Psychological Review
2004, Vol. 111, No. 3, 800–807

0033-295X/04/$12.00 DOI: 10.1037/0033-295X.111.3.800

COMMENTS

Context Noise and Item Noise Jointly Determine Recognition Memory: A Comment on Dennis and Humphreys (2001)

Amy H. Criss and Richard M. Shiffrin
Indiana University

S. Dennis and M. S. Humphreys (2001) proposed a model with the strict assumption that recognition memory is not affected by interference from other items. Instead, confusions are due to noise generated by prior contexts in which the test item appeared. This model seems disparate from existing models of recognition memory but is similar in many ways that are not superficially obvious. One difference is the order in which item and context information are used as retrieval cues. A more critical difference is the assertion that only an item's history, and not other items, affects recognition memory. Conceptual arguments along with the results of 2 experiments make a persuasive case that both types of noise affect recognition. To illustrate the approach, the authors fit experimental data with a version of the retrieving effectively from memory model (R. M. Shiffrin & M. Steyvers, 1997) incorporating both sources of noise.

Episodic recognition memory tests require participants to distinguish items presented on a recent list (targets) from those not so presented (foils). Two types of information are considered important for such a task: item and context. Item information refers to the spelling, semantics, and other information describing the item itself. Context information refers to the location, mood, and other environmental factors present with the item. In typical studies, the items on a given list have been stored in memory prior to the experiment. Thus, a match of item information verifies the existence of that test item in memory but is insufficient to verify that the item was on a particular study list. Similarly, a match of context information merely confirms the storage of something from the current context but is insufficient to determine that a particular test item had been stored. Logically, then, successful recognition requires a match of both item and context information to memory.

The first part of our comment on the bind cue decide model of episodic memory (BCDMEM; Dennis & Humphreys, 2001) emphasizes the formal similarities of the model to other extant models of episodic recognition. We point out that various models differ in the order in which item and context cues are used during retrieval. However, we argue that such differences are not easy to test and are not core assumptions of the models. The second part of our comment concerns the strong claim by Dennis and Humphreys (2001) that item noise plays no direct role in recognition memory performance. We present data from two new experiments and discuss previous work suggesting that both item noise and context noise play important roles in recognition. Finally, we present a modified version of the retrieving effectively from memory (REM) model (Shiffrin and Steyvers, 1997) that unpacks the noise component into the three types we discuss in the following section.

Figure 1 represents an episodic recognition situation in which the memory traces of different words are shown as separate traces, with each trace containing both context and item information. (However, our arguments also apply to models in which storage is composite.) In this example, a set of items was studied, and the word *casino* serves as the test item. Traces of words presented on the study list are depicted to the left of the dashed vertical line, and traces of other words are depicted to the right of this line. Within each row and within each of these sets, the traces are ordered by the similarity of their contents to the test word (higher similarity toward the left). Within each column, the traces are ordered by the similarity of their context to the test context (higher similarity toward the top). In this example, the test word *casino* is a target and hence is at the upper left. The most similar list word is *poker* and is next to the right. *Toad* is a dissimilar list word and is farther still to the right. *Gambling* is a similar word that was not on the study list but was encountered recently outside the experiment and hence is near the top of the column to the right of the dashed vertical line.

If a target is tested, the top left trace tends to match best because this trace matches both the context and the content of the probe. The second best matches tend to be the traces relatively close to the top left corner, including both those traces matching in context but not exact content (i.e., the top row) and those matching content but not exact context (i.e., the first column). Traces of words that were not presented on the study list and do not exactly match the test word (i.e., those traces to the right of the vertical line) may also match to a lesser degree. If a foil is tested, there is no trace matching both content and context, but there are secondary

This research was supported by a National Science Foundation graduate fellowship to Amy H. Criss and National Institute of Mental Health MERIT Grant 12717 to Richard M. Shiffrin. We thank Mark Steyvers and Simon Dennis for many useful discussions that helped guide this work.

Correspondence concerning this article should be addressed to Amy H. Criss or Richard M. Shiffrin, Department of Psychology, Indiana University, Bloomington, IN 47405. E-mail: acriss@indiana.edu or shiffrin@indiana.edu

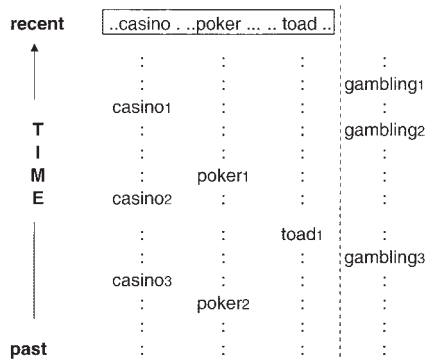

Figure 1. A schematic representation of episodic memory in which the current test word is *casino*. Each word represents a memory trace containing both context and item information. Items to the left of the dashed vertical line were presented on the study list and are ordered by similarity to the test item. Time runs along the vertical axis, with the most recent study list in the top row.

matches of the types just described. Noise in the storage and/or retrieval processes guarantees that occasionally a target will not match well (producing misses) and a foil will match unusually well (producing false alarms).

Global item-matching models refer to those models in which the majority of interference is caused by other items from the study list. These models have proved successful for a number of years, accounting for data in recognition, free recall, serial recall, cued recall, and associative recognition (Gillund & Shiffrin, 1984; Hintzman, 1988; Humphreys, Pike, Bain, & Tehan, 1989; McClelland & Chappell, 1998; Murdock, 1982, 1997; Shiffrin & Steyvers, 1997). We use the term *item-noise model* to refer to the extreme form of such a model in which the noise contributed by sources other than list items is negligible (i.e., the top row of Figure 1 produces all the noise). Although most theorists, including those cited above, had in mind some form of a global item-matching model, their models have often been implemented for simplicity as item-noise models.

Models assuming that the majority of confusion is caused by other contexts in which the test item occurred will be termed *global context-matching models*. We use the term *context-noise model* to refer to an extreme form of this class of models in which the noise contributed by anything other than traces of the test item is negligible (i.e., in Figure 1, traces in the column for the test item produce all the noise); an example of such a model is BCDMEM. To restate, the major difference between BCDMEM and other memory models lies in this strong assumption that other studied items do not contribute to the recognition decision.

Another difference between BCDMEM and global familiarity models is the order in which context and item cues are used to probe memory. In some models, these cues are used jointly in a simultaneous probe of memory (e.g., Gillund & Shiffrin, 1984; Murdock, 1997). In others, the context cue is used first to activate a set of items, and then the item cue is matched to traces in the resulting set (e.g., REM). In the remaining models, the item cue is used first to activate a set of items, and then the context cue is matched to the traces in the activated set (e.g., BCDMEM). Deciding between these orders is a most difficult problem because performance in an episodic recognition task cannot rise above chance until both cues have been used. The cuing order is not a fundamental assumption of either BCDMEM[1] or REM, and we therefore focus our efforts on the critical difference between these models: the source of noise in recognition memory.

BCDMEM is indeed distinguishable from extant models by virtue of its strong assumption that only context noise plays a role in recognition memory. It is this issue that we explore in the remaining part of this comment, and it is this assumption that we argue is inadequate.

Experiment 1

In BCDMEM, the memory probe activates all past traces of the test item but not traces of words presented on the recent list. Because none of the other list words are activated, they do not contribute noise, and confusion arises only from previous contexts in which the test item occurred. This aspect of the model seems, on the surface, at variance with empirical data showing effects of list length and similarity.

A list length effect is represented by the decline in discrimination associated with adding nominally unrelated items to the study (e.g., Gronlund & Elam, 1994). However, Dennis and Humphreys (2001) proposed that list length effects might be due to various confounds including attention drift, study–test lag, displaced rehearsal, or failure to reinstate the study context.

Similarity effects are often instantiated in recognition studies by an increase in both hits and false alarms when there is an increase in the number of studied items that are related to the test item (e.g., Shiffrin, Huber, & Marinelli, 1995; Sommers & Lewis, 1999; Zaki & Nosofsky, 2001). To explain similarity effects, Dennis and Humphreys (2001) suggested that participants invoke strategies when they notice a categorical relation among members of the study list. For example, study lists with noticeable categorical relations may encourage participants to generate a word associated to the study word, termed an *implicit associative response* (IAR; Underwood, 1965). Dennis and Humphreys proposed that an IAR results in storage of the generated word and study context as if the generated word had been presented on the list. According to BCDMEM, increased false alarms to test items that are similar to list items are due to the context stored with those test items that had been generated as IARs during the study list. Thus similarity effects, according to BCDMEM, are due to the storage of study context for items generated via IARs and not due to activation of similar list items.

These hypotheses based on IARs are not without conceptual problems. The presence of similarity effects does not seem dependent on the use of stimulus materials and designs fostering the use of IARs. Similarity effects are found in studies using categories constructed from colors (Zaki & Nosofsky, 2001) and word categories that were constructed to have low associative connections

[1] This information is based on a personal communication with Simon Dennis on September 3, 2002.

on the basis of word association norms (Shiffrin et al., 1995).[2] Another problem concerns a relation between list length effects and IARs. Dennis and Humphreys (2001) argued against the existence of true list length effects, but their IAR hypothesis seems to imply that such effects will be generated by IARs themselves. According to the IAR hypothesis, a foil tends to be mistaken for a target if it had been generated during study. The chance that this will happen depends on the number of IARs generated during study, and this number will rise with list length. Thus, an IAR mechanism will itself produce list length effects. To make these two views consistent, one must argue that IARs do not contribute significantly, except in special circumstances, such as massed study of very similar items (e.g., Roediger & McDermott, 1995). Thus although IARs almost certainly exist, it is difficult to argue they provide a general account of similarity or list length effects. To sharpen these arguments, we report a study with novel faces as stimuli, for which IARs are implausible and could not account for such effects.

Method

Participants. Sixty-seven Indiana University undergraduates participated for partial course credit.

Stimulus materials. Categories of words, related either by semantic or orthographic–phonemic similarity were used along with categories of similar faces. Each of the nine semantic categories consisted of 1 prototype word and 12 exemplar words semantically related to the prototype word. Each of the nine orthographic–phonemic categories consisted of a prototype word of three or four letters and 12 exemplars that shared a vowel sound and exactly one consonant cluster with the prototype word. The word categories are a subset of those used by Shiffrin et al. (1995), and additional details about these stimuli may be found there. The 18 face categories, half of which were male and half female, had 12 exemplar faces each and no prototype. These face categories can generally be described as clustering within category and differing between categories on the dimensions of race, age, and hair color and style. Individual faces were taken from college yearbooks and face databases, including *The Database of Faces* (AT&T Laboratories Cambridge, n.d.). Note that we use the word *prototype* in an atypical manner, referring to the word used to generate each category, not the average, central tendency, or label for the exemplars.

Procedure. The study list consisted of face–word pairs, each pair to be rated on the following question: "What is the degree of association between these two items?" For each participant, categories were randomly assigned list lengths of 2, 6, or 9 items, such that there were three categories of female faces, male faces, semantic words, and orthographic–phonemic words assigned to each list length. In addition, 20 unrelated words and 20 unrelated faces were studied for a total list length of 244 items. The word–face pairings were random with respect to category (in no case were all members of a particular face category paired with a particular word category), and the exemplars from any category were spaced throughout the entire list.

For the unexpected memory test, participants were asked to do two things: (a) to report their confidence that the test item itself had been on the study list and (b) to give their best estimate of the number of items on the study list that were similar to the test item in appearance, sound, spelling, or meaning, including the test item itself if it had been studied. The second of these items was used in place of gist instructions (see Brainerd & Reyna, 1998) because normal gist instructions would not make sense for face stimuli. There were two test blocks, one for words and one for faces, with the order of the blocks and presentation of items within blocks randomized for each participant. For each word category, the test items included the prototype, 2 studied words, and 2 related foils. For each face category, the test items included 2 studied faces and 3 related foils. In addition, 10 word foils and 10 face foils that did not belong in any category (unrelated foils) were tested for a total of 200 test trials.

Results and Discussion

Participants are reluctant to use different response criteria within a single test list, even when the items are from different categories (Morrell, Gaitan, & Wixted, 2002; Wixted & Stretch, 2000). In such a case, global item-matching models predict that increasing category length will result in an increase in the hit and false alarm rates (see Shiffrin et al., 1995). Furthermore, such models make the same predictions for faces and words. BCDMEM can predict an increase in hit and false alarm rates for the word categories on the basis of the IAR hypothesis. However, an IAR mechanism does not work for novel faces. Realizing this and other problems posed by novel materials (including the assumed representation of local codes in which each item is represented as a single node in memory and the presentation of an item perfectly activates its own node), Dennis and Humphreys (2001) restricted the domain of applicability of BCDMEM to known rather than novel items. Strictly speaking, then, our face results cannot be used to rule out BCDMEM. Nonetheless, similar results obtained for words and faces would suggest similar underlying processes and would thereby support global item-matching models.

False alarms were higher for faces than for words, $F(1, 66) = 9.14$, $p < .01$, but a repeated measures analysis of variance showed no interactions; thus, we collapsed over words and faces. False alarms to related foils rose with category length, $F(2, 132) = 16.26$, $p < .01$. False alarms for prototype words (recall that there were no prototype faces) showed a more pronounced increase with category length, $F(2, 132) = 17.59$, $p < .01$. Analysis of the hit rates revealed an uninteresting interaction between item type (word vs. face) and subtype (semantic vs. orthographic–phonemic and male vs. female) but no other interactions. Overall, faces had lower hit rates than words, $F(1, 66) = 39.28$, $p < .01$, and the apparent effect of category length for targets was not significant. Figure 2 shows the overall probability of an "old" response, *P*(old), collapsed over word and face categories, and Table 1 contains the full set of data for each condition. We cannot rule out a criterion shift on the basis of category length, but note that this must be done on a trial-by-trial basis—an assumption that seems to require some (explicit or implicit) knowledge about the amount of item noise for each particular test item. It seems strange for BCDMEM to assume that some item-noise calculation is used to

[2] Dennis and Humphreys (2001) discussed Anisfield and Knapp's (1968) finding of the directionality of false recognition. That is, false alarms to B increase when studied item A elicits B during free association, but studying B (which does not elicit A) did not increase false alarms to A. However, Anisfield and Knapp did not propose that false alarms are found only in such special circumstances. In fact, because of an experimental confound they stated "for this reason it is not possible to conclude that backward associative relations cannot produce false recognition" (p. 177). If this finding is replicated in future research without confounds, it would nevertheless be easy to accommodate in global item-matching models by including a mechanism for IARs because IARs are more likely to occur for forward than backward associations.

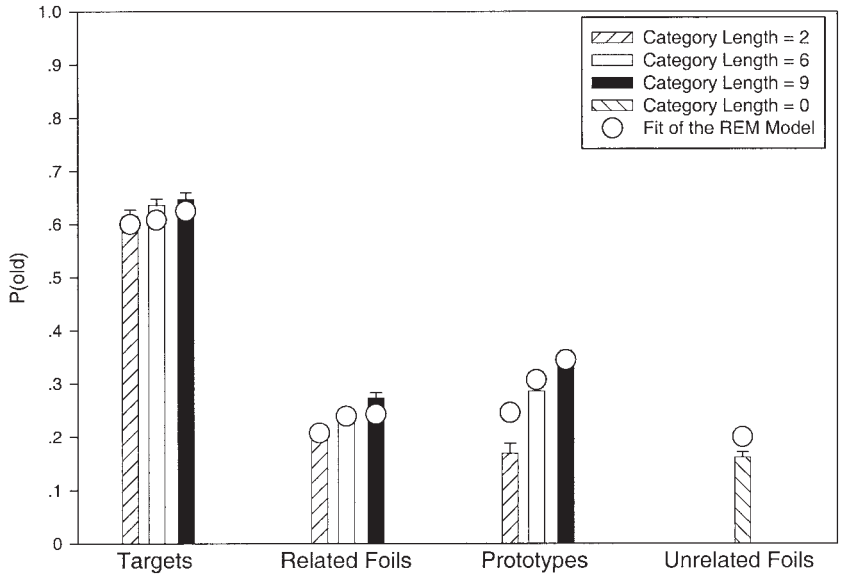

Figure 2. The probability of an "old" response, *P*(old), in Experiment 1, collapsed over all types of categories as a function of the number of category exemplars on the study list. Error bars represent one standard error above and one standard error below the mean. Open circles are fits of the modified retrieving effectively from memory (REM) model.

Table 1
Average Probability of an "Old" Response for Each Condition of Experiment 1

Test items	Category length 0	2	6	9
	Words			
Target				
Semantic		.746	.726	.724
Orthographic–phonemic		.642	.687	.719
Prototypes				
Semantic		.194	.383	.403
Orthographic–phonemic		.144	.189	.274
Related foils				
Semantic		.137	.174	.219
Orthographic–phonemic		.202	.211	.234
Unrelated foils				
Semantic	.087			
Orthographic–phonemic	.146			
	Faces			
Targets				
Female		.515	.542	.567
Male		.557	.590	.580
Related foils				
Female		.207	.269	.295
Male		.212	.294	.313
Unrelated foils				
Female	.206			
Male	.221			

change the criterion for each test item, but this same information would not be used as a basis for the decision.

In summary, although BCDMEM could assume IARs for the word categories, despite our attempts to reduce such occurrences, it has no account for the similar pattern of results for face categories. Thus, for reasons of parsimony, we prefer a global item-matching model explanation—one that assumes decisions for both words and faces are based on overall familiarity of the probe to the items in memory.

Participants were also asked to estimate the number of items on the study list that were similar to the test item. Because participants were free to respond on any scale, we collected the estimates of each participant (separately for faces and words) and transformed these to *z* scores. The *z* scores were then averaged across participants. The estimates increased with category length for all types of items: targets, $F(2, 132) = 15.48$, $p < .01$; related foils, $F(2, 132) = 20.82$, $p < .01$; and prototypes, $F(2, 132) = 15.77$, $p < .01$. Estimates to related foils showed a three-way interaction between item type (face vs. word), subtype (male vs. female and orthographic–phonemic vs. semantic), and category length. This interaction was due to the larger impact of length on semantic categories than on orthographic–phonemic categories, $F(2, 132) = 4.98$, $p < .01$. Generally, the semantic categories seemed to have a larger impact on participants' estimated number of similar items than did orthographic–phonemic categories, perhaps because of mechanisms such as IARs or greater cohesion between members of semantic categories. The estimates largely mimicked the pattern of data found for *P*(old), but the increase in estimated number of similar items with category length was more pronounced and was statistically significant for all types of test items. The combined results argue in favor of global item-matching models because such models predict our findings for both types of materials on the

basis of use of the same information (i.e., overall activation), whereas BCDMEM has no account for the face results.

Having made a case for the existence of item noise in any model of episodic memory, we now turn to an experiment in which we attempt to instantiate a structure similar to that depicted in Figure 1, thereby making context noise an experimentally varied factor.

Experiment 2

The goal of this experiment was to investigate the influence of both item and context noise within a single experiment. Context noise was manipulated by including a series of three study lists and by repeating some study items on more than one list. Item noise was manipulated by including categories of varying length on the final study list. By varying category length rather than list length, this design removes the various artifacts bedeviling list length studies, as discussed by Dennis and Humphreys (2001). We also used an incidental study design to reduce the chance that participants will purposefully produce IARs. Global item-matching models predict that traces similar to the test probe in context, item, or both will produce noise, affecting the recognition decision. Context-noise models predict that interference is a function of the previous occurrences of the test item and performance should not change as similar items are added to the study list. Both model types predict strong effects of context noise.

Method

Participants. One hundred seven undergraduates from Indiana University completed the experiment for partial course credit.

Stimulus materials. The stimulus materials were 24 categories of words including 15 semantic categories used by Shiffrin et al. (1995) and 9 similarly constructed categories.

Procedure. Each participant studied three successive lists of 92 words, separated by arithmetic tasks. Different incidental tasks (ratings of pleasantness, typicality, or personal relatedness) were used for each of the three lists. The order of the tasks was randomly assigned for each participant. A final unexpected recognition test required participants to say "old" to words presented on List 3 and "new" to all other words. We manipulated context noise by presenting words in more than one of these lists and testing words that might have been studied in various combinations of Lists 1, 2, and 3 or that might not have been studied at all. We manipulated item noise by varying the number of words studied on List 3 that were members of the category of a given test item.

On each of the first two lists, 80 words were fillers and the other 12 items were termed *critical* words. There were 3 critical words tested in each of the following 8 conditions: never presented (none), presented on just one list (1, 2, or 3), presented on two lists (1 and 2, 1 and 3, or 2 and 3), or presented on all three lists (1, 2, and 3). The third list contained additional exemplars from the category of each critical word, and the number of such exemplars was manipulated. For each group of 3 critical words, 1 was a member of a category with no other exemplars on List 3, 1 was a member of a category with 3 other exemplars on List 3, and 1 was a member of a category with 7 other exemplars on List 3. Thus, there were 24 conditions in all. The exemplars within any category were randomly spaced throughout the third list.

The unexpected recognition test began with 4 practice trials followed by 120 test trials containing an equal number of targets and foils. There were two types of targets including the 12 *critical words* that appeared on List 3 (and possibly on other lists) and 3 exemplars from each category presented on List 3 (not including those categories with length zero) termed *related targets*. In addition, there were three types of foils including the critical words that were not on List 3, called *studied foils*; the unstudied category *prototypes*; and 1 *unstudied foil* from each of the 24 categories.

Results and Discussion

Figure 3 shows that P(old) to a critical word was systematically related to the item's history. A within-participants linear contrast verified that false alarms to studied foils rose to the degree that a foil had been presented in a more recent list (or lists), $F(1, 106) = 300.20$, $p < .01$. Such effects did not occur for targets, perhaps because of a ceiling effect for repeated items. Note also that the number of repetitions of an item across lists affected response probabilities. Paired t tests verified that false alarms to items repeated on two lists was greater than the false alarm rate for items occurring only on the first list, $t(106) = 4.84$, $p < .01$, and greater than the false alarm rate for items presented only on List 2, $t(106) = 3.70$, $p < .01$. Likewise, repeating targets increased the hit rate, confirmed by within-participants linear contrast, $F(1, 106) = 19.83$, $p < .01$. In summary, context noise caused considerable interference, as measured by P(old) to words presented on prior experimental lists.

We also found an effect of item noise, as measured by the changes associated with category length (i.e., the number of category exemplars on List 3). As is apparent in Figure 4, false alarms increased with category length in all cases. These observations are confirmed by within-participant linear contrast for studied foils, $F(1, 106) = 5.59$, $p = .02$; prototypes, $F(1, 106) = 96.87$, $p < .01$; and unstudied foils, $F(1, 106) = 15.40$, $p < .01$. The false alarm rate is much higher for studied foils than for unstudied foils, but both types of test items are similarly affected by category length. The effects for targets were much smaller and not statistically significant.

In our view, the simplest account of these results is based on significant roles played by both context noise and item noise. In global item-matching models, context noise produces the increased tendency to say "old" to words presented on Lists 1 and/or 2 and therefore reduced performance for such words. Item noise produces the changes in P(old) as a function of the number of category exemplars presented on List 3.

A Modified REM Model

Consider the simple version of REM as introduced by Shiffrin and Steyvers (1997) for application to recognition memory: Each item is stored as a separate vector containing both context features and item features, in which the feature values are positive integers drawn from a geometric distribution with parameter g. During study, an incomplete and error prone copy of each item is stored in episodic memory. When study time is not varied, we let u give the probability for each feature that some nonzero value is stored during the study period. If a feature is stored, the correct value will be stored with some probability c, and a random feature value drawn from the geometric distribution will be stored with probability $1 - c$. Features not stored are coded as zeros. At test, context features are used as a probe and assumed to activate all traces of words on the recent list. Then, the item features are used as a probe

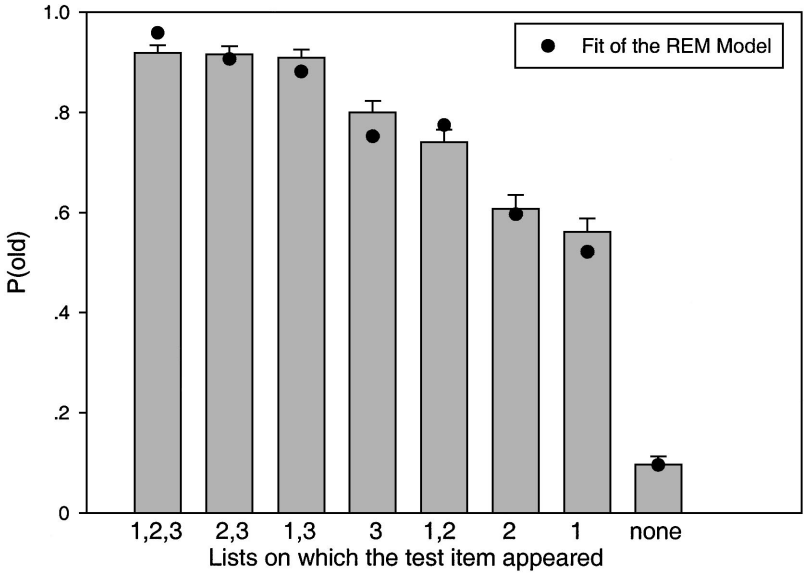

Figure 3. The context noise conditions from Experiment 2 showing the probability of an "old" response, *P*(old), as a function of the study list(s) on which the item occurred. The first four bars show results for targets, and the last four bars show results for foils. The numbers below each column give the lists in which a given test word was presented. Error bars represent one standard error above and one standard error below the mean. Solid circles are fits of the modified retrieving effectively from memory (REM) model.

and are compared with the activated memory traces. In particular, a likelihood value, λ_i, is calculated for a memory trace *i* according to the following equation:

$$\lambda_i = (1 - c)^{niq} \prod_j \left[\frac{c + (1 - c)g(1 - g)^{j-1}}{g(1 - g)^{j-1}} \right]^{njim}, \qquad (1)$$

where *niq* is the number of nonzero mismatching features and *njim* is the number of matching features with the value *j*. The term before the product represents discounting due to mismatching features between the probe and memory trace. The term after the product represents the positive evidence gained from matching features. Note this term includes both those features that match because they were copied correctly (the portion before the addition sign) and those features that were stored incorrectly but happen to match by chance (the portion following the addition sign). The parameter *u* does not enter the likelihood calculation (because a feature with a zero in the trace does not enter the calculation) but does affect the values of *niq* and *njim*.

The mean of these likelihood values gives the odds that the test item is old. If the odds value is greater than some criterion (the optimal value of 1.0 is typical), then the item is called "old," otherwise it is called "new."

This model is too simple to be used for experiments of the present type. In fact, the more complicated models presented in Shiffrin and Steyvers (1997) are also inappropriate because they fail to take both context and item noise into account in balanced fashion. Thus, we offer an alternative decision rule in which likelihood ratios and odds are calculated on the assumption that foil traces may consist of three types, and traces similar in context and traces similar in content are treated in a balanced way.

If the trace is a target, its features will tend to match the test probe on both context and item features. Otherwise, the trace is a foil, and its features will tend to match the test probe in one of the following ways: match on some context features only (C$\not$I), match on some item features only ($\not$CI), or match neither the context nor item features ($\not$C$\not$I). If these three types of foils are considered separately, the odds equation becomes

$$\Phi = \frac{1}{N} \sum_i \left[\frac{P(\mathrm{D}_i|\mathrm{O})}{\gamma P(\mathrm{D}_i|\not\mathrm{C}\not\mathrm{I}) + (1 - \gamma)\alpha P(\mathrm{D}_i|\not\mathrm{C}\mathrm{I}) + (1 - \gamma)(1 - \alpha)P(\mathrm{D}_i|\mathrm{C}\not\mathrm{I})} \right], \qquad (2)$$

where the D_i refers to the evidence, that is, the matching and mismatching feature values for the *i*th trace. The numerator represents the probability of the evidence given that the test item is old (O), and the denominator represents the probability of the evidence given that the test item is one of three types of foils. γ is the probability that noise comes from foils that do not match on item or context features, and α is the conditional probability that noise comes from foils matching item but not context features, given that something matches. *N* refers to the number of traces in episodic memory. Rearranging terms and writing this formula in terms of likelihood values calculated by Equation 1, we see that

$$\Phi = \frac{1}{N} \sum_i [\gamma \lambda_{i\mathrm{C}}^{-1} \lambda_{i\mathrm{I}}^{-1} + (1 - \gamma)\alpha \lambda_{i\mathrm{I}}^{-1} + (1 - \gamma)(1 - \alpha)\lambda_{i\mathrm{C}}^{-1}]^{-1}, \qquad (3)$$

where $\lambda_{i\mathrm{C}}$ is the likelihood value calculated from the context features of the test probe and memory trace *i* and $\lambda_{i\mathrm{I}}$ is the

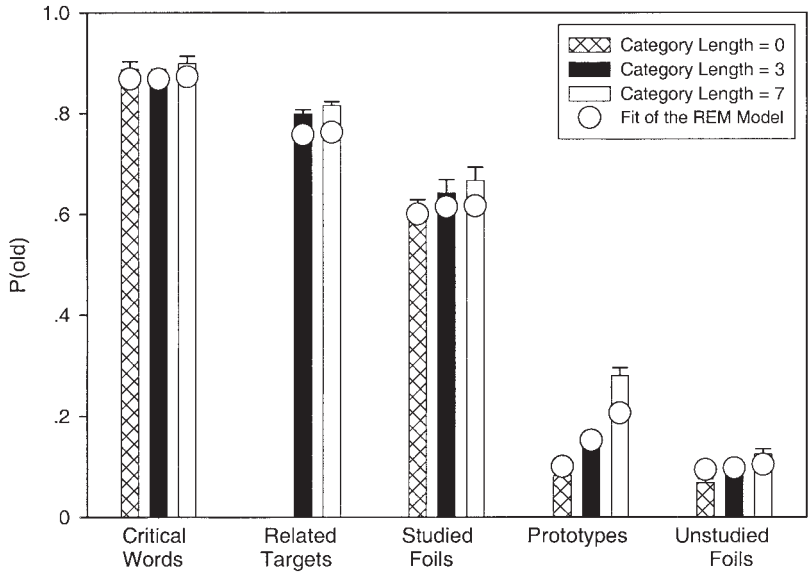

Figure 4. The item-noise conditions in Experiment 2 showing the probability of an "old" response, *P*(old), as a function of the number of additional studied exemplars from the test word category that were presented in List 3. Error bars represent one standard error above and one standard error below the mean. Open circles are fits of the modified retrieving effectively from memory (REM) model.

likelihood value calculated from the item features of the test probe and memory trace *i*. If $\alpha = 1$ and $\gamma = 0$, all noise comes from the item features and the context features are ignored. Examination of Equation 3 reveals that the odds value is high if both the context and item likelihoods are high. But if this is the case, then the first term (which combines the match from both item and context) is largely redundant and may be removed without changing the model significantly (a fact we verified through simulation). We therefore set $\gamma = 0$. Note that the sum still occurs over all memory traces, including the ones that generally mismatch in both context and item; such traces still contribute noise, but they do so through their tendency to mismatch context and item separately rather than jointly. The simulations in this article therefore calculate the odds as follows:

$$\Phi = \frac{1}{N}\sum_{i}[\alpha\lambda_{iI}^{-1} + (1 - \alpha)\lambda_{iC}^{-1}]^{-1}. \quad (4)$$

To apply the model, one must build item representations (i.e., vectors of feature values) that represent the structure of the study list. We started by fixing each vector to have 20 context features and 20 item features, drawn from the geometric distribution with parameter *g*. The item features were constructed as follows. A set of 20 item values is randomly sampled to make a category vector for each category. All items in a category are constructed by copying some proportion of the features of this category vector. For a given category, the prototype vector is constructed by copying each feature value with probability p_{cat}, and each exemplar is constructed by copying each feature value with probability p_{ex}. Any features not copied are filled with randomly sampled values. The value for p_{ex} was held constant for both experiments. For simplicity, we used the same value for p_{cat} for the different category types within study but let this value differ between studies (because Experiment 1 used both semantic and orthographic–phonemic categories but Experiment 2 used only semantic categories).

To construct the context features, we assume these features vary between the lists but remain constant throughout any single list. For Experiment 2, with three study lists, an initial context vector is chosen with random values and used as the context for List 1. These context feature values are then copied with some probability, p_{ctx}, to the context for List 2, with noncopied values chosen randomly. Then, this context vector is transformed into the context vector for List 3 by the same process with the same parameter.

The item's vectors (containing both the item and context features) are then stored as incomplete and error prone episodic traces as described above, with the parameters *u* and *c*. The value of *u*, the probability that a feature is stored, was allowed to vary across experiments, and the value of *c* was fixed on the basis of prior applications of the model.

Experiment 1 contained a single study list, and we therefore assume that the memory search at retrieval is restricted to items from that list and the item features are simply those for the particular test item. To model Experiment 2, we assume that the context features of the test probe are those for List 3 and the item features are those for the particular test item. The retrieval rules are those already described: Odds values are calculated with Equation 4, and an old response is made if that value is greater than 1.0. The parameter space was not searched exhaustively but only until a fit was obtained that reproduced the pattern of the data sufficiently well to demonstrate the merit of the approach.[3]

[3] We treated α as a free parameter, although another plausible approach would have matched α to the actual proportions of traces of each type.

The circles in Figures 2–4 show the model fit on the basis of the following parameter values. The values of $g = .4$ and $c = .9$ were fixed on the basis of prior applications of REM. The values of $\alpha = .6$ and $p_{ex} = .33$ were fit to the data but constrained to be the same for both experiments. The remaining parameters were fit to the data and allowed to vary between experiments: for Experiment 1, $p_{cat} = .7$ and $u = .19$; and for Experiment 2, $p_{ctx} = .85$, $p_{cat} = .8$, and $u = .37$. The fit certainly could be improved if we used additional parameters (or perhaps if we searched the parameter space more fully), but the fit shown is sufficient to illustrate the way the model accounts for the results with a mixture of context and item cuing.

A few notes about the parameters are worthy of mention. A higher value of p_{ctx} results in too little context change between lists, and the resultant predictions are nonmonotonic with the data. A somewhat lower value of p_{ctx} results in too much context change, predicting too few false alarms to studied foils from earlier lists. Thus, the value of p_{ctx} is tightly constrained by the data (as are the values of p_{cat} and p_{ex}). Conversely, α, the parameter governing the relative weighting of context and item likelihoods, can take on a wide variety of midrange values without dramatically changing the predictions. The reason is related to the fact that the likelihood ratios have a highly skewed distribution whereas α produces a linear mixture: Even after weighting by α, a large likelihood value tends to remain relatively large and a small likelihood value tends to remain relatively small over midrange values of α.

Reprise

The article by Dennis and Humphreys (2001) does a valuable service to the field by highlighting the importance of context noise. This factor has been present in prior models, but usually only as an afterthought (and often ignored in simulations). Indeed, in our second study the effect of context noise was much larger than that of item noise, presumably because we used foils from very similar contexts (i.e., recent lists). In more typical studies using foils never presented in the experimental session, the relative importance of context noise would be lower but still important. Thus, in this comment, we have not tried to argue against context noise but have argued that item noise also plays a role. This assumption is explicitly denied in the BCDMEM model (Dennis & Humphreys, 2001). Whereas other models allow both item noise and context noise, BCDMEM allows only context noise. We have presented relevant data and plausibility arguments in favor of the view that both types of noise play important roles in memory retrieval. It is important to emphasize that a slight modification to BCDMEM would align it closely with the other models we have mentioned. The assumption that the test probe activates only past instances of the test item (or in BCDMEM terms, only the context vector associated with the node for the test item) could be expanded to include the activation of nodes of similar words. Such an augmentation would leave order of cue utilization as the primary factor distinguishing BCDMEM from other models, and we have already noted that this is not a core assumption of any of the models under discussion and may not be readily testable. To us, models that incorporate both item and context noise seem simpler, more elegant, and better in accord with the data. One important potential consequence of this dialogue is the hope that the question will shift from a qualitative one concerning the existence of item and context noise to a quantitative one concerning the details of models that incorporate both. As one step toward that goal, we have presented a modified version of the REM model incorporating both item and context noise.

References

Anisfield, M., & Knapp, M. E. (1968). Association, synonymity, and directionality in false recognition. *Journal of Experimental Psychology, 77,* 171–179.

AT&T Laboratories Cambridge. (n.d.) *The database of faces*. Retrieved from http://www.uk.research.att.com/facedatabase.html

Brainerd, C. J., & Reyna, V. F. (1998). When things that were never experienced are easier to "remember" than things that were. *Psychological Science, 9,* 484–489.

Dennis, S., & Humphreys, M. S. (2001). The role of context in episodic recognition: The bind cue decide model of episodic memory. *Psychological Review, 108,* 452–478.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91,* 1–67.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1355–1369.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Psychological Review, 95,* 528–551.

Humphreys, M. S., Pike, R., Bain, J. D., & Tehan, G. (1989). Global matching: A comparison of the SAM, Minerva II, Matrix, and TODAM models. *Journal of Mathematical Psychology, 33,* 36–67.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 734–760.

Morrell, H., 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,* 1095–1110.

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

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review, 104,* 839–862.

Roediger, H. L., III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 803–814.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 267–287.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166.

Sommers, M. S., & Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language, 40,* 83–108.

Underwood, B. J. (1965). False recognition produced by implicit verbal responses. *Journal of Experimental Psychology, 70,* 122–129.

Wixted, J. T., & Stretch, V. (2000). The case against a criterion-shift account of false memory. *Psychological Review, 107,* 368–376.

Zaki, S. R., & Nosofsky, R. M. (2001). Exemplar accounts of blending and distinctiveness effects in perceptual old new recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1022–1041.

Received March 1, 2001
Revision received January 23, 2003
Accepted January 23, 2003 ■