, 2011), retrieval-induced forgetting (Verde, 2013), the letter frequency effect (Malmberg, Steyvers, Stephens, & Shiffrin, 2002), source recognition (Osth, Fox, McKague, Heathcote, & Dennis, 2018), effects of midazolam (Malmberg, Zeelenberg, & Shiffrin, 2004), and some implicit-memory tasks (Schooler, Shiffrin, & Raaijmakers, 2001). In Shiffrin and Steyvers's initial REM article, they presented several versions of REM, but the one called REM.1 is most often used when modeling recognition. As we review below, however, REM.1 has a simplifying assumption that renders it unable to explain some memory phenomena. Specifically, in REM.1 item repetitions are always accumulated in a single mnemonic trace, even when other study items intervene between presentations. In the present article, we give a brief overview of REM's historical underpinnings, and then explore a version that Shiffrin and Steyvers called REM.3, which, notably, lacks this simplification. We show how this version is able to account for (1) the spacing effect, (2) the finding of slightly positive list-strength effects with spaced repetitions, as opposed to massed repetitions or increased study time, and (3) list-strength effects that have been observed using very large strong-to-weak ratios.

✉ Tyler M. Ensor
tensor@csub.edu

[1] California State University, Bakersfield, CA, USA

[2] Memorial University of Newfoundland, St. John's, NL, Canada

A brief history

REM is a direct descendant of the search of associative memory (SAM) model (Raaijmakers & Shiffrin, 1980, 1981). SAM was originally applied to free and cued recall

Published online: 02 January 2020

(e.g., Huber, Tomlinson, Jang, & Hopper, 2015; Mensink & Raaijmakers, 1988, 1989; Raaijmakers & Phaf, 1999; Sirotin, Kimball, & Kahana, 2005; Tomlinson, Huber, Rieth, & Davelaar, 2009) and was later applied to recognition (Gillund & Shiffrin, 1984). The latter version of SAM was able to account for a number of benchmark findings in the recognition literature including the word-frequency effect (Allen & Garton, 1968; Glanzer & Bowles, 1976; Gorman, 1961; Schulman, 1967), the list-length effect (Strong, 1912; Underwood, 1978), and increases in recognition performance with increased study time (Ratcliff & Murdock, 1976).

SAM's recognition implementation was eventually abandoned because it could not account for two important phenomena. One of these was the strength-based mirror effect (Criss, 2006; Stretch & Wixted, 1998). In recognition, a mirror effect occurs when an increase in the hit rate is accompanied by a decrease in the false-alarm rate (Glanzer & Adams, 1990). For example, low-frequency words produce a higher hit rate and lower false-alarm rate than high-frequency words (Glanzer & Adams, 1985), so word-frequency manipulations produce a mirror effect. Although SAM can predict stimulus-based mirror effects, it cannot predict the false-alarm portion of the mirror effect with between-list strength manipulations (i.e., a higher false-alarm rate following study of a weak as compared to a strong list).

Of greater relevance to the present work is the second phenomenon: the list-strength effect. Research on the list-strength effect emerged from studies on the list-length effect, in which adding items to a list decreases the proportion of items remembered (Ebbinghaus, 1885; Ratcliff & Murdock, 1976; Roberts, 1972; Strong, 1912). In studies of the list-strength effect, a subset of items is strengthened by additional presentations, increased study time, or an elaborative-encoding task. Because the list contains both weak and strong items, it is termed a "mixed" list, and performance on this mixed list is compared to two baseline lists, one on which all items are strengthened ("pure strong") and one on which no items are strengthened ("pure weak"). This paradigm is known as the mixed-pure paradigm (Ratcliff, Clark, & Shiffrin, 1990). A positive list-strength effect is characterized by a larger strong-item advantage in mixed lists than pure lists.

It can be helpful to think of the list-strength effect in terms of what Shiffrin, Ratcliff, and Clark (1990) termed the *ratio of ratios* (R_r). This is a ratio of the strong-to-weak ratios between mixed and pure lists. Let m_{PW}, m_{PS}, m_{MW}, and m_{MS} denote memory performance on pure-weak, pure-strong, mixed-weak, and mixed-strong items, respectively. Then,

$$R_r = (m_{MS}/m_{MW})/(m_{PS}/m_{PW})$$

A list-strength effect occurs if $R_r > 1$, a null list-strength effect occurs if $R_r = 1$, and a negative list-strength effect occurs if $R_r < 1$.[1]

The list-strength effect occurs in free recall (Fritzen, 1975; Hastie, 1975; Malmberg & Shiffrin, 2005; Sahakyan, Abushanab, Smith, & Gray, 2014; Tulving & Hastie, 1972; Wixted, Ghadisha, & Vera, 1997) and some cued-recall tests (Bäuml, 1997; Verde, 2009). However, it does not occur in standard cued recall (Wilson & Criss, 2017) or item recognition (Hirshman, 1995; Murnane & Shiffrin, 1991a, 1991b; Ratcliff et al., 1990; Ratcliff, McKoon, & Tindall, 1994; Ratcliff, Sheu, & Gronlund, 1992; Yonelinas, Hockley, & Murdock, 1992). Ratcliff et al.'s (1990) initial demonstration of the null list-strength effect in recognition was surprising, as a positive list-strength effect was predicted by most models of the time, including SAM (Gillund & Shiffrin, 1984), MINERVA 2 (Hintzman, 1984, 1986, 1988), the theory of distributed associative memory (TODAM; Murdock, 1982, 1983, 1989), the composite holographic associative recall model (CHARM; Eich, 1982, 1985; Metcalfe, 1990), and the matrix model (Humphreys, Bain, & Pike, 1989; Pike, 1984). Shiffrin et al. (1990) showed that, without modification, no extant model could simultaneously predict a null list-strength effect and a positive list-length effect (for a review, see Clark & Gronlund, 1996; for debate concerning TODAM, cf. Murdock & Kahana, 1993a, 1993b, to Shiffrin, Ratcliff, Murnane, & Nobel, 1993).[2]

To accommodate the null list-strength effect in SAM, Shiffrin et al. (1990; see Shiffrin & Raaijmakers, 1992, for a review) incorporated the concept of differentiation (see Gibson, 1940; Gibson & Gibson, 1955). To that point, implementations of SAM had stored multiple presentations of the same item in separate mnemonic traces called images (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1980, 1981). Instead, Shiffrin et al. proposed that item repetitions should be accumulated in a single image. We postpone discussion of the implementation of differentiation until we have described REM. The key point for now is that storing strong items in a single image allowed SAM to predict a null list-strength effect and positive list-length effect. In SAM, adding additional images to memory increases noise, thus reducing performance. However, when repetitions are stored in a single image, noise does not increase, thereby allowing a list-length effect to occur without a list-strength effect.

[1] Note that the calculation of the R_r uses mean performance across subjects rather than computing a separate R_r for each subject. This is because the mean R_r when computed for individual subjects is inevitably influenced by outliers, such as a subject with d' scores of 2.90, 0.05, 2.85, and 2.02 on mixed-strong, mixed-weak, pure-strong, and pure-weak items, respectively ($R_r = 41.11$).

[2] There is debate concerning whether the list-length effect occurs in recognition (see Annis, Lenes, Westfall, Criss, & Malmberg, 2015; Cary & Reder, 2003; Dennis & Humphreys, 2001; Dennis, Lee, & Kinnell, 2008). We do not address this debate in the present article, as it does not directly bear on how strong items are mnemonically stored.

Retrieving effectively from memory

Although Shiffrin et al. (1990) found a way for SAM to predict the null list-strength effect, SAM was unable to simultaneously account for the strength-based mirror effect. Consequently, Shiffrin and Steyvers (1997) developed REM, a model that shares many characteristics with SAM, including differentiation (see Criss & Howard, 2015; Criss & Koop, 2015; Kılıç, Criss, Malmberg, & Shiffrin, 2017). A detailed description of REM is beyond the scope of the present article, and unfamiliar readers are encouraged to consult Shiffrin and Steyvers's original article. Here we focus on those aspects of REM critical for strength effects.

In REM, item features are represented by positive integers drawn from the geometric distribution with parameter g. Each item is made up of w features, and images are vectors of length w. Images are error-prone copies of study items. When an item is studied, information for each feature is stored in the image with probability u. If information for a feature is stored, it is copied correctly with probability c. If it is copied incorrectly, a different feature is drawn from the geometric distribution. Positions in the image vector for which no information was copied (correctly or incorrectly) are represented by 0.

Most implementations of REM set the number of features, w, to 20 and the probability that a feature will be copied correctly, c, to .7. The u and g parameters vary according to experimental variables. For example, in modeling item strengthening, Shiffrin and Steyvers (1997) had strong-item features encoded with probability .4 but weak-item features encoded with probability .28 (i.e., u_{strong} = .4 and u_{weak} = .28). The g parameter is used to vary the frequency of item features, with higher values of g producing more common features than lower values.

Finally, note that the value of g known to the subject often differs from the value of g used to draw item features. REM makes the assumption that subjects are unaware of experimental manipulations such as word frequency. Therefore, although features for low-frequency words may be drawn with g = .325 and features for high-frequency words may be drawn with g = .45, the value stored for incorrectly copied features will be drawn with g = .4 (the value of g assumed to reflect subjects' beliefs about environmental base rates). Subjects also evaluate the diagnosticity of features with reference to their beliefs about environmental base rates. We therefore distinguish between g_{draw} (the value of g used to draw target and distractor features) and g_{base} (the value of g known to subjects).

At test, probes are matched to each image in memory, with the probability that each image feature was generated by the probe and the probability that each image feature was generated by a different study item computed. Notably, a match of a common feature (e.g., 1) is far less diagnostic than a match of an uncommon feature (e.g., 8), and the recognition decision takes this into account. The evidence that a probe is a target is expressed as a ratio between the probability that the probe is a target and the probability that the probe is a distractor. Unless subjects have been instructed to use a particularly conservative or liberal criterion, a target is called "old" if the odds that it is a target exceeds 1; otherwise, it is called "new."

Like the differentiation version of SAM (Shiffrin et al., 1990), repetitions in REM accumulate in a single image. As an item becomes better learned (i.e., as its degree of differentiation increases), it becomes less confusable with other images, and thus exerts less interitem interference than weaker items. Therefore, as the strength of a target's competitors decreases, the probability of recognizing it on a recognition test decreases, and as the strength of a target's competitors increases, the probability of recognizing it on a recognition test increases. So, in the mixed-pure paradigm, mixed-weak items are better recognized than pure-weak items because, on average, mixed-weak targets have stronger competitors than pure-weak targets. In contrast, pure-strong items are better recognized than mixed-strong items because, on average, pure-strong targets have stronger competitors than mixed-strong targets. As such, REM is capable of predicting a negative list-strength effect.

Figure 1 shows the results of 1,000 simulations using the mixed-pure paradigm. For this and all subsequent simulations, hit and false-alarm rates of 1 and 0 were changed to .995 and .005, respectively, when calculating d'. Examination of the hit and false-alarm rates for the pure-weak and pure-strong lists demonstrates that REM correctly predicts the strength-based mirror effect, with the pure-strong list yielding a higher hit rate and lower false-alarm rate than the pure-weak list. REM also predicts a negative list-strength effect, with better discrimination of mixed-weak than of pure-weak items, but slightly better discrimination of pure-strong than of mixed-strong items (R_r = 0.915). Consequently, REM accounts for both of the findings that SAM could not.

Almost all implementations of REM make an important, simplifying assumption: Strengthened items are always stored in a single image. Although this simplification is reasonable for strengthening that occurs via increased study time, massed repetitions, or more elaborative encoding, it probably is not reasonable for spaced repetitions, because subjects will not always realize that a repeated item was studied earlier. Shiffrin and Steyvers (1997) recognized this, and therefore described a version of REM they termed REM.3.

REM.3 incorporates a mechanism by which spaced repetitions can be superimposed on the originally generated image. Critically, for superimposition to occur, a repetition must be recognized as having previously been studied. Algorithmically, REM.3 treats each target during the study phase like a test probe. The simulated subject matches the target to all images in memory and, if the odds that the target was previously studied exceed a given criterion, the target is superimposed on the most similar image. If the criterion is not

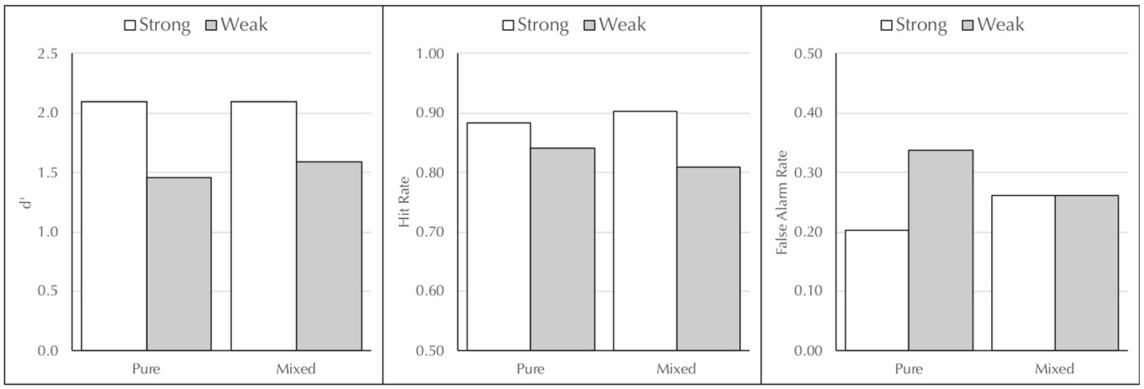

Fig. 1 REM and the mixed-pure paradigm: Predicted *d'*, hit rates, and false-alarm rates. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.35$, $w = 20$, $u_{strong} = 0.4$, $u_{weak} = 0.28$, $c = 0.7$, and $criterion_{test} = 1$.

exceeded, a new image is generated. REM.3 thus has two criterion parameters, which we will term $criterion_{study}$ and $criterion_{test}$. The former represents the threshold required for superimposition during the study phase; the latter represents the threshold required for calling a test probe "old."

Shiffrin and Steyvers (1997) noted that a number of complexities arise with REM.3, not least of which is that the time needed to run the simulations increases considerably. Given that the qualitative predictions of REM.1 and REM.3 were the same, at least for the phenomena Shiffrin and Steyvers considered, they did not pursue REM.3 further. To our knowledge, no subsequent articles implementing REM simulations have used REM.3.

Given the additional complexity inherent in REM.3, one might reasonably ask why we are pursuing it here. One factor underpinning the present investigation is that REM.3 offers more psychological realism than does REM.1. Consider a study phase from the perspective of a subject. When an item is repeated, its repetition might or might not be recognized as previously studied. These two situations are phenomenologically different, and thus computational models should treat them differently. In the case in which the repetition is recognized as being previously studied, subjects are able to retrieve the initial presentation and update it. In effect, the two episodes are bound. Conversely, if the repetition is not recognized as having previously been studied, no retrieval should take place, and a new mnemonic trace should be formed.

Our second reason for examining REM.3 is more practical: Some phenomena cannot be accommodated by REM.1, but REM.3 may be able to handle these data. Indeed, at least one phenomenon, the spacing effect, is inconsistent with REM.1. The *spacing effect* refers to the finding that spaced repetitions have a mnemonic advantage over massed repetitions; this phenomenon has been documented in a number of tests of memory (Glanzer, 1969; Greene, 1990; Hintzman, 1969; Madigan, 1969; Melton, 1967, 1970; Strong, 1916; Underwood, 1969, 1970; Verkoeijen & Delaney, 2008; Xue et al., 2011; for reviews, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010; Dempster, 1988; Ruch, 1928). Critically, if REM treats massed and spaced repetitions identically, then they will necessarily be recognized at the same rate, and therefore REM.1 cannot account for the spacing effect.

Because REM.1 predicts a negative list-strength effect in recognition, and because it predicts an increasingly negative list-strength effect as the strength of strong items increases, there is another phenomenon for which it cannot account. This was first documented by Norman (1999), who used a modified version of the mixed-pure paradigm termed the strong-interference paradigm. Norman (1999) hypothesized that previous studies had failed to detect a list-strength effect because strong items were insufficiently strengthened. Because the mixed-pure paradigm tests strong items, experimenters need to keep strong-item performance below ceiling. In the strong-interference paradigm, only weak items are tested, thereby allowing strong items to be strengthened to ceiling. The strong-interference paradigm uses two list types: a weak-interference list and a strong-interference list. Both lists contain targets and interference items, only the former of which are tested. Targets are presented the same number of times on both lists. However, interference items are given more presentations on the strong-interference list than the weak-interference list. For example, Norman (2002, Exp. 1) presented interference items six times on the strong-interference list, and once on the weak-interference list. For both lists, targets were presented once. Critically, this produced a list-strength effect, with better discrimination on the weak-interference than the strong-interference list.

The present simulations

The purpose of the present work was to assess whether REM.3 is a viable alternative to REM.1. REM.1 cannot explain the spacing effect or the results from the strong-

interference paradigm. Here we assessed whether REM.3 can predict these phenomena.

Simulation 1: Setting $criterion_{study}$

REM assumes that, when making old/new decisions on a recognition test, subjects optimize performance by making an "old" decision whenever the probability that the probe is a target exceeds the probability that the probe is a distractor. In REM.3, each item on the study list is subjected to the same evaluative process. However, is a criterion of 1 still reasonable in this case? Two factors suggest that it may not be.

First, on a recognition test, subjects are asked to make a binary decision: Did this item appear on the study list? Here, even a small amount of evidence one way or the other is sufficient to tip the proverbial scales. During a study list, in contrast, subjects are not overtly assessing whether each item was previously studied. Instead, motivated subjects are actively attempting to commit each item to memory and, presumably, recognizing that an item was studied earlier only occurs if the mnemonic trace is particularly strong. In REM.3, then, if the first presentation of an item was poorly encoded, or if, given the total length of the study list, it does not stand out, it does not make sense for the original image to be updated.

There is a second reason that a stricter value for $criterion_{study}$ is desirable. In REM.3, when a study item is identified as having previously been studied, it is superimposed on the most similar image. Superimposition can go wrong in two ways: First, the study item may not have actually appeared on the study list. This is akin to a false alarm on a test list. In this case, the updated image was actually produced by a different study item. Second, even if the item was previously studied, it may match an image generated by a different study item to a higher degree than the item that was actually studied. In both cases, the item is superimposed on the wrong image, leading to images with features from multiple targets.

To test the degree to which incorrect superimposition is a problem in REM.3, we simulated study phases on which no items were repeated. Our dependent variable of interest was the proportion of trials that resulted in superimposition. Because no items were repeated, an error was made whenever superimposition occurred.

Method

For this and all subsequent simulations, we fixed parameters to values common in the REM literature. We set w to 20, c to .7, g_{draw} to .35, and g_{base} to .4. For ease of exposition, we differentiate between two u parameters: u_1 and u_2. The u_1 parameter denotes the probability of storing features from a study item when a new image is generated; the u_2 parameter denotes the probability of storing features from a study item when superimposition occurs. Note that, during superimposition, only item features in positions where the image contains no information can be encoded.[3] In REM.1, it is standard for u_2 to be less than $2u_1$, because subjects are assumed to devote fewer resources to information that has already been learned. Here we set u_1 to .28 and u_2 to .12. This means that, when a study item is identified as new, each feature is stored with probability .28, and when superimposition takes place, each feature for which information has not yet been stored is copied to the old image with probability u_2. This matches the u_{strong} and u_{weak} values of .4 and .28 used by Shiffrin and Steyvers (1997).

We varied three parameters: g_{draw} (.3, .35, .4, .45, .5), list length (2, 4, 8, 16, 32, 64, 128, 256), and $criterion_{study}$ (1, 2, 3). We ran 1,000 simulations for each combination of these three factors, resulting in 120,000 simulated subjects.

Two algorithmic decisions bear mentioning. First, when an item was identified as having previously been studied, it was superimposed on the most similar image in memory. If a tie occurred between two or more images, the item was superimposed on the more recently generated image. Second, none of our simulated study lists contained any repeated items, even by chance. As each study item was generated, it was checked against the existing items and, if it matched any, it was replaced. These algorithmic precautions were used in this and all subsequent simulations.

Results and discussion

The simulation results are shown in Fig. 2. The probability of superimposition errors increased as word frequency (i.e., g_{draw}) increased. This makes sense, because higher-frequency distractors are more prone to false alarms than are lower-frequency distractors (Glanzer & Adams, 1985). In REM, this is because they are more likely to match an image by chance; the same process operates for study-phase items in REM.3. It is also clear that the probability of errors increases with list length. This is intuitively reasonable because, as the number of images increases, the probability of a study item mistakenly matching an image increases. This is akin to the higher false-alarm rates observed when comparing short and long lists (e.g., Gillund & Shiffrin, 1984). Of greatest importance for the present purposes is that, for all g_{draw} values and list lengths, the probability of superimposition errors decreases as $criterion_{study}$ increases.

[3] Shiffrin and Steyvers (1997) initially used three parameters: u, t_1, and t_2, where t_1 and t_2 represented the numbers of study units for weak and strong items, respectively, and u represented the probability of storing a feature for each unit of study time. In later articles, it was pointed out that these actually amount to two parameters rather than three. Since that time, the t_1 and t_2 parameters have been dropped. For example, consider the values used by Shiffrin and Steyvers: $t_1 = 7$, $t_2 = 10$, and $u = .04$. These can be collapsed into two parameters: $u_1 = t_1 \times u = .28$, and $u_2 = t_2 \times u = .4$.

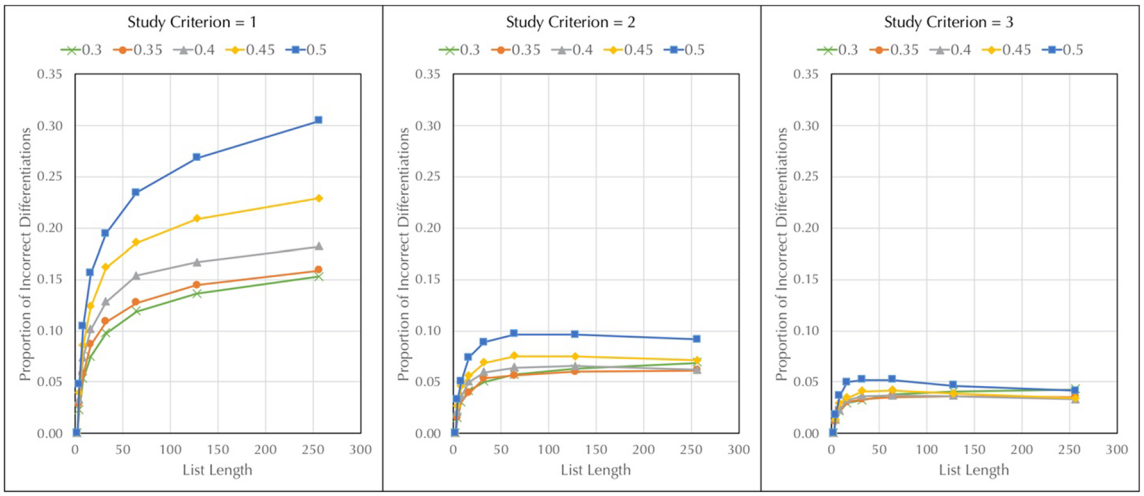

Fig. 2 Superimposition errors as a function of $criterion_{study}$, g_{draw}, and list length. The parameters were $g_{base} = 0.4$, $w = 20$, $u_1 = 0.28$, $u_2 = 0.12$, $c = 0.7$, and $u_2 = 0.12$.

None of the results from Simulation 1 are particularly surprising; indeed, they are necessary outcomes of the math underlying the model. However, what they demonstrate is that a value of 1 for $criterion_{study}$ is too liberal for a study phase. For example, collapsing across g_{draw}, the probability of superimposition errors with $criterion_{study} = 1$ was .138 for 32-item lists and .164 for 64-item lists. By increasing $criterion_{study}$ to 2, mean error rates drop to .064 and .070, respectively, and by increasing $criterion_{study}$ to 3, mean error rates drop to .039 and .041, respectively. These error rates seem more reasonable, given that subjects are usually not asked to make old/new decisions during the study phase.

As we mentioned above, these results are necessary outcomes of the model. Nevertheless, it is critical that these results mimic human data. Unfortunately, it is not possible to collect behavioral data on the probability of superimposition errors. Asking subjects to report whether each item on the study list was studied earlier renders the study phase a de facto continuous-recognition task; the subjects in such an experiment would presumably use the criterion they would use during an old/new recognition test. Therefore, Simulation 1 is necessarily speculative with regard to subjects' actual criterion placement at study. Yet the results appear reasonable, given what we know about human memory. Here, high-frequency items were more likely to be incorrectly superimposed on previously generated images than low-frequency items. This makes sense on distinctiveness grounds: High-frequency words are regularly encountered in everyday life, so distinguishing between whether the familiarity elicited by the study item stems from studying it recently or simply from encountering it recently is challenging. In contrast, low-frequency words are rarely encountered, so identifying the reason for the familiarity is easy. The same logic applies to the list-length results: As the number of items studied increases, the probability that the familiarity elicited by a study item stems from an earlier study item increases.

We systematically varied $criterion_{study}$ in the remainder of our simulations. By and large, changes in $criterion_{study}$ rarely affect the qualitative pattern of the results. However, the situations in which the qualitative pattern is affected by $criterion_{study}$ demonstrate that REM.3 better predicts human data as $criterion_{study}$ increases.

Simulation 2: The spacing effect

In REM.1, the simplifying assumption that spaced repetitions are always accumulated in a single image means that massed and spaced repetitions are identical. As we discussed in the introduction, this prevents REM.1 from predicting an advantage for spaced over massed repetitions. Given its robustness (Delaney et al., 2010), REM.3 must be able to predict the spacing effect to remain viable. The purpose of Simulation 2 was to assess whether REM.3 results in better memory for spaced (i.e., probabilistically superimposed repetitions) or massed (i.e., always superimposed) images.

Method

We simulated a 2 (strengthening method: massed vs. spaced) × 3 (list length: 16, 32, 64) design. Each study list consisted of an equal number of massed and spaced items, all of which were presented twice.

We used the following fixed parameters: $g_{draw} = .4$, $g_{base} = .35$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$. The $criterion_{study}$ parameter was varied as in Simulation 1 (1, 2, 3).

Massed repetitions were always stored in a single image, with the assumption that the immediate repetition of

an item is almost always obvious to subjects. Features from massed items were thus copied with probability $u_1 + u_2$. This is identical to how repetitions are treated in REM.1. When subjects recognized a spaced repetition as having previously been studied, it was superimposed on the most similar image, with new information copied with probability u_2. When a spaced repetition was not recognized as having previously been studied, its features were copied to a new image with probability u_1. Notice that correctly superimposed spaced items were functionally identical to massed items, with features from both copied with probability $u_1 + u_2$. Spaced and massed items only differed when the second presentation was stored in a new image rather than the originally generated image, or when incorrect superimposition occurred.

Results and discussion

Figure 3 displays the results from Simulation 2. It is evident that REM.3 correctly predicts a discrimination advantage for spaced over massed items, regardless of $criterion_{study}$.

The results of Simulation 2 demonstrate that REM.3 predicts the spacing effect. Recall that, in REM.1, repetitions are always accumulated in a single image, regardless of whether other items intervene between presentations. This renders REM.1 unable to predict the spacing effect since, algorithmically, there is no difference between massed and spaced items. This is of crucial importance, as the spacing effect is a fundamental memory effect: It is consistently observed both in and out of the laboratory (Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012; Mumford, Costanza, Baughman, Threlfall, & Fleishman, 1994), and even in nonhuman animals (Menzel, Manz, Menzel, & Greggers, 2001; Tomsic, Berón de Astrada, Sztarker, & Maldonado, 2009).

In REM.3, although massed items accumulate in a single image as in REM.1, spaced items are only stored in a single image when repetitions are recognized as having previously been studied. Therefore, in REM.3, massed and spaced items are only distinguishable when superimposition does not occur—put another way, spaced items are identical to massed items when correct superimposition takes place. Interestingly, it is this failure of superimposition that results in the spacing effect—that is, having multiple images for a spaced item results in superior memory, as compared to a single image.

Why would separate images for an item be mnemonically advantageous? Multiple images for a given item is, in effect, a case of failed differentiation—given the central role that differentiation plays in REM, a mnemonic advantage stemming from its failure may appear counterintuitive. Although, as we will discuss later, multiple-image storage produces some costs, there are some benefits for the undifferentiated item. First, multiple images allow copying errors to be corrected. Recall that, when a feature is copied from a study item to an image, it is copied correctly with probability c and incorrectly with probability $1 - c$. If a feature was copied incorrectly on one study attempt, it may be copied correctly on another study attempt. Notably, this cannot occur when superimposition occurs because, in that case, subjects focus on feature positions for which no information has been encoded. Second, multiple images allow for features to be more widely dispersed throughout the search set, thereby increasing the signal-to-noise ratio during the test phase. For subjects, this means that the recognition probe need only match one of multiple images for an "old" response to be made—that is, multiple-image storage affords multiple retrieval routes, thereby enhancing memory.

In the remaining simulations, we turned to two phenomena from studies of the list-strength effect: REM.3 needs to predict

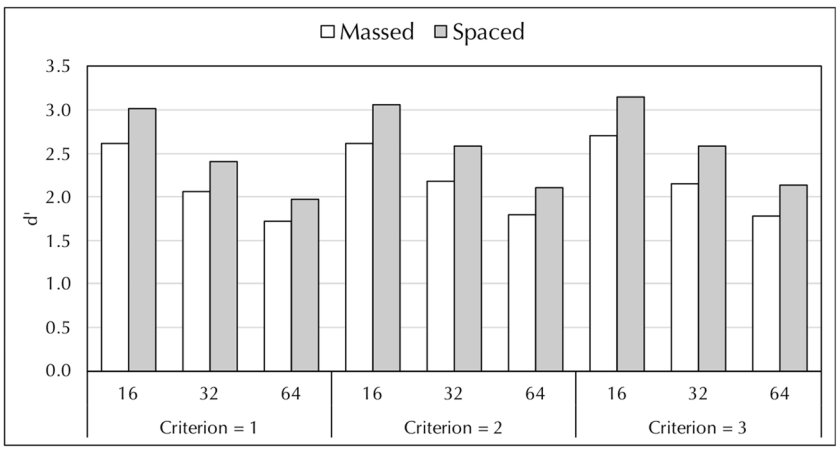

Fig. 3 Recognition of spaced and massed items. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.38$, $w = 20$, $u_1 = 0.28$, $u_2 = 0.12$, $c = 0.7$, and $criterion_{test} = 1$.

a null list-strength effect with the mixed-pure paradigm, but a positive list-strength effect with the strong-interference paradigm. Although REM.1 predicts the former result, it cannot predict the latter.

Simulation 3: The mixed-pure paradigm

As we described in the introduction, the mixed-pure paradigm generally produces a null or negative list-strength effect in recognition (i.e., $R_r \leq 1$) (Ratcliff et al., 1990). As a second test of REM.3's viability, we investigated whether it can predict this result.

Method

Each simulated subject studied five lists: a pure-weak list, on which all items were presented once; a massed pure-strong list, on which all items were presented multiple times, with strengthening accomplished through immediate repetitions; a spaced pure-strong list, on which all items were presented multiple times, with unique items intervening between repetitions; a pure mixed list, on which half of the items were presented once and half were repeated in a massed fashion; and a spaced mixed list, on which half of the items were presented once and half of the items were presented multiple times in a spaced fashion. In the spaced pure-strong list and the spaced mixed list, all items were presented once before any items were presented a second time, all items were presented twice before any items were presented a third time, and so on.

The following variables were systematically manipulated across simulations: the number of strong-item presentations (2, 3, 4, 5, 6), $criterion_{study}$ (1, 2, 3), and list length (28 vs. 84). The full design can thus be conceived of as a 2 (list type: pure vs. mixed) × 2 (strengthening method: massed vs. spaced) × 2 (list length: 28 vs. 84) × 5 (strong-item presentations: 2–6) × 3 ($criterion_{study}$: 1–3) mixed design, with list type and strengthening method manipulated within simulated subjects and the remaining factors manipulated between simulated subjects. We ran 1,000 simulations per cell, yielding 30,000 simulated subjects.

We used the following fixed parameters: $g_{draw} = .35$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$. Algorithmically, strong items were treated as in Simulation 2.

Results and discussion

Figure 4 shows d' results for the pure weak (top), pure strong (second row), mixed weak (third row), and mixed strong (bottom) conditions, and Fig. 5 shows R_r results. As expected, massed strengthening produces a slightly negative list-strength effect, with an average R_r across simulations of 0.937. Surprisingly, spaced strengthening produces a slightly positive list-strength effect, with an average R_r across simulations of 1.123. The R_r is slightly larger with an 84-item list (1.133) than with a 28-item list (1.113) and increases as $criterion_{study}$ increases (R_r = 1.083, 1.136, and 1.150 for $criterion_{study}$ values of 1, 2, and 3, respectively).

Why is the list-strength effect slightly positive with spaced strengthening? In Simulation 2, we showed that, relative to single-image storage, multiple-image storage yielded a mnemonic benefit. Yet the present simulation demonstrates that this mnemonic benefit comes with a cost: Although items stored in multiple images are remembered better than those encoded in single images, these images also produce more interitem interference than they would have if superimposition had occurred. As an image's strength increases, its degree of interference on other images decreases (i.e., differentiation). When superimposition fails, the size of the search set increases, thereby decreasing overall performance as in the list-length effect.

REM predicts that memory will decline as the size of the search set increases—that is, a list-length effect (e.g., Strong, 1912). Notice, too, that, although superimposition does not increase the size of the search set, a failure to superimpose a repetition on the previously generated image does. As such, when superimposition is made probabilistic, as in REM.3, performance suffers, particularly for weaker items.

Do these results falsify REM.3? The prevalent view in the literature is that the list-strength effect is negative or null in recognition. Yet careful examination of the R_r values reported with massed and spaced strengthening suggests that the list-strength effect tends to be null or slightly positive with spaced strengthening, and null or negative with massed strengthening. For example, Ratcliff et al. (1990, Exp. 5) manipulated strengthening method (spaced vs. massed) within subjects and between lists using the mixed-pure paradigm. They found a negative list-strength effect, with massed strengthening that did not quite reach the adopted significance level ($R_r = 0.89$, $p = .055$, two-tailed). In contrast, the R_r was slightly positive in the spaced-strengthening condition (R_r = 1.03, n.s.). Ratcliff et al. (1992) used massed strengthening in Experiment 1 and obtained an R_r of 1.04; in Experiment 2, they used spaced strengthening and obtained an R_r of 1.21, broadly consistent with the predictions of REM.3.

Table 1 shows the R_r values reported in published reports of the list-strength effect in recognition. We excluded experiments that used very fast presentation rates (Ratcliff et al., 1994; Yonelinas et al., 1992), as these are known to produce positive list-strength effects because of rehearsal borrowing (i.e., weak items are presented too quickly, so that on mixed lists, subjects rehearse strong items during weak-item presentations; for a discussion, see Yonelinas et al., 1992). We also excluded studies that used stimuli other than words or word pairs (e.g., Murnane & Shiffrin, 1991a, 1991b; Norman, Tepe, Nyhus, & Curran, 2008). Such stimuli have not been modeled

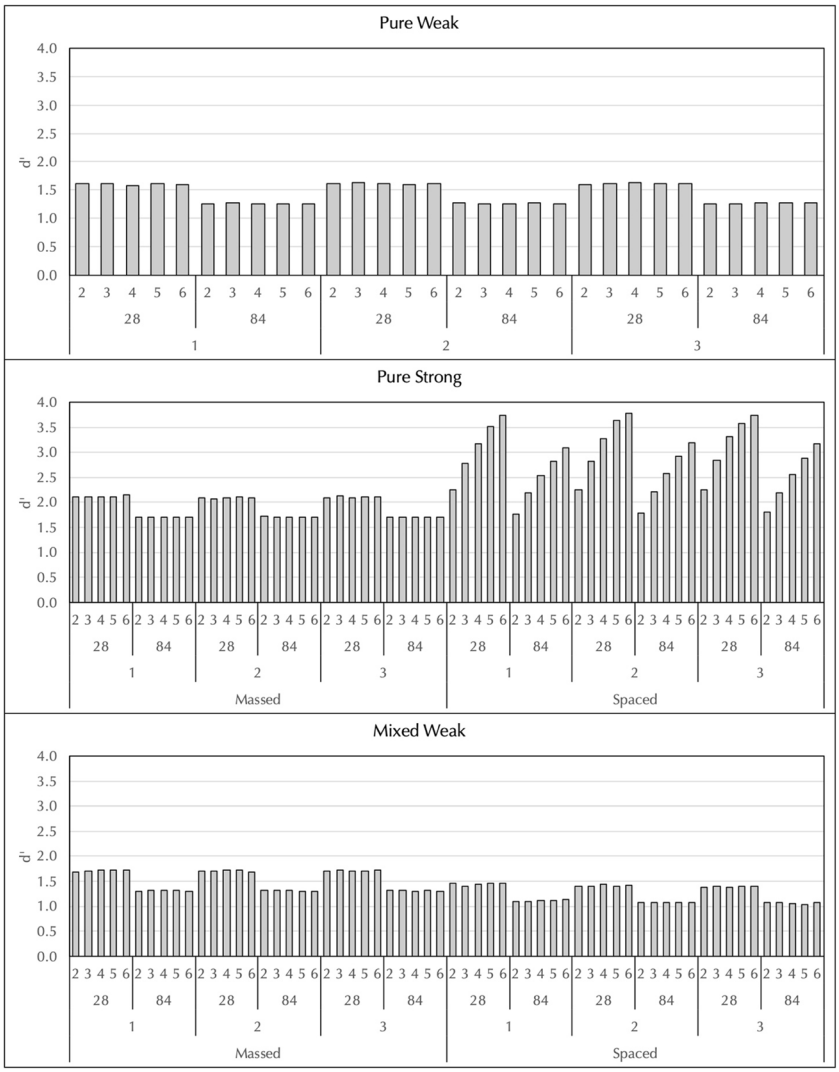

Fig. 4 Values for d' in the mixed-pure paradigm, as a function of $criterion_{study}$, degree of strengthening, and list length. The top row is pure weak, the second row is pure strong, the third row is mixed weak, and bottom row is mixed strong. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.35$, $w = 20$, $u_1 = 0.28$, $u_2 = 0.12$, $c = 0.7$, and $criterion_{test} = 1$.

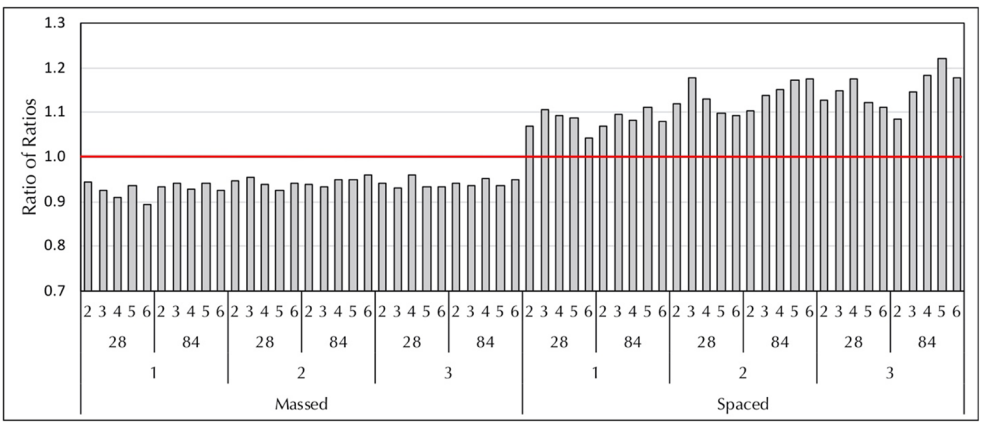

Fig. 5 Ratio of ratio (R_r) values in the mixed-pure paradigm as a function of $criterion_{study}$, degree of strengthening, and list length. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.35$, $w = 20$, $u_1 = 0.28$, $u_2 = 0.12$, c = 0.7, and $criterion_{test} = 1$.

Table 1 Ratio of ratio (R_r) values reported in previous experiments using the mixed-pure paradigm

Study	Experiment	Strengthening Manipulation	R_r
Ratcliff et al. (1990)	1	Massed	0.88
Ratcliff et al. (1990)	2	Massed	1.10
Ratcliff et al. (1990)	3	Massed	0.93
Ratcliff et al. (1990)	4A	Massed	0.77
Ratcliff et al. (1990)	4B	Massed	0.80
Ratcliff et al. (1990)	5A	Spaced	1.03
Ratcliff et al. (1990)	5B	Massed	0.89
Ratcliff et al. (1990)	6A	Spaced	1.02
Ratcliff et al. (1990)	6B	Spaced	1.08
Ratcliff et al. (1990)	6C	Spaced	0.97
Ratcliff et al. (1992)	1	Massed	1.04
Ratcliff et al. (1992)	2	Spaced	1.21

Ratcliff et al. (1990, Exp. 4) is broken into Experiments 4A and 4B. This distinction was not used in the original article, but they had a between-subjects manipulation: For one group, distractors were targets from the previous study–test cycle; for the other, distractors were novel in the experiment. Experiment 4A is the novel-distractor condition; Experiment 4B is the old-distractor condition. Ratcliff et al. (1990, Exp. 5) is broken into Experiments 5A and 5B for spaced and massed strengthening, respectively. Ratcliff et al. (1990, Exp. 6) had three spaced-strengthening conditions; we have therefore broken it into Experiments 6A, 6B, and 6C. Experiment 6A presented all strong items before all weak items on the mixed study list, Experiment 6B presented all mixed-weak items before all mixed-strong items at study, and Experiment 6C randomized the mixed study list without regard to strength.

in REM, and there is evidence that they may be more susceptible to interitem interference than are words (see Osth, Dennis, & Kinnell, 2014). Finally, we included only experiments using the full mixed-pure paradigm—that is, we excluded experiments that included only the pure-weak list or only the pure-strong list (e.g., Hirshman, 1995). Across these experiments, the mean R_r for massed strengthening is 0.916, but the mean R_r for spaced strengthening is 1.062.

It is critical to note that REM.1 cannot account for even a slightly positive list-strength effect in recognition, and therefore is unable to account for this massed–spaced discrepancy. In contrast, Simulation 3 shows that REM.3 can account for a positive list-strength effect with spaced strengthening and a negative list-strength effect with massed strengthening.

REM.3 predicts that disrupting the differentiation process will yield a slightly positive list-strength effect, but that, when differentiation is allowed to occur for all repeated items, the list-strength effect will be negative. As we described above, some evidence for this comes from an examination of list-strength effects from massed and spaced strengthening techniques. However, more direct evidence comes from a recent study by Sahakyan and Malmberg (2018). Sahakyan and Malmberg directly interfered with the differentiation process by having subjects complete a secondary task during the study phase. Crucially, this produced a list-strength effect in recognition, consistent with the argument we have presented here.

Simulation 4: Norman (2002, Exp. 1)

As we described in the introduction, Norman (1999, 2002) introduced the strong-interference paradigm as an alternative to Ratcliff et al.'s (1990) mixed-pure paradigm. In the strong-interference paradigm, only weak items are tested, thereby allowing experimenters to strengthen strong items to ceiling. Another difference between the two paradigms is that in the strong-interference paradigm, interference items are presented on both lists. The only difference is that on the weak-interference list, the interference items are given fewer presentations than on the strong-interference list.

Norman (2002) used the strong-interference paradigm to test for a list-strength effect in recognition, operationalized as better discrimination in the weak-interference list than the strong-interference list. The stimuli were unrelated, medium-frequency words. List type was manipulated within subjects, with order of conditions counterbalanced. On the weak-interference list, subjects studied five untested primacy buffers, 50 targets and 50 interference items randomly combined, and five untested recency buffers. Subjects were not made aware of the buffer/target/interference item distinction, and thus attempted to learn all items. The strong-interference list was identical to the weak-interference list except that after the five recency buffers, the 50 interference items were presented five more times. Therefore, the interference/target strength ratio was 1:1 on the weak-interference list and 6:1 on the strong-interference list. Note, as well, that interference items were all presented a second time before any were repeated a third time, all were presented a third time before any were presented a fourth time, and so on. Each repetition of the 50 interference items was randomized anew.

Two methodological precautions bear mentioning: First, to attenuate the probability of rehearsal borrowing, for each study presentation, subjects were asked whether the word's referent could fit into a banker's box. Second, to equate the study–test lag between lists, a longer distractor task was interpolated between study and test in the weak-interference list. Neither of these features were incorporated in our simulations: To our knowledge, no one has incorporated different encoding tasks in REM. In addition, because we did not include context in our REM simulations, equating study–test lag is not necessary.

Test lists consisted of the 50 targets and 50 distractors. At test, subjects rated their confidence that the probes were "old" on a six-point scale. If they believed that a probe was old (i.e., if they selected 4, 5, or 6 on the confidence scale), they then made a remember/know judgment (see Tulving, 1985).

Norman (2002) presented d', computed by changing confidence ratings of 4, 5, and 6 to "old" responses and ratings of 1, 2, and 3 to "new" responses. We did not simulate confidence-scale responding; rather, we simply had simulated subjects make old/new recognition decisions.

Norman's (2002) results produced a list-strength effect: d' was higher on the weak-interference list ($M = 2.35$, $SE = 0.12$) than on the strong-interference list ($M = 2.22$, $SE = 0.10$). The hit rate was higher in the weak-interference list ($M = .91$) than in the strong-interference list ($M = .77$). Interestingly, the false-alarm rate was also higher in the weak-interference list ($M = .22$) than in the strong-interference list ($M = .11$).[4] Hit and false-alarm rates, then, produced a concordant effect (i.e., hit and false-alarm rates increasing together) rather than a mirror effect (i.e., a decreasing false-alarm rate with an increasing hit rate).

Simulation 4A: REM.1

We began by investigating whether REM.1 could account for Norman's (2002) results. To our knowledge, this has yet to be investigated in the literature. If REM.1 is capable, then utilizing a more complex version like REM.3 would violate the principle of parsimony.

Method

The goal of the strong-interference paradigm is to strengthen interference items to ceiling on the strong-interference list. When Shiffrin and Steyvers (1997) simulated list-strength manipulations, they set u_{strong} to .4. However, this keeps strong-item performance below ceiling. Here we varied u_{strong} (.4, .6, .8) in order to simulate increasingly strengthened strong items.[5] We set u_{weak} to .28, as did Shiffrin and Steyvers. Note that, since REM.1 treats massed and spaced repetitions identically, one cannot directly manipulate the number of strong-item presentations; instead, the encoding parameter, u, serves as a proxy for the number of presentations.

Because Norman (2002) used medium-frequency words, we set g_{draw} to .38. The other parameters were fixed at values used in our previous simulations: $g_{base} = .4$, $w = 20$, $c = .7$, and $criterion_{test} = 1$ (recall that REM.1 does not have a $criterion_{study}$ parameter).

[4] We do not have standard errors for hit and false-alarm rates. This is because Norman (2002) presented means as a function of confidence scale responses.

[5] One might wonder why we are varying u rather than more directly manipulating the number of strong-item presentations, as in the REM.3 simulations. Recall, however, that in REM.1 multiple presentations of an item are simulated by increasing the probability of encoding item features. Therefore, varying u is how one strengthens items in REM.1.

Results and discussion

Figure 6 shows d', hit rates, and false-alarm rates as a function of u_{strong}. Examination of the hit and false-alarm rates reveals that REM.1 correctly predicts the concordant effect observed by Norman (2002). However, REM.1 predicts a negative list-strength effect, with better discrimination on the strong-interference list than the weak-interference list. Indeed, as u_{strong} increases, the magnitude of the strong-list advantage increases.

Simulation 4A therefore demonstrates that REM.1 cannot predict Norman's (2002) results. In particular, this is due to differentiation: As the strength of strong images increases, they become more insulated from related images, and therefore produce less interference.

Simulation 4B: REM.3

Method

In our REM.3 simulations, we only varied the $criterion_{study}$ parameter (1, 2, 3, 4). The following fixed parameters were used: $g_{draw} = .38$, $g_{base} = .4$, $w = 20$, $c = .7$, $u_1 = .28$, $u_2 = .12$, and $criterion_{test} = 1$.

Results and discussion

Figure 7 shows d', hit rates, and false-alarm rates as a function of $criterion_{study}$. Qualitatively, the results for $criterion_{study}$ values of 2, 3, and 4 replicate Norman's (2002) results, with better discrimination on the weak-interference than the strong-interference list. However, when $criterion_{study}$ is 1, REM.3 predicts equivalent discrimination between lists. Note, as well, that the hit and false-alarm rates replicate the concordant effect from Norman's experiment: Both hit and false-alarm rates are higher on the weak-interference than on the strong-interference list.

We are not concerned that a value of 1 for $criterion_{study}$ fails to produce a list-strength effect in REM.3. On the basis of the results of Simulation 1, we do not think that 1 is a reasonable value for $criterion_{study}$. The fact that higher values for $criterion_{study}$ produce results that are more consistent with the empirical data further bolsters this position.

Why is REM.3 able to replicate Norman's (2002) pattern when REM.1 could not? Consider the size of the search set between the weak- and strong-interference lists in REM.1. In both cases, there is one image for each unique study item (i.e., 110 total images, for the 10 serial-position buffers, 50 targets, and 50 distractors). The only difference is that the images for the strong-interference items are more complete copies of the study items, and thus are less likely to produce interference than the weak-interference images.

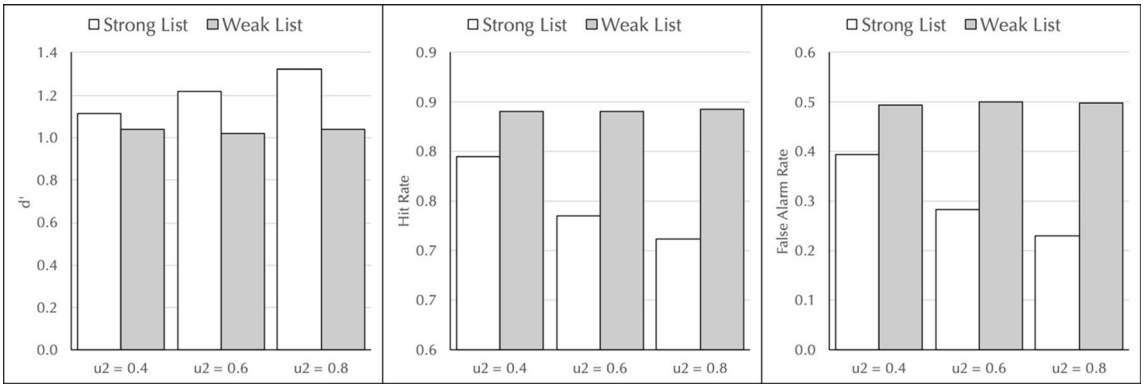

Fig. 6 Values for d', hit rates, and false-alarm rates in simulations of Norman (2002, Exp. 1) with REM.1, as a function of u_{strong}. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.38$, $w = 20$, $u_{weak} = 0.28$, and $criterion_{test} = 1$.

Next, consider the search sets in REM.3. In the weak-interference condition, there will be a maximum of 110 images (for some subjects there will be fewer, owing to incorrect superimposition; see Simulation 1). The search set for the strong-interference condition, in contrast, will be much larger: Whenever superimposition fails to occur, a new image is generated. This drastically increases interference, thus lowering performance on the targets.

Note that the explanation above makes sense when considering the experimental session from the perspective of subjects. Given the relatively long lags between repetitions in Norman's (2002) experiment, it is reasonable to assume that subjects sometimes failed to notice that an item had been studied earlier. Such an occurrence should result in multiple mnemonic traces—that is, failed differentiation.

The explanation above also sheds light on Norman's (2002) concordant effect. Prima facie, this finding is puzzling, since strength manipulations generally produce a mirror effect, with a higher hit rate and lower false-alarm rate on strong than on weak test lists (Criss, 2006; Stretch & Wixted, 1998). However, because of the larger search set in the strong-interference than in the weak-interference condition, more evidence is needed to call a probe "old," resulting in a lower hit rate. At the same time, false alarms are less likely on the strong-interference list, both because of the larger search set and because interference-item images are more differentiated than images for the weak-interference list, thereby lowering the probability of a distractor inadvertently matching one.

Simulation 5: Norman (1999, Exp. 4)

Norman (2002, Exp. 1) used relatively long study phases, with 360 trials in the study phase of the strong-interference list. To ensure that the REM.3 results from Simulation 4B were not an artifact of the large number of study trials, we also performed simulations for Experiment 4 of Norman (1999). This experiment had two between-subjects conditions, which we simulate in Simulations 5A and 5B. One condition was very similar to Norman (2002, Exp. 1), but with a shorter study list. In the other condition, distractors

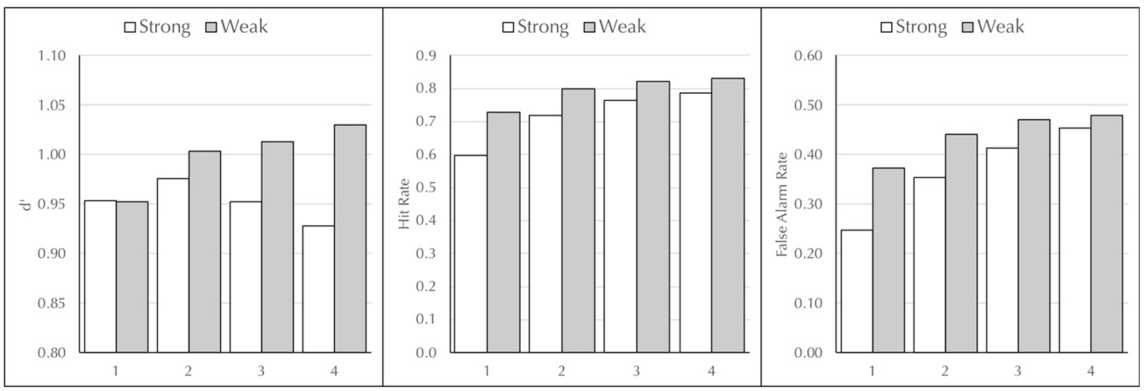

Fig. 7 Values for d', hit rates, and false-alarm rates in simulations of Norman (2002, Exp. 1) with REM.3, as a function of $criterion_{study}$. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.38$, $w = 20$, $u_1 = 0.28$, $u_2 = 0.12$, $c = 0.7$, and $criterion_{test} = 1$.

were semantically related to targets (e.g., alligator–crocodile). Both experiments revealed a list-strength effect.

Simulation 5A: Unrelated distractors

Description

The procedure of Norman (1999, Exp. 4) was very similar to that of Norman (2002, Exp. 1). The strong-interference paradigm was used, with interference items presented once in the weak-interference list and six times in the strong-interference list. However, there were 16 targets instead of 50, 48 interference items instead of 50, and three primacy and recency buffers instead of five. In addition, only eight of the 16 targets were presented at test, along with eight distractors. Unlike the confidence judgments made in Norman (2002), Norman (1999) used standard old/new recognition.

Norman (1999) uncovered a list-strength effect as measured by d': Discrimination was better on the weak-interference list (M = 2.28, SE = 0.08) than the strong-interference list (M = 1.90, SE = 0.09). The hit and false-alarm rates produced the concordant effect observed in Norman (2002): The hit rate was higher in the weak-interference list (M = .91, SE = .02) than the strong-interference list (M = .66, SE = .03), and the false-alarm rate was higher in the weak-interference list (M = .12, SE = .01) than in the strong-interference list (M = .03, SE = .01).

Method

Other than the number of targets, interference items, serial-position buffers, and distractors, the algorithm and parameters for Simulation 5A were identical to those used in Simulation 4B.

Results and discussion

Figure 8 shows d', hit rates, and false-alarm rates as a function of $criterion_{study}$. As in Simulation 4B, $criterion_{study}$ values of 2, 3, and 4 produced a list-strength effect, whereas a $criterion_{study}$ value of 1 produced a null list-strength effect. These simulations also replicated the concordant effect observed in Simulation 4B and the empirical data.

Simulation 5B: Semantically related distractors

Description

This condition was identical to the unrelated-distractors condition, save that each distractor was semantically related to a target from the study phase. The results matched those found in the unrelated-distractors condition: There was a list-strength effect, with better discrimination (as measured by d') in the weak-interference list (M = 1.87, SE = 0.08) than in the strong-interference list (M = 1.48, SE = 0.09). The concordant effect was also replicated: The hit rate was higher in the weak-interference list (M = .85, SE = .02) than in the strong-interference list (M = .58, SE = .03), and the false-alarm rate was higher in the weak-interference list (M = .18, SE = .02) than in the strong-interference list (M = .07, SE = .01).

Method

To simulate semantic relatedness in REM, we set eight features for each distractor equal to eight features of one of the targets. Otherwise, Simulation 5B was identical to Simulation 5A.

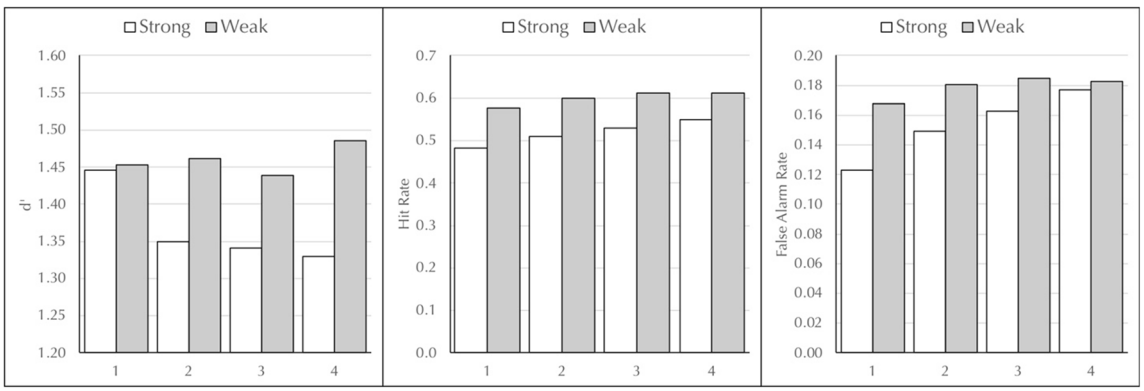

Fig. 8 Values for d', hit rates, and false-alarm rates in simulations of Norman (1999, Exp. 4, unrelated-distractors condition) with REM.3, as a function of $criterion_{study}$. The parameters were g_{base} = 0.4, g_{draw} = 0.38, w = 20, u_1 = 0.28, u_2 = 0.12, c = 0.7, and $criterion_{test}$ = 1.

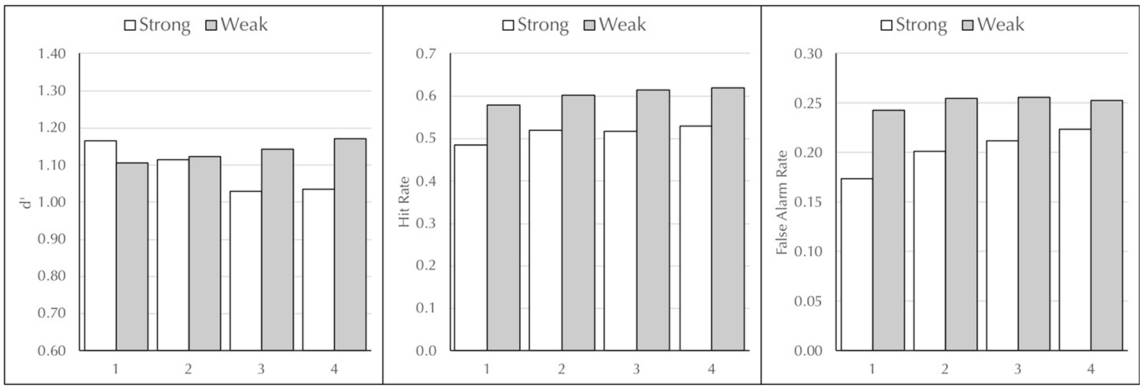

Fig. 9 Values for *d'*, hit rates, and false-alarm rates in simulations of Norman (1999, Exp. 4, semantically-related-distractors condition) with REM.3, as a function of $criterion_{study}$. The parameters were $g_{base} = 0.4$, $g_{draw} = 0.38$, $w = 20$, $u_1 = 0.28$, $u_2 = 0.12$, $c = 0.7$, and $criterion_{test} = 1$.

Results and discussion

Figure 9 shows *d'*, hit rates, and false-alarm rates as a function of $criterion_{study}$. REM.3 replicates the concordant effect, regardless of $criterion_{study}$. Similarly, values of 3 or 4 for $criterion_{study}$ produce a list-strength effect, consistent with Norman's (1999) results. However, a value of 1 for $criterion_{study}$ produces a negative list-strength effect, and a value of 2 for $criterion_{study}$ produces a null list-strength effect.

General discussion

The purpose of the present work was to assess whether REM is viable when stripped of one of its simplifying assumptions. In REM.1, item repetitions are always stored in the same image; in REM.3, repetitions are superimposed on the original image only if subjects recognize the repetition as having previously been studied. In Simulation 1, we showed that, although a value of 1 for $criterion_{test}$ is reasonable, a value of 1 for $criterion_{study}$ is not. Because subjects are not explicitly asked to make old/new judgments at study, the threshold for superimposing a study item needs to be higher. In Simulation 2, we demonstrated that REM.3 correctly predicts an advantage for spaced over massed repetitions. In Simulation 3, we examined whether REM.3 predicts a negative or null list-strength effect with the mixed-pure paradigm. REM.3 actually predicts a slightly positive list-strength effect with spaced strengthening, but examination of published reports of the list-strength effect in recognition provides empirical support for this prediction. Finally, in Simulations 4 and 5, we showed that REM.3 can account for positive list-strength effects observed with the strong-interference paradigm.

Differentiation is a critical feature of a number of models of human memory (McClelland & Chappell, 1998; Norman & O'Reilly, 2003; Shiffrin & Raaijmakers, 1992; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997). The idea makes intuitive sense: As information becomes better learned, its mnemonic trace becomes more distinctive (i.e., more dissimilar from other mnemonic traces), and it is thus less likely to interfere with other memories. Consequently, although adding items to a list depresses memory (Strong, 1912), increasing the strength of some items on a list does not, at least when item recognition (Ratcliff et al., 1990) or cued recall (Wilson & Criss, 2017) are tested. Without differentiation, REM predicts both list-strength and list-length effects; with differentiation, REM predicts a list-length effect without a list-strength effect.

REM.3 provides a more complete picture. In REM.1, the simplifying assumption that repetitions are always accumulated in a single image, regardless of the time elapsed between presentations or the number of interpolated items, renders it unable to account for the spacing effect. Obviously, if massed and spaced repetitions are identical, they will be recognized at the same rate. Here we showed that making superimposition dependent upon subjects identifying the repetition as having previously been studied produces an advantage for spaced over massed items, consistent with the spacing effect (Delaney et al., 2010).

The list-strength effect is one of the phenomena responsible for the emergence of REM. Broadly, REM.1 is consistent with experiments on the list-strength effect: Strengthening some items on a list protects the weaker items from interitem interference, thus resulting in a negative list-strength effect. Yet, although REM.1 always predicts a negative list-strength effect, this is inconsistent with null and slightly positive list-strength effects observed in the literature with spaced strengthening (Ratcliff et al., 1990; Ratcliff et al., 1992). Similarly, the prediction of an increasingly negative list-strength effect as strong-item strength increases is inconsistent with positive list-strength effects observed with the strong-interference paradigm (Norman, 1999, 2002). We have demonstrated that REM.3 is consistent with these results.

A model with a number of similarities to REM.3 is a free-recall version of REM presented by Malmberg and Shiffrin (2005). Malmberg and Shiffrin demonstrated empirically that, in free recall, a positive list-strength effect is observed with

spaced strengthening but not massed strengthening. They explained this using the one-shot hypothesis of context storage, according to which a maximum amount of context information can be stored during any given study trial (see also Burgess, Hockley, & Hourihan, 2017). Consequently, presenting an item for 2 or 6 s will result in the same amount of context information being stored. This means that spaced items accumulate more context information than massed items.

We did not implement context in our REM.3 simulations. Of course, we assume that context exists, inasmuch as subjects limit search to images generated during the study episode. Indeed, it would have been relatively straightforward to add context information to images (e.g., Malmberg & Shiffrin, 2005; Mensink & Raaijmakers, 1988, 1989). However, although we do not dispute the role of context in the free-recall list-strength effect, we question its ability to describe recognition data. It is well known that, although changing contexts between study and test produces robust effects in free recall, similar manipulations have no effect in recognition. For example, studying a list on land and being tested under water (or vice versa) depresses free recall relative to studying and testing on land or under water (Godden & Baddeley, 1975), but this pattern does not extend to recognition (Godden & Baddeley, 1980). To achieve context effects in recognition, experimenters must use very salient, item-to-item context changes, such as presenting to-be-remembered words on picture backgrounds (Hockley, 2008; Murnane, Phelps, & Malmberg, 1999). For this reason, context-based explanations in free recall often do not generalize to recognition (see Sahakyan, Delaney, Foster, & Abushanab, 2013; Sahakyan & Kelley, 2002; see also the discussion in Ensor, Guitard, Bireta, Hockley, & Surprenant, 2019).

As Shiffrin and Steyvers (1997) noted, the addition of multiple images for spaced items makes REM more complex and, as a consequence, makes simulations less feasible. Although we acknowledge the importance of parsimony in computational modeling, in this case a relatively minor change in the model substantially increases its explanatory power. Moreover, it helps us evaluate a possible reason for some previously puzzling results in the literature. Finally, the assumption that repeated items are incorporated into the existing image only if the subject notices that they are repeated is a possibility that is testable.

Acknowledgments We thank K. Malmberg for sharing his REM.1 code. This work was supported by Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) awarded to A.M.S. and I.N., and an NSERC doctoral scholarship awarded to T.M.E. Parts of this work were presented at the 2018 meeting of the Society for Computers in Psychology in New Orleans, LA.

Open Practices Statement There are no data associated with this article. The source code for the REM.3 model is available from the first author.

References

Allen, L. R., & Garton, R. F. (1968). The influence of word-knowledge on the word-frequency effect in recognition memory. *Psychonomic Science*, *10*, 401–402. doi:https://doi.org/10.3758/BF03331581

Annis, J., Lenes, J. G., Westfall, H. A., Criss, A. H., & Malmberg, K. J. (2015). The list-length effect does not discriminate between models of recognition memory. *Journal of Memory and Language*, *85*, 27–41. doi:https://doi.org/10.1016/j.jml.2015.06.001

Bäuml, K.-H. (1997). The list-strength effect: Strength-dependent competition or suppression? *Psychonomic Bulletin & Review*, *4*, 260–264. doi:https://doi.org/10.3758/BF03209403

Burgess, N., Hockley, W. E., & Hourihan, K. L. (2017). The effects of context in item-based directed forgetting: Evidence for 'one-shot' context storage. *Memory & Cognition*, *45*, 745–754. doi:https://doi.org/10.3758/s13421-017-0692-5

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*, 369–378. doi:https://doi.org/10.1007/s10648-012-9205-z

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*, 231–248. doi:https://doi.org/10.1016/S0749-596X(03)00061-5

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380. doi:https://doi.org/10.1037/0033-2909.132.3.354

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*, 37–60. doi:https://doi.org/10.3758/BF03210740

Criss, A. H. (2006). The consequences of differentiation in episodic memory: Similarity and the strength based mirror effect. *Journal of Memory and Language*, *55*, 461–478. doi:https://doi.org/10.1016/j.jml.2006.08.003

Criss, A. H., & Howard, M. W. (2015). Models of episodic memory. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), The Oxford handbook of computational and mathematical psychology (pp. 165–183). New York, NY: Oxford University Press.

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

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, *64*, 316–326. doi:https://doi.org/10.1016/j.jml.2011.02.003

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In B. H. Ross (Ed.), The psychology of learning and motivation (Vol. 53, pp. 63–147). San Diego, CA: Elsevier Academic Press. doi:https://doi.org/10.1016/S0079-7421(10)53003-2

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*, 627–634. doi:https://doi.org/10.1037/0003-066X.43.8.627

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*, 452–478. doi:https://doi.org/10.1037/0033-295X.108.2.452

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*, 361–376. doi:https://doi.org/10.1016/j.jml.2008.06.007

Ebbinghaus, H. (1885). Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie. Leipzig: Duncker and Humboldt.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, *89*, 627–661. doi:https://doi.org/10.1037/0033-295X89.6.627

Eich, J. M. (1985). Levels of processing, encoding specificity, elaboration, and CHARM. *Psychological Review*, *92*, 1–38. doi:https://doi.org/10.1037/0033-295X.92.1.1

Ensor, T. M., Guitard, D., Bireta, T. J., Hockley, W. E., & Surprenant, A. M. (2019). The list-length effect occurs in cued recall with the retroactive design but not the proactive design. *Canadian Journal of Experimental Psychology*. Advance online publication. doi:https://doi.org/10.1037/cep0000187

Fritzen, J. (1975). Intralist repetition effects in free recall. *Journal of Experimental Psychology: Human Learning and Memory*, *1*, 756–763. doi:https://doi.org/10.1037/0278-7393.1.6.756

Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review*, *47*, 196–229. doi:https://doi.org/10.1037/h0060582

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, *62*, 32–41. doi:https://doi.org/10.1037/h0048826

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67. doi:https://doi.org/10.1037/0033-295X.91.1.1

Glanzer, M. (1969). Distance between related words in free recall: Trace of the STS. *Journal of Verbal Learning & Verbal Behavior*, *8*, 105–111. doi:https://doi.org/10.1016/S0022-5371(69)80018-6

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, *13*, 8–20. doi:https://doi.org/10.3758/BF03198438

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 5–16. doi:https://doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 21–31. doi:https://doi.org/10.1037/0278-7393.2.1.21

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*, 325–331. doi:https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Godden, D. R., & Baddeley, A. D. (1980). When does context influence recognition memory? *British Journal of Psychology*, *71*, 99–104. doi:https://doi.org/10.1111/j.2044-8295.1980.tb02735.x

Gorman, A. M. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*, 23–29. doi:https://doi.org/10.1037/h0040561

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1004–1011. doi:https://doi.org/10.1037/0278-7393.16.6.1004

Hastie, R. (1975). Intralist repetition in free recall: Effects of frequency attribute recall instructions. *Journal of Experimental Psychology: Human Learning and Memory*, *1*, 3–12. doi:https://doi.org/10.1037/0278-7393.1.1.3

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of repetitions. *Journal of Experimental Psychology*, *80*, 139–145. doi:https://doi.org/10.1037/h0027133

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments & Computers*, *16*, 96–101. doi:https://doi.org/10.3758/BF03202365

Hintzman, D. L. (1986). 'Schema abstraction' in a multiple-trace memory model. *Psychological Review*, *93*, 411–428. doi:https://doi.org/10.1037/0033-295X.93.4.411

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551. doi:https://doi.org/10.1037/0033-295X.95.4.528

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 302–313. doi:https://doi.org/10.1037/0278-7393.21.2.302

Hockley, W. E. (2008). The effects of environmental context on recognition memory and claims of remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1412–1429. doi:10.1037a0013016

Huber, D. E., Tomlinson, T. D., Jang, Y., & Hopper, W. J. (2015). The search of associative memory with recovery interference (SAM-RI) memory model and its application to retrieval practice paradigms. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, and M. Steyvers (Eds.), Cognitive modeling in perception and memory: A festschrift for Richard M. Shiffrin (pp. 81–98). New York, NY: Psychology Press.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*, 208–233. doi:https://doi.org/10.1037/0033-295X.96.2.208

Kılıç, A., Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2017). Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation. *Cognitive Psychology*, *92*, 65–86. doi:https://doi.org/10.1016/j.cogpsych.2016.11.005

Lehman, M., & Malmberg, K. J. (2011). Overcoming the effects of intentional forgetting. *Memory & Cognition*, *39*, 335–347. doi:https://doi.org/10.3758/s13421-010-0025-4

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning & Verbal Behavior*, *8*, 828–835. doi:https://doi.org/10.1016/S0022-5371(69)80050-2

Malmberg, K. J., & Murnane, K. (2002). List composition and the word-frequency effect for recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 616–630. doi:https://doi.org/10.1037/0278-7393.28.4.616

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 322–336. doi:https://doi.org/10.1037/0278-7393.31.2.322

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, *30*, 607–613. doi:https://doi.org/10.3758/BF03194962

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. M. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by midazolam. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 540–549. doi:https://doi.org/10.1037/0278-7393.30.2.540

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*, 724–760. doi:https://doi.org/10.1037/0033-295X.105.4.734-760

Melton, A. W. (1967). Repetition and retrieval from memory. *Science*, *158*, 532. doi:https://doi.org/10.1126/science.158.3800.532-b

Melton, A. W. (1970). The situation with respect to the spacing of repetitions in memory. *Journal of Verbal Learning & Verbal Behavior*, *9*, 596–606. doi:https://doi.org/10.1016/S0022-5371(70)80107-4

Mensink, G.-J. M., & Raaijmakers, J. G. W. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455. doi:https://doi.org/10.1037/0033-295X.95.4.434

Mensink, G.-J. M., & Raaijmakers, J. G. W. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology*, *33*, 172–186. doi:https://doi.org/10.1016/0022-2496(89)90029-1

Menzel, R., Manz, G., Menzel, R., & Greggers, U. (2001). Massed and spaced learning in honeybees: The role of CS, US, the intertrial

interval, and the test interval. *Learning and Memory*, *8*, 198–208. doi:https://doi.org/10.1101/lm.40001

Metcalfe, J. (1990). Composite Holographic Associative Recall Model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General*, *119*, 145–160. doi:https://doi.org/10.1037/0096-3445.119.2.145

Mumford, M. D., Costanza, D. P., Baughman, W. A., Threlfall, K. V., & Fleishman, E. A. (1994). Influence of abilities on performance during practice: Effects of massed and distributed practice. *Journal of Educational Psychology*, *86*, 134–144. doi:https://doi.org/10.1037/0022-0663.86.1.134

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609–626. doi:https://doi.org/10.1037/0033-295X.89.6.609

Murdock, B. B. (1983). A distributed memory model for serial-order information. *Psychological Review*, *90*, 316–338. doi:https://doi.org/10.1037/0033-295X.90.4.316

Murdock, B. B. (1989). Learning in a distributed memory model. In C. Izawa (Ed.), Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition (pp. 69–106). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Murdock, B. B., & Kahana, M. J. (1993a). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 689–697. doi:https://doi.org/10.1037/0278-7393.19.3.689

Murdock, B. B., & Kahana, M. J. (1993b). List-strength and list-length effects: Reply to Shiffrin, Ratcliff, Murnane, and Nobel (1993). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1450–1453. doi:https://doi.org/10.1037/0278-7393.19.6.1450

Murnane, K., Phelps, M. P., & Malmberg, K. J. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General*, *128*, 403–415. doi:https://doi.org/10.1037/0096-3445.128.4.403

Murnane, K., & Shiffrin, R. M. (1991a). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 855–874. doi:https://doi.org/10.1037/0278-7393.17.5.855

Murnane, K., & Shiffrin, R. M. (1991b). Word repetitions in sentence recognition. *Memory & Cognition*, *19*, 119–130. doi:https://doi.org/10.3758/BF03197109

Norman, K. A. (1999). Differential effects of list strength on recognition and familiarity (doctoral dissertation, Harvard University). Retrieved from ProQuest Dissertations & Theses. (1999-95024-412)

Norman, K. A. (2002). Differential effects of list strength on recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1083–1094. doi:https://doi.org/10.1037/0278-7393.28.6.1083

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *110*, 611–646. doi:https://doi.org/10.1037/0033-295X.110.4.611

Norman, K. A., Tepe, K., Nyhus, E., & Curran, T. (2008). Event-related potential correlates of interference effects on recognition memory. *Psychonomic Bulletin & Review*, *15*, 36–43. doi:https://doi.org/10.3758/PBR.15.1.36

Osth, A. F., Dennis, S., & Kinnell, A. (2014). Stimulus type and the list strength paradigm. *Quarterly Journal of Experimental Psychology*, *67*, 1826–1841. doi:https://doi.org/10.1080/17470218.2013.872824

Osth, A. F., Fox, J., McKague, M., Heathcote, A., & Dennis, S. (2018). The list strength effect in source memory: Data and a global matching model. *Journal of Memory and Language*, *103*, 91–113. doi:https://doi.org/10.1016/j.jml.2018.08.002

Pike, R. (1984). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review*, *91*, 281–294. doi:https://doi.org/10.1037/0033-295X.91.3.281

Raaijmakers, J. G. W., & Phaf, R. H. (1999). Part-list cuing revisited: Testing the sampling-bias hypothesis. In C. Izawa (ed.), On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model (pp. 87–104). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search in associative memory. In G. H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory *vol.* 14 (pp. 207–262). New York, NY: Academic Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, *88*, 93–134. doi:https://doi.org/10.1037/0033-295X.88.2.93

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*, 163–178. doi:https://doi.org/10.1037/0278-7393.16.2.163

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*, 763–785. doi:https://doi.org/10.1037/0278-7393.20.4.763

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*, 190–214. doi:https://doi.org/10.1037/0033-295X.83.3.190

Ratcliff, R., Sheu, C., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, *99*, 518–535. doi:https://doi.org/10.1037/0033-295X.99.3.518

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology*, *92*, 365–372. doi:https://doi.org/10.1037/h0032278

Ruch, T. C. (1928). Factors influencing the relative economy of massed and distributed practice in learning. *Psychological Review*, *35*, 19–45. doi:https://doi.org/10.1037/h0074423

Sahakyan, L., Abushanab, B., Smith, J. R., & Gray, K. J. (2014). Individual differences in contextual storage: Evidence from the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 873–881. doi:https://doi.org/10.1037/a0035222

Sahakyan, L., Delaney, P. F., Foster, N. L., & Abushanab, B. (2013). List-method directed forgetting in cognitive and clinical research: A theoretical and methodological review. In B. H. Ross (Ed.), The psychology of learning and motivation*, vol. 59* (pp. 131–189). San Diego, CA: Elsevier. doi:https://doi.org/10.1016/B978-0-12-407187-2.00004-6

Sahakyan, L., & Kelley, C. M. (2002). A contextual change account of the directed forgetting effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 1064–1072. doi:https://doi.org/10.1037/0278-7393.28.6.1064

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. doi:https://doi.org/10.1016/j.jml.2018.04.004

Schooler, L. J., Shiffrin, R. M., & Raaijmakers, J. G. W. (2001). A Bayesian model for implicit effects in perceptual identification. *Psychological Review*, *108*, 257–272. doi:https://doi.org/10.1037/0033-295X.108.1.257

Schulman, A. I. (1967). Word length and rarity in recognition memory. *Psychonomic Science*, *9*, 211–212. doi:https://doi.org/10.3758/BF03330834

Shiffrin, R. M., & Raaijmakers, J. G. W. (1992). The SAM retrieval model: A retrospective and prospective. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), Essays in honor of William K.

Estes: From learning processes to cognitive processes (Vol. 2, pp. 69–86). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 179–195. doi:https://doi.org/10.1037/0278-7393.16.2.179

Shiffrin, R. M., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1445–1449. doi:https://doi.org/10.1037/0278-7393.19.6.1445

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166. doi:https://doi.org/10.3758/BF03209391

Sirotin, Y. B., Kimball, D. R., & Kahana, M. J. (2005). Going beyond a single list: Modeling the effects of prior experience on episodic free recall. *Psychonomic Bulletin & Review*, *12*, 787–805. doi:https://doi.org/10.3758/BF03196773

Stretch, V., & Wixted, J. T. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1379–1396. doi:https://doi.org/10.1037/0278-7393.24.6.1379

Strong, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*, 447–462. doi:https://doi.org/10.1037/h0069812

Strong, E. K. (1916). The factors affecting a permanent impression developed through repetition. *Journal of Experimental Psychology*, *1*, 319–338. doi:https://doi.org/10.1037/h0074989

Tomlinson, T. D., Huber, D. E., Rieth, C. A., & Davelaar, E. J. (2009). An interference account of cue-independent forgetting in the no-think paradigm. *Proceedings of the National Academy of Sciences*, *106*, 15588–15593. doi:https://doi.org/10.1073/pnas.0813370106

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*, 1–12. doi:https://doi.org/10.1037/h0080017

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology*, *92*, 297–304. doi:https://doi.org/10.1037/h0032367

Underwood, B. J. (1969). Some correlates of item repetition in free-recall learning. *Journal of Verbal Learning & Verbal Behavior*, *8*, 83–94. doi:https://doi.org/10.1016/S0022-5371(69)80015-0

Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society*, *12*, 89–91. doi:https://doi.org/10.3758/BF03329636

Verde, M. F. (2009). The list-strength effect in recall: Relative-strength competition and retrieval inhibition may both contribute to forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 205–220. doi:https://doi.org/10.1037/a0014275

Verde, M. F. (2013). Retrieval-induced forgetting in recall: Competitor interference revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1433–1448. doi:https://doi.org/10.1037/a0032975

Verkoeijen, P. P. J. L., & Delaney, P. F. (2008). Rote rehearsal and spacing effects in the free recall of pure and mixed lists. *Journal of Memory and Language*, *58*, 35–47. doi:https://doi.org/10.1016/j.jml.2007.07.006

Wilson, J. H., & Criss, A. M. (2017). The list strength effect in cued recall. *Journal of Memory and Language*, *95*, 78–88. doi:https://doi.org/10.1016/j.jml.2017.01.006

Wixted, J. T., Ghadisha, H., & Vera, R. (1997). Recall latency following pure- and mixed-strength lists: A direct test of the relative strength model of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 523–538. doi:https://doi.org/10.1037/0278-7393.23.3.523

Xue, G., Mei, L., Chen, C., Lu, Z., Poldrack, R., & Dong, Q. (2011). Spaced learning enhances subsequent recognition memory by reducing neural repetition suppression. *Journal of Cognitive Neuroscience*, *23*, 1624–1633. doi:https://doi.org/10.1162/jocn.2010.21532

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 345–355. doi:https://doi.org/10.1037/0278-7393.18.2.345