Cognitive Psychology 92 (2017) 65–86

Contents lists available at ScienceDirect

Cognitive Psychology

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

Models that allow us to perceive the world more accurately also allow us to remember past events more accurately via differentiation ☆

Aslı Kılıç [a,*], Amy H. Criss [b], Kenneth J. Malmberg [c], Richard M. Shiffrin [d]

[a] Department of Psychology, Middle East Technical University, Dumlupinar Blv. No. 1, Çankaya, Ankara 06800, Turkey
[b] Department of Psychology, 430 Huntington Hall, Syracuse University, Syracuse, NY 13244, USA
[c] Department of Psychology, PCD 4118G, University of South Florida, 4202 East Fowler Avenue, Tampa, FL 33620, USA
[d] Department of Psychological and Brain Sciences, Indiana University, 1101 E. 10th Street, Bloomington, IN 47405, USA

ARTICLE INFO

Article history:
Accepted 14 November 2016
Available online 28 November 2016

Keywords:
Recognition memory
Memory models
The strength based mirror effect
Output interference

ABSTRACT

Differentiation is a theory that originally emerged from the perception literature and proposes that with experience, the representation of stimuli becomes more distinct from or less similar to the representation of other stimuli. In recent years, the role of differentiation has played a critical role in models of memory. Differentiation mechanisms have been implemented in episodic memory models by assuming that information about new experiences with a stimulus in a particular context accumulates in a single memory trace and these updated memory traces become more distinct from the representations of other stimuli. A key implication of such models is that well encoded events are less confusable with other events. This prediction is particularly relevant for two important phenomena. One is the role of encoding strength on memory. The *strength based mirror effect* is the finding of higher hit rates and lower false alarm rates for a list composed of all strongly encoded items compared to a list composed of all weakly encoded items. The other is *output interference*, the finding that accuracy decreases across a series of test trials. Results from four experiments show a tight coupling between these two empirical phenomena such that strongly encoded target items are less prone to interference. By proposing a process model and evaluating the predictions of the model, we show how a single theoretical principle, differentiation, provides a unified explanation for these effects.

1. Introduction

We present a unified theoretical account of how differentiation, originally proposed for the domain of perception, operates during the encoding of new memories and the retrieval of old memories. Perceptual representations represent present experience and memory representations represent previous events. What one experiences in the present is stored in a representation that can be retrieved later. A number of theoretical approaches to memory have borrowed important assumptions from theories of perception and there are important, well-established phenomena that occur in both memory and

☆ Parts of this work were presented in partial fulfillment of Aslı Kılıç's doctoral dissertation at Syracuse University. We thank Marc Howard, Brad Wyble, and Natalie Russo for serving in the committee. We also thank Jeff Starns and William Hockley for their feedbacks on earlier versions of this work.
* Corresponding author.
E-mail address: askilic@metu.edu.tr (A. Kılıç).

http://dx.doi.org/10.1016/j.cogpsych.2016.11.005
0010-0285/

perception (e.g., Annis & Malmberg, 2013; Green & Swets, 1966; Malmberg & Annis, 2012). Here, we focus on a process known as *differentiation*, which improves the accuracy with which stimuli are perceived and events are remembered. We begin with a review of the role differentiation plays in perception and next turn our attention to role differentiation plays in memory.

1.1. Differentiation

Differentiation was brought to the study of memory via research on perception. The basic theory is that experience accumulates in the mind, which is expressed when previous exposure to a stimulus alters perception of the current experience with the stimulus (e.g., Adolph & Kretch, 2015). In a classic study, Gibson and Gibson (1955) asked participants to study a target (a spiral shape) and then make same/different judgments to series of items. Some of the items were obviously different (e.g., triangular or cloud shaped doodles) and those were easy to reject. Other stimuli were visually similar (spirals) but differed on a variety of dimensions (number of coils, width, etc.). Over multiple trials and without feedback, participants learned to discriminate the similar stimuli from the target stimulus. That is, through repetition, participants learned new information about the individual items so that what was once similar became more dissimilar. This is an example of differentiation improving perception.

The hypothesis that differentiation improves cognitive capabilities has been adopted in category learning, semantic knowledge, and episodic memory. In learning to categorize stimuli, differentiation may take the form of learning to differentially weigh relevant feature dimensions (e.g., Nosofsky, 1987) or "dimension differentiation" where once integrated feature dimensions become separate (e.g., Goldstone & Styvers, 2001) or changes in self-similarity (Nosofsky & Zaki, 2003). These same types of principles are part of the Rogers and McClelland (2008) model of semantic knowledge development, a connectionist system where the initial representation of items is similar but becomes differentiated with interleaved learning of many exemplars across many categories. They demonstrated that the notion that the learning of broad categories (e.g., plants vs. animals) sets the foundation for learning progressively more specific knowledge structures as individual items become differentiated from one another (e.g., fish vs. birds, then cardinals vs. robins).

Note that the benefits of differentiation in perceptual tasks are rooted in the effect experience has on memory, namely learning. Repeated exposures to perceptual stimuli create rich and more detailed conceptual representation of the stimuli, which allows stimuli encountered in the future to be identified and used in more nuanced manners.

In turn, memory researchers have asked how repeated exposure to similar events affects memory for the events themselves. Differentiation in episodic memory was initially leveraged to describe the pattern where similar items harmed paired-associate learning (e.g., Gibson, 1940) whereas items that were very different benefitted episodic memory (e.g., Wallace, 1965). Despite this early recognition of the potential role of differentiation in episodic memory, it was not formalized in models of memory until decades later. Differentiation was implemented in memory models by assuming that additional encoding leads to the storage of additional information in a single episodic memory trace representing the same event. This was in contrast to the classic memory theories assuming that additional encoding resulted in the storage of multiple memory traces.

In the differentiation models, as a memory trace is updated, it decreases in similarity to other, randomly similar items (e.g., Criss & McClelland, 2006; Malmberg, Holden, & Shiffrin, 2004; McClelland & Chappell, 1998; Shiffrin, Ratcliff, & Clark, 1990; Shiffrin & Steyvers, 1997). As an illustration, assume there are two empty episodic traces, each of which contains no information about the items that they represent, and hence they are exactly alike. As we allow for encoding of features representing two randomly selected items, the traces themselves become more dissimilar, hence distinguishable during retrieval. The more information is stored in each trace, the result of repetition, additional study time, or deeper encoding tasks, the more dissimilar they become.

Memory is often tested in the laboratory by having subjects study lists of items, and later presenting them with targets and foils, and subject are asked to discriminate the targets from the foils. The accuracy of the recognition task increases as the proportion of positive endorsements of targets increases (hit rate; HR) and/or the proportion of positive endorsements of foils decreases (false alarm rate; FAR). Bias is the tendency to endorse test items. As bias increases via a decrease in a criterion used to decide whether at item was studied, both HRs and FARs increase.

The effect of differentiation in such experiments is twofold: Items that were studied under strong encoding conditions and later tested are more likely to be recognized because the retrieval cues representing such targets match the memory trace that represent the occurrence better than the retrieval cues associated with weakly encoded targets match the memory trace that represent their occurrence. Hit rates are greater for strong targets than for weak targets. For instance, increasing the number of presentations of an item or manipulating the "depth" of encoding items receive, increases the HR. Second, when retrieval cues representing unstudied items (i.e., foils) are matched against traces stored under strong encoding conditions, incorrect recognition of their prior occurrence (i.e., false alarms) decreases because they mismatch the well-encoded traces more than traces stored under weak encoding conditions. False-alarm rates are lower for strong foils than for weak foils. This pattern of greater hit rates and lower false-alarm rates under strong encoding conditions is known as the *strength-based mirror effect* in the recognition memory literature lists (Cary & Reder, 2003; Criss, 2006, 2009, 2010; Glanzer & Adams, 1985; Starns, White, & Ratcliff, 2010; Stretch & Wixted, 1998).

As reviewed in Criss and Koop (2015), several converging findings support the presence of differentiation in episodic memory, including patterns of response time distributions (Criss, 2010), direct ratings of memory strength (Criss, 2009),

the interaction between test item similarity and list strength (Criss, 2006; Murnane & Shiffrin, 1991), the pattern of activations of memory-related brain activity as measured by fMRI (Criss, Wheeler, & McClelland, 2013), and differential ERPs for strong and weak foils in memory-based components (Chen, Lithgow, Hemmerich, & Caplan, 2014; Hemmer, Criss, & Wyble, 2011). However, recognition accuracy is unaffected by the strength with which *other* target items were encoded. In fact, weak targets are unaffected when studied on mixed-lists composed of both weak and strong items compared to pure-weak lists and likewise for strong items (Ratcliff, Clark, & Shiffrin, 1990). This is known as a null *list-strength effect*. By increasing the strength of some items in memory, the similarity of those traces to traces of other items decreases, and this offsets the increase in interference that might otherwise be caused by increases in the variability of encoding. Hence, differentiation can account for the unique pattern of recognition phenomena known as strength-based mirror effects and list-strength effects.

1.2. Criterion shift

In contrast to differentiation, an alternative explanation for the list-strength effect and strength-based mirror effect is a shift in the decision criterion. Within this framework, mirror effects are attributed to a complex interaction between the strength with which items are encoded and shift decision strategies that affect response bias (Starns, Ratcliff, & White, 2012; Stretch & Wixted, 1998; Verde & Rotello, 2007). Accordingly, the mean of the memory evidence distribution for targets increases with increases in encoding strength, and this tends to increase HRs, whereas the mean of the memory evidence distribution for foils is independent of encoding strength (e.g., Parks, 1950, see Fig. 1 or increases with strength in the global matching models, not pictured). Thus, there is a very special situation for which a shift in response bias should produce a mirror effect: If the increase in response bias is less than the increase in the mean of the evidence distribution for targets, then the FARs will be lower for the strong foils than the weak foils and the HRs will be higher for the strong targets than weak targets (Cary & Reder, 2003; Hirshman, 1995; Starns et al., 2012; Verde & Rotello, 2007). If, for whatever reason, the participants do not shift their response bias just right, then a mirror effect will not be observed.

1.3. Output interference

In life, stimuli are not only encountered but past events are retrieved from memory, which may also influence the strength with which past memory is encoded or the contents of the memory trace. The rapidly increasing literature on reconsolidation, for instance, suggests the possibility that the memories stored in the distant past can be modified or updated with new information when they have been retrieved. In recent research, we have focused on the consequences of testing memory (see Malmberg, Lehman, Annis, Criss, & Shiffrin, 2014 for a review). Output interference (OI) is the finding that accuracy decreases as the number of episodic test trials increases (Annis, Malmberg, Criss, & Shiffrin, 2013; Criss, Malmberg, & Shiffrin, 2011; Koop, Criss, & Malmberg, 2015; Malmberg & Annis, 2012; Malmberg, Criss, Gangwani, & Shiffrin, 2012; Murdock & Anderson, 1975; Ratcliff & Murdock, 1976).

Across several experiments, we observed that HRs decrease substantially and FARs increase slightly or remain relatively stable across test trials. To model OI, Criss et al. (2011) made some assumptions concerning the consequences of testing memory on the state of memory. Most importantly, they assumed that when memory is tested either a new trace representing the test trial is stored or a trace that was stored during the study phase was updated with information concerning the test trial. When a test item is judged to be new, a new memory trace is stored, resulting in the storage of a new memory trace for missed targets and correctly rejected foils. When an item is judged to be old, the best matching trace in memory is updated with features comprising the current test item. More complicated implementations are possible, e.g., updating the best matching memory trace only if it matches beyond some threshold, updating a trace and storing a new trace, updating based on the overall familiarity of the item (odds ratio) rather than the decision threshold, etc., but this simplified model was sufficient for the data. The main implication is that storing new traces during test is similar to increasing the study list length and causes a decrease in the HR and an increase in the FAR but updating an existing memory trace results in differentiation, which decreases the HR of subsequently tested targets and decreases the FAR. The net is a decrease in the HR and a flat FAR, exactly as observed in the empirical data.

Within our theoretical framework (REM, Shiffrin & Steyvers, 1997), there is a deep theoretical connection between the empirical findings described above. Critically, the same principle of differentiation causes the list-strength effect, strength-based mirror effect, and output interference. An alternative to the differentiation-based explanation of OI, the attention hypothesis, suggests that OI is the result of waning vigilance to the recognition task (Dennis & Humphreys, 2001; Underwood, 1978). The idea is that participants may essentially perform poorly because they are bored and/or wish to move onto other activities, and this slowly builds up as the testing proceeds. One obvious consequence of waning attention is a speed-accuracy trade-off. Accuracy may appear to decrease not because memory is worse, but instead because participants are simply responding prior to collecting sufficient evidence to make an informed decision. Hence, several models posit different mechanisms for the strength based mirror effect and OI, including a criterion-shift for the strength based mirror effect and waning vigilance for OI.

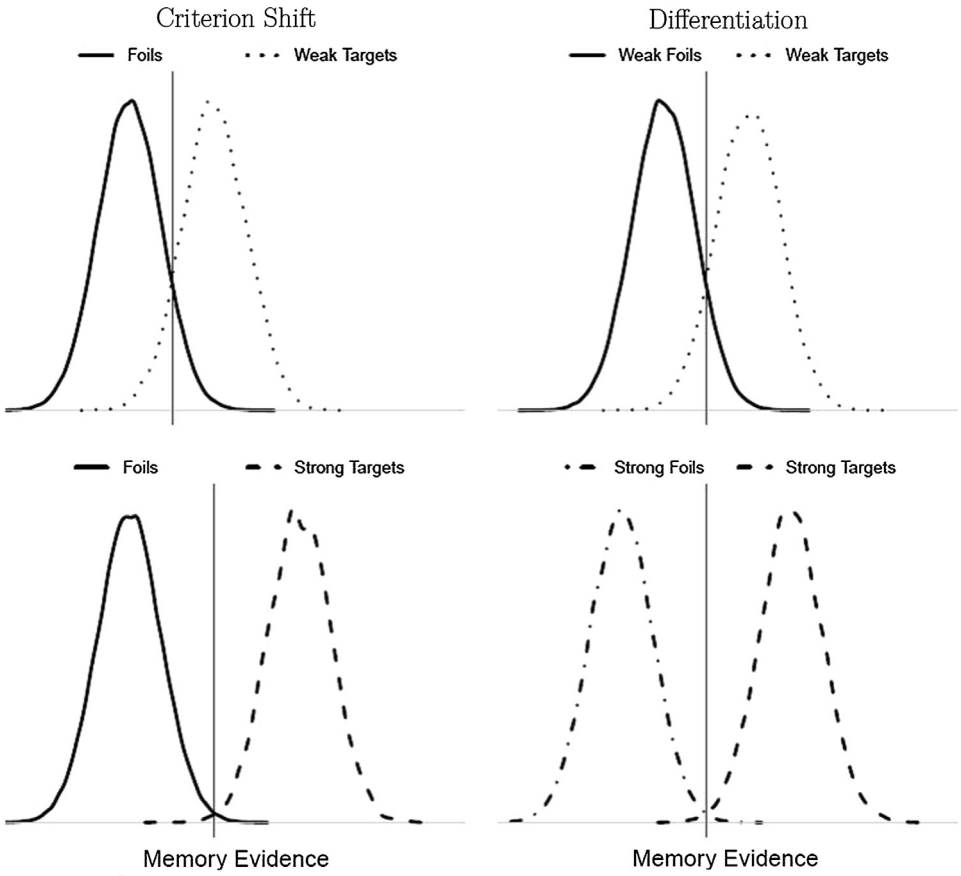

Fig. 1. Illustration of the two theoretical explanations for the strength based mirror effect.

1.4. Approach

The aim is to build on theory in the perception literature in order to relate memory and perception via common principles. Specifically, we will detail a unified theoretical framework that describes the influence of differentiation on recognition memory. Unlike prior models of differentiation that attempted to account for the effect of strengthening operations during study or the effect of prior testing on subsequent testing, the current model does both. Despite previous applications of the general REM framework to the SBME and OI, that these two phenomena were related via the differentiation mechanism had not been realized. Our aim here is to show a single differentiation mechanism accounts for these two distinct phenomena as well as the complex relationship between these two phenomena. First, we will present how the effects of strengthening operations and recognition testing are characterized in the model. Then predictions of the model were tested in four experiments, where study items were strengthened via levels-of-processing manipulation.

2. Retrieving effectively from memory model

In the full format, REM contains lexical semantic traces representing the long-term knowledge about information in the world as well as incomplete and inexact episodic traces representing a specific experience. Both traces contain item information describing features of the items and episodic traces also contain context features that refer to the specific internal and external circumstances in which the item was experienced. In the current simulations, we will focus on only the item information in the episodic traces thus only the item-noise mechanism of REM will be tested, following typical applications of REM to recognition memory (Criss & Shiffrin, 2004; Criss et al., 2011; Shiffrin & Steyvers, 1998). The length of the vector for each item is fixed at 20 and each feature value is a positive integer sampled independent of one another from the geometric distribution as follows:

$$P(v) = (1-g)^{v-1}g, \quad v = 1, \ldots, \infty, \tag{1}$$

where g is the parameter for the geometric distribution and v is the actual sampled value for each element in the vector.

One of the basic assumptions in REM is that the items are stored in episodic memory probabilistically and with error. During study, each feature is stored with some probability (u). For each feature value stored in memory, another parameter moderates the probability of correctly copying the feature value ($c = 0.70$). If a feature is not correctly copied, then a random value, sampled from the same geometric distribution is then stored in the trace. The parameter u varies because it reflects encoding strength, which is determined by the experimental design among other factors such as characteristics of the participant. In the following simulations, we fit u separately to the strong and weak conditions and further, we varied u during study and test (e.g., in accord with the testing effect see Karpicke & Roediger, 2008; Roediger & Karpicke, 2006). A higher u for strongly encoded lists indicates more stored information on which to base a decision. The result is differentiation: targets match their corresponding memory trace better and foils have more opportunities to mismatch. As a result, the mean memory evidence of a strong target distribution is greater than the mean memory evidence of a weak target distribution. The foils on the other hand are poor matches to the contents of episodic memory (Fig. 1). The mismatch grows as the memory traces are better encoded because the traces of those items are more complete and thus have more features that mismatch the test item. Features that are not stored remain zero, which simply indicates a lack of information and play no role in the decision process described next.

In REM, at retrieval, the test item is matched to all of the traces in memory and a subjective likelihood is calculated from the following equation:

$$\lambda_{(i,j)} = (1-c)^{nq_{(i,j)}} \prod_{v=1}^{\infty} \left[\frac{c + (1-c)g(1-g)^{v-1}}{g(1-g)^{v-1}} \right]^{nm(v,i,j)} \tag{2}$$

where j indexes the test item, i indexes the memory trace, c is the probability of correctly copying the feature value, nq is the number of non-zero feature mismatches, v is the feature value (sampled from the geometric distribution) and nm is the number of non-zero feature matches. To make a decision, the subjective likelihood ratios are averaged across traces stored in memory to form an odds (Φ) value:

$$\Phi_j = \frac{1}{n} \sum_{i=1}^{n} \lambda_{(i,j)}, \tag{3}$$

where n is the number of traces in the memory search set, i indexes the memory trace and j indexes the test item. If Φ exceeds the *criterion*, item j is judged to be 'old', if Φ is below the *criterion*, then the item j is judged to be 'new'. From a signal detection perspective, Φ can be considered the memory evidence and *criterion* is the threshold for endorsing an item, which has an optimal value of 1.

Criss et al. (2011) described the encoding during testing mechanism. When an item is judged to be old based on the *criterion*, the best matching episodic trace is updated (e.g., any missing features are subject to being stored according to the above described encoding mechanism). When an item is judged to be new, a new memory trace is stored. This simplified encoding during test mechanism is sufficient to predict output interference (Criss et al., 2011).

2.1. Simulation 1: the role of differentiation on list strength and output interference

In this set of simulations, the role of differentiation on list strength and OI was examined independently. Fig. 2 shows the HR and FAR from the simulations of REM where study lists are strengthened between lists (pure condition) and within lists (mixed condition). To preview, the simulations are consistent with the empirical findings of OI and an SBME that have been reported numerous times (e.g., Criss et al., 2011; Ratcliff et al., 1990; Starns et al., 2010). The primary purpose is to show the basic effects and to demonstrate that differentiation is the key factor in predicting the pattern of OI observed in the literature.

In the pure-list condition, 150 items were either all strongly or all weakly encoded. In order to control for study list-length across simulations and experiments, 75 targets were tested along with 75 foils. As mentioned earlier, encoding strength is simulated by varying u, which controls the probability of storing the traces of studied items. For the weak condition, u was set to 0.2 and for the strong condition it was 0.4. Other parameters of REM were fixed across simulations and the same parameter values were used as in Criss et al. (2011, $n = 20$, $g = 0.35$, $c = 0.7$, *criterion* = 0.72). An increase in the u parameter value shows the SBME; increase in the HRs and decrease in the FARs, as expected. REM predicts a SBME in the pure condition as the foils are compared to a list of strong traces when all the items are strengthened during study, as represented in Fig. 2. The FARs decrease as a function of list strength because the likelihood of a match between the foils and strong memory traces is relatively low compared to the match between the foils and weak memory traces.

When items were strengthened in mixed study lists, REM does not produce the SBME. In the mixed condition, 75 of the study items were strengthened ($u = 0.4$) and the other half was weakly encoded ($u = 0.2$). As the HRs from Fig. 2 shows strengthening items increased the HRs, however, strengthening within lists did not have an effect on the FARs. REM does not predict a SBME for a mixed list because the likelihood of a match between a foil and traces in memory is comparable across test lists due to similar strength levels of memory traces. This simulation shows the absence of the SBME when differentiation is eliminated during encoding of the study lists.

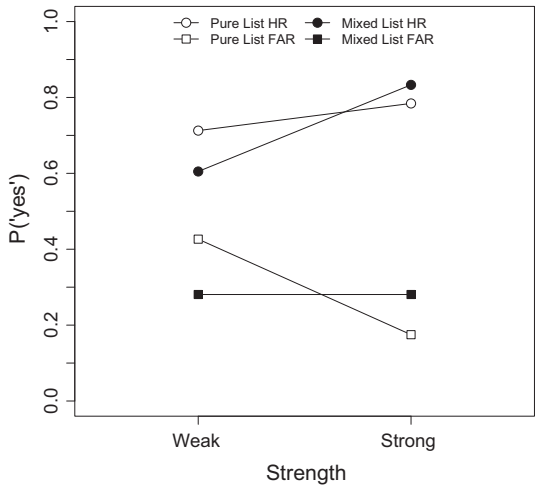

Fig. 2. Predicted probability of endorsing a test item from the list strength Simulation 1. The first panel shows the mixed study list condition where only half of the items were strengthened. In the middle panel, hit rates and false alarm rates are presented for the pure list condition where only weak items were studied. The left panel presents hit rates and false alarm rates in the pure list condition where only strong items are studied. The parameter values are as follows: $n = 20$, $g = 0.35$, $c = 0.7$, *criterion* = 0.72, $u_{strong} = 0.4$, and $u_{sweak} = 0.2$. HR stands for hit rate and FR stands for false alarm rate.

The next simulation, Fig. 3, shows the role of differentiation during test. When differentiation is eliminated during encoding of the test items, REM predicts a different pattern of OI compared to when differentiation is implemented. In the OI implementation (see Criss et al., 2011), it is assumed that when a test item is endorsed, the best matching trace is updated, which results in differentiated memory traces for subsequent test trials. Thus, it is possible to remove differentiation in OI by disabling the updating mechanism such that for every tested item (endorsed or not), a new trace is added to memory. Fig. 3 shows predicted HR and FAR as a function of test block across two implementations of OI in REM with the same parameters used above (and u = 0.20 for the study list). Here, the test positions are binned into 10 test positions. In the differentiation condition, the best matching trace is updated with $u = 0.2$ for every endorsed item. On the contrary, in the no-differentiation condition the traces of all the tested items are added to memory with $u = 0.2$. The results from the simulations show that the specific pattern of OI is different when differentiation is absent compared to being present during retrieval. Only the model with differentiation is consistent with human behavior – that is a rather large decrease in HR and a relatively flat FAR (e.g., see Koop et al., 2015).

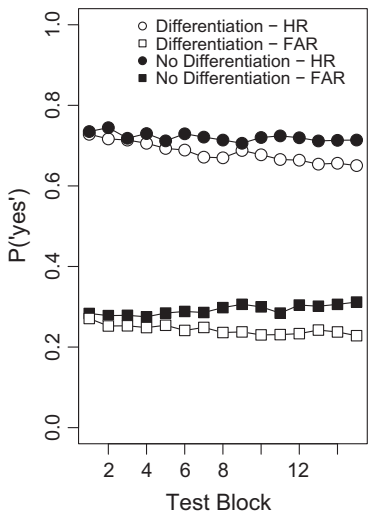

Fig. 3. Predicted probability of endorsing a test item as a function of test block from the output interference Simulation 1. In the Differentiation condition, the best matching traces of all the items that are endorsed are updated. In the No Differentiation condition all the test items are added as a new trace in memory. The parameter values are as follows: $n = 20$, $g = 0.35$, $c = 0.7$, *criterion* = 0.72, u at test = 0.35. HR stands for hit rate and FR stands for false alarm rate.

First consider the no-differentiation implementation. This is simply a list length effect – adding traces to memory increases the noise resulting in a slight increase in the FAR and slight decrease in the HR.

Now consider the implementation with differentiation. When an item is judged to be old, the best matching episodic trace is updated. Updating generally reduces the match between any future test item and the updated memory trace, thus when many items are updated there should be a reduction in both the HR and FAR across test position. The effect on HR is even more substantial when the updated memory trace does not match the test item (i.e., false alarms and an occasional hit where the wrong memory trace is updated) because the trace is contaminated with features from a different item and this will reduce the ability to accept the target that will be tested later in the list. When an item is judged to be new, a new memory trace is stored. Storing new traces generally results in a small increase in FAR and decrease in HR (e.g., a list length effect as described in the no-differentiation implementation). Specifically, when a memory trace for a foil is added (i.e., correct rejection), noise is added to the memory evidence increasing the FAR and decreasing the HR. When a second memory trace for a target is added to the list (e.g., a miss), the probability of endorsing that specific target (if it is tested later in the list) would be higher; otherwise this contributes to a decrease in the HR decrease and increase in the FAR.

2.2. Simulation 2: output interference and list strength

In this simulation, we examine the combined effect of list strength and encoding during test. The differentiation principle predicts a change in the pattern of OI as a function of encoding strength. Fig. 4 presents the predicted HRs and FARs from REM obtained by simulating 1000 participants and averaging data from the simulated participants. In this simulation, the study and test lists contain a single value of encoding strength. 150 items were studied (all strongly encoded or all weakly encoded) and 75 of those items were tested randomly along with 75 foils. Test blocks are binned into 15 test positions and the predicted HRs and FARs are plotted as a function of test block. The effect of strengthening was simulated by varying the u parameter during study. Greater values of u (e.g., 0.5) produced more complete memory traces whereas lower values (e.g., 0.2) produced weak traces which were not adequately filled with the features of the studied items. Other parameters of REM were fixed across simulations and did not differ from their conventional values ($n = 20$, $g = 0.35$, $c = 0.7$, $criterion = 0.72$, $u_{Test} = 0.35$). An increase in the u parameter value shows the SBME; increase in the HRs and decrease in the FARs, as expected.

The novel finding here is that increasing the u parameter changes the pattern of OI. It is this prediction that had not been discovered with models separately investigating OI and the SBME. Increasing encoding strength (u) reduces the degree to which the HR decreases across test block until it is nearly flat for exceptionally high learning rates (a value of 0.5 for u is greater than any value reported in the literature to the best of our knowledge). Likewise, the stronger the encoding of the target items, the less FAR decreases across test block. In fact, the FAR starts to increase a bit with high encoding due to adding memory traces for correct rejections on nearly every foil trial.

To understand the pattern of data in Fig. 4, consider the following. When the items are encoded strongly during study, fewer FARs are committed which means that more foil items are added as new memory traces (correct rejections). Adding new memory traces during the test tends to slightly increase the FAR (e.g., a list length effect), which works against the tendency for the FAR to decrease due to updating (differentiation). These factors combine with adding new traces playing a bigger role as the encoding strength of the list increases. On the relatively rare times when a foil item is endorsed despite the strongly encoded study items, there are fewer features in the memory trace that are available to be incorrectly updated with the features of the test item so the updating does not cause much interference for a target that is tested in the future. As a result, the subsequent targets are still likely to match due to their superior encoding during study. These various factors work together to change the degree of OI for targets and foils. However, the resulting discriminability over the course of the test list does not change much. These simulations show that REM predicts a different pattern of OI for different list strengths values such that with strong initial encoding of the list, the smaller the magnitude of change for both HR and FAR across test block.

2.3. Simulation 3: SMBE and OI with encoding strength varied, but strength of the tested targets fixed

In this simulation, the effect of testing conditions on OI was investigated by manipulating item strength within a study list and testing either the strong or weak targets, but not both in the same test list. The original REM formulation did not include encoding during test. That model does not predict the SBME after studying a list of items with mixed strength when the decision is solely based on the memory evidence from encoding. In such a situation, the foil items are each compared to all the episodic traces, half of which are strong and half are weak. Thus, the foils produce a comparable odds value when tested along with strong targets as when tested along with weak targets. In contrast, Starns et al. (2012) found a SBME for the strong test list vs. the weak test list following a mixed study list (cf. Kılıç & Öztekin, 2014). They concluded that (a) the differentiation principle is not necessarily required to account for the SBME and (b) the observed SBMEs for both mixed and pure study list conditions could possibly result from changes in criterion placement. This simulation addresses whether the implementation of encoding via differentiation during test results in a different predicted outcome.

The predictions in Fig. 5 were obtained from simulating 1000 participants. The study list consists of 75 strong and 75 weak items. At test, in the strong condition, only the strong targets were tested along with randomly intermixed foils and in the weak condition, only the weak targets were used to construct the test list. In the simulations, the u parameter was set to 0.2 for the weak condition in all of the simulations. The u parameter was varied from 0.3 to 0.5 for the strong condition.

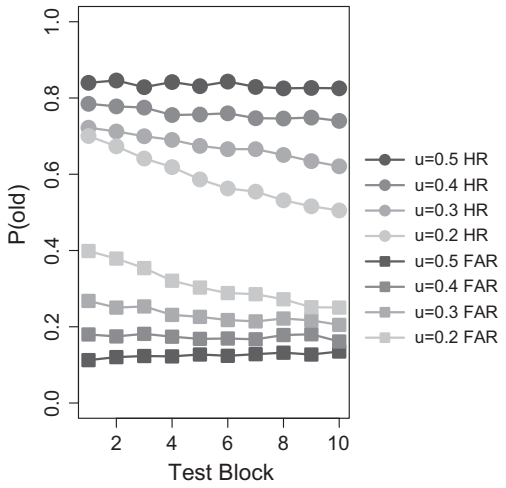

Fig. 4. The results of Simulation 2 showing the predicted probability of endorsing a test item as a function test block when items are strengthened across lists. The parameter values are as follows: $n = 20$, $g = 0.35$, $c = 0.7$, *criterion* = 0.72, u at test = 0.35. HR stands for hit rate and FR stands for false alarm rate.

The other parameter values were the same as the ones used in Simulation 1 ($n = 20$, $g = 0.35$, $c = 0.7$, *criterion* = 0.72, $u_{Test} = 0.35$).

In Fig. 5, HRs and FARs are plotted as a function of test block. The words in each label (Strong, Weak) indicate the strength of the target being tested and the *uS* parameter indicates the learning rate of the strong items during encoding. For example Weak HR (uS = 0.3) shows the HR for a test list composed of foils and weak targets (encoded with u = 0.2 during study) that were studied along with medium strength strong targets (encoded with u = 0.3 during study). HR and FAR were affected by the overall strength of the list, that is increasing the learning parameter for strong targets decreases the FAR and increases the strong HR. However, the FARs were minimally affected by the strength of the test list (e.g., compare Strong FR, $u_{strong} = 0.5$ versus Weak FR, $u_{strong} = 0.5$). Thus, in order to account for a SBME for mixed study but pure test lists in REM, we'd need to incorporate a strategic criterion shift presumably induced by the instructions. More importantly, the pattern of OI is not affected by the degree of encoding during study when the study list was of mixed strength. In other words, the changing pattern of HR and FAR in Fig. 4 for pure strength study lists is not observed here. This stems from the similar levels of FARs across strong and weak conditions and from the presence of poorly encoded memory traces in all conditions. The degree to which memory traces are contaminated with features from other items (i.e., when a false alarm is committed) is approximately the same regardless of whether strong or weak targets are tested and the contaminated traces tend to be the weakly encoded traces. Thus, unlike strongly encoded lists, there aren't sufficient correct rejections that would temper the decrease in the FAR across test block.

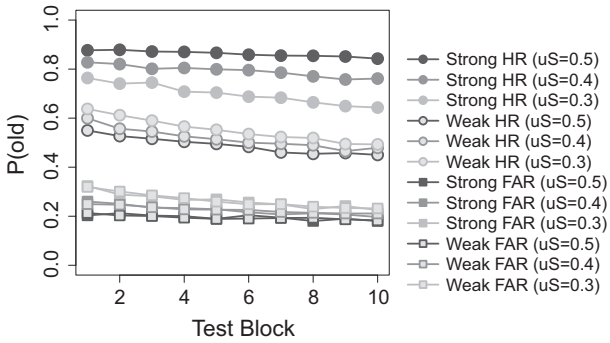

Fig. 5. The results of Simulation 3, plotting the predicted probability of endorsing a test item as a function test block when items are strengthened within lists, but tested in pure lists. Three study list conditions were simulated in all of which u for the weak encoding condition was 0.2. For the strong encoding condition, u was 0.3, 0.4 and 0.5. Strong HR is the hit rate and strong FR is the false alarm rate obtained from the test list that is composed of strong targets and foils. Weak HR is the hit rate and weak FR is the false alarm rate obtained from the test list that is composed of weak targets and foils. The u_S values in parentheses indicate the encoding condition of the strong items. For example, weak HR ($us = 0.5$) is the hit rate of the weak test list in which the weak targets ($u = 0.2$) are studied along with items that were encoded with $u = 0.5$. The other parameter values are as follows: $n = 20$, $g = 0.35$, $c = 0.7$, *criterion* = 0.72, u at test = 0.35.

We test these predictions in the following set of Experiments with the goal of better understanding the effects of encoding strength during study and test on episodic recognition memory. No existing model accounts for this full set of data; here we present a unified differentiation-based account.

3. Experiment 1

Simulation 2 showed that strengthening items during study results in different patterns of OI during recognition testing. To investigate the relationship between list strength and OI, we manipulated strength between lists in a pure-study paradigm. In the strong-list condition, subjects studied words with the instructions to make a semantic judgment, and in the weak-list condition subjects studied words with the instruction to make an orthographic judgment. The SBME data were analyzed as a function of test trial position.

3.1. Methods

3.1.1. Participants

Thirty-four undergraduate students from the Syracuse University Research Participation Pool took part in the experiment. Twenty-five of the participants were female. Six participants who had close to chance level accuracy ($d' < 0.5$) were excluded from the subsequent analyses.

3.1.2. Materials

The word pool consisted of nouns selected from MRC Psycholinguistics Database with a range of Kucera and Francis (1967) written frequency between 4 and 400 (M = 39.77), and number of letters between 4 and 8 (Coltheart, 1981). 2929 words made up the word pool when multiple forms of the same word (e.g., CHILD and CHILDREN) were excluded.

3.1.3. Procedure and design

Participants completed 12 cycles of study and test in two sessions with 6 rounds of study-test in each session. The two sessions were scheduled on two consecutive days, so the sessions were approximately twenty-four hours apart. In each study cycle, participants were presented with 150 words and a levels-of-processing task was administered to manipulate the encoding strength (Craik & Lockhart, 1972). In each session, half of the cycles (3) were strong and the other half (3) were weak. The strong and weak cycles were randomly ordered for each participant. For the strong study lists, participants were asked to make a semantic judgment (“Does the word have a pleasant meaning?”) and for the weak study lists, the judgment was orthographic (“Does the word contain the letter ‘e’?”). The study trials were self-paced as the participants responded by pressing the ‘z’ or the ‘/?’ keys on the keyboard and a 100 ms inter stimulus interval followed each response. The test list was constructed from 75 words that were randomly selected from the study list and 75 new words. For each test trial, participants were asked to make an ‘old/new’ recognition judgment by pressing ‘z’ for ‘old’ response and ‘/?’ for ‘new’ response. The experiment was a 2 (Strength) $\times$ 5 (Test Block) within-subjects design. The responses in each study-test cycle were pooled with regards to the strength condition resulting in 6 study-test cycles for each strength condition. Later, test positions were binned into 5 test blocks which contain 180 items (150 trials $\times$ 6 cycles = 900 trials, 900/5 test blocks = 180 observations per test block) in each for each participant. On average half of these 180 items were targets and the other half were foils.

3.2. Results and discussion

Table 1 presents recognition discriminability across strength and test blocks in d' units. A 5 (test block) $\times$ 2 (strength) repeated measures ANOVA on d' showed a main effect of strength, $F(1,27) = 141.09$, $p < 0.001$, $\eta_p^2 = 0.84$, where items that were studied in the semantic condition increased d' when compared with encoding through orthographic judgments. Similarly, the main effect of test position was significant, $F(4,108) = 27.52$, $p < 0.001$, $\eta_p^2 = 0.51$, suggesting that performance decreases over the course of testing. A 5 (test block) $\times$ 2 (strength) repeated measures ANOVA on HR and FAR revealed a main effect of strength. When the HRs and FARs are examined simultaneously, see Fig. 6, a SBME is observed with lower FARs, $F(1,27) = 67.033$, $p < 0.001$, $\eta_p^2 = 0.71$, and higher HRs, $F(1,27) = 148.422$, $p < 0.001$, $\eta_p^2 = 0.85$, for the strong lists. In addition to the SBME, a decrease across test position is observed for both HRs, $F(4,108) = 102.65$, $p < 0.001$, $\eta_p^2 = 0.80$, and FARs, $F(4,108) = 19.280$, $p < 0.001$, $\eta_p^2 = 0.42$. As the test block by strength interaction shows, this effect was more prominent for the weak items both in the HRs, $F(4,27) = 6.87$, $p < 0.001$, $\eta_p^2 = 0.20$, and the FARs, $F(4,27) = 5.13$, $p < 0.001$, $\eta_p^2 = 0.16$.

In addition to frequentist statistics, Bayes factors have been presented to quantify the ratio of evidence for the reported interaction (strength $\times$ test position) effects on HR and FAR (Rouder, Morey, Speckman, & Province, 2012). We used *BayesFactor* package in R with default settings to calculate the Bayes factors (Morey & Rouder, 2015). A Bayes factor ANOVA on HR suggested that the data are 7.86 times more likely to be observed under the interaction model compared to the only main effects model. Similarly, a Bayes factor of 2.78 revealed that the interaction model is marginally preferred over the main effects model for the FAR data.

Table 1
Recognition discriminability (d') in weak and strong test conditions in Experiments 1–4.

Strength	Test block				
	1	2	3	4	5
Exp 1					
Weak	0.92 (0.09)	0.71 (0.07)	0.70 (0.07)	0.63 (0.06)	0.51 (0.08)
Strong	2.01 (0.13)	1.83 (0.12)	1.68 (0.11)	1.70 (0.13)	1.57 (0.12)
Exp 2					
Weak	1.13 (0.10)	1.02 (0.09)	0.88 (0.09)	0.84 (0.09)	0.75 (0.07)
Strong	1.84 (0.11)	1.74 (0.11)	1.57 (0.11)	1.49 (0.09)	1.52 (0.11)
Exp 3					
Weak	1.03 (0.78)	0.76 (0.05)	0.70 (0.07)	0.63 (0.05)	0.59 (0.07)
Strong	1.74 (0.12)	1.64 (0.11)	1.55 (0.11)	1.38 (0.11)	1.36 (0.12)
Exp 4					
Weak	1.00 (0.11)	0.88 (0.09)	0.78 (0.10)	0.70 (0.07)	0.73 (0.08)
Strong	1.83 (0.14)	1.72 (0.13)	1.63 (0.11)	1.62 (0.13)	1.60 (0.11)

Note: Standard errors are presented in parentheses. On average, d′ declines with a slope of 0.07 across 5 test blocks in all four experiments and strength conditions. Additionally, the difference in d′ between strong and weak conditions is 0.85 when averaged across experiments and test blocks.

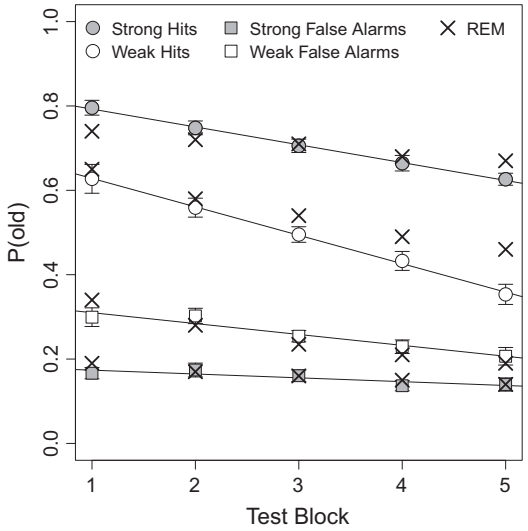

Fig. 6. The hit rates and the false alarm rates as a function of the strength condition and the test position in Experiment 1. The points and the squares represent the data averaged across participants and × represents the predictions from REM. Error bars are within-subject 95% CI (Loftus & Masson, 1994).

As predicted by Simulation 1 (shown in Fig. 4), the change in HRs and FARs across test list was less prominent when the test items were already differentiated during study due to being strongly encoded. Fig. 6 also presents the predicted HRs and FARs from a simulation of REM with the parameter values presented in Table 2. Additionally, predicted d′s are presented across strength and test block conditions in Table 3. REM can capture the qualitative pattern in the data observed both in the HRs and the FARs with slight and meaningful changes in the parameters. Although the model predicts a decrease in accuracy as a function of test blocks, the predicted decrease is not as prominent as the decrease observed in data. More importantly, REM suggests that differentiation provides a unified explanation for the two empirical findings: the SBME and OI.

4. Experiment 2

Here we consider a mixed-encoding study list, that is where half of the items are weakly encoded and half are strongly encoded. A model with encoding restricted to study predicts an increase in the HRs for strongly vs. weakly encoded items, but a FAR that does not change much with strength of the tested targets. Even when the test conditions are varied such that only strong or only weak targets are tested, such a model predicts little difference in the FARs. This prediction is due to the identical encoding conditions for the two cases (see Fig. 5 and Simulation 3). Presumably, when participants are *not informed* of the strength of the targets that will appear on the test list, they would also be less likely to adopt a different criterion for the two conditions. Thus, in this experiment, where participants are not informed about the strength of the targets, we do not expect to observe a SBME due to similar levels of FARs, and as a result, we do not expect to observe differences in the pattern of OI for HR and FAR that depend on the strength of the tested targets.

Table 2
The parameter values used in the REM simulations of Experiments 1–3.

Parameter	Value	Description
N	20	Vector length of the items
g	0.35	Feature frequency (geometric distribution parameter)
C	0.7	Probability of correctly copying a feature
u_{strong}	0.36	Probability of storing a feature of a strong item
u_{weak}	0.21	Probability of storing a feature of a weak item
u_{test}	0.40	Probability of storing a feature of the best matching trace at test
Criterion		Threshold for endorsing an item
Exp 1 and 2	0.75	
Exp 3: Strong	0.80	
Exp 3: Weak	0.70	

Note: In the REM simulations of Experiments 1–3, same parameters were used except the criterion parameter. The criterion parameter varied in Experiment 3 across test list strength conditions as the participants were told about the test list condition and thus let to adapt a different criterion for those conditions.

Table 3
Predicted d' values from REM simulations of Experiments 1–4.

Strength	Test block				
	1	2	3	4	5
Exp 1					
Weak	0.85	0.82	0.81	0.80	0.77
Strong	1.52	1.49	1.48	1.49	1.44
Exp 2					
Weak	0.89	0.88	0.84	0.83	0.81
Strong	1.44	1.47	1.45	1.48	1.43
Exp 3 and 4					
Weak	0.87	0.85	0.85	0.81	0.78
Strong	1.49	1.48	1.48	1.51	1.49

4.1. Methods

4.1.1. Participants

Thirty-two students from the same pool as Experiment 1 participated. Thirteen of the participants were female. The exclusion criterion described in Experiment 1 was employed resulting in a total of twenty-six participants.

4.1.2. Materials

The lists were constructed from the same word pool used in Experiment 1.

4.1.3. Procedures and design

The procedures were very similar to Experiment 1 except that the words were strengthened *within* list during study (rather than between lists). In each of 12 study cycles, the participants were presented with 150 words and were asked to make a semantic judgment (“Does the word have a pleasant meaning?”) for 75 of the words and an orthographic judgment (“Does the word contain the letter ‘e’?”) for the other 75. The order of the encoding tasks was random and the participants responded by pressing the ‘z’ or the ‘/?’ keys on the keyboard. In each of two sessions, half of the cycles (3) were randomly selected for the strong test list condition, meaning that only the strongly encoded targets along with 75 foils were tested. The remaining cycles were assigned to the weak test list condition and only the weakly encoded target words were tested along with 75 foils. Participants were not informed about the type of test list (strong or weak). The experiment was a 2 (Strength) × 5 (Test Block) within-subjects design. All other details are identical to Experiment 1.

4.2. Results and discussion

As in Experiment 1, a 2 (Strength) × 5 (Test Block) repeated measures ANOVA on d' revealed a main effect of strength, $F(1,25) = 58.42$, $p < 0.001$, $\eta_p^2 = 0.71$, suggesting that accuracy increased when items were encoded semantically and a main effect of test block, $F(4,100) = 18.94$, $p < 0.001$, $\eta_p^2 = 0.44$, showing a decrease in accuracy as a function of test block (see Table 1). Fig. 7 presents HRs and FARs for each test block and strength condition. HRs were greater for strong items than for weak items, $F(1,25) = 91.13$, $p < 0.001$, $\eta_p^2 = 0.79$, but manipulating the strength of targets on the test list did not elicit a difference in FARs, $F(1,25) = 0.18$, $\eta_p^2 = 0.007$. The lack of a SBME is consistent with the predictions of differentiation models and further show that a criterion shift does not occur when targets differ in strength between lists.

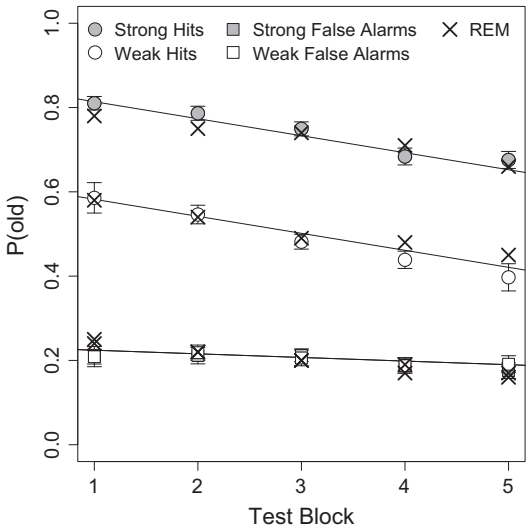

Fig. 7. The hit rates and the false alarm rates as a function of the strength condition and the test position for Experiment 2. The points and the squares represent the averaged hit rates and the averaged false alarm rates over participants. × represents the predicted hit rates and the predicted false alarm rates from REM. Error bars are within-subject 95% CI.

Note that the root cause of a criterion shift is not well specified; some propose the criterion is set on the basis of the study items (e.g., Hirshman, 1995) and others propose that the criterion is set on the basis of the first few test items (e.g., Benjamin & Bawa, 2004). The pattern of data observed here are not consistent with a model based on estimating memory from the first few test trials because such a model predicts a difference in FAR depending on the strength of the targets being tested.

Typical OI effects were found: HRs decreased with increases in the number of items tested, $F(4,100) = 56.80$, $p < 0.001$, $\eta_p^2 = 0.70$, and the decrease in the FARs was very small in magnitude, $F(4,100) = 3.76$, $p < 0.01$, $\eta_p^2 = 0.13$. In REM, when the FARs are comparable across test conditions, the magnitude of OI observed in HRs were also comparable (see Simulation 3, Fig. 5). The results from a mixed-study SBME paradigm supported this finding as the interaction between OI and the SBME was not significant in HR, $F(4,100) = 1.74$ and in FAR, $F(4,100) = 0.18$.

Consistent with the frequentist ANOVA on HR, a Bayes factor ANOVA showed that the main effect model was preferred by a factor of 11.60 over the interaction model, suggesting that test strength and test position have an additive effect on HR. A Bayes factor ANOVA on FAR revealed that the data are 6.53 times more likely to be observed under the test position only model compared to the main effects (strength + test position) model, and 209.19 times more likely compared to the interaction model. Further, REM was able to capture the pattern of the empirical HRs and FARs with the same parameter values used in Experiment 1 (see Table 2). The data and model fits are presented in Fig. 7. As before (and as in the original paper describing OI in REM), the decrease in d′ as a function of test blocks was less pronounced in predicted d′ values (Table 3), especially in the strong test condition. The reason for this could be due to the parameter values used in the current simulations (see Section 8 for a discussion).

5. Experiment 3

This experiment is identical to Experiment 2, with the exception that participants were *informed* prior to the beginning of the test list of the encoding task used to study the targets. Starns et al. (2010, 2012) showed the SBME in a mixed list paradigm where participants were explicitly informed, and argued that the SBME observed in such a paradigm was evidence that the SBME could be observed even when differentiation was controlled. In this experiment, we sought to further evaluate the pattern of OI when differentiation due to encoding strength is controlled. If participants adjust their response bias in response to being informed about the strength of the targets resulting in an SBME, as predicted by Starns et al., then we should find a SBME in this experiment. If the FAR changes, then the important question is whether the pattern of OI is affected.

5.1. Methods

5.1.1. Participants

Thirty-one Syracuse University undergraduates took part in the experiment in exchange for course credit. Fifteen of the participants were female. Applying the exclusion criterion described in Experiment 1 resulted in a sample size of twenty-two participants.[1]

[1] The qualitative pattern of the data does not change when 29 participants (d′ > 0.1) are included in the analysis.

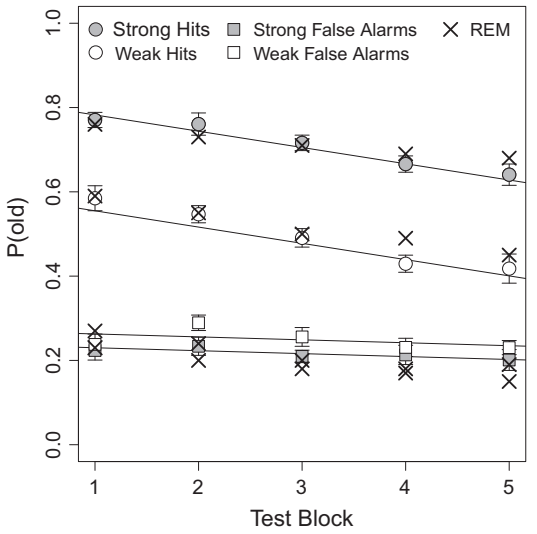

Fig. 8. The hit rates and the false alarm rates as a function of the strength condition and the test position in Experiment 3. The squares and the points represent the false alarm rates and the hit rates averaged over participants. Error bars are the within-subjects 95% CI. × represents the predicted hit rates and the predicted false alarm rates from REM.

5.1.2. Materials

The words were sampled from the same pool used in Experiments 1 and 2.

5.1.3. Procedures and design

The procedures were exactly the same as in Experiment 2 except that the participants were informed at the beginning of the test of the type of targets that would appear on the test. At the beginning of the experiment, participants were instructed to pay attention to the information about the test that would be provided after each study list. Before the weak test lists, participants saw "You will be tested only on the words for which you decided whether they contain the letter 'e'." For the strong test lists, they saw "You will be tested only on the words for which you made a pleasantness judgment." The information was on the screen for 5 s in both cases.

5.2. Results and discussion

Similar to results from Experiment 2, A 5 (test block) × 2 (strength) repeated measures ANOVA on d' showed a main effect of strength, $F(1,21) = 61.19$, $p < 0.001$, $\eta_p^2 = 0.77$, and a main effect of test block, $F(4,84) = 15.44$, $p < 0.001$, $\eta_p^2 = 0.46$. Accuracy was greater when participants were tested on strong targets compared to the accuracy of weak targets and accuracy decreased as the number of items tested increased (see Table 1).

Fig. 8 presents the probability of endorsing a test item as a function of test strength condition and test block along with the REM fits. A 5 (test block) × 2 (strength) repeated measures ANOVA on HRs showed that strengthening targets during study increased accuracy $F(1,21) = 48.57$, $p < 0.001$, $\eta_p^2 = 0.84$, and the strength of the target items did not effect FARs when participants were informed about the nature of the test list, $F(1,21) = 3.57$, $p = 0.07$, $\eta_p^2 = 0.14$. The results showed OI in the form of a decrease in HRs and FARs as a function of the test position, $F(4,84) = 42.48$, $p < 0.001$, $\eta_p^2 = 0.67$, $F(4,84) = 4.42$, $p < 0.01$, $\eta_p^2 = 0.18$, respectively.

A Bayes factor ANOVA on HR replicated the results from the frequentist ANOVA. Data are 17.18 times more likely to be observed under the main effect model compared to the interaction model. Similarly, for FAR, main effect model was preferred over the interaction model by a factor of 10.79. Interestingly, the Bayes factor ANOVA revealed that, the main effects model that includes the strength effect was preferred by a factor of 87.73 compared to the test position only model. More importantly, consistent with the mixed-list simulations of REM (Simulation 3), a decrease in HR was largely responsible for OI and the pattern was comparable across testing conditions.

As presented in the mixed test list simulations and as found in Experiment 2, REM does not predict the SBME on the basis of differentiation due to the encoding-at-test mechanism. However, to account for the slight tendency for a strength effect in FARs, a criterion shift based on the strength of the tested targets was implemented in REM. The data are well fit by the same parameters used for both prior experiments (Table 1) with the addition of *criterion* parameter fit to 0.70 for the weak test list and 0.80 for the strong test list. The model captured the lower HRs for the weak targets than strong targets. REM also accounted for the slight strength effect on the FARs and the slight decrease in false alarms across test position.

In this experiment, we have evidence of a smaller effect that than observed by Starns et al.: FARs qualitatively match their data but there is not consensus among the statistics (see also Kılıç & Öztekin, 2014). A potentially critical difference between the experimental designs was that we manipulated strength via levels-of-processing wherein the strong items were encoded by a deep processing and the weak items were encoded by a shallow processing task, whereas Starns et al. manipulated item repetition. Thus, the information that participants received at the beginning of our test was perhaps not as intuitive as number of repetitions. Participants are likely aware, from a lifetime of experience, that additional encoding time helps memory; however they are potentially less likely to be aware that a judgment of pleasantness in comparison to judging the letter 'e' leads to different levels of accuracy. This, of course, assumes that participants do not have meta-cognitive information about their memory following the level-of-processing manipulation. Thus, the smaller and less reliable effects of strength on FAR are likely because there was high variability among participants in their willingness to change the criterion. Taken together, the absence of the SBME in a mixed list paradigm is likely due to the unreliable strength effect on the FAR even when the participants were informed. Another difference in the designs between the current procedure and Starns et al. procedure was that Starns et al. required participants to press 5 for the strong test lists and press 1 for the weak test lists to begin the test. In the following experiment, we used a similar procedure to increase the strength effect on FAR by requiring participants to verify the information about test list strength, which they receive prior to the test.

6. Experiment 4

In Experiment 3, the study list was mixed in strength and the test list included only strong or weak items. Prior to testing participants were informed of the target strength, and consequently, they were expected to set a more stringent criterion for the strong test lists. However, the unreliable strength effect on FAR suggested that the shift in the criterion placement was not as prominent even when participants were informed of the test conditions. The goal of Experiment 4 was to strengthen the manipulation for criterion shifts.[2] In order to do so, during test instructions participants were required to press a certain key depending on the strength condition to proceed with the test. This ensured that participants paid attention to the information about the targets on the test list.

6.1. Methods

6.1.1. Participants

Thirty-nine Syracuse University undergraduates took part in the experiment in exchange for course credit. Applying the exclusion criterion described in Experiment 1 resulted in a sample size of thirty-two participants.[3]

6.1.2. Materials

The words were sampled from the same pool used in Experiments 1–3.

6.1.3. Procedures and design

The procedures were identical to Experiment 3 with the exception that participants were tested in a single set of 6 study-test cycles, and prior to the test, they were asked to press the key 'E' for the letter task and the key 'P' for the pleasantness task to proceed with the test.

6.2. Results & discussion

A 5 (test block) × 2 (strength) repeated measures ANOVA on d' showed a main effect of strength, $F(1,31) = 98.33$, $p < 0.001$, $\eta_p^2 = 0.76$, and a main effect of test block, $F(4,124) = 74.434$, $p < 0.001$, $\eta_p^2 = 0.13$. Similar to the results of the previous experiments, accuracy was greater when participants were tested on strong targets compared to the accuracy of weak targets. Similarly, accuracy decreased towards the end of the test list (see Table 1).

Fig. 9 presents the probability of endorsing a test item as a function of test strength condition and test block. A 5 (test block) × 2 (strength) repeated measures ANOVA on HRs showed that strengthening targets during study increased accuracy $F(1,31) = 86.54$, $p < 0.001$, $\eta_p^2 = 0.74$, and a 5 (test block) × 2 (strength) repeated measures ANOVA on FARs showed a significant strength effect, $F(1,31) = 10.31$, $p < 0.01$, $\eta_p^2 = 0.25$. Similarly, the results showed OI in the form of a decrease in HRs and FARs as a function of the test position, $F(4,124) = 24.5$, $p < 0.001$, $\eta_p^2 = 0.44$, $F(4,124) = 4.115$, $p < 0.01$, $\eta_p^2 = 0.18$, respectively.

As expected based on the previous results, a Bayesian ANOVA on HRs revealed that the main effects model is 33.49 times more likely to be preferred over the interaction model. Similarly, the Bayesian ANOVA showed that FARs in Experiment 4 are 45.72 more likely to be observed under the main effects model compared to the interaction model. The strength effect is also

[2] We thank Jeff Starns for this suggestion during the review process.

[3] The qualitative pattern of the data does not change when all of the participants are included in the analysis.

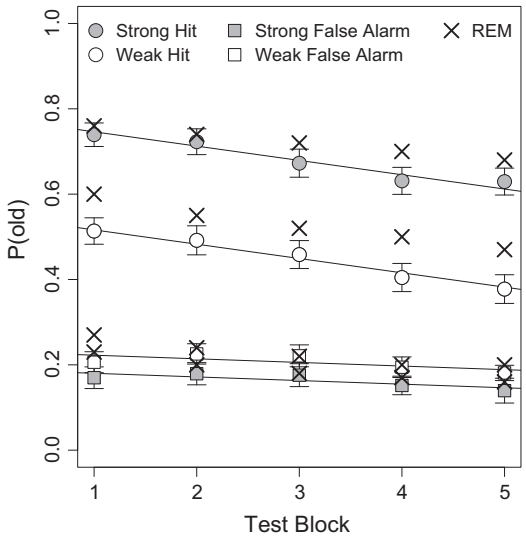

Fig. 9. The hit rates and the false alarm rates as a function of the strength condition and the test position in Experiment 4. The squares and the points represent the false alarm rates and the hit rates averaged over participants. Error bars are the within-subjects 95% CI.

found to be more prominent compared to the evidence for strength effect in Experiment 3. FARs are observed to be 13544.86 times more likely under the both strength and test position model compared to the test position only model.

The results from this experiment showed that requiring participants to verify the information that they received regarding the composition of the test list encouraged them to ever so slightly shift their response criterion, increasing FARs. More importantly, when differentiation was controlled during encoding but a SBME is elicited by criterion shifts, we fail to observe a change in the pattern of OI for HRs and FARs.

7. General discussion

Differentiation, adapted from the perception literature (Gibson, 1940; Gibson & Gibson, 1955; Murnane & Shiffrin, 1991; Ratcliff et al., 1990), is proposed as a core principle in episodic memory. This paper proposed an integrated theoretical account based on differentiation for two memory effects, namely the strength based mirror effect and output interference. Because the SBME and OI result from the same theoretical principle – differentiation – there is a tight coupling between the two empirical findings. This is in contrast to other explanations for these findings that emphasize criterion shifts for the former and waning attention for the latter. Simulations from REM show that strengthening a list of items interacts with test position such that strong target items and foil items are both less susceptible to interference that results from encoding during testing when the list is encoded well. The behavioral results from Experiment 1 were compatible with the predictions from REM. When items were studied in pure lists, the pattern of OI was dependent on the strength of the encoded list. However, when items were strengthened in mixed lists, interference due to encoding at test was comparable across strength conditions (Experiment 2). Although much less prominent compared to the effect in the pure paradigm, the SBME was observed in the mixed list paradigm when participants were encouraged to adopt a different criterion for the strong test condition (Experiment 4). However, unlike the pure paradigm, the OI curves were approximately parallel for strong and weak lists. Next, we discuss how the differentiation principle accounts for the SBME and OI simultaneously and further discuss the alternative explanations of these two effects.

7.1. SBME and OI

In an extension of the REM model, OI was implemented as encoding of test trials (Criss et al., 2011). When a test item is endorsed, the best matching trace is updated and when an item is rejected, a new trace is encoded in memory. This is sensible because when a target item is judged to be old, the best matching trace is most likely to be the trace stored during encoding of that test item. Accordingly, the subsequent foils would be less likely to match those traces that have been updated and the FAR would decrease. Similarly, when a foil item is judged to be old, the best matching trace would be updated incorrectly and that would decrease both the FARs and the HRs for subsequent test trials. Thus, updating traces in memory results in differentiation such that the traces become less similar to the subsequent test items.

When items are strengthened during study, more complete traces are encoded in memory. The more complete encoding of targets differentiates the traces, as different traces become less similar to each other, which results in a higher overall HR

for the strong lists. In addition, stronger encoding conditions produce a greater tendency for traces representing the target to be updated when old judgments are given, reducing the chance that the trace will be selected for updating for subsequent test trials (from both other target items and foils). On balance, therefore, differentiation results in a less prominent decrease in the HRs and flatter FARs over the course of testing for a strongly encoded list compared to a situation where items are weakly encoded during initial study. That is, the form of OI depends on the accuracy of the encoded memory traces.

One interesting question is that whether a change in response bias and resulting differences in the FARs across strength conditions produce the relationship between the SBME and OI. The results from Simulation 3 in which REM was applied to a mixed study list/pure test list paradigm, showed no interaction between list strength and test block along with a null strength effect on FARs, consistent with the behavioral results reported in Experiment 2. Similarly, when participants studied a mixed list but were informed of the test list strength, there was no interaction even when the tendency to shift criterion produced a SBME. The predictions of REM suggest that when differentiation is constant during study, the SBME does not interact with OI even when the SBME is observed due to criterion shifts.

Although a shift in criterion placement does not produce the specific pattern of HRs and FARs in mixed study lists, more liberal criterion placement in pure study lists might facilitate the difference in the pattern of OI across strength conditions in pure study list conditions. The results from Simulation 2 suggest that the decrease in HR and FAR as a function of test position in weak lists is more prominent when the criterion is placed more liberally (e.g., 0.72 vs. 1 on Fig. 10). Future studies can further investigate these predictions by experimentally manipulating criterion placement in pure study conditions.

It is important to note that the empirically observed interactions on HR and FAR reported here could be removed due to a transformation of probability measures such as HR and FAR (see Wagenmakers, Krypotos, Criss, & Iverson, 2012). However, a critical point is of that paper is that interpretations absent a process model should be interpreted with caution. Indeed, we have a process model, REM, which we use to interpret these interactions and buffer against the concerns raised by Wagenmakers et al. (2012).

7.2. OI: interference and waning vigilance

A decrease in memory evidence as a function of test position manifested as a decrease in HR accounted for in REM by assuming encoding and we assumed that encoding during test is more effective than encoding during study (Karpicke & Roediger, 2008; Raaijmakers & Shiffrin, 1981a, 1981b; Roediger & Karpicke, 2006). This pattern was present and consistent in all three experiments.

An alternative explanation for the decrease in accuracy over the course of testing could be waning vigilance. The waning vigilance hypothesis, if operationalized as a change in speed-accuracy tradeoffs, could be further tested by applying the diffusion model (DM, Ratcliff, 1978). As a dynamic signal detection framework, the diffusion model provides parameter estimates for boundary separation, which measures the speed-accuracy tradeoffs. Boundary separation is defined as the criterion for evidence accumulation. For example, if participants are instructed to give fast responses, they tend to compromise accuracy. Thus, it is possible that accuracy decreases over the course of testing due to faster and inaccurate responses, rather than due to interference. Actually, the DM analysis of the current data showed that boundary separation decreases across test blocks, which suggests that the decrease in d′ observed in data could be a result of an additive effect of interference (measured by the drift rate parameter) and waning vigilance (measured by the boundary separation parameter, Kılıç, 2012). However, the DM analysis of the current data showed that boundary separation decreases at the very end of the test list, but not much before, while the decrease in accuracy is consistent across the test list (e.g., not restricted to the final trials). Additional evidence against the vigilance hypothesis comes from testing the foil items. If attention wanes and this results in OI, then foils presented toward the end of testing should be less likely to be remembered than foils tested toward the beginning of testing. However, accuracy in memory for foils does not depend on test position (Criss et al., in preparation). Converging evidence comes from the excellent fits of REM to HR and FAR data despite no additional parameters representing fluctuations in attention. As we illustrated and as demonstrated in the original implementation of the OI model (Criss et al., 2011), the predicted decrease in d′ is less than the observed decrease in d′. This suggests that additional studies are required to further investigate the possibility of multiple causes of (or lack of) OI in overall discrimination.

7.3. SBME: criterion shift and differentiation

Differentiation, proposed as a core principle in perception and episodic memory, presents a comprehensive and theoretically consistent account for multiple empirical findings such as the null list-strength effect, the SBME, OI and the complex relationship between the last two. However, others argue that metacognitive processes could produce the SBME observed in these studies (Cary & Reder, 2003; Hirshman, 1995; Starns et al., 2012; Stretch & Wixted, 1998; Verde & Rotello, 2007). The hypothesis is that when participants are aware of the fact that they would be tested on strong targets, they become more conservative and less likely to endorse a test item. That results in a decrease in the FARs. One possible explanation, the one adopted by Starns et al., is that participants change the placement of their criterion in response to information about the perceived difficulty of the test. The results from Experiment 2 showed that mixed study-pure test design does not elicit a SBME when participants are not explicitly informed about the test conditions. When they are informed as in Experiment 3, they show an unreliable SBME in FARs (see also Kılıç & Öztekin, 2014), whereas when participants were required to pay

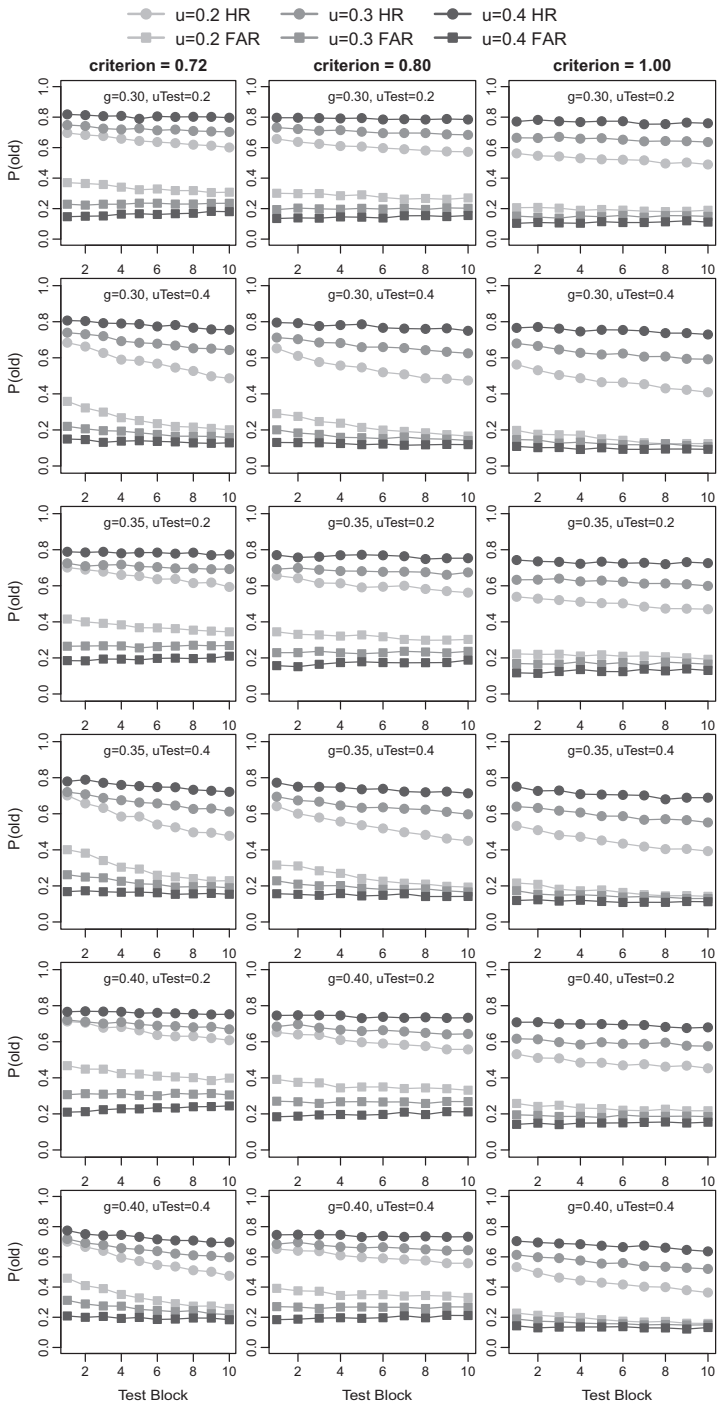

Fig. 10. Predicted probability of endorsing a test item as a function of test block across a set of parameters. Each panel presents the simulated HRs and FARs when items are encoded with different u values (0.2, 0.3, 0.4). The columns plot simulations with varying values of *criterion* (left: 0.72, middle = 0.80, right = 1). Additionally, the simulations with three different values of g (0.30, 0.35, 0.40) and two encoding strength levels at test (0.20, 0.40) are presented in different rows of the grid. The other parameter values are $n = 20$ and $c = 0.7$.

attention to the information they receive about the targets that they will be tested on, they adopt a more stringent criterion for tests of strong items (Experiment 4) as observed in Starns et al. (2010, 2012) studies.

Critically, when participants are tested with lists composed entirely of targets or entirely of foils, they do not set different criteria compared to the standard recognition tests in which the lists consist of half targets and half foils (Cox & Dobbins,

2012). In fact, HR and FAR are nearly identical in standard tests and test absent foils or absent targets. These findings were replicated in a recent study by Koop et al. (2015), further showing that participants responded differently only when they were provided with feedback. Presumably, the accurate feedbacks enabled the participants to adapt to the all-targets or all-foils condition. On the other hand, random feedback resulted in lower accuracy in the all-targets or all-foils-lists contrary to the null feedback effect in the standard recognition test. An interesting question for future research is why participants do not change their criterion in response to the *actual*, *experienced* difficulty of the test (e.g., even in the uninformed condition).

An alternative explanation for the SBME in the mixed list paradigm is that participants are able to narrow their search of memory based on the contextual information provided by the test instruction, to the items studied in the relevant encoding task. For example, when given instructions that memory will be tested for items studied in the pleasantness task, participants may be able to use those context features in the memory probe. That is, participants may constrain their retrieval based on the source (Jacoby, Shimizu, Daniels, & Rhodes, 2005; Jacoby, Shimizu, Velanova, & Rhodes, 2005). This same principle would result in a SBME pattern for the mixed study-pure test paradigm. However, the ability of participants to actually implement such a retrieval strategy is unknown with evidence both for (e.g., Annis et al., 2013; Jacoby, Shimizu, Daniels, et al., 2005; Jacoby, Shimizu, Velanova, et al., 2005) and against (e.g., Stretch & Wixted, 1998) such a mechanism.

Prior research showed that the semantic judgments strengthened the qualitative details by encoding more distinctive features of items rather than increasing the quantitative information as in strengthening by repetition (Gallo, Meadow, Johnson, & Foster, 2008). However, if the SBME observed in the pure-list condition (Experiment 1) is due to encoding of distinctive features then a similar SBME should have been observed in the mixed list paradigm regardless of being informed of the test strength condition. Pending further research, we conclude that differentiation contributes to the SBME. Criterion shifts only contribute when strategically implemented by participants and often in response to experimenter instructions (e.g., see Koop & Criss, 2016).

8. On the generality of the REM predictions

We've presented simulations of REM and tested those simulations, in each case showing that the observed pattern of HR and FAR is as predicted by the model. We used an "off the shelf" set of parameter values for these simulations in order to demonstrate the robustness of these predictions (i.e., we did not search around for parameter values that would result in the particular set of data that we observed). Of course, this raises questions about the generality of the predictions.

REM is a highly constrained process model such that some parameters are either fixed or vary only in specific circumstances (e.g., the number of features and the *c* parameter, see Malmberg, Zeelenberg, & Shiffrin, 2004 for an example). Some parameters are constrained by the stimuli (the *g* parameter) rather than the performance. Through (combined) decades of fitting the model, we have a strong informed prior about what are reasonable (and unreasonable) parameter values.

In Simulation 4, we show that the qualitative pattern of how HR and FAR changes across test position in response to changes in list strength are observed across a range of a reasonable parameter values that are typical of those reported in the literature. A total of 10,000 sample parameters were obtained by a random drawn from uniform distributions of the following parameter values: the number of features was fixed at 20 and c was fixed at 0.7, g ranged from 0.2 to 0.6, u_w ranged from 0.1 to 0.3, u_s ranged from $u_w + 0.1$ to $uw + 0.3$, u_T ranged from 0.2 to 0.4, *criterion* ranged from 0.7 to 1.3 for pure lists, and $criterion_W$ ranges from 0.7 to 1.1, $criterion_S$ ranges from $criterion_W + 0.1$ to $criterion_W + 0.2$ for mixed lists. 100 participants for each of the 10,000 randomly selected set of parameters were simulated. In Fig. 10, we show a sample of individual simulations. In Fig. 11, we report the distribution of differences in slopes across all samples. For each 100 simulated participants, the slope of the decrease in HR and FAR across test positions was obtained from the best fitting linear line. Then, we calculated the difference between the slopes in weak and strong conditions and plot those values in Fig. 11. Values greater than 0 imply that the slope of the decrease in weak HRs or FARs are greater than the slope of the decrease in strong HRs or FARs across test positions. Fig. 12 shows the slope of d′ across test position. The pattern we observed in the data is predicted by a range of reasonable parameters.

In Simulation 5, we take a sample across the full parameter space, including reasonable and unreasonable values. When we say unreasonable values, we mean that the values do not reflect what an experimental psychologist would do in a laboratory experiment (e.g., a g value of 0.7 would be like having a study list populated with words like: the, and, a, they) or do not reflect human behavior (e.g., a u value of 0.7 would reflect extraordinarily high memory accuracy). Under reasonable parameter values, (see, Figs. 4, 10 and 11) REM predicts a less prominent decrease across test blocks for HR and FAR from a strongly encoded study list than a weakly encoded study list in pure study conditions. However, this is not predicted when items are encoded in mixed strength lists (see Figs. 5, 6, 10 and 11). Here, we consider the full range of parameter values (g, u_T, and u_w ranged from 0.01 to 0.69, u_s ranged from $u_w + 0.1$ to $uw + 0.3$, *criterion* ranged from 0 to 2). As seen in Fig. 13, some parameter values predict the finding that we observed – a difference in slopes across strength conditions that exceed 0, when items are strengthened in pure lists. However, there are parameter values that generate predictions contrary to what we observed, namely a difference in slopes that is below 0 which indicates greater decrease in HR and FAR across test positions in strongly encoded lists. Similarly, the difference in slopes of OI is centered around 0 when items are strengthened in mixed lists. However, different from the results of Simulation 5, the range of the slope differences are extended which indicates that with some unreasonable parameter values, a difference in the slopes across strength conditions is expected in either direction.

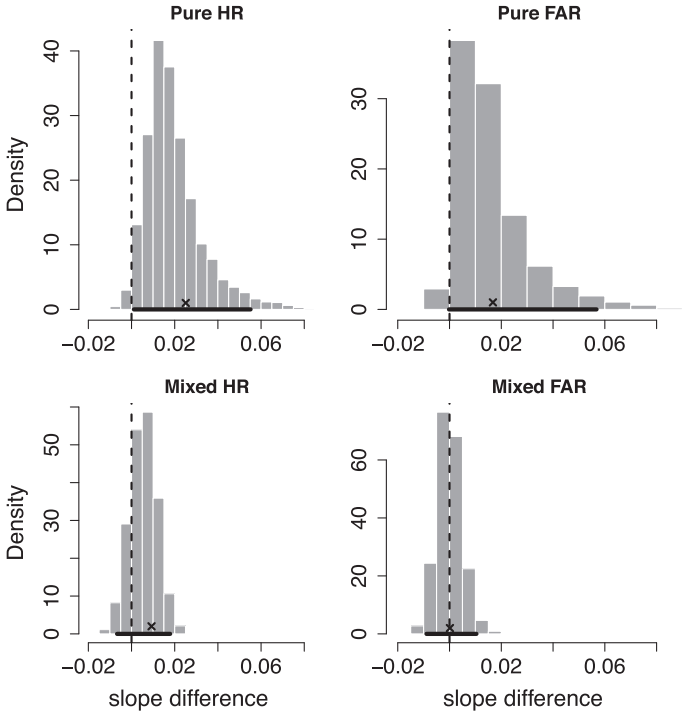

Fig. 11. The distribution of the difference in the slopes of OI in HR and FAR between weak and strong list conditions. Bold lines represent 95% High Density Interval for the slope differences across a set of parameter values drawn from a uniform distribution (Simulation 4). See the text for the ranges of the reasonable parameter values. Greater values of slope difference suggest that the decrease in HRs or FARs across test blocks is greater for weak lists compared to that of strong lists. The slope difference values in the top two panels are from the simulations of pure lists and the bottom panels are from the simulations of mixed lists. × represents the slope differences observed in Experiment 1 for the pure conditions and in Experiment 3 for the mixed conditions.

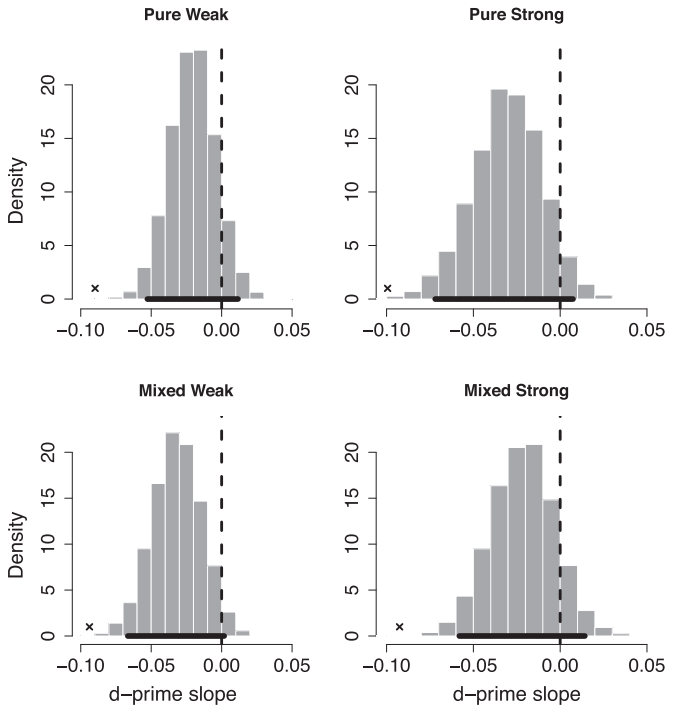

Fig. 12. The distribution of slopes of OI in d′ across strength conditions in Simulation 4. Bold lines represent 95% High Density Interval for the OI slope of d′. See the text the ranges of the reasonable parameter vales. × represents the slope differences observed in Experiment 1 for the pure conditions and in Experiment 3 for the mixed conditions.

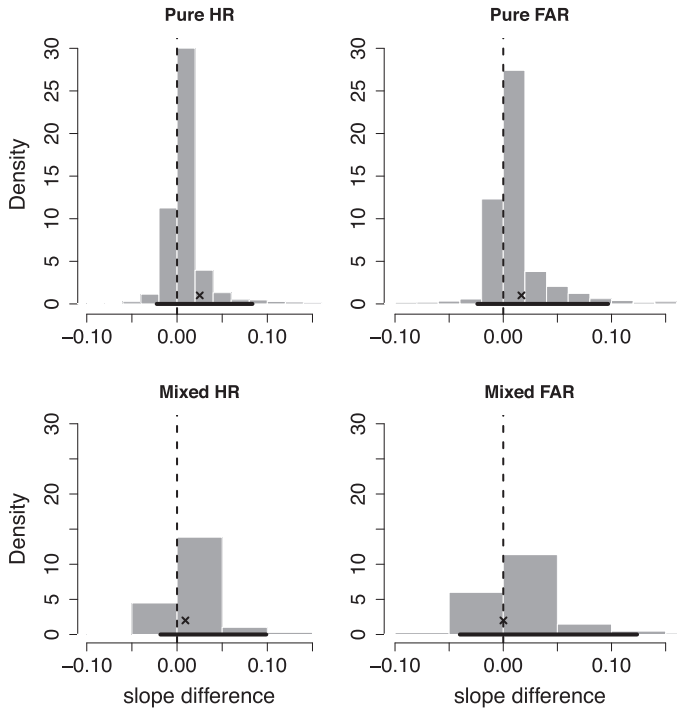

Fig. 13. The distribution of the difference in the slopes of OI in HR and FAR between weak and strong list conditions. Bold lines represent 95% High Density Interval for the slope differences across an extended set of parameter values (both reasonable and unreasonable values) drawn from a uniform distribution (Simulation 5). See the text for the ranges of the parameter values. Values of slope differences that exceed 0 suggest that the decrease in HRs or FARs across test blocks is greater for weak lists compared to that of strong lists. On the contrary, values of slope differences that are below 0 suggest that the decrease in HR or FAR across test blocks is greater for strong lists compared to that of weak lists. × represents the slope differences observed in Experiment 1 for the pure conditions and in Experiment 3 for the mixed conditions.

Thus, the empirical findings we observe are significant—they reveal data that are predicted by REM with reasonable parameter values. These simulations also show that the empirical findings further constrain the model and confirm that some parameter values do not predict the observed pattern of data (namely those parameter values that are unreasonable).

9. Conclusion

Differentiation describes how the accumulation of experience improves cognitive abilities. The general theory of differentiation has been successfully used to explain perceptual learning, category formation, semantic knowledge development, and episodic memory. Here we showed how differentiation provides a unified explanation for strength and interference effects in recognition memory. In contrast, other theories of memory propose multiple different mechanisms to account for these findings. The empirical and modeling results presented here suggest a single elegant theoretical framework for episodic memory. More broadly, we suggest that differentiation should be under consideration as a potential universal law of cognition (e.g., Chater & Brown, 2008).

Acknowledgments

This research was supported by National Science Foundation – United States Grant # 0951612 to AHC.

References

Adolph, K. E., & Kretch, K. S. (2015). Gibson's theory of perceptual learning. In J. D. Wright (Ed.). *International Encyclopedia of the Social and Behavioral Sciences* (2nd ed.) (Vol. 10, pp. 127–134). New York: Elsevier.

Annis, J., & Malmberg, K. J. (2013). A model of positive sequential dependencies in jufgments of frequency. *Journal of Mathematical Psychology, 57*(5), 225–236.

Annis, J., Malmberg, K. J., Criss, A. H., & Shiffrin, R. M. (2013). Sources of interference in recognition testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1365–1376.

Benjamin, A. S., & Bawa, S. (2004). Distractor plausibility and criterion placement in recognition. *Journal of Memory and Language, 51*(2), 159–172. http://dx.doi.org/10.1016/j.jml.2004.04.001.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language, 49*(2), 231–248. http://dx.doi.org/10.1016/S0749-596X(03)00061-5.

Chater, N., & Brown, G. D. A. (2008). From universal laws of cognition to specific cognitive models. *Cognitive Science, 32*, 36–67.
Chen, Y. Y., Lithgow, K., Hemmerich, J. A., & Caplan, J. B. (2014). Is what goes in what comes out? Encoding and retrieval event-related potentials together determine memory outcome. *Experimental Brain Research, 232*(10), 3175–3190.
Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology, 33A*(4), 497–505. http://dx.doi.org/10.1080/14640748108400805.
Cox, J. C., & Dobbins, I. G. (2012). The striking similarities between standard, distractor-free, and target-free recognition. *Memory & Cognition, 39*, 925–940. http://dx.doi.org/10.3758/s13421-011-0090-3.
Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. http://dx.doi.org/10.1016/S0022-5371(72)80001-X.
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. http://dx.doi.org/10.1016/j.jml.2006.08.003.
Criss, A. H. (2009). The distribution of subjective memory strength: List strength and response bias. *Cognitive Psychology, 59*(4), 297–319. http://dx.doi.org/10.1016/j.cogpsych.2009.07.003.
Criss, A. H. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(2), 484–499. http://dx.doi.org/10.1037/a0018435.
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. 112–115). Florence, KY: Psychological Press.
Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64*(4), 316–326. http://dx.doi.org/10.1016/j.jml.2011.02.003.
Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55*(4), 447–460. http://dx.doi.org/10.1016/j.jml.2006.06.003.
Criss, A. H., Salomão, C., Malmberg, K. J., Aue, W. R., Kılıç, A., & Claridge, M. (in preparation). Release from output interference in recognition memory.
Criss, A. H., & Shiffrin, R. M. (2004). Context noise and item noise jointly determine recognition memory: A comment on Dennis & Humphreys (2001). *Psychological Review, 111*(3), 800–807.
Criss, A. H., Wheeler, M. E., & McClelland, J. L. (2013). A differentiation account of recognition memory: Evidence from fMRI. *Journal of Cognitive Neuroscience, 25*(3), 421–435.
Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–477.
Gallo, D. A., Meadow, N. G., Johnson, E. L., & Foster, K. T. (2008). Deep levels of processing elicit a distinctiveness heuristic: Evidence from the criterial recollection task. *Journal of Memory and Language, 58*(4), 1095–1111. http://dx.doi.org/10.1016/j.jml.2007.12.001.
Gibson, E. J. (1940). A systematic application of the concepts of generalization and differentiation to verbal learning. *Psychological Review, 47*(3), 196–229. http://dx.doi.org/10.1037/h0060582.
Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62*(1), 32–41. http://dx.doi.org/10.1037/h0048826.
Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13*(1), 8–20. http://dx.doi.org/10.3758/BF03198438.
Goldstone, R. L., & Styvers, M. (2001). The sensitization and differentiation of dimensions during category learning. *Journal of Experimental Psychology: General, 130*, 116–139.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Hemmer, P., Criss, A. H., & Wyble, B. (2011, November). Assessing a neural basis for differentiation accounts of recognition memory. In *Poster presented at the Psychonomic Society meeting*. Seattle, WA.
Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 302–313. http://dx.doi.org/10.1037/0278-7393.21.2.302.
Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review, 12*(5), 852–857. http://dx.doi.org/10.3758/BF03196776.
Jacoby, L. L., Shimizu, Y., Velanova, K., & Rhodes, M. G. (2005). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language, 52*(4), 493–504. http://dx.doi.org/10.1016/j.jml.2005.01.007.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968. http://dx.doi.org/10.1126/science.1152408.
Kılıç, A. (2012). *Output interference and strength based mirror effect in recognition memory (Doctoral Dissertation)*. Retrieved from SURFACE Psychology-Dissertaions.
Kılıç, A., & Öztekin, I. (2014). Retrieval dynamics of the strength based mirror effect in recognition memory. *Journal of Memory and Language, 76*, 158–173. http://dx.doi.org/10.1016/j.jml.2014.06.009.
Koop, G. J., & Criss, A. H. (2016). The response dynamics of recognition memory: Sensitivity and bias. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 42*(5), 671–685.
Koop, G. J., Criss, A. H., & Malmberg, K. J. (2015). The role of mnemonic processes in pure-target and pure-foil recognition memory. *Psychonomic Bulletin & Review, 22*, 509–516.
Kucera, F., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.
Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*(4), 476–490. http://dx.doi.org/10.3758/BF03210951.
Malmberg, K. J., & Annis, J. (2012). On the relationship between memory and perception: Sequential dependencies in recognition memory testing. *Journal of Experimental Psychology: General, 141*(2), 233–259. http://dx.doi.org/10.1037/a0025277.
Malmberg, K. J., Criss, A. H., Gangwani, T. H., & Shiffrin, R. M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science, 23*(2), 115–119. http://dx.doi.org/10.1177/0956797611430692.
Malmberg, K. J., Holden, J. E., & Shiffrin, R. M. (2004). Modeling the effects of repetitions, similarity, and normative word frequency on judgments of frequency and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 319–331.
Malmberg, K. J., Lehman, M., Annis, J., Criss, A. H., & Shiffrin, R. M. (2014). Consequences of testing memory. In B. Ross (Ed.), *Psychology of learning & motivation* (Vol. 61, pp. 285–313).
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*(2), 540–549.
McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*(4), 724–760. http://dx.doi.org/10.1037/0033-295X.105.4.734-760.
Morey, R. D., & Rouder, J. N. (2015). *BayesFactor* (Version 0.9.10-2)[Computer software].
Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage and retrieval of item information. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 145–194). Hillsdale, New Jersey: Erlbaum.
Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*, 855–874.
Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(1), 87–108.
Nosofsky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(6), 1194–1209.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981a). Search of associative memory. *Psychological Review, 88*, 93–134.
Raaijmakers, J. G. W., & Shiffrin, R. M. (1981b). Order effects in recall. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 403–415). Hillsdale, NJ: Erlbaum.
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review, 85*(2), 59.
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. http://dx.doi.org/10.1037/0278-7393.16.2.163.
Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review, 83*(3), 190. http://dx.doi.org/10.1037/0033-295X.83.3.190.
Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255. http://dx.doi.org/10.1111/j.1467-9280.2006.01693.x.
Rogers, T. T., & McClelland, J. L. (2008). A précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences, 31*, 689–714.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*, 356–374.
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. http://dx.doi.org/10.1037/0278-7393.16.2.179.
Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*(2), 145–166. http://dx.doi.org/10.3758/BF03209391.
Shiffrin, R. M., & Steyvers, M. (iffrin and Steyvers 1998). The effectiveness of retrieval from memory. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 73–95). Oxford, U.K.: Oxford University Press.
Starns, J. J., Ratcliff, R., & White, C. N. (2012). Diffusion model drift rates can be influenced by decision processes: An analysis of the strength-based mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. http://dx.doi.org/10.1037/a0028151.
Starns, J. J., White, C. N., & Ratcliff, R. (2010). A direct test of the differentiation mechanism: REM, BCDMEM, and the strength-based mirror effect in recognition memory. *Journal of Memory and Language, 63*(1), 18–34. http://dx.doi.org/10.1016/j.jml.2010.03.004.
Stretch, V., & Wixted, J. (1998). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1379–1396. http://dx.doi.org/10.1037/0278-7393.24.6.1379.
Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society, 12*(2), 89–91.
Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory & Cognition, 35*(2), 254–262. http://dx.doi.org/10.3758/BF03193446.
Wagenmakers, E. J., Krypotos, A. M., Criss, A. H., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40*, 145–160.
Wallace, W. P. (1965). Review of the historical, empirical, and theoretical status of the Von Restorff phenomenon. *Psychological Bulletin, 63*(6), 410.