Cognitive Psychology 149 (2024) 101617

Contents lists available at ScienceDirect

Cognitive Psychology

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

Retrieving effectively from source memory: Evidence for differentiation and local matching processes

Sinem Aytaç [a,*], Aslı Kılıç [b], Amy H. Criss [a], David Kellen [a]

[a] Department of Psychology, Syracuse University, Syracuse, NY, USA
[b] Department of Psychology, Middle East Technical University, Ankara, Turkey

ARTICLE INFO

Keywords:
Retrieving Effectively from Memory
Source memory
Strength-based mirror effect
Null list-strength effect
Output interference
Differentiation

ABSTRACT

The ability to distinguish between different explanations of human memory abilities continues to be the subject of many ongoing theoretical debates. These debates attempt to account for a growing corpus of empirical phenomena in item-memory judgments, which include the *list strength effect*, the *strength-based mirror effect,* and *output interference*. One of the main theoretical contenders is the Retrieving Effectively from Memory (REM) model. We show that REM, in its current form, has difficulties in accounting for source-memory judgments – a situation that calls for its revision. We propose an extended REM model that assumes a local-matching process for source judgments alongside source differentiation. We report a first evaluation of this model's predictions using three experiments in which we manipulated the relative source-memory strength of different lists of items. Analogous to item-memory judgments, we observed a null list strength effect and a strength-based mirror effect in the case of source memory. In a second evaluation, which relied on a novel experiment alongside two previously published datasets, we evaluated the model's predictions regarding the manifestation of output interference in item and lack of it in source memory judgments. Our results showed output interference severely affecting the accuracy of item-memory judgments but having a null or negligible impact when it comes to source-memory judgments. Altogether, these results support REM's core notion of differentiation (for both item and source information) as well as the concept of local matching proposed by the present extension.

1. Introduction

Episodic memory is defined as a memory for an event experienced in a particular context (Tulving, 1983). This definition suggests that episodic memory calls for the storage of at least two types of information: (1) of the event itself, and (2) of the context in which said event took place. In the case of a typical memory experiment where participants study lists of words, each word item would correspond to an event, whereas context would refer to the characteristics surrounding its occurrence, such as the list in which it was studied (e.g., List 1, List 2), or its perceptual characteristics (e.g., color, font). The ability to remember contextual information is commonly referred

* Corresponding author at: Department of Psychology, Syracuse University, 900 South Crouse Ave., Syracuse, NY 13244, USA.
E-mail address: aytac.sinem@gmail.com (S. Aytaç).

https://doi.org/10.1016/j.cogpsych.2023.101617
Received 30 July 2022; Received in revised form 7 November 2023; Accepted 9 November 2023
Available online 5 January 2024

to as *source memory*, in distinction from the ability to remember the item itself – *item memory* (Batchelder & Riefer, 1990; Johnson et al., 1993; Lindsay, 2008).[1]

Source memory plays an important role in everyday life: Suppose that you see somebody in the university cafeteria, and you are sure that you have met this person before but are unsure as to where exactly. There are several possibilities, such as the cafeteria itself, the library, or the psychology department where you spend most of your days. Given that you are seeing this person at the cafeteria, it is reasonable to indulge the possibility of having met them there before. But let us suppose that you end up figuring out (correctly) that the cafeteria is not where this previous encounter took place. In this example, the ability to recognize this person as someone that you have met before was based on the *item information*, whereas the cafeteria served as a *source cue* in the sense it offered a possible context for this previous encounter.

The investigation of episodic memory has led to the identification of a large body of empirical phenomena and the development of numerous formal models attempting to provide a theoretical account for them (e.g., Anderson & Milson, 1989; Davelaar et al., 2005; Dennis & Humphreys, 2001; Gillund & Shiffrin, 1984; Hintzman, 1988; Howard & Kahana, 2002; Humphreys et al., 1989; McClelland & Chappell, 1998; Murdock, 1997; Osth & Dennis, 2015; Raaijmakers & Shiffrin, 1980; 1981; Shiffrin & Steyvers, 1997). But although contextual or source information plays a central role in many theoretical accounts (e.g., Anderson & Bower, 1972; Davelaar et al., 2005; Dennis & Humphreys, 2001; Howard & Kahana, 2002; Mensink & Raaijmakers, 1988; Sederberg et al., 2008), there has been a focus on item-memory judgments at the expense of source-memory judgments (for notable exceptions, see Glanzer et al., 2004; Osth et al., 2018; Starns & Ksander, 2016). To the point that one of the most prominent candidate models in the literature, REM (Retrieving Effectively from Memory; Shiffrin & Steyvers, 1997), which will be the focus of the present work, is currently unable to provide an account for source judgments that is commensurate with its achievements when it comes to item memory (e.g., Criss, 2006; Criss et al., 2011; Criss & Shiffrin, 2005; Diller et al., 2001; Kılıç et al., 2017; Malmberg et al., 2004; Malmberg & Shiffrin, 2005, see also Osth et al., 2018).

The goal of the present work is to contribute towards bridging this gap. After reviewing a number of relevant concepts and empirical findings, we will identify the shortcomings of REM in its current form and propose a revision that addresses them. We will then conduct a first evaluation of this proposal using data from three novel experiments that manipulate *source strength* in the same vein as earlier manipulations of item strength, whose results motivated the initial development of REM (Ratcliff et al., 1990; Shiffrin et al., 1990; Shiffrin & Steyvers, 1997). Lastly, we will scrutinize the general explanatory power of the theoretical account provided by the revised REM – local matching – by testing its predictions regarding the negligible appearance of *output interference* in source judgments. These predictions will be tested using one novel experiment and two previously published datasets.

1.1. Strengthening item and source memory

Manipulations of study events (i.e., strengthening), either by manipulating exposure time or via repetition, have long been at the center of theoretical debates due to their ability to distinguish between different candidate explanations (e.g., Glanzer et al., 2009; Kellen & Klauer, 2015; Shiffrin et al., 1990). These experimental manipulations are designed to yield two classes of studied items – weak and strong. Lure items can also be classified as weak or strong depending on the characteristics that they share with their studied counterparts (e.g., perceptual, such as being presented in the same colors as weak/strong items) or the context in which they are tested (e.g., lures included at test after a study list of exclusively weak or strong items).

One of the key findings coming out of these strengthening manipulations is known as the *strength-based mirror effect*: Relative to weak items, the testing of lists of strong items results in a greater proportion of targets being recognized alongside a reduced proportion of recognized lures (Benjamin, 2001; Cary & Reder, 2003; Criss, 2006, 2009, 2010; Criss et al., 2014; Glanzer & Adams, 1985; Kılıç & Öztekin, 2014; Kılıç et al., 2017; Starns et al., 2010, 2012; Stretch & Wixted, 1998). There is general agreement that an increase in learning opportunities should improve the recognition of targets. But in the case of lures, there are disagreements on how to best explain the decrease in their recognition rates: The *criterion-shift account* argues that the increase in study opportunities affects metacognitive processes. Specifically, criteria are more stringent when judging test lists comprised of strong items than weak items (Cary & Reder, 2003; Stretch & Wixted, 1998; Verde & Rotello, 2007; Starns et al., 2010, 2012). In contrast, the *differentiation account* proposes that the additional study opportunities allow strong targets to be better differentiated from other items in general. One of the outcomes of this differentiation process is that lures become more distinct and less likely to be recognized (see Shiffrin & Steyvers, 1997; Criss, 2006, 2009, 2010; Criss et al., 2013; Kılıç et al., 2017; Koop et al., 2019).

Another key finding concerns the interactive effects (or lack thereof) between strong and weak items – the *null list-strength effect* (Hirshman, 1995; Murnane & Shiffrin, 1991; Ratcliff et al., 1990; Shiffrin et al., 1990; Yonelinas et al., 1992). This effect establishes that the propensity to recognize weak items is unaffected by their intermixing with strong items. This null effect contrasts with the case of free recall, where it is found that strong items benefit from their intermixing with weak items, whereas weak items are negatively affected (e.g., Malmberg & Shiffrin, 2005; Ratcliff et al., 1990; Wilson & Criss, 2017; see also Tulving & Hastie, 1972). Since its

[1] Throughout this paper, we will use the terms 'context' and 'source' interchangeably. However, we acknowledge that there are theoretical reasons to enforce a fine-grained distinction between the two, for instance between "list context" and "source context" (e.g., Osth et al., 2018). Our references to either "context" or "source" map onto the latter.

establishment, the null list-strength effect has been extensively studied, with numerous candidate accounts being proposed (e.g., Cary & Reder, 2003; Dennis & Humphreys, 2001; Murdock & Kahana, 1993; Shiffrin & Steyvers, 1997; see also, Osth and Dennis, 2014).

Although most investigations have focused on the effect of strengthening in terms of item judgments (was this item studied or not?), a small number of studies have turned their focus to its impact on *source judgments*. For instance, both Dobbins and McCarthy (2008) and Glanzer et al. (2004) reported higher source accuracy for deeply processed words (strong items) than shallowly processed ones (weak items). More recently, Starns and Ksander (2016) showed the strengthening of items through repetition also leads to increases in accuracy in both item and source judgments (for similar findings, see also Dobbins & McCarthy, 2008, Experiment 1; Glanzer et al., 2004, Experiment 1; Starns et al., 2013; Osth et al., 2018).

Starns and Ksander (2016) also investigated the effect of items occurring under more than one context or source. They found that the strengthening of one of these sources (through repetition) had a negative impact on the recognition of the other, non-strengthened sources. These results contrast with Dobbins and McCarthy's (2008) earlier report, where their findings did not suggest any negative impact for weak sources encountered earlier in the study phase (see their Table 5). Importantly, both findings were also observed by Kim et al. (2012), which suggests that the negative impact observed by Starns and Ksander (2016) might be due to a poorer encoding of sources encountered later during study. More recently, Osth et al. (2018) reported null list strength effects in source judgments, the only exception being a study in which source judgments were *not* preceded by item judgments. Osth et al. attributed this discrepancy to the expectation of mnemonic evidence being unavailable for items that would not have been recognized in the first place (see Batchelder & Riefer, 1990; Hautus et al., 2008; Klauer & Kellen, 2010).

1.2. The Retrieving Effectively from Memory model

Osth et al. (2018) relied on their source memory results to motivate the revision of a global matching model proposed earlier (Osth & Dennis, 2015). The present work follows along similar lines, although it turns its focus on a different theoretical contender – the REM model (REM; Shiffrin & Steyvers, 1997). REM describes episodic memory in terms of memory traces representing our experiences in the world. Each trace is represented as a vector, with each element therein referring to a unique feature. For example, suppose one had breakfast yesterday with oatmeal, milk, apple, and honey. In this case, each food ingredient is represented as an individual vector that stores its properties (e.g., the apple is a fruit, red, and sweet). In typical implementations of REM, it is assumed that each vector is comprised of twenty feature values and that each feature value *v* is a positive integer randomly sampled from a geometric distribution with parameter *g*:

$$P(v) = (1-g)^{v-1}g, \quad v = 1, 2, \cdots, \infty \tag{1}$$

When a memory trace is created, each feature is assumed to be stored with the probability *u* parameter. The vector elements associated with non-stored features all have a value of zero. With probability *c*, this storage is accurate. But when this is not the case, which is expected to occur with complementary probability *1-c*, a random value is assumed to be stored instead (this random value is sampled from the same geometric distribution). Going back to our earlier example, the feature of being "red" for the apple may not be stored in memory at all or incorrectly stored, e.g., "green". Altogether, this storage process is expected to yield memory traces comprised of correct, incorrect, and absent feature information.

According to REM, retrieval is based on a global matching process in which a *probe item* is compared to the existing episodic traces. This matching process quantifies the degree to which the features in the probe are the same as or different from those found in each trace out of a total of N traces. For a probe item *j*:

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

with *i* indexing episodic traces, and *v* feature values, respectively. The number of non-zero features that mismatch is *nq,* whereas the number of non-zero features that match is *nm*. Features that do not contain information (i.e., features with a value of zero) are not considered. These matches are averaged across traces, which yields an odds ratio Φ:

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

This ratio captures the relative support that the probe item was previously encountered. When Φ is higher than the decision criterion (typically set to 1), the item is endorsed as "old". Conversely, if Φ is lower than the criterion, the probe item is not endorsed (i.e., it is judged to be "new").

1.2.1. The list-strength effects in REM

The original implementation of REM assumes that the mnemonic benefits from additional study opportunities (e.g., repetition of items during study) come from the storage of currently absent features. Although this assumption implies that incorrectly stored

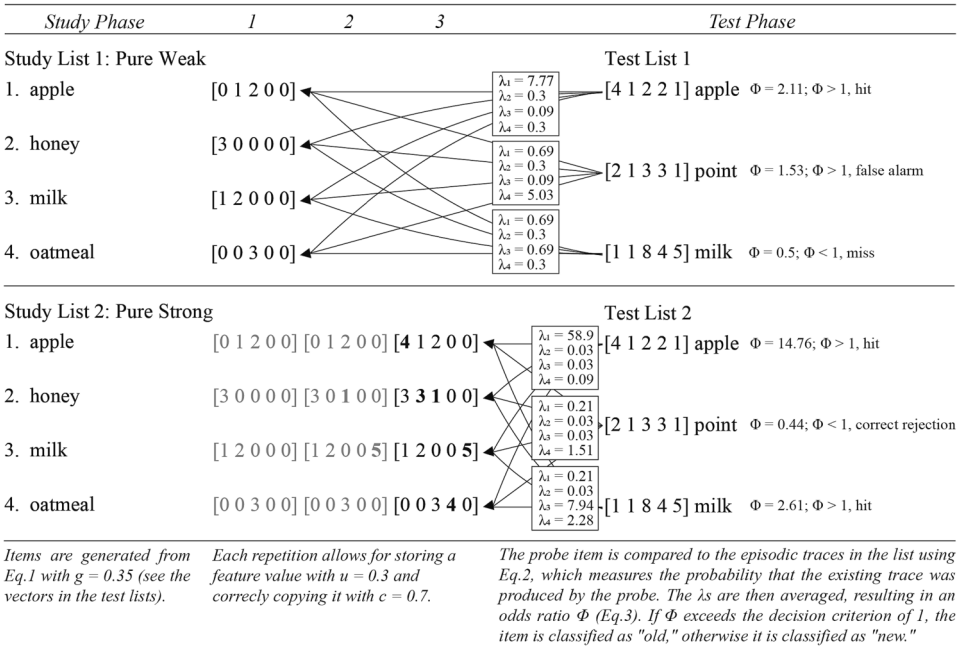


Fig. 1. *Differentiation Process During Study in REM.* Two lists are illustrated: (1) a list in which items are presented once and (2) a list in which items are presented three times. As items are repeated, memory traces become more similar to their corresponding targets and less similar to other items. This results in a greater match between a target and its memory trace, increasing the probability of a correct response (i.e., a hit), and a lower match between a probe and any other item, decreasing the probability of an incorrect response (i.e., a false alarm).

features will remain so (i.e., there is no updating), the end result is nevertheless a more complete and accurate representation of items, with memory traces becoming less similar to each other. This process, commonly referred to as *differentiation*, is illustrated in Fig. 1: As the number of study opportunities increases, so does the probability that items presented at test (targets and lures) are correctly judged. This is because lures become less similar to the existing traces, whereas targets become increasingly similar to the trace representing their previous occurrence and dissimilar from all other traces.

To make this process clear, let us walk through the example in Fig. 1:

i. Assume we have the exact same list of words, such as *apple, honey, milk,* and *oatmeal*, presented once (Study List 1: Pure Weak) or three times (Study List 2: Pure Strong).
ii. Each word is represented as a vector with five features for simplicity (sampled using Eq. (1)). After studying, these vectors contain positive integers or zeros, indicating available or missing information, respectively. The features are stored – correctly or incorrectly – probabilistically based on the *u* and *c* parameters.
iii. For items presented in the study phase, the model assumes the same kind of item-recognition judgment that takes place during the test phase. Now, let us suppose that a presented item is identified as previously studied. In that case, the best-matching trace (the one with the highest likelihood ratio) is updated. When items are repeated at study (e.g., Study List 2), each time they are presented constitutes an opportunity for additional features to be stored, as highlighted in bold in the figure.
iv. During the test phase, the test probes, including targets and foils like *apple, point,* and *milk,* are compared to each item trace, generating a likelihood ratio for each comparison (e.g., λ_1 is the likelihood ratio from comparing the probe with the first item *apple*, and so on). These likelihood ratios are averaged (Φ) and compared to the criterion of 1 for "old"-"new" decisions.
v. Additional study opportunities for target items result in stronger matches between their probes and their respective memory traces (e.g., the word "apple" produces $\lambda_1 = 7.77$ vs. $\lambda_1 = 58.9$ after its single vs. repeated presentations) and a weaker match with other traces (e.g., the word "apple" produces $\lambda_1 = 0.3$ vs. $\lambda_1 = 0.03$ after the single vs. repeated presentations of the word "honey").

1.2.2. Output interference in REM

The explanatory power of REM's trace updating processes and the differentiation that follows are not limited to strengthening effects. Given that these processes are also set to take place during testing, where target and lure items are re/encountered, they yield new testable predictions. Among these is the effect known as *output interference*, which is observed in item recognition as a decrease in recognition of targets throughout the test, while false endorsement of lures either slightly increases or remains unchanged (Annis et al., 2013; Criss et al., 2011; Criss et al., 2017; Kılıç et al., 2017; Koop et al., 2015; Malmberg et al., 2012).

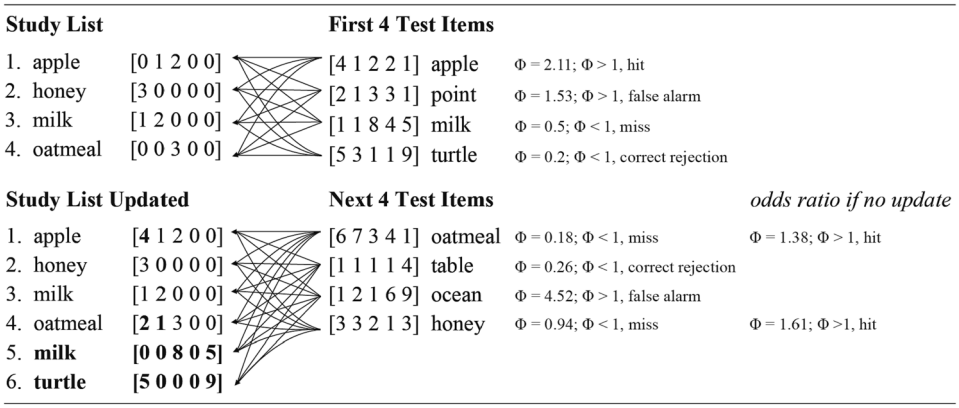


Fig. 2. *Differentiation Process During Test in REM*. This figure presents how REM integrates learning during the test phase: (i) memory traces are updated following old judgments, and (ii) new memory traces are stored after new judgments. Please note that, for simplicity, the figure illustrates updates after four items. However, in both the model and the simulations reported in this paper, updates occur after every decision. Also, note that the odds ratios for the targets "oatmeal" and "honey" are provided as reference points if there were no updates.

Criss et al. (2011) suggested that incorporating the differentiation process into the REM model during testing can explain the output interference observed in item recognition. As depicted in Fig. 2, when a probe item is judged to be old, the best-matching trace gets updated using the information provided by the probe. However, if the probe is a lure, this leads to an incorrect update of an existing trace representing a target's previous encounter, causing impaired recognition of that target later in testing. On the other hand, when a probe item is judged to be new, a new memory trace forms, increasing the number of items to compare and the overall noise.

Again, let us walk through the example in Fig. 2:

i. Assume the exact same word list used in the previous example. The target word "apple", when presented at the test, has a Φ of 2.11, surpassing the criterion of 1. Consequently, the best-matching trace is updated – in this instance, the target trace "apple" happens to be the best-matching trace and is updated with the storage of an additional feature, the integer "4", highlighted in bold.
ii. Continuing with the presentation of a lure item, "point", its comparison yields a Φ of 1.53, once again surpassing the criterion. But since this is a lure item, the best-matching trace is the one associated with the word "oatmeal". This, in turn, leads to an incorrect updating of this memory trace, which eventually leads to the word "oatmeal", when presented at the test, not being recognized.
iii. Lastly, let's examine the target item, "milk", and the lure item, "turtle", each judged to be new at the test. Given these judgments, both items are stored as new traces, shown as bold vectors under "Study List Updated". The inclusion of these new traces increases the length of the list of memory traces, introducing additional noise. One of its consequences is failure to recognize target items tested later on, such as "honey".

1.2.3. The question of context retrieval

An alternative version of REM included in its original proposal (REM.4; see Shiffrin & Steyvers, 1997) introduced the concept of *context features* that vary as time passes, alongside a threshold used to discriminate a list of items encoded in one context from lists learned in other contexts (context threshold). These context features are represented in *context vectors* appended to the vectors already postulated by REM to represent item features. Any two appended vectors stand as the mnemonic representation of a specific item encountered in a specific temporal context.

According to this REM version, retrieval follows a *two-step process* that begins with the activation of memory traces as a function of the similarity between their context features and the context probe. This is followed by a matching process that is circumscribed to the items whose context-feature activation was above the context threshold. Fig. 3 illustrates how context (e.g., breakfast in the morning) is appended to each item.[2] Returning to the example at the beginning of the section, let us ask whether one had an apple at breakfast. According to REM (or REM4, to be more specific), the context features (in this case, "breakfast") would first be used to activate the images of food eaten at breakfast, which would then be compared with the item probe "apple". However, note that this first activation is imperfect: On one hand, it is possible that memory traces representing the food eaten at breakfast might not pass the first context threshold. On the other, traces representing food encountered in different contexts (e.g., yesterday's breakfast) may be erroneously activated if their context is similar enough to the probe context. Altogether, this REM model expects attempts to recognize events that took place in a given context to be formed as a function of memory traces that include events that took place in said context but also events that took place elsewhere (for further details, see Shiffrin & Steyvers, 1997).

[2] For simplicity, this toy example assumes that both vectors are comprised of five features only.

Item | Context	Step 1	Step 2
apple | breakfast	[0 1 2 0 0][5 0 1 1 0]	[0 1 2 0 0][5 0 1 1 0]
oatmeal | breakfast	[0 0 3 0 0][0 0 6 1 0]	[0 0 3 0 0][0 0 6 1 0]
milk | breakfast	[1 2 0 0 0][0 3 1 0 0]	[1 2 0 0 0][0 3 1 0 0]
honey | breakfast	[3 0 0 0 0][0 0 1 1 2]	[3 0 0 0 0][0 0 1 1 2]
eggs | yesterday's breakfast	[7 3 0 0 0][5 3 0 0 1]	[7 3 0 0 0][5 3 0 0 1]
	[5 3 1 1 2] breakfast	[4 1 2 2 1] apple

Fig. 3. *Two-Step Recognition Memory Account in REM.* The first step is the activation of memory traces based on context features (e.g., breakfast). The second step is the item recognition process in which the probe item (e.g., apple) is matched to the item traces that passed the threshold.

The two-step process postulated by this REM variant was originally motivated by a desire to demonstrate how the model could be more efficient in retrieving items that occurred in a specific context (e.g., the study list, see Shiffrin & Steyvers, 1997, pp. 155). However, this proposal is insufficient in the sense that it does *not* define how the context(s) of an individual item would be retrieved from memory. In short, REM, in its current form, is unable to account for source-memory judgments.

1.2.4. Retrieving Effectively from Source Memory

For REM to describe source-memory judgments, we first need to establish how source information is stored. We begin by assuming that – similar to their 'item' counterparts – there is an imperfect storage of source features, which can take on incorrect values or be absent altogether (i.e., take on value zero). For simplicity, it is assumed that this storage process is assumed by the same probabilities u and c. The resulting source vector is appended to the item vector. Moreover, we will also assume that encountering the same item multiple times across different sources results in multiple source vectors being appended to the same item vector (see Fig. 4).

When an item probe is presented for item recognition, *the model ignores the source features, using only the item features to determine if the item was studied in the most current list.* As discussed earlier, this is achieved by comparing the features in the probe with those contained in the item traces (see Eq. (2)). In turn, when an item is endorsed as "old", the model uses the source features *of the best-matching trace* to determine its source (see Fig. 4). This is achieved by comparing the features in the source probe with those in the source traces of the item as follows:

$$\lambda(i,s,r) = (1-c)^{nq}(i,s,r)\prod_{v=1}^{\infty}\left[\frac{c+(1-c)g(1-g)^{v-1}}{g(1-g)^{v-1}}\right]^{nm(v,i,s,r)} \tag{4}$$

where i indexes the best-matching item trace, s the source trace(s), r the source probe, and v is the feature value in the source memory trace. Features that contain no information do not contribute to the decision process, as in the case of "old"-"new" judgments. If an item was studied in more than one source, the resulting likelihood ratios λ are averaged and converted into an odds ratio:

$$\Phi r = \frac{1}{N_s}\sum_{s=1}^{N_s}\lambda_{(s,r)} \tag{5}$$

where N_s is the number of source memory traces appended to the item trace. Like in item recognition, if the odds ratio Φ is higher than a criterion, the source is endorsed; otherwise, it is rejected.

When items are encountered repeatedly, it is assumed that there is an evaluation of whether they were previously encountered and, if so, whether this encounter took place in the same source (see Eqs. (2)–(5)). If the just-encountered source is not deemed to be novel, then the best-matching source trace is *updated*. Otherwise, a new source trace is appended to the item trace (see Fig. 4). But if an item is deemed to be novel (e.g., not a repetition), a new item trace and associated source trace are introduced.

Once again, let us walk through the example in Fig. 4:

i. Consider two items studied under different conditions: one involving multiple sources, like *apple* studied in multiple contexts, such as *breakfast*, *lunch*, and *dinner*, and the other involving a single context, like *oatmeal* studied in the context of *breakfast*.
ii. Whenever an item is introduced during the study phase, the model initially evaluates the item probe against the item traces previously stored in memory. It then proceeds to compare the source probe with the source(s) linked to the best-matching item trace. In this example, each repetition of the item, whether within the same context or a different one, results in an update of the correct memory trace. This update involves the probabilistic incorporation of new features into the vector – indicated by bold integers. Similarly, if an item is repeated within the same context, the corresponding source trace is probabilistically updated. On the contrary, when an item is repeated in new contexts, these contexts are appended to the existing item trace.

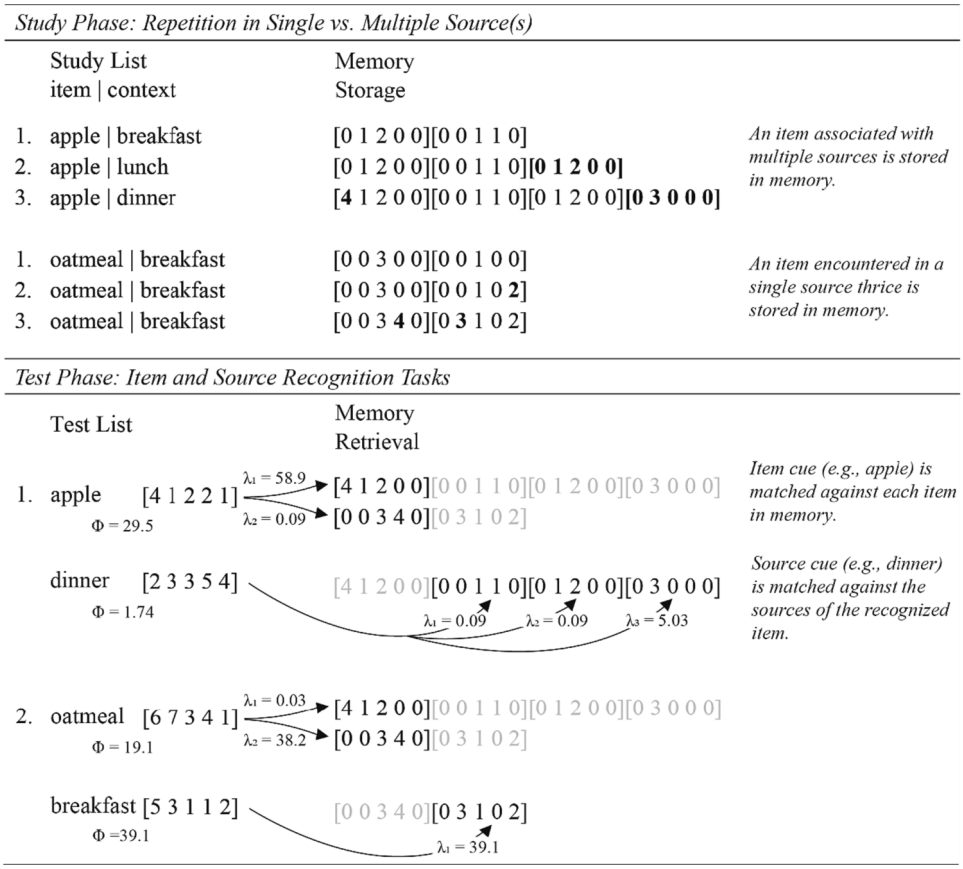


Fig. 4. *Retrieving Effectively from Source Memory*. The figure illustrates the Retrieving Effectively from Source Memory model using two simplified examples: one involves a single-source study of an item (such as oatmeal), and the other involves a multiple-source study of another item (such as apple).

iii. Note that if the reiterated context is not sufficiently similar to what has been stored, an update might not occur. In that case, the reiterated context could be appended to the item vector as if it were a novel context. Similarly, if a novel context closely resembles the stored source traces, it could erroneously update one of the existing context traces. The example is simplified as everything goes correctly.

iv. Transitioning to the test phase, the figure depicts two tasks: (1) item recognition and (2) source recognition, in which an item is presented in one of the potential sources. In the initial task, an item probe is presented without an associated context. The model assesses this item probe by comparing it to the existing item memory traces, similar to the original REM model, without considering context traces. In the subsequent task, the model evaluates the source probe by comparing it to the source vector(s) of the best-matching item trace. In either task, the resulting odds ratios are compared with a set criterion for the judgments related to item and source.

The aforementioned processes of updating and introducing new source traces imply the occurrence of *source differentiation* (analogous to the differentiation found in item judgments), a prediction that can be empirically tested. In the first series of experiments, we tested this prediction by means of study-repetition manipulations that vary 'source strength' while keeping 'item memory strength' constant, analogous to some of the previous studies that motivated the development of REM (Ratcliff et al., 1990; Shiffrin et al., 1990). These experiments will enable us to evaluate the empirical adequacy of source differentiation as well as the alternative—although not mutually exclusive—explanation provided by a criterion-shift account (see Cary & Reder, 2003; Stretch & Wixted, 1998; Verde & Rotello, 2007; Starns et al., 2010, 2012).

2. Experiment 1

The first experiment manipulated source strength while keeping constant the number of times each item was studied. Specifically, one list of items was presented three times with the same source, whereas another list was presented two times with one source and once with another. In a third list, each item was presented once with three different sources. To be clear, all the items in a given list were repeated thrice regardless of the number of sources in which they were studied.

2.1. Methods

2.1.1. Participants

Fifty-two undergraduates from Middle East Technical University (METU) participated in exchange for partial course credit. All participants were native Turkish speakers with normal color vision and normal or corrected-to-normal visual acuity. Ethical approval for the study was granted by the METU Human Subjects Ethics Committee (2017-SOS-172). Prior to the experiment, each participant provided written informed consent. Two participants were removed because of not complying with the instructions: One gave no responses, and the other always pressed "yes" when making source recognition judgments. After removing these two individuals, the data of 50 participants (M age = 21.4, SD age = 1.77) were used in the final analysis. 60% of the participants were females, and 74% were right-handed.

2.1.2. Materials

Words were randomly sampled from Turkish Word Norms (Tekcan & Göz, 2005) after removing color words such as "blue", "yellow", or "black" and words having less than four and more than seven letters. The computer screen was divided into four, and each quadrant was assigned a different color to create source information.

2.1.3. Procedure

The experiment consisted of eight study-test cycles, each including 24 words at study. None of these words was re-used across cycles. Between each study and test phase, participants completed a digit-sum task, where they were presented with a sequence of random digits, which they were then required to add on at a time.

Three distinct conditions were used in the study phases, with their order randomized across participants. In the first condition, defined as a single-source repetition condition (3–0-0 condition), items were repeated three times in a single source. That is not to mean that all items presented in this condition were shown in one source; rather, after all items were first presented in a randomly assigned source, they were presented two more times in the same source, for a total of three presentations (see Fig. 5A). In the two-source repetition condition (2–1-0 condition), items were randomly assigned to two sources and randomly presented once in one source and twice in the other. In the final three-source repetition condition (1–1–1 condition), items were presented in three different sources. The 3–0-0 and 1–1–1 conditions were assigned to two study-test cycles. In contrast, the 2–1-0 condition was assigned to four study-test cycles. The rationale behind this will become clear below. These manipulations of source repetition are expected to introduce variation in source strength (with greater values for repeated sources) while keeping item strength equalized (as all items are studied thrice).

Each test list included 30 words: 24 targets and six lures. Participants first completed a yes–no item recognition task where they were required to indicate if the presented word was studied. Source judgments were elicited for the items recognized as old. One-fourth of the targets were presented in a new source – *source foils* – and three-fourths in their studied source – *source targets*. Source foils were randomly selected from one of the sources that the item had never been shown in the study. On the other hand, source targets were selected for each condition as follows: In the 3–0-0 condition, targets were re-presented in the only source in which they were studied before, whereas in the 2–1-0 and 1–1–1 conditions, one of the sources in which items were presented was randomly selected.

In the study phase, items included in the 2–1-0 and 1–1–1 conditions were presented in two and three different sources, respectively. Since the items in the 1–1–1 condition had been displayed an equal number of items (i.e., once) per source, source targets were randomly selected from one of the three sources. On the other hand, the 2–1-0 condition had sources in which items were repeated twice or once, and the current study tested both sources. As mentioned earlier, the 2–1-0 condition was assigned to four study-test cycles—twice as many as the other conditions. The reason behind this allocation was to test twice-repeated sources separately in two of these cycles, and the once-repeated sources in the other two cycles (i.e., test lists were always pure; see Fig. 5B). Note that all these cycles had separate study lists, meaning that no item-source pairing was repeated across cycles. The twice repeated sources from the 2–1-0 condition played a critical role in our assessment of how source strength increases. In turn, the sources presented once provided a point of comparison with their counterparts in the 1–1–1 condition.

2.2. Results and discussion

2.2.1. Item recognition

One-way repeated-measures ANOVA revealed that neither the probability of correctly endorsing a studied item (i.e., hit rate) nor the probability of falsely accepting a new item as old (i.e., false alarm rate) in item recognition was different across conditions, $F(3,147) = 0.39$, $p = .76$ and $F(3,147) = 1.28$, $p = .28$, respectively. Likewise, no significant difference was observed in the sensitivity of discriminating old items from new items (i.e., d', see Kellen & Klauer, 2018) across conditions, $F(2.76,135.24) = 2.35$, $p = .08$ (see Fig. 6A). These results suggest that additional learning benefits item memory irrespective of the number of different sources that deliver the information. In other words, obtaining information from multiple sources improves memory as much as obtaining information repeatedly from a single source (see also Starns & Ksander, 2016).

2.2.2. Source recognition

As previously mentioned, participants were asked to make source judgments for any item they had recognized at the test, regardless of whether the item was in fact a target or foil. However, it only makes sense to talk about source accuracy for targets. The hit rate was defined as the proportion of correct endorsements of targets presented at the test in one of their original sources. In turn, the false-alarm rate was defined as the proportion of incorrect endorsements of targets presented in a new source at the test.

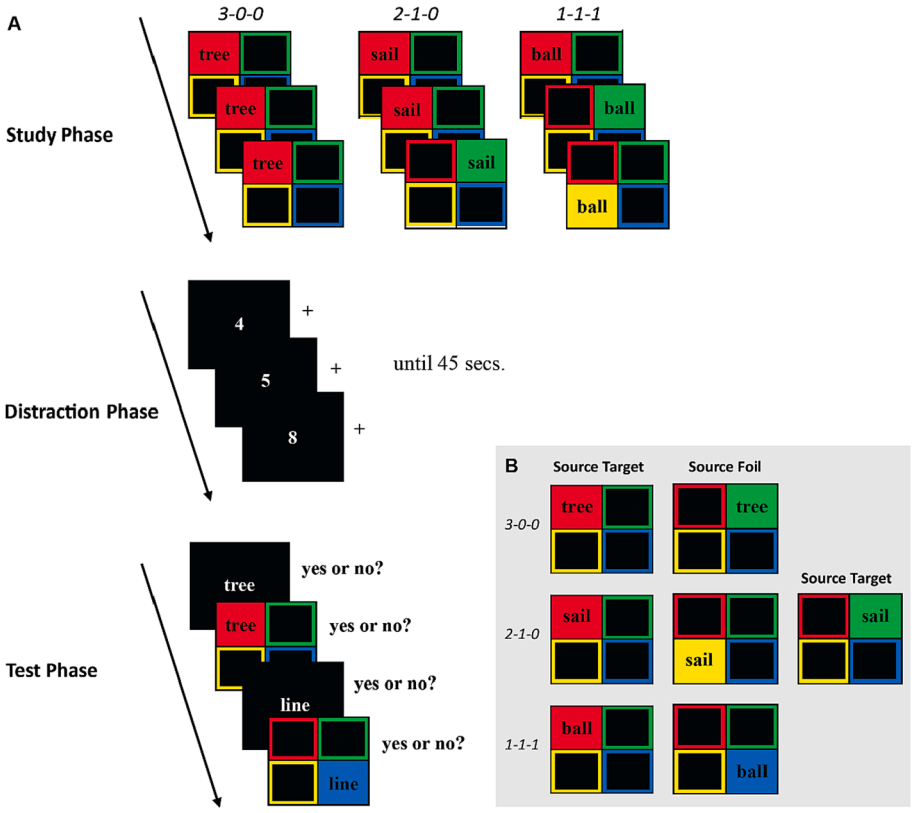


Fig. 5. *Illustration of Study-Test Cycle in Experiment 1.* All sources were used an equal number of times in each condition, and during the repetitions, all of the items were presented before an additional repetition occurred. Items were displayed for two seconds with 250 msec interstimulus-interval at the study. Figure B on the gray background illustrates example source targets and foils for each condition. Please note that the figure is simplified such that both sources that presented the same item once or twice are illustrated as source targets in the 2–1-0 condition. In the 2–1-0 example, *sail* was presented twice in the red box and once in the green box. However, both sources of the same item were never tested in the experiment. Instead, the condition was assigned to four study-test cycles, of which half tested twice-repeated sources while the other half tested once-repeated sources for different item lists. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

One-way repeated measures ANOVA showed a significant difference for hit rates in source memory across conditions, F (3,147) = 62.15, $p < .001$, $\eta^2 = 0.56$. Post-hoc comparisons using a *t*-test with Bonferroni correction further indicated that the repetition of items thrice in the same source generated higher source recognition compared to the repetition of items twice in the same source ($M = 0.09$, 99.2 % *CI* = [0.04, 0.13], $p < .001$, $d = 0.71$)[3]or once, either in the 2–1-0 condition ($M = 0.23$ [0.17, 0.29], $p < .001$, $d = 1.54$) or the 1–1–1 condition ($M = 0.19$ [0.14, 0.24], $p < .001$, $d = 1.50$). Similarly, sources were better recognized when presented twice relative to single-source presentations in the 2–1-0 ($M = 0.15$ [0.09, 0.20], $p < .001$, $d = 1.06$) and 1–1–1 conditions ($M = 0.11$ [0.05, 0.16], $p < .001$, $d = 0.81$). Finally, there was no significant difference between the recognition of sources presented once in the 2–1-0 and the 1–1–1 conditions ($M = 0.04$ [-0.01, 0.09], $p = .15$) (see the left-hand side of Fig. 6B).

False alarm rates in source memory significantly differed across conditions as one-way repeated measures ANOVA revealed, F (3,147) = 25.64, $p < .001$, $\eta^2 = 0.34$. According to the post-hoc comparisons using a *t*-test with Bonferroni correction, the probability of falsely recognizing a new source as old was lower for the 3–0-0 condition than the twice-repeated sources ($M = -0.18$ [-0.27, -0.09], $p < .001$, $d = -0.74$) and once-repeated sources in the 2–1-0 condition ($M = -0.22$ [-0.31, -0.13], $p < .001$, $d = -0.95$) or the 1–1–1 condition ($M = -0.28$ [-0.39, -0.17], $p < .001$, $d = -0.99$). In addition, there were fewer incorrect endorsements of source foils when the test list contained the sources that presented items twice from the 2–1-0 condition relative to the 1–1–1 condition ($M = -0.10$ [-0.18, -0.02], $p = .007$, $d = -0.49$). More importantly, there was no significant difference in the endorsements of source foils when the test list contained the sources that presented items twice or once from the 2–1-0 condition ($M = -0.04$ [-0.13, 0.05], $p = 1$). Finally, the results did not reveal any significant difference in the acceptance of new sources when items were studied once in a source – and two more times in another source – in the 2–1-0 condition versus once in three different sources in the 1–1–1 condition ($M = 0.06$ [-0.03, 0.15], p

[3] Due to the multiple comparisons conducted in Experiment 1 and the associated Bonferroni corrections, we report 99.2% confidence intervals. Everywhere else we will report intervals with 95% coverage. Also, note that in the present comparisons *M* and 99.2% *CI* represents the mean difference and the confidence interval for the difference in the means, respectively.

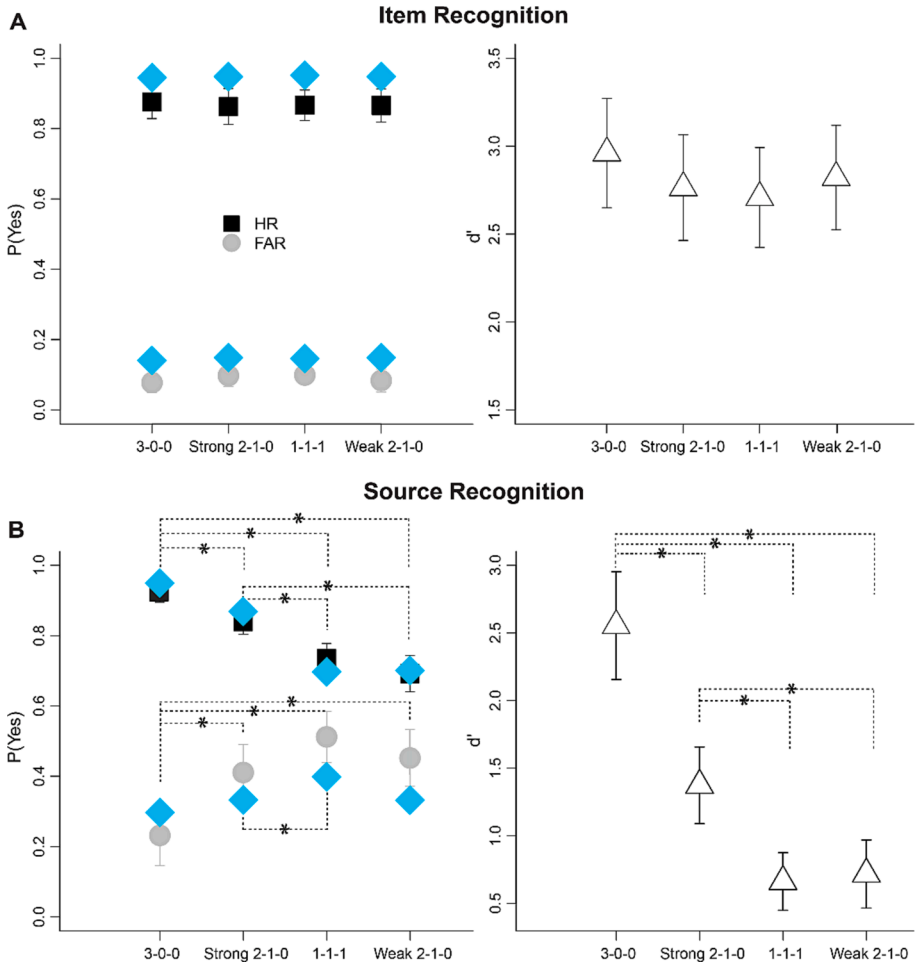


Fig. 6. *Illustration of Results from Experiment 1.* The figures demonstrate hit rates, false alarm rates, and d-prime as a function of strength for item recognition on the top and source recognition on the bottom. Strong 2–1-0 refers to the test list containing strong sources, while weak 2–1-0 refers to the test list containing weak sources from the 2–1-0 condition. The squares and the circles represent the averaged hit and false alarm rates over participants, respectively. The triangles represent the sensitivity to discriminate targets from foils (d-prime) across participants. Blue diamonds represent the predicted hit and false alarm rates from the REM model. (* = $p < 0.05$). Error bars represent 99.2 % confidence intervals. Please refer to Table 1 for the descriptive statistics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

= .42) (see the left-hand side of Fig. 6B).

These findings indicate that the increase in source repetitions improved the recognition of studied sources while diminishing the incorrect endorsement of new sources. This pattern suggests the presence of a strength-based mirror effect in source memory similar to what is commonly found in item recognition (e.g., Kılıç & Öztekin, 2014; Kılıç et al., 2017). The lack of differences in false alarms between twice and single-presented sources in the 2–1-0 condition is also relevant given that, from a differentiation model perspective (Shiffrin & Steyvers, 1997), it could be assumed that the source foil is compared with memory traces representing both strong and weak sources, which would virtually produce no difference based on different test lists. However, these results can also be accommodated by a criterion shift (e.g., Stretch & Wixted, 1998; Starns et al., 2010): Given that participants were not informed about the content of test lists, they might have adopted similar decision criteria regardless of being tested with strong versus weak sources, which would lead to equivalent false alarm rates. We will return to this issue later on.

Turning our attention to source discriminability, one-way repeated measures ANOVA indicate that it differs across conditions, $F(2.48,121.42) = 83.83$, $p < .001$, $\eta^2 = 0.63$. Post-hoc comparison using a *t*-test with Bonferroni correction showed that thrice-presented sources produced higher discriminability than their twice-presented ($M = 1.18$ [0.75, 1.62], $p < .001$, $d = 1.05$) and once presented counterparts ($M = 1.84$ [1.43, 2.25], $d = 1.75$ and $M = 1.89$ [1.44, 2.35], $p < .001$, $d = 1.62$). Additionally, the discriminability was higher for the test lists containing strong versus weak sources from the 2–1-0 condition ($M = 0.66$ [0.33, 0.99], $p < .001$, $d = 0.79$) due to the difference in hit rates. Similarly, the twice-presented sources had greater discriminability than once-presented sources in the 1–1–1 condition ($M = 0.71$ [0.42, 1.01], $p < .001$, $d = 0.94$). On the other hand, there was no significant difference in discriminability between once-presented sources from the 2–1-0 and 1–1–1 conditions, M = -0.06 [-0.34, 0.23], $p = 1$ (see the right-hand side of

Table 1
Descriptive statistics for source recognition in experiment 1.

	HR			FAR			d′		
	M	SD	99.2% CI	M	SD	99.2% CI	M	SD	99.2% CI
3–0-0	0.93	0.09	[0.89, 0.96]	0.23	0.24	[0.14, 0.33]	2.55	1.14	[2.11, 2.99]
Strong 2–1-0	0.84	0.10	[0.80, 0.88]	0.41	0.23	[0.32, 0.50]	1.37	0.81	[1.06, 1.69]
Weak 2–1-0	0.69	0.15	[0.64, 0.75]	0.45	0.23	[0.36, 0.54]	0.72	0.72	[0.44, 0.99]
1–1–1	0.74	0.12	[0.69, 0.78]	0.51	0.21	[0.43, 0.59]	0.66	0.61	[0.43, 0.90]

Note. Strong 2–1-0 refers to the sources in which items were repeated twice, while weak 2–1-0 refers to the sources in which items were shown once from the 2–1-0 condition. M refers to the mean, SD to the standard deviation, and CI to the 99.2% confidence interval. HR refers to hit rate, FAR to false alarm rate, and d′ to discriminability index from Signal Detection Theory.

Fig. 6B).

Overall, the findings show that the more times items are studied in the same source, the greater the discriminability between studied sources and new sources. Another critical finding from the current experiment was the comparable memory performance for the sources encountered once in the 2–1-0 and 1–1–1 conditions. Note that the 2–1-0 condition was a mixed study list design containing both strong and weak sources. On the other hand, the 1–1–1 condition was a pure study list design in which each source was weak. The results suggesting similar source memory performances with similar hit and false alarm rates can be interpreted as a null list-strength effect in source memory.

3. Experiment 2

Experiment 2 further investigated the strength-based mirror effect in source memory by focusing on the 3–0-0 and 1–1–1 conditions. Specifically, a mixed study list design was implemented in which half of the study items were presented in a single source, and the other half were presented in multiple sources.

3.1. Methods

3.1.1. Participants

Fifty undergraduates (*M* age = 20.98, *SD* age = 1.62) from METU participated in the study for partial course credit. 80% of the participants were female, and 88% were right-handed. All participants were native Turkish speakers with normal color vision and normal or corrected-to-normal visual acuity.

3.1.2. Materials

The same word pool and source information as in Experiment 1 were used.

3.1.3. Procedure

The procedure was similar to Experiment 1 except for the number of conditions and the list types. Specifically, this experiment contained study lists in which half of the study items were presented three times in a single source (strong source items) and the other half in three different sources (weak source items), as shown in Fig. 7. Although the study lists intermixed weak and strong sources, the test lists were always pure, with either strong or weak sources being tested. Each test list contained 15 items: Twelve targets plus three foils. Each test item was first shown at the center of the screen until an item-recognition judgment was made. Any item recognized as old, whether previously studied or not, was then presented in one of four possible sources. Participants were then asked to endorse the source if the item had been studied there before, or otherwise reject it. Nine targets were shown in one of their original sources, with the remaining being shown in a source in which they had not been studied before.

3.2. Results and discussion

Item recognition results replicated the ones from Experiment 1. Specifically, the benefits from repeated study for the accuracy of item-memory judgments were not affected by the number of sources encountered in these repetitions. A paired-sample *t*-test revealed no difference across conditions for hit rates, t (49) = -0.24, p = .82; for false alarm rates, t (49) = -0.35, p = .73; and for d', t (49) = -0.77, p = .45 (see Fig. 8A). Moreover, paired-sample t-tests indicated better recognition success for strong over weak sources, both in terms of the recognition of encountered sources, t (49) = 10.12, M = 0.19 [0.16, 0.23], $p < .001$, d = 1.43, as well as in terms of the incorrect acceptance of new sources, t (49) = -2.38, M = -0.06 [-0.11, -0.01], p = .021, d = -0.34. These differences imply a greater discriminability for strong sources, t (49) = 9.64, M = 1.11 [0.88, 1.34], $p < .001$, d = 1.36 (see Fig. 8B).

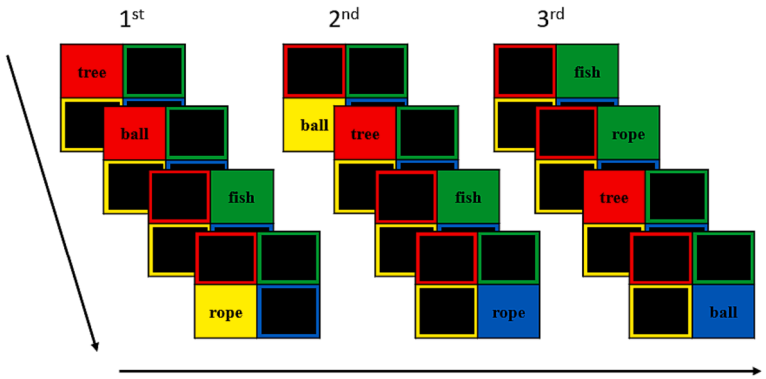


Fig. 7. *Illustration of a Study Phase for Experiments 2 and 3.* The figure illustrates the first, second, and third repetitions of a word list, with half of the items repeated in a single source and the other half repeated in multiple sources. Additional repetitions were introduced after all the words were initially presented.

4. Experiment 3

This experiment investigated the impact of participants' knowledge regarding the composition of the test lists on the placement of response-criteria. More specifically, the goal was to better understand the role that changes in response criterion might have in accounting for the results observed in the previous experiment. In the previous experiment, the test lists were always pure, i.e., comprised of items with sources studied once or thrice. It is possible that participants noticed this at the early stages of each test phase and adjusted their criterion accordingly (see Verde & Rotello, 2007). Like Experiment 2, this experiment tested items with sources studied once and thrice, but this time participants were informed of the composition of the test lists.

4.1. Methods

4.1.1. Participants

Fifty-four undergraduates from METU participated in exchange for partial course credit. Four participants were excluded from the analysis due to not following the instructions: Two participants were excluded due to not making source-memory judgments, and another two were excluded due to pressing the same response key for all source-memory judgments. In the end, the data from 50 individuals (*M* age = 21.98, *SD* age = 2.18) were retained for analysis. 56% of the participants were female, and 88% were right-handed. All participants were native Turkish speakers with normal color vision and normal or corrected-to-normal visual acuity.

4.1.2. Materials

The materials were identical to Experiment 1.

4.1.3. Design and procedure

The procedure was exactly the same as Experiment 2, except that the participants were informed about test-list strength in Experiment 3. For the test lists that included the items shown in a single source, participants were instructed as follows: "*You will be tested on the words repeated at the SAME location of the computer screen. Press '1' to proceed.*". For the test lists that included the items repeated in multiple sources, participants were instructed as follows: "*You will be tested on the words repeated at DIFFERENT locations of the computer screen. Press '3' to proceed.*".

4.2. Results and discussion

Item recognition results replicated the results of both Experiments 1 and 2. A paired-sample *t*-test has revealed no statistical difference in hit rates, $t\,(49) = -0.53$, $p = .60$, in false alarm rates, $t\,(49) = -0.40$, $p = .69$, or in d′, $t\,(49) = 0.11$, $p = .92$ between the 3–0-0 and 1–1–1 conditions in item recognition (see Fig. 8C).

Source recognition results show a more robust strength-based mirror effect than those reported in Experiment 2. Sources that provided items thrice were recognized better compared to those that provided items only once, as paired-sample *t*-test revealed, $t\,(49) = 10.73$, $M = 0.22$ [0.18, 0.26], $p < .001$, $d = 1.52$. New sources were falsely identified less often for items studied repeatedly in a single source than items studied in several sources, $t\,(49) = -5.56$, $M = -0.17$ [-0.23, -0.11], $p < .001$, $d = -0.79$. Compared to the previous results, a larger effect size is noticeable for the difference in the endorsements of source foils in Experiment 3 ($d = 0.79$) than in Experiment 2 ($d = 0.34$). On the other hand, the effect size is still smaller than Experiment 1 ($d = 0.99$). This suggests that the difference in the errors on the tests containing weak versus strong sources increased substantially when people were informed of the content of the test. However, the difference is still smaller for the mixed-study lists than the pure-study ones. Overall, the 3–0-

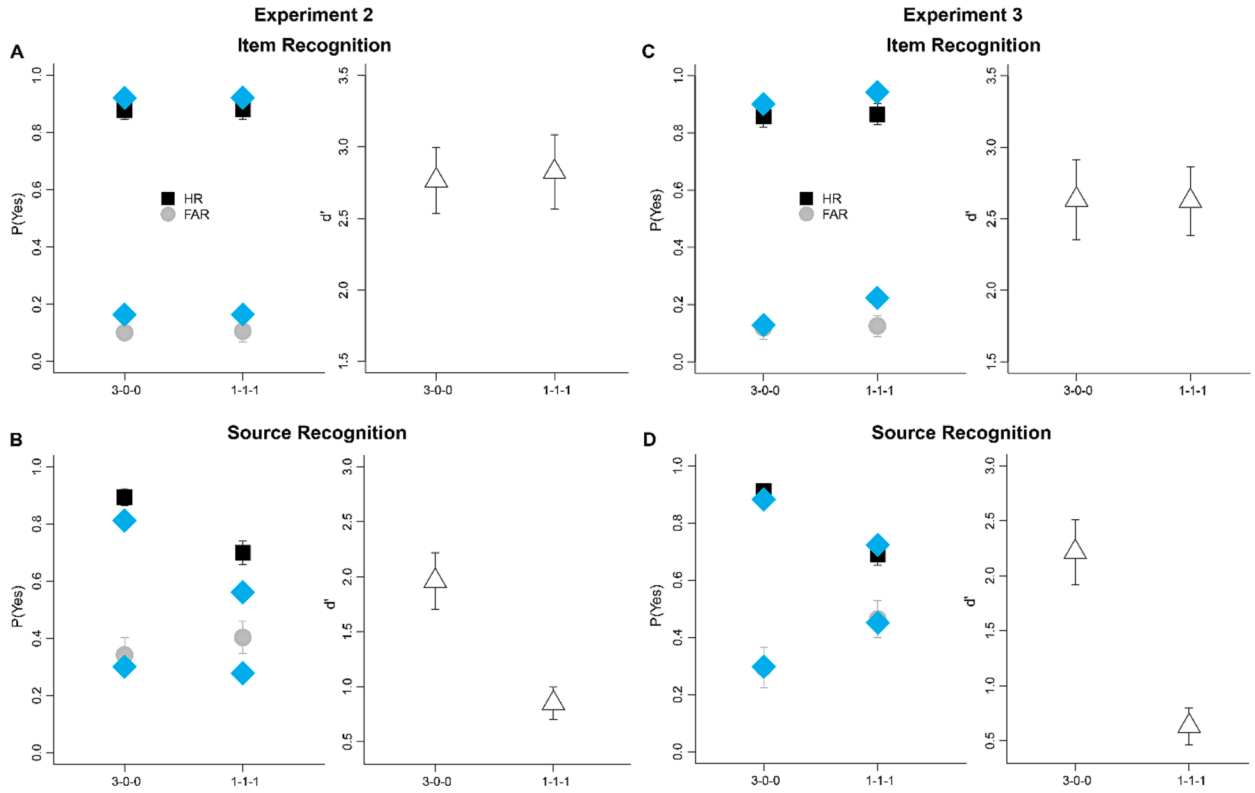


Fig. 8. *Illustration of Results from Experiments 2 and 3*. The figures demonstrate hit rates, false alarm rates, and d-prime as a function of strength for item recognition on the top and source recognition on the bottom. The squares and the circles represent the averaged hit and false alarm rates over participants, respectively. The triangles represent the sensitivity to discriminate targets from foils (d-prime) across participants. Blue diamonds represent the predicted hit and false alarm rates from the REM model. (* = $p < 0.05$). Error bars represent 95 % confidence intervals. Please refer to Table 2 for the descriptive statistics. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Descriptive statistics for source recognition in experiments 2 and 3.

	HR			FAR			d′		
	M	SD	95% CI	M	SD	95% CI	M	SD	95% CI
Experiment 2									
3–0-0	0.89	0.11	[0.86, 0.92]	0.34	0.22	[0.28, 0.4]	1.96	0.91	[1.7, 2.22]
1–1–1	0.7	0.14	[0.66, 0.74]	0.4	0.2	[0.35, 0.46]	0.85	0.52	[0.7, 1]
Experiment 3									
3–0-0	0.91	0.09	[0.89, 0.94]	0.29	0.25	[0.22, 0.37]	2.21	1.04	[1.92, 2.51]
1–1–1	0.69	0.13	[0.65, 0.73]	0.46	0.23	[0.4, 0.53]	0.63	0.59	[0.47, 0.8]

Note. M refers to the mean, SD to the standard deviation, and CI to the 95% confidence interval. HR refers to hit rate, FAR to false alarm rate, and d′ to the discriminability index from Signal Detection Theory.

0 condition had higher source discriminability than the 1–1–1 condition, $t(49) = 12.15$, $M = 1.58$ [1.32, 1.84], $p < .001$, $d = 1.72$ (see Fig. 8D).

5. Model simulations

In this set of simulations, we followed our experimental designs and created forty-eight targets along with four unique sources. The features representing the properties of items and sources were randomly sampled from a geometric distribution with parameter $g = 0.35$ (see Eq.1). Both item and source vectors contained twenty features each ($l = 20$).

To simulate the study phase, we set the storage parameter u to 0.3 for item memory and 0.1 for source memory. The rationale behind using different u parameters for item and source was the observation that people tend to have greater difficulty with source

judgments than item judgments (e.g., Kellen et al., 2014). However, we set the storage-accuracy parameter c to 0.7 for both item and source memory, therefore establishing source memory traces as incomplete and error-prone as their item-memory counterparts.

The presentation of items in the study phase was implemented in the model through the application of the recognition process formalized by Eqs.(2)–(3). Specifically, each item presentation initiated an evaluation of whether that item was presented before or not (i.e., if it was a repetition). This evaluation was based on a comparison of said item's features with those of the items encountered so far. In the case of successful recognition – when the item's overall match surpassed a criterion of 1 – the best-matching trace was updated by replacing its empty features with integers (note that this updating is performed by a probabilistic process, described earlier, with encoding parameters u and c, and therefore, some features might not be replaced). Otherwise, a new item trace was stored along with a source trace. A very similar process was assumed for sources: the source features of the best-matching item trace are compared with the just-presented item's. If this comparison surpasses the established criterion of 1, then the source trace is updated. If not, then a new source trace is appended. These evaluation processes were assumed to take place until the end of the study phase.

For the test phase, we once again followed our experimental design and considered the evaluation of the forty-eight targets (with the same or different sources) along with twelve lures. Similar to the study phase described above, for every item presented in the test, the model first evaluated whether the item was old and then – if it was – whether the source was original or new by comparing the odds ratios with a criterion of 1 (as shown in Fig. 4). We considered one-thousand synthetic participants and reported the model's predictions averaged across said participants in the following section.

5.1. Model predictions

As illustrated in Fig. 6A, the model fits reflect the finding that repetition, whether in a single or multiple source(s), improves item recognition. According to the model, each repetition creates an opportunity for storing an unlearned item feature in the trace, leading to stronger memories.

In the case of source memory, the model also captures the pattern that repetitions increase correct-recognition probabilities for studied sources while decreasing false-recognition probabilities for new sources – a strength-based mirror effect (see Fig. 6B). According to the model, each repetition in the same source enables an update of that source trace. This update leads to a more complete trace, further differentiating it from the other traces. Consequently, presenting the item in its original source increases the match between the features of the source probe and its corresponding trace, leading to higher recognition of the source. On the other hand, presenting the item in a new source decreases the match, leading to decreased false recognition of this new source. Moreover, the model predicts no difference in the probability of falsely recognizing new sources for items studied in the context of weak and strong sources, which is in line with the data (see the lower left panel of Fig. 6). To better understand how the model makes this prediction, consider a scenario where a word was studied thrice across two sources, twice with source A and once with source B. Because the model compares the source cue with both source memory traces associated with this word, no difference is expected in terms of the false-alarm rates.

Turning our attention to experiments 2 and 3, we simulated model predictions for lists in which half of the items were repeated in a single source (3–0-0), whereas the other half were studied in three different sources (1–1–1). The parameter values adopted were the same as in the previous simulations (see Table 3). As shown on the left-hand side of Fig. 8, the model predicts higher recognition rates for strong sources than weak sources, which is consistent with the data. Additionally, the model suggests almost no difference in the incorrect judgments of source foils between the 3–0-0 and 1–1–1 conditions. This is generally consistent with the data since the difference between these two groups was small (Cohen's $d = 0.34$).

In the case of Experiment 3, participants were informed of the content of the upcoming test lists, which elevated the observed differences in source false-alarm rates between strength conditions (Cohen's $d = 0.79$). This elevation might be due to people adjusting their criteria in response to the information provided (e.g., Starns et al., 2010; Kılıç & Öztekin, 2014; Kılıç et al., 2017). We implemented this possibility of criterion shift by setting the criterion to 1.2 for the test list that included items from the 3–0-0 condition, and

Table 3
Parameter Values Used in the REM Simulations.

Parameter	Value	Description
l	20	Vector length
g	0.35	Feature frequency
c	0.7	Probability of correctly copying a feature
u_{item}	0.3	Probability of storing an item feature
u_{source}	0.1	Probability of storing a source feature
Criteria		Criteria for endorsing a probe
Exp 1 & 2	1	
Exp 3: 3–0-0	1.2	
Exp 3: 1–1–1	0.8	
Submodel: 3–0-0	1.9, 0.7	Criteria for item and source memory, respectively
Submodel: 1–1–1	1.8, 0.3	
Submodel: Strong 2–1-0	1.9, 0.4	
Submodel: Weak 2–1-0	2, 0.3	

Note. The submodel represents the simulation of the REM submodel without differentiation but with varying criteria, illustrated in Fig. 9B.

to 0.8 for the test list that included items from the 1–1–1 condition. The model fits the source memory data as illustrated on the right-hand side of Fig. 8. Although this shift improved model predictions for source judgments, it negatively impacted the predictions for item-memory judgments. This tradeoff is due to the simplifying assumption that the same criterion applies to both item and source judgments. We expect this discrepancy to disappear if we conduct an exhaustive search for best-fitting parameters across tasks, conditions, and individuals.

Altogether, the model predictions produced here showcase how REM, through its assumption of differentiation, can account for the strength-based mirror effects and null list-strength effects found in source memory. However, it would be a mistake to interpret these results as somehow implying that no other processes, such as *criteria shifting,* could be playing a major role. They very well may. After all, it is possible that participants adopted a different criterion depending on the strength of the sources encountered in the study phase (see Hirshman, 1995; Starns et al., 2010, for a discussion on item recognition; Starns et al., 2013). This possibility raises questions regarding the actual role of differentiation: If the observed effects can be accounted for by criterion shifting, then is differentiation even necessary to begin with? The model simulations discussed in the section below address this question.

5.2. *Exploring alternative explanations*

To assess the necessity of differentiation to explain the present data (with the previous simulations establishing its sufficiency), we began by considering a REM submodel that does *not* include it. Instead, it assumes that the repeated study of items and sources always

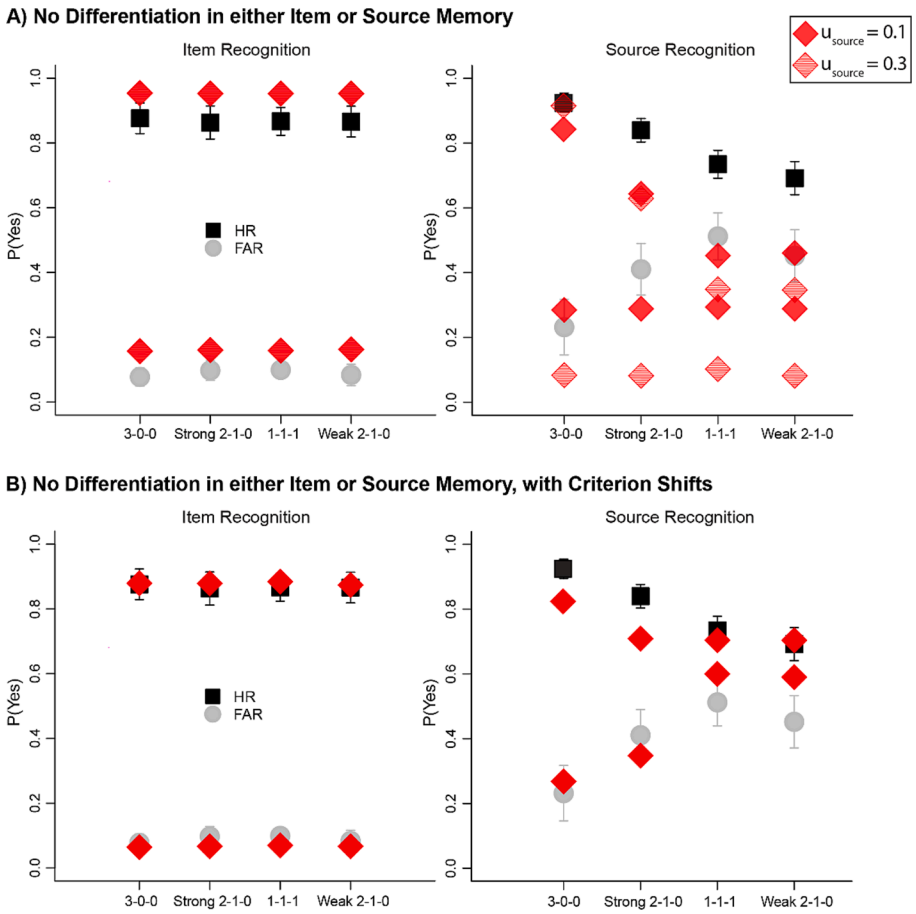


Fig. 9. *Predictions of a REM Submodel without Differentiation.* The figures illustrate the predictions of a submodel that stores new traces with each repetition (instead of updating the existing traces) for both item and source memory; in other words, no differentiation occurs in either item or source memory. Solid shapes represent the model's predictions when the *u* parameter is set to 0.1 for source memory, while dashed shapes represent those when the *u* parameter is set to 0.3 for source memory. The top figure represents the submodel's predictions with a constant criterion (*criterion* = 1), while the bottom figure represents the submodel's predictions with varying criteria (see Table 3).

leads to the creation of new traces. Specifically, the submodel assumes no differentiation for *both* item and source memory. We considered its predictions when the probability of storing a source feature is the same as in the REM model discussed earlier ($u_{source} = 0.1$; see Table 3), as well as when it takes on a higher value ($u_{source} = 0.3$). As shown in Fig. 9A, the submodel's predictions are not consistent with the source memory data. This submodel grossly underestimates the proportion of hits in source memory with varying repetitions. Note that the submodel stores three separate item traces for the same item, so the odds that any of these three traces align with the item probe are essentially the same. What this means is that the best-matching item memory trace is basically a random draw from one of these traces. This is unproblematic in the case of item-memory judgments or source judgments in which only one source was encountered multiple times (e.g., the 3–0-0 condition) but not in cases where the same item was encountered across multiple sources. This is particularly clear when the probability of storing source features is higher (e.g., $u_{source} = 0.3$). Also problematic is the submodel's inability to capture the differences in the false recognition of new sources across conditions. Varying the storage parameter u_{source} can affect the overall false recognition rates but not overcome the fact that the same rate is predicted across the different strength conditions.

The submodel considered so far assumed the same response criteria across conditions. We relaxed this restriction by allowing response criteria to vary (i) across strength conditions and (ii) across item and source memory tasks. The best-fitting criteria (i.e., those that minimize squared errors) were found through a grid search (see Table 3). As can be seen in Fig. 9B, the submodel is able to capture the main qualitative trends in the data, which suggests that a pure criterion-shift account can provide an empirically-adequate account.[4]

Based on these results, one might be tempted to infer that there are limited grounds for the idea of differentiation at the level of source judgments, as proposed by our REM extension. After all, a pure criterion-shift account appears to do a good-enough job.[5] We identify two reasons why such a move would be premature: First, there is the question of how plausible a criterion-shift account is, to begin with, given how much we know about people's reluctance towards shifting criteria. There is a large body of work showing that people often fail to adjust their decision criterion according to changes in memory strength within the same test (e.g., Stretch & Wixted, 1998; Morrell et al., 2002; Starns et al., 2006; Starns et al., 2010, Experiment 4; Singer & Wixted, 2006, Experiments 1–2; Verde & Rotello, 2007, Experiments 1–4). Instead, criteria placement seems to be largely driven by stable individual-level predispositions (Kantner & Lindsay, 2012; 2014; see also Miller & Kantner, 2020; but see Starns et al., 2010). In the three experiments reported here so far, item and source judgments were interleaved, which means one would have to expect participants to constantly readjust their criterion across judgments, which flies in the face of the aforementioned studies. Nevertheless, it is worth noting that these studies primarily concentrate on item recognition judgments rather than source judgments in which weak and strong items are cued differently to prompt criterion adjustments (e.g., Stretch & Wixted, 1998; Verde & Rotello, 2007). Indeed, it is possible that source judgments are sensitive to cues that vary across trials, specifically the items themselves. The memory strengths of items that differ from trial to trial may automatically influence participants' decision criteria for source judgments.[6] However, in the present study, item strength remained constant across different source-strength conditions, making it unlikely to use item strengths as cues to determine the source criteria.

Second, and most importantly, there is the fortunate fact that the underdetermination between these two competing accounts can be sidestepped by leveraging their differential generalizability: In the case of criterion shifting, we are dealing with an account that is closely tied to the data at hand, such that very little, if anything, can be generalized from it to different settings. In contrast, the processes of differentiation and local matching are expected to manifest themselves in numerous ways beyond the list-strength phenomena considered up to this point. In other words, their influence should be observable elsewhere. What this means is that we can conduct further tests that can speak to the relative plausibility of the REM account proposed here. This is what we will pursue in a later section, where we discuss and evaluate predictions made by the proposed REM extension regarding the occurrence of *output interference* in item and source judgments. To be perfectly clear, this strategy is not intended to suggest that criterion shift accounts are somehow not viable or should be dismissed altogether. All that is being put to test here are the explanatory virtues of the differentiation and the local-matching accounts proposed here.

5.2.1. On the necessity of local matching

Turning our focus to the necessity of a local-matching process for source judgments, we examined a REM submodel that postulates a global-matching process instead. This latter process introduces two critical changes to the model's functioning. The first change is that

[4] For completeness sake, we also considered a REM submodel that only postulated differentiation in item memory. This model, when allowing criteria to vary, was also able to account for the results. However, conceptually, this model is of questionable status given REM's framing of differentiation as a core aspect of memory.

[5] Another possible explanation could be developed by appealing to a recall-to-reject-like mechanism (Rotello & Heit, 2000; Rotello et al., 2000). According to said mechanism, the decrease of false alarms for targets tested in a non-studied source could be explained by the retrieval of their correct source, retrieval that is expected to be more likely for items studied multiple times under the same source. Although plausible at a first glance, we find this hypothesis to be questionable for two main reasons. First, it is predicated on a dual-processing account (e.g., Yonelinas & Parks, 2007) for which there is no critical evidence for (e.g., Starns et al., 2012; 2014; Wixted, 2007; but see Ma et al., 2022). Of course, one could reconcile recall-like processes with a memory-strength account (e.g., Kellen et al., 2021; Wixted, 2007) but such a move would ultimately compromise the ability to distinguish this alternative explanation from the kind proposed by models such as REM. Second, we note that the empirical evidence for said recall-like processes comes from studies where its occurrence is *sufficient* for determining the status of the test item., e.g., recalling source A implies that source B is incorrect. However, this is *not* the case in our studies, where items were encountered across several different sources. In other words, recalling a given source by itself does imply the dismissal of other sources as they are not mutually exclusive.

[6] We thank Adam Osth for raising this point.

during study, the model compares an item-source probe in its entirety (i.e., not just the item or source portions separately) with all of the item-source traces currently available. This comparison ultimately leads to one of two possible outcomes: the updating of an existing item-source memory trace or alternatively, the introduction of a new item-source memory trace. The second change is that source judgments during test are made on the basis of the same global-matching process that was just described. The latter change does not extend to item judgments, given that, in our experiment design, items were presented at the test without an associated source.

As shown in Fig. 10, this submodel's predictions diverge from the observed data in two notable ways. First, this submodel underestimates the proportion of hits in item recognition, which results from incorrect updating of item traces during study due to the match of the source probe. Second, it substantially overestimates the proportions of both hits and false alarms in source recognition. The culprit here is the aforementioned global matching process postulated for source judgments during test. Specifically, it impairs the model's ability to effectively evaluate the familiarity or novelty of source probes at the test. This is due to the probes' alignment with one-fourth of the source traces appended to other item traces. This is what causes the high acceptance of both old and new sources at the test. Although minor, another concerning issue worth noting is that the global-matching process often leads to erroneous updating of source traces during repeated study.

This submodel represents only one out of the many possible ways in which a global matching account can be implemented. The reason why we focused on this specific submodel is that it is the only one among the many variants pursued that did not call for arbitrary decisions on key aspects of its algorithmic implementation. For example, we too often faced situations in which the best-matching item trace (after the item judgment) and the best-matching item-source trace (after the source judgment) *were not the same*. These situations required a decision on which trace to be updated – the initial trace, the second one, or both. These complexities, as well as the unimpressive results that followed, discouraged us from going further down this garden of forking paths.

6. Output interference

As previously discussed in the introduction, the differentiation process postulated by REM entails a phenomenon known as *output interference*, which consists of continuing degradation of memory-judgment accuracy throughout a test phase (see Annis et al., 2013; Criss et al., 2011; Criss et al., 2017; Kılıç et al., 2017; Koop et al., 2015; Malmberg et al., 2012). According to REM, output interference arises from the interplay of two different types of events: (i) the incorrect updating of memory traces when endorsing an item as previously studied, which can happen with lures but also with targets (in both cases, the best-matching trace being updated referring to a different item) and (ii) the encoding of new memory traces whenever a target or a lure are not endorsed as previously studied (see Fig. 2).

Remarkably, in the case of source judgments, the REM extension considered here expects output interference to be virtually absent. The reason is that source judgments involve a *local-matching process* in which the source of a recognized item is identified by comparing source probes *only* with the source information appended to the best-matching memory trace. To see this more clearly, let us consider the example illustrated in Fig. 11: For test items recognized as previously studied, the source information appended to the best-matching item trace is compared to the different possible sources (e.g., sources A and B). The model then selects the source with the highest likelihood (in the case of ties, one source is selected randomly). Each source judgment (either correct or incorrect) is followed by an updating of the item information in the best-matching trace as well as of the source information appended to it. On the other hand, when a test item is rejected, a new memory trace is created. As shown at the bottom of Fig. 11, this updating process

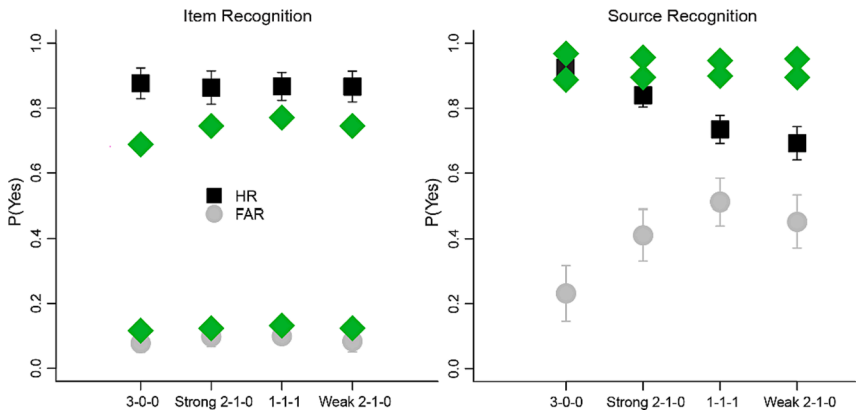
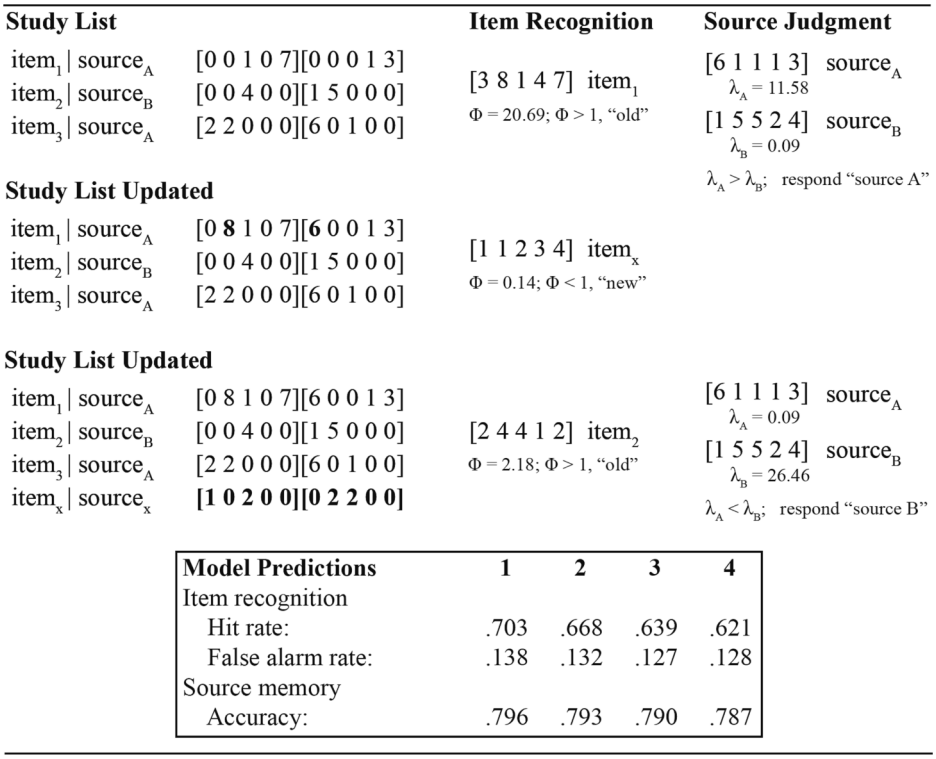


Fig. 10. *Predictions of a REM Submodel without Local-Matching Process.* The figure illustrates the predictions of a submodel that replaces the local-matching process for source judgments with a global-matching process. Green diamonds represent this model's predictions when the global match of item-source pairs is assumed during the study and during source judgments at the test. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



Model Predictions	1	2	3	4
Item recognition				
Hit rate:	.703	.668	.639	.621
False alarm rate:	.138	.132	.127	.128
Source memory				
Accuracy:	.796	.793	.790	.787

Fig. 11. *Differentiation Process During Test in the Extended REM Model.* The model simulation included forty-eight targets randomly assigned to two unique sources and forty-eight lures for 10,000 synthetic participants. Thereby, each test block from 1 to 4 included twenty-four items per synthetic participant. We used the same parameters as in the previous simulations and set the criterion to 1 during the item recognition test. However, note that the qualitative predictions obtained for item and source memory hold for other parameter values as well.

predicts output interference for item-memory judgments but not for source-memory judgments, which means that the accuracy of the former decreases across test blocks, whereas the accuracy of the latter does not. There is, however, a minute decrease in performance: in the example illustrated in Fig. 11, the difference in source accuracy between test blocks 1 and 4 is 0.009. This difference is due to the unlikely cases where one of the new traces being introduced throughout the test phase happens to be the best matching trace later on.

To test these predictions, we considered three datasets, one original and the other two previously published (Fox & Osth, 2022, Experiment 3a-3b).[7] In these experiments, participants made item recognition judgments (Old vs. New?) followed by source judgments (Source A or B?). To sidestep the ongoing debate regarding the possibility of above-chance source judgments for unrecognized items (e.g., Bell et al., 2017; Chen et al., 2018; Cook et al., 2006; Fox & Osth, 2022; Kurilla & Westerman, 2010; Malejka & Bröder, 2016; Starns et al., 2008), we decided to focus on source judgments for items that were recognized as previously studied.

7. Experiment 4

7.1. Methods

7.1.1. Participants

Forty-one undergraduate students (*M* age = 21.04, *SD* age = 1.34) from METU participated in the study to earn partial course credit in their psychology courses. 46% of the participants were female, and 61% were right-handed. All participants were native Turkish speakers, had normal color vision, and normal or corrected-to-normal visual acuity.

7.1.2. Materials

The materials included words randomly selected from Turkish Word Norms (Tekcan & Göz, 2005) that contain 873 words after removing color words such as blue, yellow, or black. All words were randomly assigned to study/test cycles and sources for each participant. Source information was manipulated with colors such that words were studied in a yellow or blue box.

[7] We didn't consider the other experiments conducted by Fox and Osth (2022) due the fact that their designs, which included the manipulation of testing order or blocking of item and source judgments, limited their commensurability with our own studies.

7.1.3. Procedure

Participants were first informed about the experimental procedure with instructions on a computer screen. Then, they were asked to summarize instructions without warning to ensure every detail was clear. Participants also completed a practice phase, a brief version of the actual experiment. The experiment consisted of nine study-test cycles, in each of which 48 words were studied either in yellow or blue box, as illustrated in Fig. 12. After the study, a distraction phase was run for 30 seconds, in which random digits were presented one at a time for a cumulative summation task. At the test, all targets, along with the same number of lures, were presented randomly one by one. Participants were first asked if the test probe was old or new. If they judged the probe as old, they immediately proceeded to the source identification task in which they were asked to identify the source of the probe.

8. Data from Fox & Osth (2022; experiments 3a and 3b)

Fox and Osth (2022; Experiment 3a and 3b) conducted a study in which participants studied words either in the bottom-left corner of the screen or the top-right one (colored with green or yellow, counterbalanced for participants) and then proceeded to test in which they made an initial item recognition judgment followed by source judgments. Experiment 3a involved a 6-point confidence scale, while Experiment 3b involved binary yes–no judgments. We collapsed the confidence judgments from Experiment 3a to the binary item and source judgments.

9. Results

A one-way repeated measures ANOVA was conducted to determine if there were significant changes in item-recognition hit rates, false alarm rates, and source judgments across test blocks. Significant decreases in hit rates have been detected across all three datasets, $F\,(3,120) = 48$, $p < .001$, $\eta^2 = 0.55$, $F\,(3,456) = 47.24$, $p < .001$, $\eta^2 = 0.24$, and $F\,(3,465) = 87.76$, $p < .001$, $\eta^2 = 0.36$. Although there was no significant change in false-alarm rate in Experiment 4 ($F\,(3,120) = 1.54$, $p = .207$), the other two datasets showed small increases ($F\,(3,456) = 18.52$, $p < .001$, $\eta^2 = 0.11$; $F\,(3,465) = 12.83$, $p < .001$, $\eta^2 = 0.08$). Finally, Experiment 4 did not reveal any changes in source judgments throughout testing, $F\,(3,120) = 1.38$, $p = .252$, but the two other results indicated a small decrease in source-judgment performance ($F\,(3,456) = 6.58$, $p < .001$, $\eta^2 = 0.04$; $F\,(3,465) = 5.71$, $p < .001$, $\eta^2 = 0.04$).[8]

10. Model predictions for output interference

Our REM simulations followed the experimental designs reported above and considered the evaluation of forty-eight targets randomly assigned to two unique sources along with forty-eight lures. All the parameters are set to the same values noted earlier in the paper (see Table 3), except for the criterion at the test. The criterion was set to 1.4 for item recognition for the simulation of Experiment 4 since the empirical data demonstrated that people were biased to respond as new. Similarly, the criteria were set to 0.5 and 0.8 for the simulations of Fox & Osth's (2022) data, respectively. The same storage parameters used in the study phase ($u_{\text{item}} = 0.3$, $u_{\text{source}} = 0.1$) are in force in the test phase. As before, for each simulation, we considered one thousand synthetic participants.

As illustrated in Fig. 13, the model accurately describes the decrease in hit rates and predicts the lack of meaningful changes in false alarm rates across test blocks. Similarly, the model predictions for the source memory judgments indicate the absence of any meaningful amount of output interference throughout the test, which is consistent with the data. Overall, the model's success in describing the overall data suggests that, in contrast with their item-recognition counterparts, source memory judgments appear to be driven by a local matching process. As testing progresses, the differentiation process unfolds, wherein item judgments result in either updating existing memory traces or creating new ones. This dynamic reduces the ability to accurately recognize items, as evident from the observed decrease in hit rates. However, despite the diminishing recognition of targets during testing, once an item is recognized, source judgment occurs by comparing source probes solely with the source information appended to the memory trace of the recognized item. Similar to item judgments, as source judgments are made correctly or incorrectly, the memory traces representing the sources are updated. However, incorrect updating does not create a meaningful amount of interference in source-memory judgments, given that the match is restricted to the best-matching trace (i.e., it is localized).

11. Discussion

The differentiation account proposed by REM has implications beyond the strength effects discussed earlier. It also leads to predictions regarding the presence of output interference. This phenomenon is explained by the incorrect updating of existing memories and encoding of new memories during the test phase, which negatively impacts the accurate recognition of subsequent test items. Interference occurs in item recognition due to the involvement of a global matching process – every time a probe is presented at test, it is compared against all item traces in memory, including newly encoded items and items updated incorrectly in the previous test trials.

[8] One concern that arises when analyzing these datasets is the possibility of contaminant responses such as guesses, specifically 'source guesses' following 'item guesses'. We provide a further analysis in the Supplementary Materials by jointly fitting item and source memory judgments with a hierarchical-Bayesian implementation of the high threshold source memory model (Batchelder & Riefer, 1990; Bayen et al., 1996; Kellen et al., 2014; Klauer & Kellen, 2010). The results obtained from this model-based analysis support the hypothesis that there is little or no output interference for source judgments.

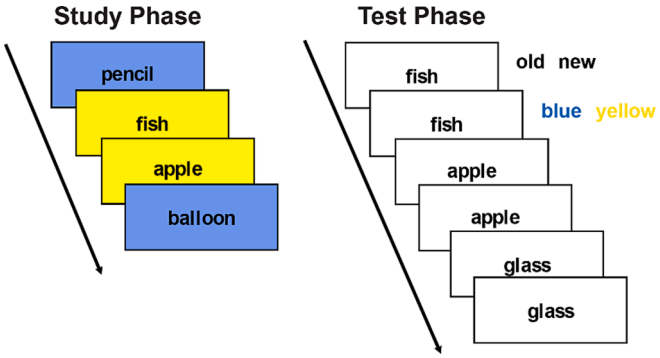


Fig. 12. *Illustration of Study-Test Phase for Experiment 4*. During the study phase, words in specific colors remained on the screen for two seconds with 100 ms inter-stimulus interval. During the test phase, participants made source judgments only for the recognized items, and the words stayed on the screen for both tasks until the response was provided or seven seconds were up. If the participants did not respond in seven seconds, they received a warning indicating that a response must be given within seven seconds, upon which the current trial terminated and the next appeared.

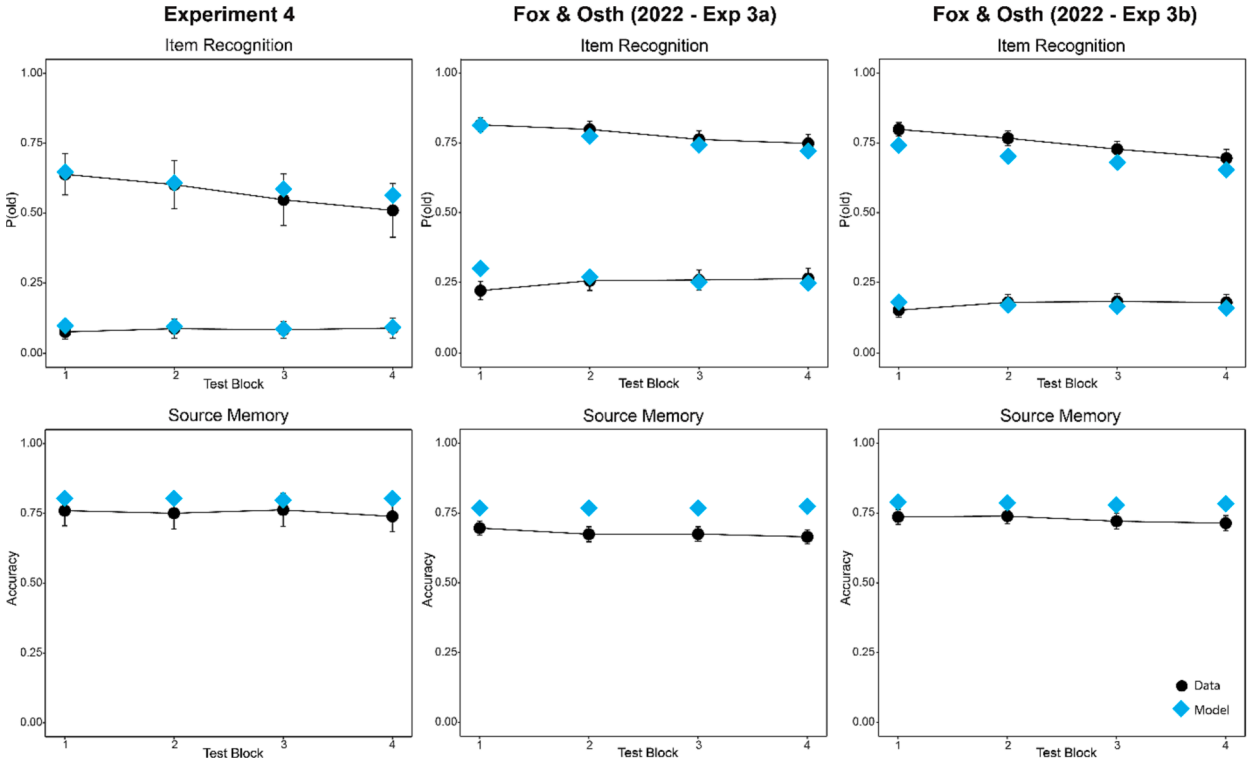


Fig. 13. *Illustration of the Data and the REM Model's Predictions Regarding the Accuracy Throughout the Test for Item and Source Judgments*. Black circles represent the data, while blue diamonds represent the model predictions. The top figures present the probability of old responses (i.e., p(old)) to old items at the top and new items at the bottom across test blocks. The bottom figures represent the probability of correct source judgments (i.e., accuracy) for the recognized items throughout the test. Each test block from 1 to 4 includes twelve targets and twelve lures, which gives one-hundred-and-eight observations of each item type per participant in Experiment 4, and seventy-two in Fox & Osth (2022, Experiments 3a-3b). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

On the other hand, the model implements a local matching process in source memory, which is the main component explaining the stability in source accuracy throughout testing. Although learning continues during testing through processes such as source-trace updating, it does not affect the accuracy of subsequent source judgments, as sources are identified using the source information appended to the best-matching item trace.

Similar to our previous analysis of strength effects, we do not dismiss the possibility of an alternative account of the output-interference results reported here. For instance, the global matching proposed by Osth et al. (2018) establishes different kinds of "noise", namely 'item noise', 'context noise', and 'background noise', the latter being the main source of interference. Given that

background noise is deemed to be the major source of interference in source judgments (see Osth et al., 2018), we expect them to be minimally affected by output interference. However, this account excludes the 'context drift' necessary to explain output interference in item judgments (see Osth & Dennis, 2015; Osth et al., 2018, for the implementations of context drift). To the best of our understanding, the presence of context drift should affect both item and source judgments, which would place this model at odds with the data. However, given that part of the work reported here included the development of a REM extension that can handle source judgments, we do not want to deny the same possibility to competitor models nor suggest that having to undergo such a development process somehow speaks against them. Developing these extensions and comparing their empirical adequacy is a matter for future research.

12. General discussion

The present work began as an exploration of how one of the most prominent computational accounts of human memory, the REM model (Shiffrin & Steyvers, 1997), could be applied to the domain of source judgments. Our first evaluation was that the model's native capabilities were unsuited for this domain, which led to the development of an extension that assumes differentiation for source information in addition to item information, alongside a local-matching process.

We proceeded by testing the empirical adequacy of this model extension by investigating the effects of additional study (i.e., strengthening) on source memory when items can be encountered across multiple sources. Our findings showed that the benefits of repeated learning from the same source take the form of a *mirror effect*: an increased correct endorsement of studied sources alongside a decreased incorrect endorsement of new sources. This result complements the previous reports that alluded to its presence (Glanzer et al., 2004; Dobbins & McCarthy, 2008; Starns & Ksander, 2016; Osth et al., 2018). We also found that learning an item multiple times from a single source does not harm the correct recognition of another source that was also encountered. A similar finding was reported in an earlier study by Osth et al. (2018); however, their design only relied on two sources and confounded item and source strengthening. The present work addresses these issues (by using multiple sources and decoupling item memory from source memory strengthening) and shows that this null list-strength effect is a robust phenomenon.

Since its inception, REM has successfully accounted for list-strength effects in item memory by appealing to a differentiation process that effectively changes the amount of noise associated with studied items as a function of how well they are learned. For instance, 'item noise' (the mismatch between a probe and the different item-memory traces) is smaller among pure strong lists than pure weak lists. This difference results in greater mnemonic discriminability (e.g., as quantified by SDT index d') for pure strong lists relative to pure weak lists, manifested in terms of higher hit rates alongside lower false-alarm rates. In mixed lists, in which half of the items are strong and the other half weak, one expects an intermediate amount of noise. In comparison with their pure-list counterparts, decreases in hit and false alarm rates are expected for the weak, whereas increases are expected for the strong items. However, these changes are such that the levels of mnemonic discriminability (d') for weak and strong items are the same for pure and mixed lists – a *null* list-strength effect (for further details, see Figure 2 of Kılıç et al., 2017 and the associated discussion).

For the case of source memory, we extended REM in a way that preserved its original assumption that item and source information are represented by different traces appended to each other. We argued that each additional source in which the item is studied becomes appended to the updated item trace. In turn, each repetition in the same source leads to differentiation in item and source memory. When elicited, source judgments are based on the matches between the source probes and the source traces appended to the best-matching item trace. These assumptions are sufficient to account for the strength-based source mirror effects and null-list source strength effects reported here.

One of the advantages of keeping the REM extension in line with the original model's tenets is that it allows us to explore their generalizability. Specifically, the extension of differentiation and updating processes to the domain of source memory allows us to evaluate their influence beyond scenarios involving differentially strengthened items. In the present work, we considered the phenomenon known as output interference, which, according to REM, is a byproduct of the ongoing differentiation and updating that takes place during testing (Annis et al., 2013; Criss et al., 2011; Criss et al., 2017; Kılıç et al., 2017; Koop et al., 2015; Malmberg et al., 2012). We found, in contrast to item memory judgments, that source memory judgments are virtually unaffected by output interference. This difference, which was empirically corroborated across multiple datasets, is due to the local-matching process assumed to underlie source memory judgments. However, the model in its current form arguably cannot explain source priming effects (e.g., Hicks & Starns, 2006) because of its reliance on the single memory trace with the highest activation to make a source judgment. One possible solution, to be addressed in future research, is to allow source judgments to consider more than a single best-matching trace, perhaps multiple traces weighted as a function of their match.

In closing, it is sensible to set aside the support for the present REM extension and identify alternative accounts that could be considered in future comparisons. Within the REM framework, there is an alternative '*ensemble extension*' recently proposed by Osth et al. (2018), which assumes that the association of an item to a source creates a set of ensemble features, which are appended as a third vector to the conjoined item and source vectors in the model (for a precursor implementation in associative recognition, see Criss & Shiffrin, 2005). This extension was motivated by configural compound effects found when participants learn associative information (e.g., Cox & Criss, 2017; Criss & Shiffrin, 2005; Dosher & Rosedale, 1997). Also worth noting is the alternative assumption that different sources have non-overlapping features, which was originally pursued by Cox and Shiffrin (2012). It is possible that such an assumption could replace the appending of source traces established in the present REM extension. Beyond REM variants, the most prominent candidate explanation is the global matching model proposed by Osth and Dennis (2015), which was recently updated by Osth et al. (2018) to account for source judgments. This model assumes distinct representations for item, source, and list information, which are conjoined during the study phase. At test, these conjunctive representations of item and list context are combined with each

source cue and subsequently compared with the existing memory traces. Source-identification judgments are made by comparing the difference in the matching strengths of different source probes with a predefined decision criterion. But unlike the case of item judgments, where a target probe is expected only to match its respective memory trace, source matching will occur for as many items as were studied in that source. In other words, the model assumes a global matching of sources that is distinct from the local matching proposed here. Future work should target predictions that distinguish between different local and global matching accounts.

Declaration of Competing Interest

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

Data availability

Data is available here: https://osf.io/4xz78/?view_only=0fc91b15bd47427abf235a5f17e3ba08.

Acknowledgments

Experimental portions of this work were presented in partial fulfillment of Sinem Aytaç's master's thesis at Middle East Technical University. The authors are grateful to Klaus Oberauer, Adam Osth, and Jeff Starns for their valuable comments. David Kellen was supported by an NSF CAREER Award (ID 2145308).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2023.101617.

References

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*(2), 97–123. https://doi.org/10.1037/h0033773
Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*(4), 703–719. https://doi.org/10.1037/0033-295X.96.4.703
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*(5), 1365–1376. https://doi.org/10.1037/a0032188
Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*(4), 548–564. https://doi.org/10.1037/0033-295X.97.4.548
Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 197–215. https://doi.org/10.1037/0278-7393.22.1.197
Bell, R., Mieth, L., & Buchner, A. (2017). Emotional memory: No source memory without old-new recognition. *Emotion, 17*(1), 120–130. https://doi.org/10.1037/emo0000211
Benjamin, A. S. (2001). On the dual effects of repetition on false recognition. *Journal of Experimental Psychology: Learning Memory and Cognition, 27*(4), 941–947. https://doi.org/10.1037/0278-7393.27.4.941
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. https://doi.org/10.1016/S0749-596X(03)00061-5
Chen, X. R., Gomes, C. F. A., & Brainerd, C. J. (2018). Explaining recollection without remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(12), 1921–1930. https://doi.org/10.1037/xlm0000559
Cook, G. I., Marsh, R. L., & Hicks, J. L. (2006). Source memory in the absence of successful cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 828–835. https://doi.org/10.1037/0278-7393.32.4.828
Cox, G. E., & Criss, A. H. (2017). Parallel interactive retrieval of item and associative information from event memory. *Cognitive Psychology, 97*, 31–61. https://doi.org/10.1016/j.cogpsych.2017.05.004
Cox, G. E., & Shiffrin, R. M. (2012). Criterion setting and the dynamics of recognition memory. *Topics in Cognitive Science, 4*, 135–150. https://doi.org/10.1111/j.1756-8765.2011.01177.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*, 461–478. https://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*, 297–319. https://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. https://doi.org/10.1037/a0018435
Criss, A. H., Aue, W., & Kılıç, A. (2014). Age and response bias: Evidence from the strength-based mirror effect. *Quarterly Journal of Experimental Psychology, 67*(10), 1910–1924. https://doi.org/10.1080/17470218.2013.874037
Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language, 64*(4), 316–326. https://doi.org/10.1016/j.jml.2011.02.003
Criss, A. H., Salomão, C., Malmberg, K. J., Aue, W., Kılıç, A., & Claridge, M. (2017). Release from output interference in recognition memory: A test of the attention hypothesis. *Psychology, 1*, 1–9. https://doi.org/10.1080/17470218.2017.1310265
Criss, A. H., & Shiffrin, R. M. (2005). List Discrimination in Associative Recognition and Implications for Representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1199–1212. https://doi.org/10.1037/0278-7393.31.6.1199
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. https://doi.org/10.1162/jocn_a_00292
Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review, 112*(1), 3–41. https://doi.org/10.1037/0033-295X.112.1.3

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*(2), 452–478. https://doi.org/10.1037//0033-295X.108.2.452

Diller, D. E., Nobel, P. A., & Shiffrin, R. M. (2001). An ARC–REM model for accuracy and response time in recognition and recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(2), 414–435. https://doi.org/10.1037/0278-7393.27.2.414

Dobbins, I. G., & McCarthy, D. (2008). Cue-framing effects in source remembering: A memory misattribution model. *Memory and Cognition, 36*(1), 104–118. https://doi.org/10.3758/MC.36.1.104

Dosher, B. A., & Rosedale, G. S. (1997). Configural processing in memory retrieval: Multiple cues and ensemble representations. *Cognitive Psychology, 33*, 209–265. https://doi.org/10.1006/cogp.1997.0653

Fox, J., & Osth, A. F. (2022). Does source memory exist for unrecognized items? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(2), 242–271. https://doi.org/10.1037/xlm0001111

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

Glanzer, M., Hilford, A., & Kim, K. (2004). Six regularities of source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(6), 1176–1195. https://doi.org/10.1037/0278-7393.30.6.1176

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review, 16*, 431–455. https://doi.org/10.3758/PBR.16.3.431

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

Hautus, M. J., MacMillan, N. A., & Rotello, C. M. (2008). Toward a complete decision model of item and source recognition. *Psychonomic Bulletin and Review, 15*(5), 889–905. https://doi.org/10.3758/PBR.15.5.889

Hicks, J. L., & Starns, J. J. (2006). Remembering source evidence from associatively related items: Explanations from a global matching model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(5), 1164–1173. https://doi.org/10.1037/0278-7393.32.5.1164

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528–551. https://doi.org/10.1037/0033-295X.95.4.528

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(2), 302–313. https://doi.org/10.1037/0278-7393.21.2.302

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*(3), 269–299. https://doi.org/10.1006/jmps.2001.1388

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review, 96*(2), 208–233. https://doi.org/10.1037/0033-295X.96.2.208

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory and Cognition, 40*, 1163–1177. https://doi.org/10.3758/s13421-012-0226-0

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin and Review, 21*, 1272–1280. https://doi.org/10.3758/s13423-014-0608-3

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review, 122*(3), 542–557. https://doi.org/10.1037/a0039251

Kellen, D., & Klauer, K. C. (2018). Elementary signal detection and threshold theory. In E. J. Wagenmakers, & J. T. Wixted (Eds.), *The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (Vol. 5, pp. 161–200). Wiley. https://doi.org/10.1002/9781119170174.epcn505.

Kellen, D., Singmann, H., & Klauer, K. C. (2014). Modeling source-memory overdistribution. *Journal of Memory and Language, 76*, 216–236. https://doi.org/10.1016/j.jml.2014.07.001

Kellen, D., Winiger, S., Dunn, J. C., & Singmann, H. (2021). Testing the foundations of signal detection theory in recognition memory. *Psychological Review, 128*(6), 1022–1050. https://doi.org/10.1037/rev0000288

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

Kılıç, A., & Öztekin, I. (2014). Retrieval dynamics of the strength based mirror effect in recognition memory. *Journal of Memory and Language, 76*, 158–173. https://doi.org/10.1016/j.jml.2014.06.009

Kim, K., Yi, D., Raye, C. L., & Johnson, M. K. (2012). Negative effects of item repetition on source memory. *Memory and Cognition, 40*, 889–901. https://doi.org/10.3758/s13421-012-0196-2

Klauer, K. C., & Kellen, D. (2010). Toward a complete decision model of item and source recognition: A discrete-state approach. *Psychonomic Bulletin & Review, 17*, 465–478. https://doi.org/10.3758/PBR.17.4.465

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. https://doi.org/10.3758/s13423-014-0703-5

Koop, G. J., Criss, A. H., & Pardini, A. M. (2019). A strength-based mirror effect persists even when criterion shifts are unlikely. *Memory and Cognition, 47*, 842–854. https://doi.org/10.3758/s13421-019-00906-8

Kurilla, B. P., & Westerman, D. L. (2010). Source memory for unidentified stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 398–410. https://doi.org/10.1037/a0018279

Lindsay, D. S. (2008). Source monitoring. In H. L. Roediger, III (Ed.), *Cognitive psychology of memory. Vol. 2 of Learning and memory: A comprehensive reference*, 4 vols. (J. Byrne, Editor), pp. 325-347. Oxford: Elsevier.

Ma, Q., Starns, J. J., & Kellen, D. (2022). Bias effects in a two-stage recognition paradigm: A challenge for "pure" threshold and signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(10), 1484–1506. https://doi.org/10.1037/xlm0001107

Malejka, S., & Bröder, A. (2016). No source memory for unrecognized items when implicit feedback is avoided. *Memory & Cognition, 44*(1), 63–72. https://doi.org/10.3758/s13421-015-0549-8

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. https://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 old-new recognition and judgments of frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 319–331. https://doi.org/10.1037/0278-7393.30.2.319

Malmberg, K. J., & Shiffrin, R. M. (2005). The "one-shot" hypothesis for context storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 322–336. https://doi.org/10.1037/0278-7393.31.2.322

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. https://doi.org/10.1037/0033-295X.105.4.734-760

Mensink, G. J., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review, 95*(4), 434–455. https://doi.org/10.1037/0033-295X.95.4.434

Miller, M. B., & Kantner, J. (2020). Not all people are cut out for strategic criterion shifting. *Current Directions in Psychological Science, 29*(1), 9–15. https://doi.org/10.1177/0963721419872747

Morrell, H. E., Gaitan, S., & Wixted, J. T. (2002). On the nature of the decision axis in signal-detection-based models of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1095–1110. https://doi.org/10.1037//0278-7393.28.6.1095

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104*(4), 839–862. https://doi.org/10.1037/0033-295X.104.4.839

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 689–697. https://doi.org/10.1037/0278-7393.19.3.689

Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(5), 855–874. https://doi.org/10.1037/0278-7393.17.5.855

Osth, A. F., & Dennis, S. (2014). Associative recognition and the list strength paradigm. *Memory & Cognition, 42*, 583–594. https://doi.org/10.3758/s13421-013-0386-6

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260–311. https://doi.org/10.1037/a0038692

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

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology, 104*, 106–142. https://doi.org/10.1016/j.cogpsych.2018.04.002

Raaijmakers, J. G., & Shiffrin, R. M. (1980). SAM: A theory of probabilistic search of associative memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (Vol. 14, pp. 207–262). New York: Academic Press.

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

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*(2), 163–178. https://doi.org/10.1037/0278-7393.16.2.163

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory and Cognition, 28*(6), 097–922. https://doi.org/10.3758/bf03209339

Rotello, C. M., Macmillan, N. A., & Tassel, G. V. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language, 43*, 67–88. https://doi.org/10.1006/jmla.1999.2701

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review, 115*(4), 893–912. https://doi.org/10.1037/a0013396

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. https://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 and Review, 4*(2), 145–166. https://doi.org/10.3758/BF03209391

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory and Cognition, 34*(1), 125–137. https://doi.org/10.3758/BF03193392

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition, 36*, 1–8. https://doi.org/10.3758/MC.36.1.1

Starns, J. J., Hicks, J. L., & Marsh, R. L. (2006). Repetition effects in associative false recognition: Theme-based criterion shifts are the exception, not the rule. *Memory, 14*(6), 742–761. https://doi.org/10.1080/09658210600648514

Starns, J. J., & Ksander, J. C. (2016). Item strength influences source confidence and alters source memory zROC slopes. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42*, 351–365. https://doi.org/10.1037/xlm0000177

Starns, J. J., Pazzaglia, A. M., Rotello, C. M., Hautus, M. J., & Macmillan, N. A. (2013). Unequal-strength source zROC slopes reflect criteria placement and not (necessarily) memory processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1377–1392. https://doi.org/10.1037/a0032328

Starns, J. J., Ratcliff, R., & McKoon, G. (2012). Evaluating the unequal-variance and dual-process explanations of zROC slopes with response time data and the diffusion model. *Cognitive Psychology, 64*, 1–34. https://doi.org/10.1016/j.cogpsych.2011.10.002

Starns, J. J., Rotello, C. M., & Hautus, M. J. (2014). Recognition memory zROC slopes for items with correct versus incorrect source decisions discriminate the dual process and unequal variance signal detection models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*(5), 1205–1225. https://doi.org/10.1037/a0036846

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*, 18–34. https://doi.org/10.1016/j.jml.2010.03.004

Starns, J. J., White, C. N., & Ratcliff, R. (2012). The strength-based mirror effect in subjective strength ratings: The evidence for differentiation can be produced without differentiation. *Memory and Cognition, 40*, 1189–1199. https://doi.org/10.3758/s13421-012-0225-1

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. https://doi.org/10.1037/0278-7393.24.6.1379

Tekcan, A. I., & Göz, I. (2005). *Türkçe kelime normları (Turkish word norms)*. Istanbul: Bogazici Universitesi Yayinevi.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92*(3), 297–1204.

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory and Cognition, 35*(2), 254–262. https://doi.org/10.3758/BF03193446

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

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*(1), 152–179. https://doi.org/10.1037/0033-295X.114.1.152

Yonelinas, A. P., Hockley, W. E., & Murdock, B. B. (1992). Tests of the list-strength effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 345–355. https://doi.org/10.1037/0278-7393.18.2.345

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin, 133*(5), 800–832. https://doi.org/10.1037/0033-2909.133.5.800