The list length effect in recognition memory: an analysis of potential confounds

Angela Kinnell · Simon Dennis

Published online: 17 November 2010
© The Psychonomic Society 2010

Abstract The list length effect in recognition memory refers to the finding that recognition performance for a short list is superior to that for a long list. The list length effect is consistent with the predictions of item noise models, but context noise models predict no effect. Recently, it has been argued that if potential confounds are controlled, the list length effect is eliminated. We report the results of two experiments in which we looked at the role of attention and the remember-know task in the detection of the list length effect. We conclude that there is no list length effect when potential confounds are controlled and that it is the design used to control for attention that is most vital.

Keywords List length effect · Recognition memory · Memory models · BCDMEM · REM

The sources of interference that determine performance remain an unresolved issue in our understanding of recognition memory. The task of recognition involves determining whether an item appeared in a given context. On logical grounds, then, prime candidates for sources of interference are the other items that appeared on the list and the other contexts in which the item appeared (Humphreys, Wiles & Dennis, 1994). The item noise approach assumes that it is the other items on the study list that interfere with one's ability to recognize a test probe (Criss & Shiffrin,

2004a). There are numerous mathematical models of recognition memory that adopt this approach, including the global matching models (GMMs), such as the Theory of Distributed Associative Memory (TODAM; Gronlund & Elam, 1994; Murdock, 1982), Minerva II (Hintzman, 1986), the Matrix model (Pike, 1984), and Search of Associative Memory (SAM; Gillund & Shiffrin, 1984), as well as the Retrieving Effectively from Memory model (REM; Shiffrin & Steyvers, 1997) and the Subjective Likelihood model (SLiM; McClelland & Chappell, 1998). Alternatively, interference could arise from the other contexts in which an item has been encountered in the past (Dennis & Humphreys, 2001), and any interference from other items may be negligible (Criss & Shiffrin, 2004a). Context noise models are much fewer in number than item noise models and include the Bind Cue Decide Model of Episodic Memory (BCDMEM; Dennis & Humphreys, 2001) and the model of Anderson and Bower (1972). Furthermore, it is possible that both context and item noise play substantive roles in recognition performance (Cary & Reder, 2003; Criss & Shiffrin, 2004a).

Initial evidence in favor of the context noise approach came from the inability to find a list strength effect. The GMMs predict that if one increases the strength of some items on a list by increasing either the duration or the number of presentations, performance on unstrengthened items should decrease, because the amount of item interference increases (Shiffrin, Ratcliff, & Clark, 1990). However, this does not occur (Ratcliff, Clark & Shiffrin, 1990). Subsequently, item noise models have incorporated some form of differentiation mechanism, so that strengthening an item not only increases its strength, but also simultaneously decreases its similarity to other items, so that there is no net change in interference (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). Consequently,

A. Kinnell (✉)
School of Psychology, University of Adelaide,
Adelaide, SA 5005, Australia
e-mail: angela.kinnell@gmail.com

S. Dennis
Department of Psychology, Ohio State University,
Columbus, OH, USA

the lack of a list strength effect is no longer indicative of the context noise approach, and attention has shifted to the impact of increasing the length of the list.

Item noise models predict that as one increases the number of items on the study list, one should compromise performance, since the amount of interference should increase. The existence of the list length effect in recognition has been well documented (e.g., Bowles & Glanzer, 1983; Cary & Reder, 2003; Gronlund & Elam, 1994; Murdock & Kahana, 1993; Murnane & Shiffrin, 1991; Shiffrin, Ratcliff, Murnane, & Nobel, 1993; Strong, 1912; Underwood, 1978), and as a result, its existence has been somewhat ubiquitously accepted in the literature (Cary & Reder, 2003; Dennis & Humphreys, 2001).

However, closer inspection of a number of published studies reveals contradictory findings. Schulman's (1974) results indicated that memory for a particular word was unaffected by the number of words that followed it in study. Buratto and Lamberts (2008) conducted a study involving both list length and list strength manipulations. They did not identify a significant effect of list length on recognition performance. Jang and Huber (2008) tested participants on a series of lists, using the list-before-the-last paradigm of Shiffrin (1970). Participants were tested using both a free recall task and a recognition memory task. Jang and Huber found that the length of the study list (6 vs. 24 items) did not significantly affect recognition performance. Murnane and Shiffrin (1991) found no significant effect of list length in their Experiment 3 when short-list performance was compared with the equivalent portion of the long list. Dennis and Humphreys (2001) argued that previous studies that had identified the list length effect had failed to control for four possible confounds: retention interval, attention, displaced rehearsal, and contextual reinstatement. When they controlled for these confounds, they found no significant difference in recognition performance between a 24-word and a 72-word list and argued that interference does not generate a list length effect in recognition memory.

Four potential confounds of the list length effect

Retention interval

The first potential confound of the list length effect that was outlined by Dennis and Humphreys (2001) is retention interval. The retention interval is the duration of time elapsed between a word being presented at study and the subsequent testing of that item. More time is required to view all of the items on a long list than is needed for the short list, meaning that there is a longer retention interval for long-list items. This confound can be controlled by equating the average retention interval of the short and long

lists, using either a retroactive or a proactive design (both designs are illustrated in Fig. 1). In the retroactive design, the short list is followed by a period of filler activity, such that the duration of the short list and filler activity combined is equal to the duration of the long list. Only the items at the start of the long list (the same number as is in the short list) are tested, such that target items from each list have had the same average retention interval (Cary & Reder, 2003; Dennis & Humphreys, 2001). The proactive design is the converse of this, with a period of filler activity preceding the short list. In this case, only the items at the end of the long list are tested, in order to equate the retention interval between lists.

Attention

Another possible confound of the list length effect, first raised by Underwood (1978), is attention. It is likely that participants will tire over the course of the long list to a greater extent than for the short list and, thus, pay comparatively less attention to the items. This is more problematic in the proactive design (Cary & Reder, 2003; Dennis & Humphreys, 2001; Underwood, 1978), in which it is the final items of the long list that are tested and performance on these items is compared with performance on the short list. In the retroactive design, all the targets appear at the beginning of the respective test lists, and the attention paid to the target items in each list should not differ significantly. Having participants perform an encoding task that requires a response during study (such as a pleasantness-rating task; Cary & Reder, 2003; Dennis & Humphreys, 2001) allows for the assumption that all items will have been processed to some level, regardless of fatigue. It should be noted, however, that there may be no way to completely eliminate attentional lapses in the proactive condition, and the larger the difference in length between the short and the long lists, the more likely it is that this confound will play a role in the list length finding.

Fig. 1 Retroactive and proactive experimental designs. Shading indicates the portion of the study list included as targets at test

Displaced rehearsal

Displaced rehearsal refers to differences in the pattern of item rehearsal between the short and long lists and may also confound the list length effect finding. This problem arises when retention interval is controlled and only some long list items are tested, whereas all the short-list items are included as targets. In this case, any rehearsal of short-list items will be beneficial to performance, since all the studied items are included as targets at test; therefore, any rehearsal is advantageous. However, there is no such guarantee with rehearsal of long-list items, since only a subset of the studied items are tested. Thus, rehearsal of nontested items would detract from the rehearsal of tested items, reducing performance on those targets. This would favor performance for the short list and could result in a list length effect (Cary & Reder, 2003; Dennis & Humphreys, 2001).

Furthermore, the issue of displaced rehearsal is exacerbated in the retroactive condition, wherein the period of filler activity follows the short list. This period, despite not being intended as a time of rehearsal, may nonetheless be used by participants as an opportunity for the rehearsal of short-list items. In the long list, items are continually presented, providing less opportunity for rehearsal. This would again favor performance in the short list and, possibly, give rise to the list length effect.

Displaced rehearsal can be controlled in a number of ways. First, it is important to ensure that the filler task is more interesting and stimulating for participants than the study task, thereby encouraging them to focus on the filler, rather than rehearse the items (Cary & Reder, 2003; Dennis & Humphreys, 2001). A second strategy is to include the recognition test as incidental, although this is problematic when a within-subjects design is used, since both tests cannot be incidental (Cary & Reder, 2003).

Contextual reinstatement

An influential view proposes that one component of context can be thought of as a set of elements that vary randomly with the passage of time (Estes, 1955; Mensink & Raaijmakers, 1989). The closer in time two events occur, the more similar their contexts will be. As a result, the current context at test is likely to differ from the current context at study as a function of the amount of time that has elapsed. The more similar the active contextual elements present at test are to those that were encoded during study, the better the performance will be.

At test, item noise models require a representation of the study context in order to retrieve all the list items associated with that context. Context noise models use the test probe to cue the retrieval of all previous contexts in which that item has been seen before and then compare the retrieved

vector with the representation of the study context. Therefore, the reinstatement of the study context at test is important in both item and context noise models of recognition memory. The more accurate this reinstated study context is, the better the recognition performance is likely to be. It is also possible that participants will not attempt to reinstate an earlier study context at test and, rather, will use the current end-of-list context. Because context varies with the passing of time, there will be more scope for variability in the long list, which would negatively impact performance for that list, given that some of its items would have been studied less recently.

In addition, when retention interval controls are implemented, the issue of contextual reinstatement can become a problem, particularly in the retroactive design, when a period of puzzle filler activity follows the short, but not the long, list. The puzzle activity following the short list involves a clear change in context, through both the passage of time and the change of activity. As a result, when it is time for the test list, it is clear that the puzzle context is inappropriate for the memory test, and a reinstatement of the study context is likely to occur. There is no break at the end of the long list before the beginning of the test list and, thus, no clear demarcation that a change in context has occurred. In this situation, participants may rely on an end-of-list context, which may be different from the start-of-list context. In the retroactive condition, it is the first items of the long list that are tested. As such, reinstating the end-of-list context is unlikely to benefit performance for early items. Considered together, these factors would favor performance on the short list, where the study list context is reinstated more accurately following the puzzle activity.

To control for this confound, contextual reinstatement in both length conditions can be encouraged by including an extended period of filler activity after both the long and short lists, in addition to the period of filler activity already included as a control for retention interval (see Fig. 2). This value has typically varied between studies, from just 9 s (Gronlund & Elam, 1994) to 8 min (Dennis & Humphreys, 2001).

Fig. 2 Experimental design showing the filler as a control for retention interval (in both retroactive and proactive designs) and the inclusion of an additional filler as a control for contextual reinstatement. Shading indicates the portion of the study list included as targets at test

Past and current attempts to eliminate list length effect confounds

Following the work of Dennis and Humphreys (2001), Cary and Reder (2003) conducted three experiments that investigated the list length effect. **Experiment 1** was the basic list length design with none of the controls outlined by Dennis and Humphreys implemented. List lengths were 16, 32, 48, and 64 words, with a test list immediately following each. A statistically significant effect of list length was identified.

Experiment 2 was identical to **Experiment 1**, with the exception of the addition of a 5-min word search puzzle following each study list and before the subsequent test list. This was done to decrease performance but could also be seen to function as a control for contextual reinstatement. No other confounds were controlled, and again, a statistically significant effect of list length resulted.

In **Experiment 3**, controls were implemented for all four of the potential list length effect confounds outlined by Dennis and Humphreys (2001). Contrary to their findings, however, a list length effect was identified. On the basis of this evidence, Cary and Reder (2003) argued that there is a list length effect in recognition memory and that Dennis and Humphreys's design was not strong enough to detect it.

A further analysis of Cary and Reder's (2003) published results suggests that their finding of a list length effect under all conditions may not be so clear. We compared the magnitude of the list length effect in each of Cary and Reder's three experiments. In **Experiments 1 and 2**, the present analysis involved only the 16-word (short) and 64-word (long) lists, in order to match the 1:4 list length ratio of the 20-word (short) and 80-word (long) lists in **Experiment 3**. A *t*-test for unequal samples was carried out to analyze the differences between the *d'* scores for the short and long lists in each of the three experiments.

Analysis revealed that the difference between short and long list *d'* in **Experiment 1** was not statistically significantly different from the same comparison in **Experiment 2**, $t(46) = 0.40$, $p > .05$, two-tailed. This result is unsurprising, given the similarity in experimental design. Interestingly, the difference between short- and long-list *d'* in each of these experiments was significantly different from the *d'* difference in **Experiment 3** [$t(68) = 2.99$, $p < .05$ (two-tailed) for **Experiment 1** vs. **Experiment 3**, and $t(56) = 3.04$, $p < .05$ (two-tailed) for **Experiment 2** vs. **Experiment 3**]. These results are illustrated in Fig. 3. Each bar represents the difference in short- and long-list *d'* for each experiment. It is evident that despite the existence of a statistically significant list length effect in **Experiment 3**, the magnitude of this effect is different from that of the significant effects identified in the previous two experiments. As was previously noted, **Experiment 3** involved the introduction

Fig. 3 Differences in *d'* scores between the short and long lists in each of Cary and Reder's (2003) three experiments. Bars represent 95% confidence intervals of the differences between the means

of controls for the four potential list length effect confounds identified by Dennis and Humphreys (2001). It is clear in the present analysis that employing these controls reduces the magnitude of the list length effect by a statistically significant amount from the original experiments. The fact that a list length effect is still identified may highlight the difficulty in controlling for the possible confounds.

Furthermore, there were a number of differences between the list length studies of Dennis and Humphreys (2001) and Cary and Reder's (2003) **Experiment 3**. To begin with, a different list length manipulation was used. Dennis and Humphreys used both a 1:2 (40:80 word) ratio and a 1:3 (24:72 word) ratio, whereas Cary and Reder's **Experiment 3** had a 1:4 (20:80 word) ratio. It is more likely that item interference will be detected in the latter experiment, given the stronger manipulation of list length, but there is also more potential for the list length confounds to play a role. The greater the ratio, the more likely there is to be differences in attention paid to short and long lists. A larger list length ratio also results in a longer period of filler activity following the short list and more opportunity to rehearse list items. Finally, a larger list length ratio would magnify any differences in contextual drift.

There was also a difference in the analysis of the results, with Cary and Reder (2003) combining the results of the retroactive and proactive conditions. It is therefore unclear whether both designs contributed to the list length effect or whether it was primarily the proactive condition, where attentional lapses are more likely to occur.

Cary and Reder (2003) also employed the remember-know (RK) paradigm in their study, as opposed to the traditional yes/no recognition paradigm. It was the first study to use the RK task to investigate the list length effect. Originally developed by Tulving (1985), the RK paradigm has been used as a means of investigating an individual's conscious experience and awareness in the recognition task

(Dunn, 2004). The RK task is easily incorporated into the standard yes/no recognition paradigm. The most common method is a two-step procedure. After making a “yes” response, indicating that they have recognized the test probe from the study list, participants are given the additional step of deciding whether that decision was based on a remember or a know judgment. A remember response is said to signify that the participant can consciously recollect the experience of seeing the remembered word during study (Gardiner, 1988). Know responses are thought to be given when there is no such recollection, with the decision based primarily on a general feeling of familiarity with the test probe (Gardiner & Richardson-Klavehn, 2000; Knowlton & Squire, 1995).

The inclusion of the RK task in the study design could potentially confound the list length effect. Remember responses, in that they are based on recollection of the study experience, have been said to involve a recall-like process (Diana, Reder, Arndt & Park, 2006). Participants are no longer just asked whether or not they recognize a particular word, but rather, they are asked to recall elements of the study event. As Diana et al. noted, the use of the RK paradigm may alter the task requirements, such that participants may rely on recollection under those conditions more than they would with the yes/no recognition paradigm. Thus, the use of the RK task may induce recall. Since the list length effect is widely accepted to occur in recall, this may help to explain why Cary and Reder (2003) found a positive list length effect where Dennis and Humphreys (2001) did not.

There was also a difference in the contextual reinstatement control used in each of the experiments. Dennis and Humphreys (2001) included an 8-min period of puzzle filler before each test list in their second experiment, whereas Cary and Reder's (2003) control for contextual reinstatement was a 2-min period of algebra problem solving before each test list. It is possible that 2 min was not a sufficient amount of time to encourage participants to reinstate the study context at test, rather than rely on the end-of-list context. If this was the case, the list length effect may still be identified, despite the controls.

Dennis, Lee and Kinnell (2008) addressed this issue. Their study contained a condition that encouraged contextual reinstatement after both the short and long lists (filler condition) and one that facilitated contextual reinstatement only after the short list (no-filler condition). In both cases, controls were implemented for the other three potential confounds. Dennis et al. found a significant effect of list length in the no-filler condition, the condition in which contextual reinstatement was facilitated only after the short list, by means of puzzle filler activity. Conversely, no significant effect of list length was identified in the filler condition when controls for Dennis and Humphreys's

(2001) four confounds were implemented. On this basis, it seems that failure to control for contextual reinstatement could confound the list length effect finding.

The present experiments continued to explore the role of potential confounds of the list length effect. Specifically, the role of attention and the use of the RK task in the detection of the list length effect in recognition memory was investigated.

Experiment 1 - attention

The aim of this experiment was to examine the influence of attention on the detection of the list length effect. More specifically, we wanted to investigate whether there are differences in performance depending on whether a retroactive or a proactive design is adopted. In addition, the difference between a condition in which participants were required to perform a pleasantness-rating task at study and a condition in which there was no such requirement was investigated. The pleasantness-rating task has been used in previous studies in an attempt to control for differential lapses in attention (Cary & Reder, 2003; Dennis & Humphreys, 2001).

Method

Participants

Participants were 160 psychology students from the University of Adelaide. Each received either course credit or a payment of \$12 in exchange for their participation. All gave informed consent.

Design

This experiment had a $2 \times 2 \times 2 \times 2$ factorial design, with the factors being list length (short or long), word frequency (low or high), attention task (pleasantness rating or read only), and design (retroactive or proactive). List length and word frequency were within-subjects factors, whereas attention task and design were between-subjects manipulations.

The word frequency manipulation was included as a check of the power of the experimental design. The ability to detect a significant word frequency effect in this experiment would indicate that any failure to find a list length effect would not be because the power of the experiment was too poor to detect any effects.

Materials

The stimuli for this experiment were 140 five- and six-letter words from the Sydney Morning Herald Word Database

(Dennis, 1995). Half of the words were of high frequency (100–200 occurrences per million), and half were of low frequency (1–4 occurrences per million). All the lists had the same number of five- and six-letter and high- and low-frequency words. All the words were randomly assigned to lists, with no participant seeing the same word twice, except for targets.

Procedure

Participants were first given an overview of the study and were introduced to the filler activity that would be used throughout the experiment. A computerized sliding tile puzzle was used as the filler task. An image of a fractal was split into 12 pieces of equal size and then scrambled. The participants' task was to rearrange the pieces and return the image to its original form.

Participants studied one short (20-word) and one long (80-word) list, the same list lengths as those in Cary and Reder's, (2003) study. Each study word appeared for 3,000 ms. Test lists were made up of 20 targets and 20 distractors. All the lists had half high-frequency and half low-frequency words. All the words were presented in lowercase letters in the center of a computer screen in white font on a blue background.

Participants were split equally into two attention task conditions. In the pleasantness-rating condition, participants were asked to rate the pleasantness of each word on the study list on a 6-point Likert scale (1, *least pleasant*; 6, *most pleasant*) by clicking the appropriate button while that word was being displayed on screen. Participants were told that if they missed rating one of the words within the 3,000 ms, they should rate the next word instead. In the read-only task condition, participants simply read the words of the study list as they appeared on the screen. No response was required.

Within each condition, the design of the lists was either retroactive or proactive in nature. Participants were again divided equally into these conditions. In the retroactive design, the short list was followed by a 3-min period of sliding tile puzzle filler, and the first 20 words of the long list were included as targets at test. In the proactive design, there was 3 min of puzzle filler before the beginning of the short list, and the last 20 words of the long list were tested.

Participants were given 15-s notice before the onset of the test list, which was in the form of the yes/no recognition paradigm. Each word was presented in the middle of the screen above two response buttons marked “yes” and “no.” Participants were instructed to respond “yes” if they recognized the word from the study list and to respond “no” if they did not recognize that word, by clicking on the appropriate button. The test list was self-paced, and a response was recorded for each test word. The targets were

the entire study list (short list), the first 20 words of the long study list (retroactive design), or the last 20 words of the long list (proactive design).

An 8-min period of sliding tile puzzle filler activity was included before each test list. This was done in an attempt to offset potential differences in contextual reinstatement of the short- and long-list study contexts at test.

The experiment was counterbalanced for order; within each condition, half of the participants began with the short list, and the other half began with the long list.

Results and discussion

The analysis of the present results presents a fundamental difficulty. In circumstances in which both the null and alternative hypotheses have theoretical implications, standard approaches to null hypothesis testing, such as an analysis of variance (ANOVA), are not applicable (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Although we appreciate that most readers will be more familiar with ANOVA and, so, we provide this analysis, we also employ the Bayesian analysis introduced by Dennis et al. (2008), which is designed to address within-subjects designs that submit to a signal detection analysis. Under this approach, results are reported as a pair of probabilities—for example, (.05, .81). The first number presented is the probability that at least 90% of participants favor an error-only model; that is, their pattern of responses is better accounted for by a model in which there is no effect, just chance variation, analogous to the null hypothesis. The second number is the probability that at least 90% of participants are better accounted for by an error-plus-effect model; that is, their pattern of responses is better accounted for by a model in which there is both an effect and chance variation, analogous to the alternative hypothesis. Interpretation is straightforward and is not prone to the subtle mistakes that commonly plague the interpretation of p values. Note that the two probabilities do not sum to 1, because there is some probability that neither of these alternatives is the case. For instance, it might be that half of the subjects show an error-only pattern and half show an error-plus-effect pattern.

The Bayesian method has a number of advantages over null hypothesis testing (Dennis et al., 2008). For our purposes, the most critical of these is that evidence can be accumulated in favor of the error-only hypothesis. The first number in the parentheses is the probability that we should favor this hypothesis. In addition, rather than focus on a difference in the means of two distributions over participants, the method makes a statement about the patterns of responses that individuals tend to show. Whereas the comparison of means can lead one to make generalizations to a population when only a minority of participants are

affected, the Dennis et al. method cannot be misused in this way. Furthermore, the method can be used iteratively and with arbitrarily small sample sizes, does not require edge corrections to avoid infinite d' s, and is inherently sensitive to the number of targets and distractors that were presented to the participants. For a more detailed discussion of the method and its advantages, see Dennis et al.

List length

Figure 4 shows the d' results as a function of list length for all four conditions, and Table 1 shows the corresponding hit and false alarm rates. As recommended by Snodgrass and Corwin (1988), the edge correction applied to all hit and false alarm rates for use in d' calculations was made by adding a value of .5 to the hit and false alarm counts and adding 1 to the number of target and distractor items. For all analyses, $F < 1$ and $p > .05$, unless otherwise stated.

A $2 \times 2 \times 2 \times 2$ (length $\times$ frequency $\times$ task $\times$ design) repeated measures ANOVA yielded a nonsignificant effect of list length on d' , $F(1,156) = 2.71$, $p = .1$, and the hit rate, $F(1,156) = 1.21$, $p = .27$. However, there was a statistically significant effect of list length on the false alarm rate, $F(1,156) = 11.01$, $p = .001$, $\eta_p^2 = .07$.

For comparison with the results of Cary and Reder (2003), a $2 \times 2 \times 2$ (length $\times$ frequency $\times$ design) repeated measures ANOVA was carried out for the pleasantness task condition. Analysis revealed a nonsignificant interaction between list length and design on d' , $F(1,78) = 2.09$, $p = .15$, the hit rate,

Fig. 4 d' values for each of the four attention conditions. There was a nonsignificant list length effect when the retroactive design was used and a positive list length effect when the proactive design was used. Bars represent 95% within-subjects confidence intervals. Asterisks indicate statistically significant differences

$F(1,78) = 3.70$, $p = .06$, and the false alarm rate. Cary and Reder obtained the same result and, on that basis, collapsed the retroactive and proactive conditions. Note that such a procedure relies on the inference that a nonsignificant interaction implies equality across conditions, which, as we will see, is not necessarily the case. In the present analysis, the conditions will remain separated.

Four planned comparisons were also carried out on each of the four subgroups in this experiment: pleasantness ratings in the retroactive condition, read only in the retroactive condition, pleasantness ratings in the proactive condition, and read only in the proactive condition. Both the one-way ANOVAs and Bayesian analyses¹ examined the effect of list length collapsed across word frequency.

Retroactive Pleasantness Condition The retroactive pleasantness condition provided controls for all four potential confounds of the list length effect: retention interval, attention, displaced rehearsal, and contextual reinstatement. Repeated measures ANOVAs showed a nonsignificant effect of list length on d' , the hit rate, and the false alarm rate, $F(1,39) = 3.95$, $p = .054$. Similarly, the Bayesian analysis of the d' values found in favor of the error-only model (.81, .01). Thus, there was no significant list length effect found in this condition, although the effect on the false alarm rate was close to significance.

Retroactive Read Condition The retroactive read condition controlled for the potential influence of retention interval, displaced rehearsal, and contextual reinstatement. There was no control for attention implemented. Repeated measures ANOVAs in this condition yielded nonsignificant effects of list length on d' , $F(1,39) = 3.06$, $p = .09$, and the false alarm rate. There was, however, a statistically significant effect of list length on the hit rate, $F(1,39) = 9.95$, $p = .003$, $\eta_p^2 = .20$. It should be noted, however, that in this condition, performance on the long list was superior to that on the short list, meaning that this result is significant in the direction opposite to that previously identified in the literature. Bayesian analysis of the d' values again found in favor of the error-only model (.78, .06) and a null list length effect.

Proactive Pleasantness Condition The proactive pleasantness condition involved controls for the four potential list length effect confounds. However, it remains the case that by the end of the long list, participants may not be paying as much attention to the study words as they were at the start of the list, and so the potential for an attention-induced list

¹ The Bayesian results were obtained by implementing the unequal variance signal detection model ($\tau = 0.8$), a burn-in period of 10^3 samples, followed by the collection of 10^4 samples.

Table 1 Hit and false alarm rates for each of the four attention conditions

	Hit Rate				False Alarm Rate			
	Short List		Long List		Short List		Long List	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Retroactive Pleasantness	.84	.14	.87	.11	.14	.13	.18	.16
Retroactive Read	.72	.17	.82	.15	.19	.16	.21	.16
Proactive Pleasantness	.89	.11	.87	.14	.12	.12	.17	.11
Proactive Read	.78	.16	.74	.13	.17	.13	.22	.14

length effect remains. In this condition, repeated measures ANOVAs yielded statistically significant effects of list length on both d' , $F(1,39) = 11.55$, $p = .002$, $\eta_p^2 = .23$, and the false alarm rate, $F(1,39) = 6.72$, $p = .013$, $\eta_p^2 = .15$. There was no significant effect on the hit rate. In contrast to the ANOVA d' results, the Bayesian analysis found in favor of the error-only model (.68, .13).

Proactive Read Condition Finally, the proactive read condition provided controls for retention interval, displaced rehearsal, and contextual reinstatement. There was no control for differential lapses in attention, and, as was noted in the previous section, the use of the proactive design may have exacerbated this. A repeated measures ANOVA in this condition yielded a significant effect of list length on d' , $F(1,39) = 8.26$, $p < .001$, $\eta_p^2 = .17$. There was, however, a nonsignificant effect of list length on both the hit rate, $F(1,39) = 2.40$, $p = .13$, and the false alarm rate, $F(1,39) = 3.65$, $p = .06$, although this was close to significance. The Bayesian analysis of d' values was ambiguous for this condition (.46, .27).

Word frequency

A $2 \times 2 \times 2 \times 2$ (length $\times$ frequency $\times$ task $\times$ design) repeated measures ANOVA yielded a significant effect of word frequency on d' , $F(1,156) = 232.79$, $p < .001$, $\eta_p^2 = .60$, in the overall data. Furthermore, planned comparisons were carried out on the word frequency data in each of the four conditions, collapsing across list length. A strong word frequency effect was identified under all conditions (see Fig. 5), using both the standard ANOVA analysis and the Bayesian analysis. These findings place a lower bound on the power of the experiment.

The retroactive pleasantness condition involved controls for all the confounds outlined by Dennis and Humphreys (2001), and the use of the retroactive design made it less likely that attention would play a part in a spurious list length effect finding. It should be noted that, in this condition, no list length effect was identified. At the other end of the scale, however, the proactive read condition involved no control for attention under circumstances

(proactive design) in which inattention was likely. A positive list length effect was identified in this condition. In addition, it should be noted that in the retroactive read condition, long-list performance was superior to short-list performance. This result is the first example of superior recognition performance for the long list of which we are aware and a reversal of the traditional list length finding, which sees short-list performance exceed long-list performance. A significant reversal of the effect would be problematic for existing models to account for.

The results of this experiment suggest that it is the retroactive versus proactive distinction that is most influential in the detection of the list length effect, rather than the nature of the study task. This is critical to the comparison with Cary and Reder's (2003) Experiment 3, in which the retroactive and proactive design conditions were collapsed for analysis. When the present results were collapsed in the same manner, a positive list length effect was identified. It appears that the design of the experiment is important and that it is the proactive condition that drives the effect.

Experiment 2 – the remember–know task

The aim of Experiment 2 was to investigate the influence of the RK task at test on the list length effect finding, while controlling for the four potential confounds of the list length effect. It was thought that the RK task may induce recall and that a positive list length effect would be identified in that condition as it is in recall, whereas there would be a null list length effect under yes/no task conditions (a condition equivalent to the retroactive pleasantness condition in Experiment 1).

Method

Participants

Participants were 80 first-year psychology students from the University of Adelaide who participated in exchange for course credit. All gave informed consent.

Fig. 5 A significant word frequency effect was identified in both hit and false alarm rates in each condition. Bars represent 95% within-subjects confidence intervals. Asterisks indicate statistically significant differences

Design

A $2 \times 2 \times 2$ factorial design was used in this study. The factors were list length (short or long), word frequency (low or high), and test task (yes/no task or RK task). List length and word frequency were within-subjects variables, and test task was a between-subjects comparison. The word frequency manipulation was again included as a check on the power of the experiment.

Materials

In this experiment, 140 words were used as stimuli. They were chosen on the basis of the same criteria as those in Experiment 1.

Procedure

The procedure of this experiment largely followed that of Experiment 1, with a few exceptions. Only the retroactive design was used, and pleasantness ratings were included in both conditions. The results of Experiment 1 and previous literature suggest that the use of the retroactive design and pleasantness rating-task are best for controlling for potential

confounds of the list length effect. Thus, the procedure followed that of the retroactive pleasantness condition of Experiment 1.

In the yes/no task condition, the test list took the same form as in Experiment 1. In the RK task condition, however, an extra step was added to the test task. Upon answering “yes” to a probe word, participants were shown a new screen and were asked to indicate whether they had made a “remember” or a “know” judgment by clicking on the appropriate button with the mouse. These options remained on screen until a response was made. The difference between the two responses was explained to participants prior to the beginning of the experiment. This was based on explanations given by Cary and Reder (2003), which, in turn, were based on those of Knowlton and Squire (1995). On completion of the experiment, participants were asked to give examples of both a remember and a know judgment, to ensure that they had comprehended the instructions.

The controls for retention interval and contextual reinstatement were implemented as in Experiment 1. For retention interval, the retroactive design was used, the short list was followed by 3 min of sliding tile puzzle activity, and it was the first 20 words of the long list that were used

as targets in the test list. To offset potential differences in contextual reinstatement, an additional 8 min of puzzle activity was included before each test list. Lists were counterbalanced for order.

Results and discussion

List length

Figure 6 shows the d' results as a function of list length for the yes/no and RK task conditions, and Table 2 shows the corresponding hit and false alarm rates. For all analyses, $F < 1$ and $p > .05$, unless otherwise stated.

A $2 \times 2 \times 2$ (length $\times$ frequency $\times$ task) repeated measures ANOVA showed no significant interaction between list length and test task, $F(1,78) = 1.32$, $p = .25$, on d' . A similar pattern was identified in both the hit and false alarm rates. There was a significant effect of test task on the false alarm rate, $F(1,78) = 4.89$, $p = .03$, $\eta_p^2 = .06$.

In addition, two planned comparisons were carried out on each of the test task conditions separately. Both the repeated measures ANOVAs and the Bayesian analyses were carried out to examine the effect of list length while collapsing across word frequency.

Yes/No Task Condition Controls for the four potential confounds of the list length effect were implemented in the yes/no task condition. Repeated measures ANOVAs yielded nonsignificant effects of list length on d' , the hit rate, and the false alarm rate. Similarly, the Bayesian analysis found strongly in favor of the error-only model (.84, .03). These are null list length effect findings.

RK Task Condition The RK task condition also involved controls for the four potential confounds of the list length

Fig. 6 d' values for the yes/no task and RK task conditions. There was not a significant list length effect in either condition. Bars represent 95% within-subjects confidence intervals

effect. Three repeated measures ANOVAs yielded nonsignificant effects of list length on d' , $F(1,39) = 1.21$, $p = .28$, the hit rate, and the false alarm rate, $F(1,39) = 1.24$, $p = .27$. Again, the Bayesian analysis favored the error-only model (.69, .22) and suggested a null list length effect.

Analysis of RK data

A 2×2 (list length $\times$ whether a word was studied) repeated measures ANOVA was conducted on all the remember responses given by participants in the RK task condition. There was no significant effect of list length on the number of remember responses given (see Fig. 7 for the proportions of remember and know responses). There was an anticipated effect of whether or not a particular word was studied on the number of remember responses, $F(1,39) = 520.75$, $p < .001$, $\eta_p^2 = .93$. Given that remember responses should be given only if the participant has a recollection of the word's appearing in the study list, the strength of this effect is expected. There was also a nonsignificant interaction between list length and whether or not a particular word was studied on the number of remember responses given. It should be noted that this interaction was significant in the study of Cary and Reder (2003).

The results for the know responses followed a pattern similar to that for the remember responses. Again, the effect of list length on the number of know responses was nonsignificant. Whether or not a word was studied had a significant effect on the number of know responses elicited, $F(1,39) = 35.89$, $p < .001$. Finally, the interaction between list length and whether or not a word was studied was nonsignificant.

Word frequency

Figure 7 shows hit rates and false alarm rates as a function of frequency for the yes/no and RK task conditions.

A $2 \times 2 \times 2$ (length $\times$ frequency $\times$ task) repeated measures ANOVA yielded a significant effect of word frequency on d' in the overall data, $F(1,78) = 35.81$, $p < .001$, $\eta_p^2 = .31$. In addition, planned comparisons were carried out on the word frequency data for both the yes/no task and RK task conditions separately, collapsing across list length. The word frequency effect was found in all the conditions, using both the ANOVA and Bayesian methods. However, the effect of word frequency on hit rate in the yes/no task condition was marginally significant, $p = .08$, consistent with several other studies that identified disruptions to the effect in the hit rate under a variety of encoding conditions (e.g., Criss & Malmberg, 2008; Criss & Shiffrin, 2004b; Glanc & Greene, 2007; Hirshman & Arndt, 1997).

Despite the initial hypothesis that the RK task would yield a positive list length effect, the results failed to

Table 2 Mean hit and false alarm rates for the Yes/No Task and RK Task conditions in Experiment 2

	Hit Rate				False Alarm Rate			
	Short List		Long List		Short List		Long List	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Yes/No Task Condition	.85	.15	.85	.13	.14	.14	.15	.13
RK Task Condition	.85	.14	.84	.14	.08	.11	.10	.11

support this, and a significant list length effect was not identified in either the RK task or the controlled yes/no task condition. It is also interesting to note that if remember responses are taken to be generated by recall-like processes, a list length effect in the number of remember responses given would be the expected finding. This was not the case in this experiment.

General discussion

We have presented the results of two experiments that investigated the list length effect in recognition memory. Dennis and Humphreys (2001) proposed that when four potential confounds (retention interval, displaced rehearsal, attention, and contextual reinstatement) were controlled for, there was a null list length effect in recognition memory. A subsequent partial replication by Cary and Reder (2003), however, did not support this assertion. Our objective has been to explore the role of the confounds outlined by

Dennis and Humphreys and to identify the differences between their list length experiment and that of Cary and Reder.

First, we investigated the role of attention in the detection of the list length effect. Underwood (1978) suggested that lapses in attention could explain poorer recognition performance on long lists than on shorter lists. The results of Experiment 1 indicated that attention played a significant role in the detection of the list length effect. Most important, a list length effect was identified when the proactive condition was used as a control for retention interval, but there was no effect when the retroactive design was used. This is not unexpected, given that the proactive design involves the last 20 words of an 80-word list being tested as targets and this performance is compared to that for a 20-word short list. In this case, the amount of attention paid to each block of words is likely to differ and give rise to the list length effect. In the retroactive design, however, all the targets in both test lists were presented at the beginning of both the long- and short-list study blocks, where there should be no differences in attention.

In their third experiment, Cary and Reder (2003) included both retroactive and proactive designs. In their analysis, the data were collapsed across these conditions into one, on the basis of the finding of a nonsignificant interaction between list length and study design. The results of Experiment 1 suggest that collapsing the data in this way may have been problematic. We also found a nonsignificant interaction between list length and design, and when we collapsed across the study design condition, as Cary and Reder did, we also identified a positive list length effect when the pleasantness-rating task was used at study. However, as has already been noted, a positive list length effect was identified in the proactive design only. This finding suggests that the nonsignificance of the interaction effect should not be used as justification for collapsing across conditions and that this may have altered the interpretation of Cary and Reder's findings.

The results also indicated that the attention task used at study does not have a large influence on the list length effect finding. Nevertheless, the retroactive pleasantness condition—that is, the most controlled condition—had the smallest effect size. In addition, the retroactive read condition had the first case, to our knowledge, of a reverse

Fig. 7 A significant word frequency effect was identified in all comparisons but the hit rate in the yes/no task condition. The means were in the expected direction. Bars represent 95% within-subjects confidence intervals. Asterisks indicate statistically significant differences. Hatched shading indicates the proportion of know responses

list length effect, with long-list performance superior to that for the short list (although this effect was only marginally significant). These findings support the claim of Dennis and Humphreys (2001) that it has been a failure to adequately control for potential confounds that has resulted in past list length effect findings.

Experiment 2 focused on the RK task that Cary and Reder (2003) included as part of the test list in their experiments. Dennis and Humphreys (2001) used only the yes/no recognition paradigm. It was thought that the use of the RK task may have induced recall strategies in participants. No list length effect was found with either the RK task or the yes/no task condition. Furthermore, no effect of list length on the number of remember responses given was found. If these responses are based on a recall-like process, as some have suggested (cf. Diana et al., 2006), a positive list length effect would be expected, as occurs in recall. The fact that this was not the case in the present experiment suggests that the inclusion of the RK task did not induce recall (see Cohen, Rotello, & MacMillan, 2008; Dunn, 2004) and was not responsible for the list length finding of Cary and Reder.

The present results support the assertion by Dennis and Humphreys (2001) that the history of positive list length effect findings may have been influenced by a failure to control for four possible confounds: retention interval, attention, displaced rehearsal, and contextual reinstatement. When controls for these confounds are implemented, there is little evidence for a list length effect. More specifically, using the retroactive design to control for retention interval and the inclusion of an extended (8-min) period of filler activity before the test list of both long and short lists leads to equivalent performance across list lengths.

There are at least two potential objections to the conclusions we have drawn. First, it may be the case that the use of a within-subjects design induced order effects that compromised the list length finding. Second, although accuracy seems to show no effect of list length, it may be that there are effects in reaction times. In the following sections, we address these two possibilities, before discussing the implications for models of recognition memory.

A between-subjects analysis

In order to provide maximal power, all the experiments reported in this article and the majority of studies that have looked at the list length effect in the literature have used a within-subjects design, with each participant seeing both short and long lists counterbalanced for order.

There are at least two potential problems with the within-subjects design, however. First, it means that by the start of the second test list, all the participants will have seen the

same number of study items within the experimental session, regardless of whether they are nominally in the short or the long condition. This could potentially serve to reduce the size of any list length effect. Second, participants who experience the long–short ordering may be more susceptible to attention lapses with the second study list than are participants who experience the short–long ordering, particularly in the retroactive condition, in which the majority of items have yet to appear when participants are studying the critical test items. Any such effect would systematically favor the long list and potentially counteract length effects due to interference.

One way to check whether the within-subjects design has confounded the results is to use only the data from the first list studied by each participant in the analysis. Table 3 shows both the original within-subjects ANOVA results and the between-subjects reanalysis of the effect of list length on d' for each of the six conditions, as well as descriptive statistics. Using an alpha level of .05, the conclusions remain unchanged, although the result in the retroactive pleasantness condition becomes marginally significant.² Similarly, effect sizes in the between-subjects analysis failed to reach the criterion for a small effect of .1 suggested by Cohen (1988) in all but the proactive conditions (see Fig. 8).

Taking the first list of each participant necessarily comes at the cost of a loss of experimental power, since the sample size is halved and the comparison is between subjects. However, the retroactive pleasantness condition in Experiment 1 and the yes/no task condition in Experiment 2 were identical to one another. When the combined data from these two conditions were analyzed, no effect of list length on d' was found, $F(1,78)=2.21$, $p=.14$, and the effect size remained very small (see Fig. 8).

It would be useful to apply the Bayesian analysis of Dennis et al. (2008) to the between-subjects analysis. However, at this stage, a between-subjects version of the model has not been devised.

Reaction time analyses

The analyses conducted to this point have focused on discriminability. However, it is also possible that differences between short and long lists would emerge in reaction times. Table 4 shows the within- and between-subjects analyses of median reaction times for the correct (hits and correct rejections) and incorrect (false alarms and misses) responses in each of the experiments, as well as descriptive statistics. Using an alpha level of .5, there is one difference

² The effect is eliminated completely if the most extreme participant is removed.

Table 3 Descriptive statistics and ANOVA results of the effect of list length on d' for both experiments using both the within-subjects and between-subjects (first list only) analyses

	Within Subjects Analysis		Between Subjects Analysis	
	Short List	Long List	Short List	Long List
Experiment 1 - Attention				
Retroactive Pleasantness	2.21 (.79) $F(1, 39) = .38, p = .54$	2.13 (.82)	2.57 (.69) $F(1,38) = 3.43, p = .07$	2.08 (.95)
Retroactive Read	1.60 (.92) $F(1,39) = 3.06, p = .09$	1.88 (.85)	1.97 (.69) $F(1,38) = .01, p = .93$	1.99 (.79)
Proactive Pleasantness	2.52 (.79) $F(1,39) = 11.55, p = .002^*$	2.16 (.69)	2.90 (.58) $F(1,38) = 17.13, p = .0002^*$	2.05 (.70)
Proactive Read	1.86 (.98) $F(1,39) = 8.26, p = .007^*$	1.48 (.67)	2.40 (.92) $F(1,38) = 21.58, p = .00004^*$	1.30 (.52)
Experiment 2 – The Remember–Know Task				
Yes/No Task	2.28 (.86) $F(1,39) = .43, p = .51$	2.19 (.80)	2.36 (.95) $F(1,38) = .05, p = .82$	2.30 (.67)
RK Task	2.53 (.74) $F(1,39) = 12.21, p = .28$	2.41 (.72)	2.75 (.57) $F(1,38) = 1.29, p = .26$	2.51 (.76)

Note. Standard deviations are in parentheses

* $p < .05$

for correct responses in the within-subjects comparison for the RK task. However, there is no effect in the corresponding between-subjects comparison, and if one employed a Bonferroni correction, the alpha level would be .0021, so there is little evidence for any difference in reaction times.

Implications for existing models of recognition memory

The absence of the list length effect in recognition memory presents difficulties for item noise models and lends support

to context noise models, which do not predict such an effect.

The set of models known as the global matching models (GMMs), including TODAM (Gronlund & Elam, 1994; Murdock, 1982), SAM (Gillund & Shiffrin, 1984), the Matrix model (Pike, 1984), and Minerva II (Hintzman, 1986), all predict a list length effect in recognition memory. Although these models differ from each other, they all predict the effect in a similar way. As their name suggests, GMMs involve a global matching process. The test probe cues the retrieval of all study list items from memory. Each of these items is then compared with the test probe. The results of these comparisons are then summed to produce a global level of activation, which is compared against a criterion in order to decide whether to produce a “yes” or a “no” response. Although the effect of increasing list length on the means of the signal and distractor distributions is equivalent, the inclusion of more items results in an increase in the variance of these distributions, performance drops, and a list length effect is produced (Clark & Gronlund, 1996).

Furthermore, it seems doubtful that these models can be easily modified to avoid the list length prediction. In the present experiments, performance was neither near ceiling nor near floor, and yet a fourfold increase in item interference had little effect. Given that the defining mechanism of the item noise models is item interference, it is difficult to see how they could accommodate these results without significant modification.

More recent item noise models such as REM (Shiffrin & Steyvers, 1997) are in a somewhat better position to account for negligible length effects. Although the REM model predicts list length effects for the same reason as the GMMs, under typical parameterizations, these effects are

Fig. 8 Effect sizes for between-subjects analyses of the six conditions plus the combined data from the retroactive pleasantness condition of Experiment 1 and the yes/no task condition of Experiment 2. RK = remember–know

Table 4 ANOVA results of the effect of list length on median reaction times (in milliseconds) for both experiments using both the within-subjects and between-subjects (first list only) analyses

		Within-Subjects Analysis		Between-Subjects Analysis	
		Short List	Long List	Short List	Long List
Experiment 1 - Attention					
Retroactive Pleasantness	Correct	1,092 (155)	1,106 (158)	1,067 (150)	1,124 (153)
		$F(1, 39) = 0.35, p = .56$		$F(1,38) = 1.39, p = .25$	
	Incorrect	1,404 (674)	1,372 (477)	1,391 (753)	1,492 (556)
		$F(1,37) = 0.05, p = .82$		$F(1,36) = 0.22, p = .64$	
Retroactive Read	Correct	1,104 (185)	1,123 (153)	1,099 (191)	1,107 (119)
		$F(1,39) = 0.62, p = .44$		$F(1,38) = 0.02, p = .88$	
	Incorrect	1,322 (339)	1,358 (466)	1,399 (277)	1,420 (466)
		$F(1,38) = 0.23, p = .64$		$F(1,38) = 0.03, p = .86$	
Proactive Pleasantness	Correct	1,104 (186)	1,137 (195)	1,174 (221)	1,100 (142)
		$F(1,39) = 3.76, p = .06$		$F(1,38) = 1.62, p = .21$	
	Incorrect	1,946 (2489)	1,568 (788)	2,594 (3,530)	1,484 (399)
		$F(1,37) = 1.56, p = .22$		$F(1,36) = 1.95, p = .17$	
Proactive Read	Correct	1,110 (162)	1,152 (206)	1,087 (124)	1,166 (190)
		$F(1,39) = 3.52, p = .07$		$F(1,38) = 2.41, p = .13$	
	Incorrect	1,533 (583)	1,460 (457)	1,712 (692)	1,365 (444)
		$F(1,37) = 0.90, p = .35$		$F(1,36) = 3.45, p = .07$	
Experiment 2 – The Remember–Know Task					
Yes/No Task	Correct	1,133 (195)	1,144 (248)	1,127 (181)	1,149 (290)
		$F(1,39) = 0.14, p = .71$		$F(1,39) = 0.08, p = .79$	
	Incorrect	1,591 (1,127)	1,470 (369)	1,864 (1,586)	1,468 (407)
		$F(1,36) = 0.37, p = .55$		$F(1,36) = 1.17, p = .29$	
RK Task	Correct	1,166 (156)	1,223 (235)	1,194 (165)	1,226 (270)
		$F(1,39) = 4.71, p = .04^*$		$F(1,39) = 0.19, p = .66$	
	Incorrect	1,938 (1,090)	1,763 (780)	2,151 (875)	1,788 (1,016)
		$F(1,35) = 1.97, p = .17$		$F(1,36) = 1.39, p = .25$	

Note. Standard deviations are in parentheses

* $p < .05$

quite small. The reason is that the feature generation mechanism employed by REM effectively ensures that items that are not assumed to have any similarity structure will tend to be very dissimilar. Individual lower probability features are sufficient to distinguish items from each other. Probabilistic encoding of these features at study means that critical low-probability features may or may not appear and that performance will be primarily influenced by this encoding, rather than by interference from other items. Thus, there are parameterizations of REM under which items become so dissimilar to one another that item noise plays no role, thereby accommodating the present results. However, this still suggests that the item noise assumption is unnecessary.

In context noise models such as BCDMEM (Dennis & Humphreys, 2001), a test word is used to cue previous contexts in which that word has been encountered. If one of the retrieved contexts matches the reinstated study context, a “yes” response will result. The greater the number of contexts in which an item has been seen, the greater the interference, and the poorer the recognition performance.

This happens regardless of the length of the study list, since other list items are not considered during retrieval. Thus, BCDMEM does not predict a list length effect and is consistent with the results of the experiments presented here.

Conclusions

In summary, the experiments presented in this article suggest that there is no list length effect as a consequence of interference in recognition memory. This finding is in contrast to those of the majority of previous studies, which have involved a manipulation of list length. However, the results are consistent with several examples in the literature in which a significant list length effect was not identified. It appears that previous list length findings have occurred as a consequence of a failure to control for a number of confounding variables. In particular, as Underwood (1978) suggested, it seems that attention is a critical factor. When we employed a retroactive design, in which differential

lapses in attention should not be an issue, we found no effect. When using a proactive design in which test items come from the end of the long list, where attention is most likely to wane, we found an effect. By contrast, the use of the RK procedure, which we anticipated might induce recall and, hence, a list length effect, seemed to have little impact on the results.

The failure to find a list length effect as a consequence of interference is consistent with context noise models of recognition memory but challenges the majority of existing mathematical models of recognition memory, which assume that the primary source of interference arises from the other items that appeared at study. Even if one is unwilling to accept that there is no item interference at all in recognition for words, the present experiments demonstrate that the contribution of this kind of interference is small.

References

- Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79, 97–123.
- Bowles, N. L., & Glanzer, M. (1983). An analysis of interference in recognition memory. *Memory & Cognition*, 11, 307–315.
- Buratto, L. G., & Lamberts, K. (2008). List strength effect without list length effect in recognition memory. *The Quarterly Journal of Experimental Psychology*, 61, 218–226.
- Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, 49, 231–248.
- Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, 3, 37–60.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Cohen, A. L., Rotello, C. M., & MacMillan, N. A. (2008). Evaluating models of remember-know judgments: Complexity, mimicry and discriminability. *Psychonomic Bulletin & Review*, 15(5), 906–926.
- Criss, A. H., & Malmberg, K. J. (2008). Evidence in favor of the early-phase elevated-attention hypothesis: The effects of letter frequency and object frequency. *Journal of Memory and Language*, 59, 331–345.
- Criss, A. H., & Shiffrin, R. M. (2004a). Context noise and item noise jointly determine recognition memory: A comment on Dennis and Humphreys (2001). *Psychological Review*, 111, 800–807.
- Criss, A. H., & Shiffrin, R. M. (2004b). Pairs do not suffer interference from other types of pairs or single items in associative recognition. *Memory & Cognition*, 32, 1284–1297.
- Dennis, S. (1995). *The Sydney Morning Herald word database*. Retrieved from <http://psy.uq.edu.au/CogPsych/Noetica>.
- Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108, 452–478.
- Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, 59, 361–376.
- Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, 13, 1–21.
- Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, 111, 524–542.
- Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological Review*, 62, 145–154.
- Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, 16, 309–313.
- Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 229–244). Oxford: Oxford University Press.
- Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, 91, 1–67.
- Glanc, G. A., & Greene, R. L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition*, 35, 365–371.
- 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. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, 93, 411–428.
- Hirshman, E., & Arndt, J. (1997). Discriminating alternative conceptions of false recognition: the cases of word concreteness and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1306–1323.
- Humphreys, M. S., Wiles, J., & Dennis, S. (1994). Towards a theory of human memory: Data structures and access processes. *The Behavioral and Brain Sciences*, 17, 655–666.
- Jang, Y., & Huber, D. E. (2008). Context retrieval and context change in free recall: Recalling from long-term memory drives list isolation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 112–127.
- Knowlton, B. J., & Squire, L. R. (1995). Remembering and knowing: Two different expressions of declarative memory. *Journal of Experimental Psychology: General*, 121, 699–710.
- McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective likelihood approach to the effects of experience in recognition memory. *Psychological Review*, 105, 724–760.
- Mensink, G. M., & Raaijmakers, J. G. W. (1989). A model for contextual fluctuation. *Journal of Mathematical Psychology*, 33, 172–186.
- Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, 89, 609–626.
- Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 689–697.
- Murnane, K., & Shiffrin, R. M. (1991). Interference and the representation of events in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 855–874.
- Pike, R. (1984). A comparison of convolution and matrix distributed memory systems. *Psychological Review*, 91, 281–294.
- Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 163–178.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237.
- Schulman, A. I. (1974). The declining course of recognition memory. *Memory & Cognition*, 2, 14–18.
- Shiffrin, R. M. (1970). Forgetting: Trace erosion or retrieval failure? *Science*, 168, 1601–1603.
- Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 179–195.
- Shiffrin, R. M., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1445–1449.

- Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4, 145–166.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117, 34–50.
- Strong, E. K., Jr. (1912). The effect of length of series upon recognition memory. *Psychological Review*, 19, 447–462.
- Tulving, E. (1985). How many memory systems are there? *The American Psychologist*, 40, 385–398.
- Underwood, B. J. (1978). Recognition memory as a function of length of study list. *Bulletin of the Psychonomic Society*, 12, 89–91.