Item Noise versus Context Noise: Using the List Length Effect to Investigate the Source of Interference in Recognition Memory

Angela Kinnell

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

School of Psychology

The University of Adelaide

August, 2009

Chapter 1

Interference in Recognition Memory

Recognition is the identification of a stimulus as having been encountered before. It is important in our everyday lives as it is the type of memory that allows us to identify people we have met before, restaurants we have eaten in and films we have seen, as well as many other things. In an experimental setting, the question is often not whether the stimulus has ever been seen before, but rather is constrained to whether it has been seen in a specific context, the study episode (Mandler, 1980).

The most common method for testing recognition memory in an experimental setting is the yes/no or old/new paradigm. Participants are given a list of items, usually words, to study. In the ensuing test session, participants are presented with more items, some of which have been seen before (targets) as well as new items (distractors). They are instructed to respond "yes" (or "old") to items that appeared in the preceding study list and "no" (or "new") to items that appeared for the first time during the test list (Snodgrass & Corwin, 1988). In this way, recognition can be distinguished from recall wherein the participant is required to retrieve and reproduce the stimulus encountered during study. Recognition requires only identification that the stimulus was seen in the study context, no generation is required.

Recognition memory is rarely perfect and many current theories propose that this forgetting is based on interference (Norman, Tepe, Nyhus & Curran, 2008). However, the source of this interference has been the subject of much research interest and indeed, current debate. The first, and most dominant approach, is that of item noise in which interference is

deemed to arise from the other items that appear on the study list. Another alternative is the context noise approach in which it is the previous contexts in which an item has been seen that is the source of interference. The item and context noise approaches, in their purest forms, are opposing ends of an interference continuum (Criss & Shiffrin, 2004a). It is also possible that interference arises from some combination of other items and previous contexts.

A real life example of the item noise approach to recognition would be a question such as "did you see Carole yesterday?". An item noise model of recognition memory would answer this question by first retrieving the names of all people who were seen in the context of yesterday. If one of those people retrieved was Carole, the response would be that yes, Carole was seen yesterday. Interference to this memory trace comes from all other people that were seen the previous day. In order to answer the same question, a context noise model would first retrieve all of the previous instances or contexts in which Carole was seen. If one of these previous occurrences was yesterday, then the question can be answered, yes, Carole was seen yesterday. Interference comes from all of the previous contexts in which Carole has been seen.

A critical distinction between the item and context noise approaches concerns their predictions regarding the list length effect in recognition memory. The list length effect refers to the finding that recognition performance for items from a short list is superior to that of items that were part of a long list at study. This finding has been well documented in the literature (e.g. Bowles & Glanzer, 1983; Cary & Reder, 2003; Gronlund & Elam, 1994; Murdock & Kahana, 1993a; Murnane & Shiffrin, 1991; Ratcliff, Clark & Shiffrin, 1990; Shiffrin, Ratcliff, Murnane & Nobel, 1993; Strong, 1912; Underwood, 1978). However, there have also been several studies that have not identified the list length effect (e.g. Dennis & Humphreys, 2001; Dennis, Lee & Kinnell, 2008; Jang & Huber, 2008; Murnane &

Shiffrin, 1991; Schulman, 1974). The debate about the existence of the list length effect will be discussed further in Chapter 3. However, it is important to note that the list length effect finding is consistent with the predictions of item noise models but is not accounted for by context noise models. This issue is the focus of the present thesis and the existence, or non-existence, of the list length effect will be used to help identify the source of interference in recognition memory.

1.1 Item Noise Models

The item noise approach is based upon the idea that it is the other items in the context of interest that interfere with one's ability to recognise a test probe. Most mathematical models of recognition memory are item noise models. These include global matching models (GMMs) such as the Theory of Distributed Associative Memory (TODAM; Gronlund & Elam, 1994; Murdock, 1982), Minerva II (Hintzman, 1984), the Matrix model (Pike, 1984) and Search of Associative Memory (SAM; Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981) as well as the Retrieving Effectively from Memory model (REM; Shiffrin & Steyvers, 1997) and the Subjective Likelihood Model (SLiM; McClelland & Chappell, 1998). These are some of the most often cited recognition memory models and have been applied to single item and associative recognition data, as well as recall, categorisation and serial order data (Clark & Gronlund, 1996).

Figure 1 illustrates the general format of item noise models in an experimental context. The study list context is first used to retrieve all items from that list to memory, in this case, “bear”, “table” and “car”. The test item, “bear” in the example, is then compared to each of the retrieved study items and the signals are combined. The stronger the resulting

overall signal, the more likely the participant is to give a "yes" response indicating that the test item was present on the study list. These models involve a global matching process meaning that all list items are involved in the recognition process and any interference is contributed by the other study list items. Some specific examples of item noise models and the mechanisms by which they produce interference will be provided next.

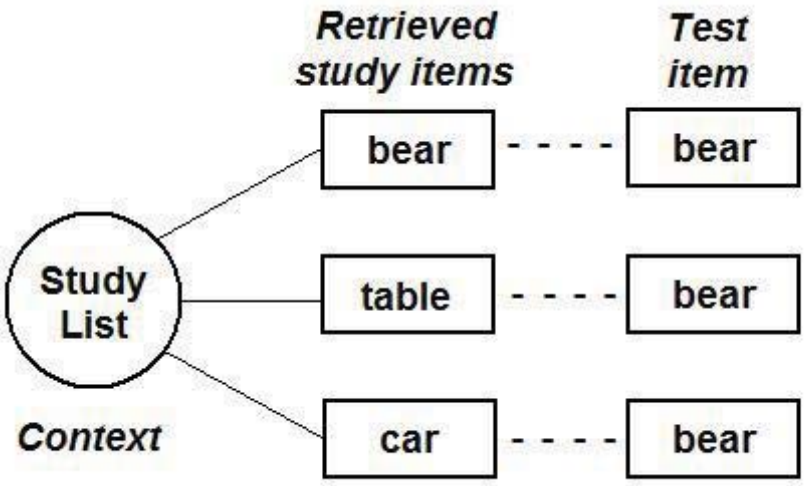

Figure 1. The item noise approach

1.1.1 Minerva II

Minerva II (Hintzman, 1984) is an example of an item noise model. It is a separate storage model in which events are stored in individual memory traces. All studied items are represented by a vector made up of a random sequence of -1s, 0s or 1s. The more completely encoded the memory trace, the fewer zeroes will be in the vector. At test, the vector for the test item is compared to each study item vector in memory by computing the dot product. The overall similarity of the test probe and an item in memory is given by:

$$S_i = \sum_{i=1}^{N} p_j T_{ij} / N_i$$

Where p_j is the *j*th feature of the test probe, T_{ij} is the counterpart of this feature in the *i*th item in memory and N_i is the total number of features for which both the test probe and the item in memory are nonzero. The activation level of each memory trace (A_i) is given by:

$$A_i = S_i^3$$

Finally, the familiarity value, or intensity of this activation for each test probe is calculated in a global matching fashion by summing the activation level for all items:

$$I = \sum_{i=1}^{M} A_{(i)}$$

Where *I* is the overall intensity of the trace and *M* is the number of traces in memory. If this summed level of intensity is above a certain threshold, the individual will respond "yes" indicating that they recognise the word from the study list. If this activation fails to reach the threshold, a "no" response will ensue.

Minerva II predicts a list length effect. A longer study list means that more items must be matched to the test cue. The difference between the means of the target and distractor distributions does not change, but each new item introduces more variability. This

additional variance in longer lists results in poorer performance and a significant list length effect that is predicted by Minerva II and the GMMs (Clark & Gronlund, 1996). Other study list items are necessarily a part of the decision process in that they are involved in the global matching process used to calculate the intensity of the trace. In Minerva II interference comes from the other list items.

In recent years, Minerva II and the GMMs have lost favour because they predict a significant effect of list strength on recognition performance while the data show otherwise. Ratcliff et al. (1990) found that recognition performance for weak items on a mixed list (both weak and strong items) was comparable to weak items in a pure (weak only) list. That is, the strengthening of other list items, either through repetition or increased study duration, did not have an impact on the recognition performance for weak items. They termed this effect the list strength effect, however, to avoid confusion and because it is a nonsignificant effect, it will be referred to here as the null list strength effect.

As with the list length effect, Minerva II and the GMMs predict a list strength effect as a consequence of variance. When an item in memory has been strengthened, the variance of the match between that item and the test probe increases as a result. Strengthening multiple items on a list (in a pure strong or mixed list) increases the variance of the global match causing an increase in the overlap of the target and distractor distributions and poorer recognition performance. The list strength effect is therefore predicted by Minerva II in that there is less variance in the pure weak list than in the mixed list which leads to superior performance for weak items in the former (Clark & Gronlund, 1996).

The studies of Ratcliff et al. (1990) which identified consistent null list strength effects are problematic for Minerva II and the GMMs. The problem arises in that the mechanism used to predict the list length effect necessarily predicts a positive list strength

effect also (Clark, 1999). Two of the GMMs, SAM and TODAM, were modified to enable them to predict the null list strength effect using differentiation and a continuous memory assumption respectively. The differentiation hypothesis assumes that as an item is strengthened, the trace for that item becomes more distinguishable from other items in memory. The memory trace itself is strengthened and no new traces are stored (Clark & Gronlund, 1996; Criss, 2006; Criss & McClelland, 2006). As its name suggests, the continuous memory assumption posits memory as a continuous series of events throughout a lifetime and is not confined to the studying of lists in the experimental session (Murdock & Kahana, 1993a; 1993b). Two assumptions are made, first, that memory is filled prior to the beginning of the experimental session and is not set to zero, and that the memory vector is not set to zero at the completion of a list or the experimental session, but that memory is continuous and ongoing (Clark & Gronlund, 1996; Murdock & Kahana, 1993a; 1993b).

Thus, the null list strength effect finding is not a consequence of variance differences between lists, since there are no lists, just continuous memory with approximately stable variance. However, it should also be noted that as the continuous memory assumption eliminates the list strength effect in recognition memory, in doing so, it also eliminates the list length effect, which is also based on variance. As Shiffrin et al. (1993) note, this is problematic for TODAM as well as the other GMMs. Minerva II has neither a differentiation hypothesis nor the continuous memory assumption and cannot predict the null list strength effect using its standard parameters (Clark & Gronlund, 1996).

1.1.2 Retrieving Effectively from Memory Model (REM)

In light of the problems faced by the GMMs regarding the null list strength effect finding, several new models were developed to incorporate a null list strength effect and a positive list length effect, as well as a positive word frequency effect. The word frequency effect is one of the most replicated findings in the recognition memory literature. This effect is evidenced in better recognition memory performance for words of low normative frequency than high frequency words and has been identified in a large number of studies (e.g. de Zubicaray, McMahon, Easton, Finnigan & Humphreys, 2005; Glanzer & Adams, 1990; Joordens & Hockley, 2000; Malmberg, Steyvers, Stephens & Shiffrin, 2002; Reder et al., 2000; Underwood & Freund, 1970; but see Criss & Malmberg, 2008; Criss & Shiffrin, 2004b; Glanc & Greene, 2007; Hirshman & Arndt, 1997). The word frequency effect is a mirror effect, so called as a result of the mirror pattern of that emerges in the hit and false alarm rates for low and high frequency words. The general finding is of a higher hit rate and lower false alarm rate for low frequency words (Glanzer, Adams, Iverson & Kim, 1993).

The REM model was developed by Shiffrin and Steyvers in 1997 and borrowed certain features from the GMMs. The major differences between REM and the GMMs, such as Minerva II, were REM's use of a likelihood based decision rule and its implementation of differentiation to account for the null list strength effect.

The REM model represents images in memory as vectors of feature values, zeroes represent features about which no information is known, while all positive numbers indicate some knowledge of the feature. The environmental base rate for each feature, *g*, also differs, for example, high frequency words have a higher base rate than do low frequency words.

When an item is studied, an incomplete copy of the study vector is stored in memory

as an episodic vector. There is a certain probability, *u*, that a value will be stored for each feature of a studied item per unit of time. If a value is stored then there is also a certain probability, *c*, that it is a correct copy of that feature that will be stored.

At test, a probe vector is compared to the stored episodic vectors for each item in the list. Each feature value is compared in parallel and either a match or a mismatch is recorded. A likelihood ratio is then calculated for the probability that the pattern of matches and mismatches would have occurred if the test item was a target rather than a distractor, while taking the environmental base rate into consideration. Bayes rule is used to determine the odds ratio (ϕ) that the test probe is a target versus a distractor. When the odds are greater than one it suggests that the item is a target and a "yes" response results based on the following equation:

$$\phi = \frac{1}{n}\sum_{j=1}^{n}\lambda_j$$

Where λ is the likelihood ratio calculated as follows:

$$\lambda_j = (1-c)^{n_{jq}} \prod_{i=1}^{\infty}\left[\frac{c+(1-c)g(1-g)^{i-1}}{g(1-g)^{i-1}}\right]^{n_{ijm}}$$

Where n_{jq} is the number of nonzero mismatching features in image *j* and n_{ijm} is the number of nonzero matching features in image *j*.

REM predicts a list length effect in recognition memory. As the length of the study list increases, there is a greater probability that the extra words will match with the test probe

by chance (Shiffrin & Steyvers, 1997). In addition, the calculation of the odds ratio (ϕ) involves dividing by the number of items on the list which results in a prediction of superior performance for shorter lists and the list length mirror effect – a higher hit rate and lower false alarm rate for items that were part of a short list at study. The source of interference in this process is the other study list items.

REM is able to produce the null list strength effect via the differentiation assumption that was introduced in SAM. The strengthening of some items on the study list means that more features are stored for these items in REM. The additional features stored help to make the strong items more identifiable. There are more matches for each strong study item and it is therefore less likely that any item other than a target would have a strong match with the test probe by chance. Strengthening an item increases the likelihood ratio for an image of the test item and decreases the ratio for images of all other items (Shiffrin & Steyvers, 1997), thus performance for weak items is unaffected by their presence on the mixed list.

REM is also able to produce the word frequency effect. High frequency words have higher frequency features than do low frequency words. The higher environmental base rate for high frequency words results in lower and more common feature values being stored for these items. Thus, the matching feature values will also be lower in value and less indicative that the test probe is a target which reduces the likelihood ratio. A lower likelihood ratio leads to more "no" responses, and a lower hit rate for high frequency words. When the test probe is a high frequency distractor, it is likely that some of its features will match by chance with features of items stored in memory. This chance matching would increase the likelihood ratio and result in more "yes" responses. More "yes" responses for distractors contributes to the higher false alarm rate for high rather than low frequency words. In combination, the word frequency effect is produced, with a lower hit rate and higher false alarm rate for high

frequency words (Shiffrin & Steyvers, 1997).

1.2 Context Noise Models

Alternatively, interference could arise from the other contexts in which an item has been encountered in the past (Dennis & Humphreys, 2001), with any interference from other items 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).

In the case of episodic memory, context can mean a number of different things. Dennis and Humphreys (2001) refer to two different types of context; processing context and temporal context. Processing context relates to the conditions surrounding the processing of particular items, such as rating the pleasantness of items or determining whether the item is a word or a nonword. In contrast, temporal context refers to the context at a specific time and is always changing, such that two contexts will be more similar the shorter the time between them.

Figure 2 provides an example of the general context noise process in an experimental setting. In this case, the test item, “bear”, is used to retrieve all previous contexts in which that item has previously been seen, for example, at the zoo, on the study list and in a bedroom. The stronger the global match between the retrieved contexts and the reinstated study context, the more likely it is that the participant will respond “yes”, indicating that the test item was present on the study list. Thus, item and context noise models differ in the way in which the test probe cues memory. Item noise models use the study list context to cue the retrieval of study list items, while context noise models use the test item to cue the retrieval

of previous contexts. In a similar manner to that in which item noise models involve a global match across all study items, context noise models involve a global match across all previous contexts that an item has been seen in. It is the previous contexts which are the source of noise, with more similar and more recently encountered contexts generating the most interference. BCDMEM will be discussed in greater detail as an example of a context noise model.

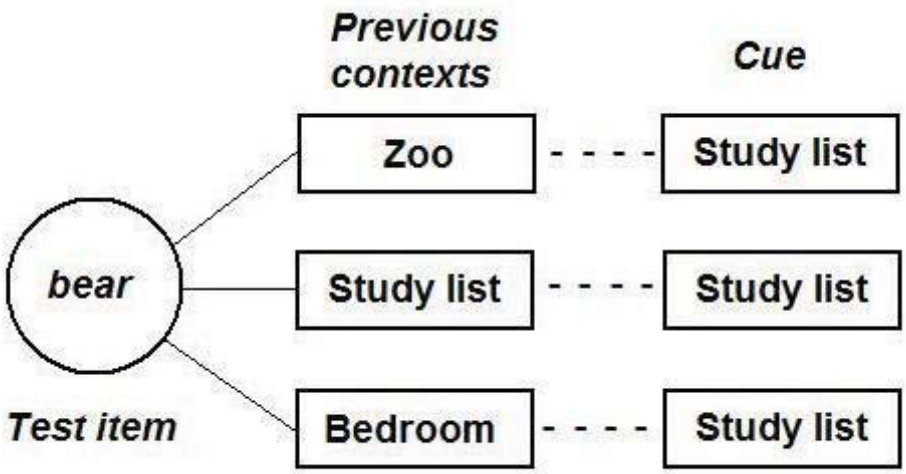

Figure 2. The context noise approach

1.2.1 Bind Cue Decide Model of Episodic Memory (BCDMEM)

In BCDMEM (Dennis & Humphreys, 2001), memory is made up of layers of nodes (see Figure 3). Some nodes are active (value = 1) and some are inactive (value = 0). A word is represented by a single node in the input layer. The output layer contains a pattern of activity unique to the current study context. Each node in the input layer is connected with varying weight (strength) to the output layer. When an item is studied, both the input and output layers are activated and the relative weights of the links between input layer and

output layer nodes are established with probability *r*, which is the rate of learning.

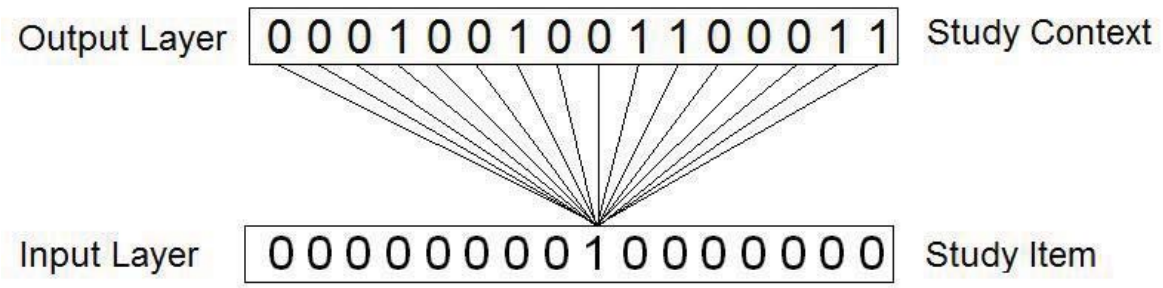

Figure 3. BCDMEM at study

At test, the probe is activated in the input layer (see Figure 4). This reinstated input layer node is then used to retrieve a composite vector at the ouput layer which is made up of all previous contexts in which the test item has been encountered. When the test item is a target, the output layer will include some nodes corresponding to the study context, as well as some nodes established in other contexts. When the test item is a distractor, any weights between the input and output layers are from previous contexts, although there may be some similarity between these previous contexts and the study context in which case some weights may be present between the test item and the study context. The strength of the weights between the input and output layers determines which nodes in the output layer are activated at test. The reinstated study context vector is matched to the composite vector that has been retrieved. The reinstated study context vector may be an imperfect reproduction of the original study context and errors can arise as a result. The more contexts, the more similar these contexts and the more recent these contexts, the more nodes will be activated and the greater the context noise (*p*) and risk of error.

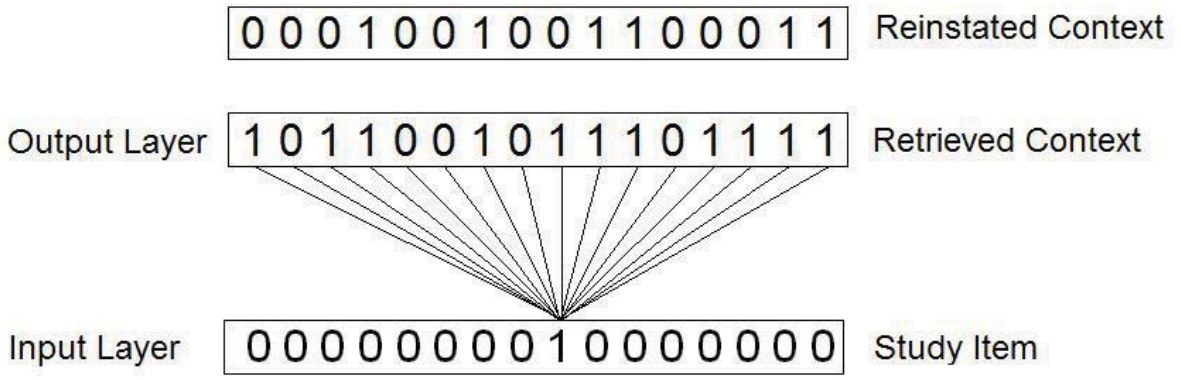

Figure 4. BCDMEM at test

Like REM, BCDMEM uses Bayes rule to calculate the odds ratio. An odds ratio greater than one will result in a "yes" response while an odds ratio less than one will result in a "no" response. The odds ratio is given by:

$$\frac{P(c_i' \text{ and } m_i|old)}{P(c_i' \text{ and } m_i|new)} = ([1-s+d(1-r)s]/[1-s+ds])^{n00}(1-r)^{n10} \\ ([p(1-s)+d(r+p-rp)s]/[p(1-s)+dps])^{n01} \\ ([(r+p-rp)/p])^{n11}$$

Where c'_i is the reinstated study context and m_i is the retrieved study context vector for the same item, *s* is sparsity and represents the probability that a feature value of the study context vector is a one rather than a zero and *d* is the probability that a feature in the study context vector that had a value of one will have a value of zero in the reinstated study context vector, that is, a measure of how well the study context is reinstated. *n00* refers to a match of zero feature values in the reinstated study context vector and the retrieved context vector

respectively, *n11* refers to a match of one feature values in the two vectors. There can also be two types of mismatches, *n10* and *n01*.

BCDMEM does not predict a list length effect in recognition memory. Other study list items, while involved in the likelihood calculation, do not affect performance as they do not necessitate a change to the model's parameters as a function of list length. The only exception to this is if there is a retention interval difference which might differentially compromise the participant's ability to reinstate the study context (see Dennis & Humphreys, 2001). The similarity of previous contexts to the reinstated study context is unaffected by the number of items present on the study list and thus there is no item noise and no difference in memory expected as a consequence of list length.

BCDMEM predicts a null list strength effect for the same reason it predicts a null list length effect. Other items on the study list, weak or strong, are not involved in the retrieval process and do not introduce interference. Weak item performance is therefore equivalent regardless of whether the items were presented in a pure weak or mixed list.

Finally, BCDMEM is also able to produce the word frequency effect. By definition, high frequency words have been encountered in more contexts than have low frequency words. There is therefore more interference (context noise) associated with high frequency words than low frequency words which leads to poorer performance for the former.

1.3 Combined Interference Models

The item and context noise approaches to recognition memory respectively state that either other items or other contexts are primarily responsible for the interference observed in recognition memory, with minimal interference from the other source. Nevertheless, it is

conceivable that interference arises through a combination of other items and other contexts. Criss and Shiffrin (2004a) modified the REM model (Shiffrin & Steyvers, 1997) to allow for interference from both other list items and previous contexts. In particular, they modified the decision mechanism and the way in which both the odds ratio and the likelihood ratio are calculated in REM. Targets have some item and context features in common with the test probe while a distractor can either share some context features or some item features with the test probe, or fail to match it on either.

Another model that involves interference from both other items and previous contexts is the Source of Activation Confusion model (SAC; Reder et al., 2000). According to this model, recognition may be based on two sources of information; recollection or familiarity. Memory is made up of interconnected nodes each with a pre-experimental level of activation. The level of pre-experimental activation decreases in line with a power function. When learning a study list in the recognition paradigm, the node corresponding to the studied word, the concept node, is activated and there is an increase in the current strength of the node. This level of activation falls away quickly with time and returns towards the pre-experimental level of activation.

In addition, when an item is studied, an event node is created and both this node and a context node, representing the study list, are connected to the concept node. When the word is tested, the concept and context nodes are activated and this activation spreads to all associated nodes in the network (see Figure 5).

SAC uses this spread of activation to give either "yes" responses based on recollection, or on familiarity. When the strength of activation in the event node is greater than the event threshold, a "yes" response will be given. Decisions of this kind are said to be based on recollection. Alternatively, in the absence of recollection, a "yes" response may

also be given when the activation of the concept node is greater than the concept threshold. This decision is based on familiarity.

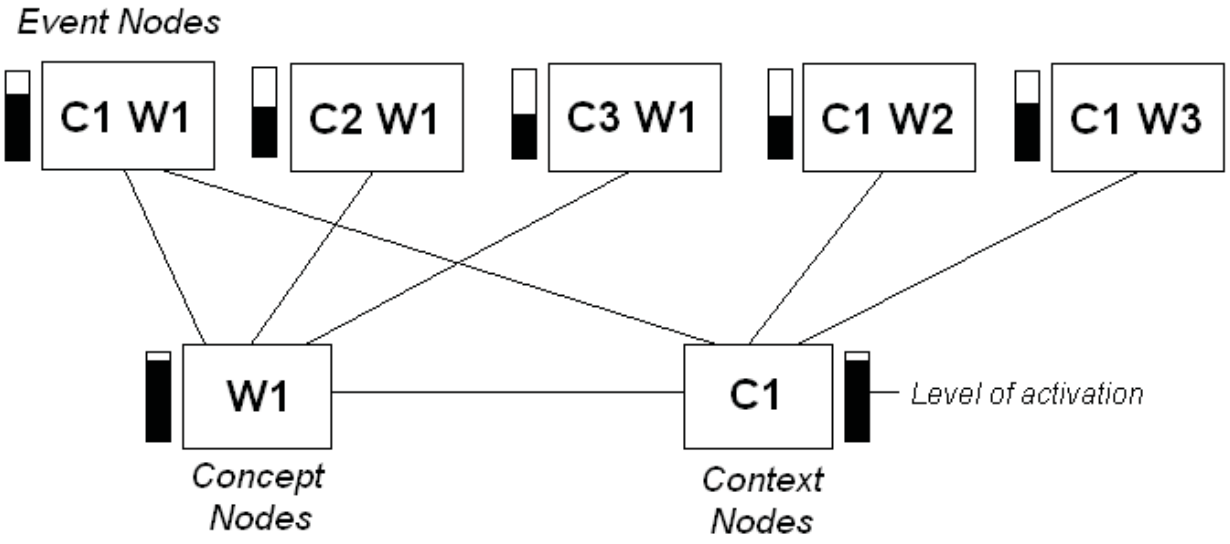

Figure 5. An illustration of the SAC model showing event, concept and context nodes. At test, activation spreads from the concept and context nodes to the event nodes. (Note that C1 refers to context one and W1 refers to word one).

Interference from previous contexts can also influence the recollection decision in SAC. This interference is generated through the connections between the concept node and the event nodes representing previous occasions on which the item has been encountered. The more contexts in which the item has been encountered, that is, the higher its frequency, the less activation will spread to the relevant study event node. This makes a "yes" response less likely and reduces performance regardless of other list items.

The majority of mathematical models of recognition memory are item noise models

which predict a list length effect in recognition memory. Some of the GMMs, namely SAM (Gillund & Shiffrin, 1984; Raaijmakers & Shiffrin, 1981) and TODAM (Gronlund & Elam, 1994; Murdock, 1982), as well as REM (Shiffrin & Steyvers, 1997) can accommodate the null list strength effect finding. Context noise models, like BCDMEM (Dennis & Humphreys, 2001) predict a null list strength effect and a null list length effect. An alternative to these models are models which involve interference from both other items and previous contexts. The present thesis attempts to differentiate between the item and context noise approaches to interference. The existence or non-existence of the list length effect can be used to help achieve this goal. Just as the null list strength effect findings resulted in a change to the mathematical models of recognition memory, so too would a nonsignificant list length effect finding. The list length effect will now be discussed in more detail.

Chapter 2

The List Length Effect in Recognition Memory

A critical difference between the item and context noise approaches is their predictions regarding the influence of the length of the study list on recognition performance. The aim of this chapter is to review the current evidence for this effect and the variables that may mediate it. Many studies which have manipulated the length of the study list have found that performance for items that were part of a short list at study is superior to that of items that were part of a longer list at study (e.g. Bowles & Glanzer, 1983; Cary & Reder, 2003; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991; Ohrt & Gronlund, 1999; Strong, 1912; Underwood, 1978). This is characterised by a higher hit rate, lower false alarm rate, and higher d' for the former. The list length effect finding is now well established and almost ubiquitously accepted in the literature to the extent that its existence has been termed a "touchstone for any model of recognition memory" (Gronlund & Elam, 1994, p.1355).

However, there are also a number of published studies which have reported a nonsignificant effect of list length (e.g. Dennis & Humphreys, 2001; Dennis et al., 2008; Jang & Huber, 2008; Murnane & Shiffrin, 1991; Schulman, 1974). In order to reconcile the apparent inconsistency in the literature, Dennis and Humphreys (2001) argued that previous studies which had identified the list length effect had failed to control for four variables that may be confounded with list length and have produced the effect. These variables were retention interval, attention, displaced rehearsal and contextual reinstatement. Retention interval refers to the amount of time elapsed between an item being studied and then the testing of that item. Longer lists will have a longer average retention interval. The amount of

attention paid by a participant to the stimuli may vary between lists, with lapses more likely in long lists. Problems with displaced rehearsal arise as a result of implementing controls for retention interval. It may be possible for participants to rehearse items from the short list prior to test where this is not possible following the long list. Finally, contextual reinstatement can be an issue, again as a consequence of retention interval controls. It may be that prior to the test list, participants reinstate the study context following the short list but that this is not deemed to be necessary following the long list. An investigation of these potential confounds may help resolve the inconsistent findings reported in the literature and these confounds will now be discussed in greater detail.

2.1 Potential Confounds of the List Length Effect

A survey of the literature revealed that there have been 26 conditions/experiments in 16 published papers that have manipulated the length of the study list. Of these experiments, 17 have identified a significant effect of list length on recognition performance, while the remainder reported nonsignificant effects. The results of these experiments on recognition performance are summarised in Table 1. It should be noted that this table is not an attempt at a meta-analysis, but rather, is used to summarise the literature and illustrate which of the four potential confounds were controlled in each experiment and whether a significant effect of list length was identified. While a meta-analysis would be useful in this case, the vast majority of studies have not reported effect sizes or the standard error or a standard deviation which could be used to calculate the effect size, making this impossible (such studies include Bowles & Glanzer, 1983; Dennis & Humphreys, 2001; Jang & Huber, 2008; Ratcliff & Murdock, 1976; Schulman, 1974; Strong, 1912; Underwood, 1978). More specific detail

Table 1

Summary of list length experiments indicating which of the four possible confounds were controlled and whether or not a significant effect of list length was identified.

Study	Retention Interval (Retroactive or Proactive)	Attention	Displaced Rehearsal	Duration of Contextual Reinstatement filler (seconds)	Significant Effect of List Length?
Strong (1912)	-	-	*	-	Yes
Schulman (1974)	**Retro**	-	✔	**100s**	No
Ratcliff & Murdock (1976) Expt 3	-	-	*	-	Yes
Underwood (1978)	**Pro**	-	✔	-	Yes
Bowles & Glanzer (1983)	**Retro & Pro**	-	✔	-	Yes+
Murnane & Shiffrin (1991) Expt 1	**Pro**	-	*	**30s**	Yes
Murnane & Shiffrin (1991) Expt 2	**Pro**	-	*	**30s**	Yes
Murnane & Shiffrin (1991) Expt 3	**Retro**	-	✔	**30s**	No
Murnane & Shiffrin (1991) Expt 4a	**Retro**	-	✔	**30s**	Yes
Murnane & Shiffrin (1991) Expt 4b	**Pro**	-	*	**30s**	Yes
Gronlund & Elam (1994) Expt 1 & 2	**Retro**	-	✔	**9s**	Yes
Clark & Hori (1995) Expt 1	-	-	*	**45s**	Yes
Ohrt & Gronlund (1999) Expt 1	**Pro**	-	*	**9s**	Yes
Nobel & Shiffrin (2001) Expt 1	-	-	*	**26s**	Yes
Nobel & Shiffrin (2001) Expt 3	-	-	*	**26s**	Yes
Dennis & Humphreys (2001) Expt 1a	**Retro**	✔	✔	**480s**	No
Dennis & Humphreys (2001) Expt 1b	**Pro**	✔	✔	**480s**	No
Dennis & Humphreys (2001) Expt 2	**Retro**	✔	✔	**240s**	No
Cary & Reder (2003) Expt 1	-	-	*	-	Yes
Cary & Reder (2003) Expt 2	-	-	*	**300s**	Yes
Cary & Reder (2003) Expt 3a & 3b	**Retro & Pro**	✔	✔	**120s**	Yes+
Criss & Shiffrin (2004c) Expt 1	-	✔	*	-	No
Buratto & Lamberts (2008)	**Retro**	✔	✔	-	No
Jang & Huber (2008) Expt 2	-	-	✔	✔	No
Dennis et al. (2008) Expt 1a	**Retro**	✔	✔	-	Yes
Dennis et al. (2008) Expt 1b	**Retro**	✔	✔	**480s**	No

✔ Controls for this confound were implemented.

- No controls for this confound were implemented.

* No controls for this confound were implemented. However, the control was unnecessary in this case because either retention interval was not controlled or the retroactive design was not used.

+ Results were reported collapsing across the retroactive and proactive designs.

regarding the studies contained in this table will be given below. It is clear from the table that prior to the publication of Dennis and Humphreys' work in 2001, no study had controlled for all four of the potential confounds that they outlined. Each of these confounds will now be discussed.

2.1.1 Retention Interval

The first potential confound suggested by Dennis and Humphreys (2001) was 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 viewing short list items, meaning that there is a longer retention interval for long list items.

The confounding effect of retention interval can be controlled by equating the average retention interval of the short and the long lists using either a retroactive or proactive experimental design. In the retroactive design, the short list is followed by a period of filler activity such that the duration of the short list and the filler activity combined is equal to the duration of the long list. Only the words at the beginning of the long lists (the same number as is in the short list) are included in the test as targets so that the target words 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 words at the end of the long list are tested in order to equate the retention interval between lists. Both designs are illustrated in Figure 6.

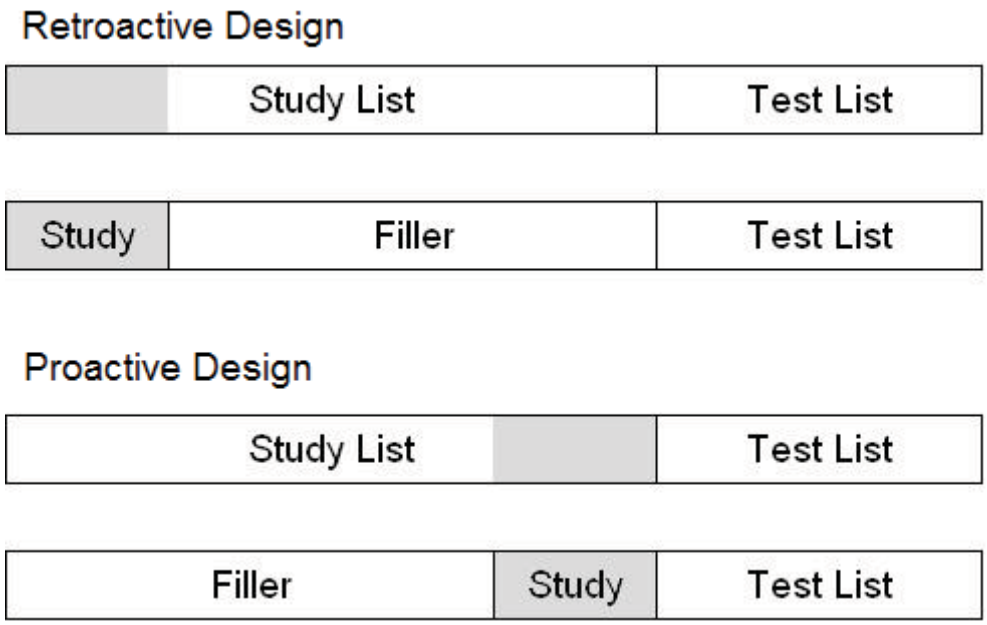

Figure 6. Retroactive and proactive experimental designs. Grey shading indicates the portion of the study list from which targets are obtained.

As shown in Table 1, 17 of the 26 list length conditions/experiments included controls for retention interval. Of these experiments, nine used the retroactive design, six used the proactive design and two used both designs between subjects but collapsed the results of both together in the analysis.

It should be noted that retention interval can be differentiated from study-test lag which may also confound the list length effect (Murdock & Kahana, 1993b). Where retention interval refers to the intervening time period between an item appearing at study and at test, study-test lag refers to the number of intervening items between a particular item appearing at study and test (Murdock & Kahana, 1993b; Waugh & Norman, 1965). Thus, it is possible to control for both retention interval and study-test lag, as was done by Ohrt and Gronlund (1999). The use of the proactive design is in itself a control for study-test lag in

that the number of intervening items between the study of an item on either a short or long list and the testing of that item are, on average, equal. However, in the retroactive design, the study-test lag is greater for the long list items that will be tested (i.e. those that were presented at the beginning of the study list) than for short list items. Thus, study-test lag could be an additional confound of the list length effect.

2.1.2 Attention

Underwood (1978) was first to suggest that differential lapses in attention between lists of different lengths may be the cause of observed differences in recognition performance. It is likely that participants tire as they view the continuous presentation of items in a study list and this is likely to occur to a greater extent when viewing a long list. In the recognition paradigm, particularly when the proactive design is used, the differences in attention between lists can become more pronounced (Cary & Reder, 2003; Dennis & Humphreys, 2001). In the proactive design the final words of the long list are compared to performance on the short list which comes after a period of filler activity. The performance comparison in this case is made between the short list, where attention is likely to be paid to each item, and the end of the long list where attention is waning. This comparison is biased to favour performance on the short list and may produce a list length effect. This is not problematic in the retroactive design as all targets appear at the beginning of the respective study lists where differences in attention would be minimal.

The confounding effect of differential lapses in attention can be reduced by having participants perform an encoding task that requires a response during study (Cary & Reder, 2003; Dennis & Humphreys, 2001). The inclusion of this task allows for the assumption that

all items will have been processed to some level, regardless of fatigue. Another approach is to split the long list into separated blocks with the aim being to maintain attention within each block at the same level as in the short list (Dennis & Humphreys, 2001). It should also be considered that it may be impossible to completely eliminate attentional lapses in the proactive condition. Despite the concerns raised by Underwood (1978), no list length experiment included explicit controls for attention prior to those of Dennis and Humphreys (2001).

2.1.3 Displaced Rehearsal

Displaced rehearsal may also confound the list length effect finding in two primary ways. First, when retention interval is controlled and only some long list items are tested while all short list items are included as targets, any rehearsal of short list items will be beneficial to performance. There is no such guarantee with rehearsal of long list items, as not all of the studied and rehearsed items will be tested. A related issue is that when there is rehearsal, short list items will be rehearsed, on average, more than long list items because they are fewer in number. This would favour performance for the short list and could result in a list length effect finding (Cary & Reder, 2003; Dennis & Humphreys, 2001).

Second, the issue of displaced rehearsal is exacerbated when the retroactive design is used and a period of filler activity follows the short list. This period may provide an opportunity for the rehearsal of short list items. In the long list, words are continually presented leaving no time for rehearsal. This would again favour performance on the short list and give rise to the list length effect.

The effects of displaced rehearsal can be controlled in a number of ways. The use of

the proactive design is the primary way in which this can be done as it limits the potential for rehearsal of items from either list. However, if the retroactive design is used, 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 study items (Cary & Reder, 2003; Dennis & Humphreys, 2001). The nature of this filler task has varied between experiments and has included doing arithmetic (Murnane & Shiffrin, 1991, Experiment 3), solving algebraic equations (Cary & Reder, 2003, Experiment 3), performing addition (Gronlund & Elam, 1994), playing a video game task (Burrato & Lamberts, 2008) or solving puzzles (Dennis & Humphreys, 2001, Experiment 2). A second strategy is to include the recognition test as incidental (Dennis & Humphreys, 2001), although this is problematic when using a within subjects design as both tests cannot be incidental (Cary & Reder, 2003). Finally, the effect of displaced rehearsal on the results can be minimised by either testing a subset of study words of both lists (Underwood, 1978) or analysing the results of only those targets that appeared at the beginning of both the short and long list study blocks (Cary & Reder, 2003, Experiment 3). That is, analysing only half as many targets as short list words and the equivalent number from the long list. In that situation there is a possibility that, for both lists, participants would rehearse some items that would not be tested later. In addition, it is likely to be the items at the end of the short list that would be rehearsed. Analysis of only the first block of targets would limit the differential effects of rehearsal (Cary & Reder, 2003).

It should be noted that it is difficult to ascertain how well these controls for displaced rehearsal in the retroactive condition are working. These are suggestions based on prior research and attempt to limit rehearsal after the short list to the same extent that the continuing presentation of new words in the long list also prevents rehearsal of the critical

starting items of that list. However, it is assumed that the use of an engaging filler task should mitigate the impact of displaced rehearsal.

Table 1 shows that 14 of the list length conditions/experiments included some control for displaced rehearsal. The remaining 12 experiments did not require the control to be implemented since they either included no control for retention interval or used the proactive design.

2.1.4 Contextual Reinstatement

Reinstating the study context at test is important in both item and context noise models of recognition memory. In item noise models, the test item acts as a cue to reinstate the study list context which is then used to retrieve all items that appeared on the study list from memory. To the extent that the study context is accurately reinstated there will be better retrieval of items which would lead to better recognition performance. In context noise models, the test word cues the retrieval of all previous contexts in which the item has been seen. The task instructions ask the participant if the test probe was present on the study list and thus cue the reinstatement of the study context (Humphreys, Bain & Burt, 1989). The retrieved contexts are then compared to this reinstated study context and the stronger the global match between them, the higher the level of activation with the result being a "yes" response. If the reinstated study context is an inaccurate representation of the actual study context, there will be reduced activation and poorer recognition performance. In addition, if the reinstated context is diffuse and matches multiple previous contexts, then it may spuriously match a distractor.

However, reinstating the study context at test is not straightforward, as temporal

context varies with the passing of time (Estes, 1955; Howard & Kahana, 2002; McKenzie & Tiberghien, 2004; Mensink & Raaijmakers, 1988; 1989) and this can have an impact on the reinstated study context. When the test list immediately follows the study list, there may be a tendency for participants to cue with the current end of list context with which only the final study items are associated (Dennis & Humphreys, 2001). The best recognition performance will come when the study context as a whole is reinstated at test, rather than a reliance on the current end of list context. Given that there is more scope for variability of the study context in the long list than there is in the short list, relying on the end of list context could lead to a list length effect finding, with better performance on the short list where the study context is unlikely to have changed significantly throughout the list.

When retention interval controls are implemented, contextual reinstatement has more potential to confound the list length effect finding. This is particularly so in the retroactive design when a period of filler activity follows the short but not the long list. That is, there is no break at the end of the long list before the beginning of the test list. In this situation, participants may rely on an end of list context for the long list rather than reinstating the entire list context. Having just been engaged in the study list context, participants may decide that they already have the study context available to them and are likely to rely on this rather than go through the process of reinstating the entire list context. This end of list context is likely to differ from the start of list context. In the retroactive condition, it is the first words of the long list that are tested. As such, reinstating the end of list context is unlikely to benefit performance for early words. In the case of the short list, the presence of the filler activity prior to the test list means that the participant has just been engaged in a filler context rather than a study context (Dennis & Humphreys, 2001). At test, there is no end of list context to rely upon and it is necessary to reinstate the entire study list context.

Thus, it is likely that a test of short list items involves comparison with the entire reinstated study list while a test of long list items involves comparison with an end of list context. Given that it is the items at the beginning of the long list that are tested in the retroactive design, this favours performance on the short list.

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 (Dennis & Humphreys, 2001). This filler is in addition to the puzzle activity following the short list as a control for retention interval (see Figure 7). Including this filler ensures that contextual reinstatement of the entire study context is encouraged following both the long and the short lists. Eighteen of the list length conditions/experiments included some period of filler after both lists which could be taken as a control for contextual reinstatement, although the duration of this filler activity has varied greatly from nine seconds (Gronlund & Elam, 1994) to eight minutes (Dennis & Humphreys, 2001, Experiment 1).

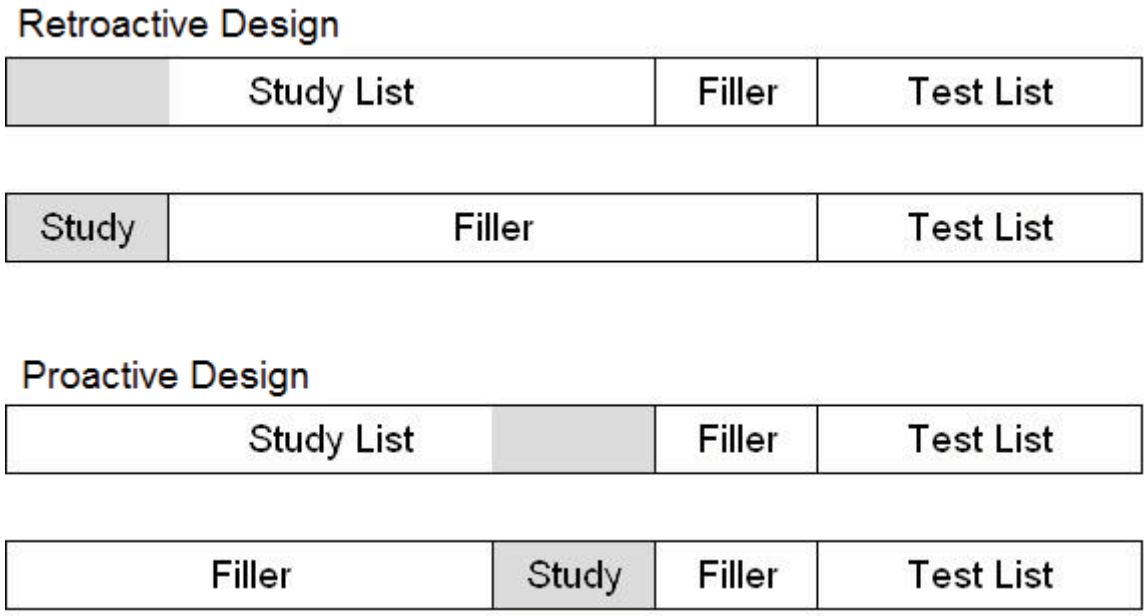

Figure 7. Experimental design showing the filler as a control for retention interval (in both retroactive and proactive designs) and the inclusion of additional filler as a control for contextual reinstatement. Grey shading indicates the portion of the study list from which the targets are obtained.

The preceding discussion has highlighted that the use of the retroactive or proactive design as a control for retention interval is influential on the effects of the other potential confounds of the list length effect. The retroactive design controls for the effect of differences in the retention interval and differential lapses in attention between the short and long lists (see Table 2). However, using this design also means that displaced rehearsal, contextual reinstatement and study-test lag are able to contribute to differences between the short and long lists. The proactive design also controls for retention interval while preventing differential effects of displaced rehearsal, contextual reinstatement and study-test lag based on list length. The only potential confound which the proactive design fails to control for is

differential lapses in attention which are likely to be more profound when this design is used (Underwood, 1978).

Table 2

The influence of the retroactive and proactive retention interval controls on the other potential list length effect confounds

	Retention Interval	Study-Test Lag	Attention	Displaced Rehearsal	Contextual Reinstatement
Retroactive	✔	-	✔	-	-
Proactive	✔	✔	-	✔	✔

✔ This potential confound is controlled when this design is used..

- This potential confound is not controlled when this design is used.

On this basis, the proactive design is preferable to the retroactive design in terms of eliminating most potential confounds of the list length effect. However, returning to Table 1, of the nine experiments that did not identify a significant list length effect, six used the retroactive design. In all but one of these cases (Buratto & Lamberts, 2008), controls were also implemented for displaced rehearsal and contextual reinstatement. Only one condition which used the proactive design did not identify a significant effect of list length (Dennis & Humphreys, 2001, Experiment 1), and this condition also included controls for the other three potential confounds. In addition, of the ten experiments that did identify a significant list length effect and controlled for retention interval, three used the retroactive design, five used the proactive design and two combined the results from each design together in the analysis. If the proactive design provided better control for potential confounds of the list length effect it would be anticipated that the majority of experiments which failed to identify the effect

would have used the proactive design. The fact that this is not the case suggests that the attention confound is a much more significant problem than the other confounds.

The previous discussion has highlighted that, as argued by Dennis and Humphreys (2001), retention interval, attention, contextual reinstatement and displaced rehearsal have the potential to confound the list length effect finding if not controlled, as does study-test lag (Murdock & Kahana, 1993b). In addition, the distinction between the retroactive and proactive designs as controls for retention interval may mediate the impact of some of the other potential confounds, particularly the influence of attention. Specific details of previous list length experiments will now be discussed with reference to the controls implemented for each of these potential confounds.

2.2 List Length Studies

In the early 1900s, Strong (1912) undertook what was the first experiment designed to look explicitly at the role of list length on recognition performance. Participants were presented with four series of advertisements of varying lengths. The number of advertisements in each series was either 5, 10, 25, 50, 100 or 150 items. Participants made their recognition decision at test by sorting the advertisements into four piles corresponding to a confidence scale. There were piles for 100 per cent, 75 per cent and 50 per cent confidence that the advertisement was seen at study. The fourth pile was for those advertisements deemed to be new. No controls were implemented for the four potential confounds highlighted by Dennis and Humphreys (2001). Strong analysed the percentage of correct and incorrect recognition judgements and found a decrease in the former and an increase in the latter as the length of the series increased. That is, recognition performance

was poorer when the series involved more advertisements; a list length effect.

Underwood (1978) compared recognition performance across four list lengths; 24, 40, 60 and 80 words. The last 20 words of each list were tested in a forced choice recognition paradigm. Thus, the proactive design was used. The forced choice recognition task involves the simultaneous presentation of both a target and a distractor at test with the participants required to select the target. The number of errors increased significantly as the length of the study list increased, leading Underwood to suggest that differential lapses in attention may be the cause. This hypothesis is given further weight by the fact that the use of the proactive design controls for the other potential confounds listed by Dennis and Humphreys (2001).

In the study of Bowles & Glanzer (1983), performance on a list of 120 words was compared in a between subjects design with a list of 240 words. They used both the retroactive and the proactive design to control for retention interval. In the retroactive design, the 120 words were followed with the presentation of 120 numbers, as part of a counting task, to equate the retention interval of the short and long lists. Using a two alternative forced choice recognition test, a statistically significant effect of list length was identified on recognition accuracy. Similarly, there was a significant effect of list length on the mean response latency for correct responses. The effect on incorrect responses was not significant, however this may be attributed to the fact that this analysis involved only a small number of observations. In addition, the results from the retroactive and proactive designs were not reported separately meaning that differences between the two designs could result in the significant list length effect findings.

There was renewed interest in the list length effect in the early 1990s. The null list strength effect finding of Ratcliff et al. in 1990 was not consistent with the predictions of the existing mathematical models of memory. There was an effort to modify some of these

models to accommodate the null list strength effect and list length manipulations were often included in experiments that also looked at the effect of list strength.

Murnane & Shiffrin (1991) carried out a series of experiments designed to investigate the list strength effect with a list length manipulation also included. Each experiment included both in session testing and end of session testing. During in session testing, participants were presented with single words and asked to respond "old" or "new" with regard to whether the test word had appeared on the previous list. In end of session testing, the "old" or "new" response was given in relation to whether the test word had appeared on any of the study lists. It is the in session testing which is the comparison of interest for the list length effect. The lists of interest in terms of the length comparison were the pure weak list (10 sentences) and the long list (150 words presented as 30 sentences). Experiments 1 and 2 differed only in the blocking of items at study. They included a 30 second control for contextual reinstatement only. Performance on the pure weak list was then compared to performance on the entire long list, performance on the first part of the long list and performance on the last part of the long list. In both experiments, a significant effect of list length was found between the pure weak list and each of the long list comparisons. When the comparison was between the short list and the end of the long list, retention interval was effectively controlled using the proactive design.

Experiment 3 differed from the first two in that the retention interval was equated for the long list and the short list with the retroactive design. Results indicated no significant effect of list length when performance on the short list was compared to the equivalent part of the long list (a comparison as in the retroactive design) and similarly when compared to those items that were studied at the end of the long list. However, there was a significant list length effect for the comparison between the pure weak (short) list and the entire long list. Their

fourth experiment included both the retroactive and proactive designs as a within experiment manipulation. There was a list length effect both when the retention interval was equated and when it was not. In this case, the differing results may be attributable to insufficient controls for the potential confounds, particularly contextual reinstatement which was controlled using a 30 second period of arithmetic. As we shall see, this may not be of adequate duration to ensure that participants reinstate the study context after both the long and short lists (Dennis et al., 2008).

Gronlund & Elam (1994) conducted two studies involving lists of 10 and 82 items. The first experiment was a multi-list design with each participant studying between 64 and 112 lists (half of each length) over several days. In the second experiment participants studied only one list, either short or long. Both experiments used the retroactive design and used an addition task as the filler activity following the short list and to control for displaced rehearsal. A further nine seconds of addition were included to control for contextual reinstatement although, as with Murnane and Shiffrin's experiments, the duration of this task may not have been sufficient to encourage contextual reinstatement. The responses for each test list were obtained from a six-point confidence rating scale. In each experiment, a receiver operating characteristic (ROC) analysis revealed a list length effect. When confidence ratings are provided at test, it allows for the construction of an ROC curve to be constructed by plotting the hit rate against the false alarm rate for each level of confidence (Wixted, 2007).

Clark and Hori (1995) carried out a forced choice associative recognition experiment based on the method of Clark, Hori and Callan (1993). Associative recognition involves the presentation of pairs of items at study (e.g. AB, CD, EF). At test, participants are presented with a combination of targets, which are intact pairs as presented on the study list (e.g. AB),

and distractors, which are rearranged pairs (e.g. CF). Some studies also include pairs of never presented items at test (e.g. GH). The correct response for intact pairs is "yes" and a "no" response is correct for rearranged and new pairs In Clark and Hori's between subjects design, participants studied either a short list (34 word pairs) or a long list (100 pairs). The test condition was also manipulated between subjects. Half of the participants were in the overlap (OLAP) condition in which the intact and rearranged pairs shared a common word and half were in a non-overlapping (NOLAP) condition in which the items that made up the intact and rearranged pairs were unique. Participants were instructed to form an associative link between the two items presented together at study, however this was not enforced and no responses were recorded. A 45 second period of mental arithmetic was included before the onset of the test list to control for contextual reinstatement. No controls were in place for any of the other potential confounds. Clark and Hori found a significant effect of list length on recognition performance, with the magnitude of the effect greater in the NOLAP condition.

Ohrt & Gronlund (1999) compared performance on 12 and 82 word lists with each participant viewing multiple lists. The proactive design was used with a nine second period of arithmetic as a control for contextual reinstatement. Half of the lists were followed by a yes/no recognition task and the remainder were followed by a free recall test. A significant effect of list length was identified in both test types and again, the period allocated to encourage contextual reinstatement after both lists may have been insufficient.

Three associative recognition experiments, two of which contained list length manipulations, were carried out by Nobel and Shiffrin (2001). Experiment 1 involved lists of 10 and 40 pairs of words in a within subjects design, while lists in Experiment 3 were comprised of 10 and 20 pairs of words. There was a 26 second period in which participants completed an arithmetic task prior to each test list which could act as a control for contextual

reinstatement but again the duration of this task is unlikely to have adequately controlled for this potential confound (Dennis et al., 2008). In Experiment 1, test lists were either free response (participants were free to respond at their leisure) or used the signal-to-respond procedure of Reed (1973) with ten lags varying from 100ms to 4500ms. In each case one word from a presented pair was presented as the test probe. In both conditions there was a significant list length effect in the accuracy data and in the free response condition there was also a significant effect of list length on the mean response time for hits and the median response time for false alarms. In Experiment 3, one condition involved an associative recognition task involving intact and rearranged pairs and the other was a cued recall task. A significant effect of list length was identified on both d' and the hit rate in the associative recognition condition.

Cary and Reder (2003) conducted three studies which investigated the list length effect. The first experiment compared performance across 16, 32, 48 and 64 word lists. No controls were implemented and a significant list length effect was identified. In their second experiment, Cary and Reder (2003) sought to reduce recognition performance. The experiment was as the first with the exception of reduced stimulus presentation time and the introduction of a five minute word search puzzle before each test list. Thus, there was a control for contextual reinstatement. A significant list length effect was identified. Experiment 3 was a partial replication of Dennis and Humphreys' (2001) first experiment. Attempts were made to control for all four potential confounds. Both the retroactive and proactive designs were used within subjects and a pleasantness rating task was used to control for attention. Displaced rehearsal was controlled by introducing an algebra filler task and analysing the data from the first block of targets presented at test for each list. Finally, a two minute period of filler task was included before each test list to control for contextual

reinstatement. Again, a significant list length effect was identified with the data from the retroactive and proactive designs collapsed together in the analysis.

2.2.1 The Effect of List Length on Response Latency

While the majority of list length studies have analysed only the accuracy data, others have also investigated the effect of list length on response latency and some have analysed both. The analysis of response times is important to consider in that there may be a trade off between the accuracy of the response and the speed of the response. This can be manipulated experimentally such that participants are instructed to either respond accurately, and as such take their time to respond, or to respond as quickly as possible regardless of the accuracy of their response (e.g. Mulligan & Hirshman, 1995). In the absence of these instructions, as is the case in most studies, participants may vary in terms of whether they favour accurate or fast responses. All previous studies which have looked at the effect of list length on response latency have identified a significant effect, but very few potential confounds were controlled (see Table 3). The present thesis will consider the effect of list length on both recognition accuracy and response latency.

Table 3

Summary of list length experiments indicating which of the four possible confounds were controlled and whether or not a significant effect of list length was identified (in response latency data).

Study	Retention Interval Control (Retroactive or Proactive)	Attention	Displaced Rehearsal	Duration of Contextual Reinstatement Filler (seconds)	Significant Effect of List Length?
Sternberg (1966)	*	*	*	-	Yes
Reed (1976)	*	*	*	-	Yes
Juola et al. (1971) Expt 1	-	-	*	-	Yes
Atkinson & Juola (1973; 1974)	-	-	*	-	Yes
Ratcliff & Murdock (1976) Expt 3	-	-	*	-	Yes
Monsell (1978) Expt 1 & 2	*	*	*		Yes
Bowles & Glanzer (1983)	**Retro & Pro**	-	✔	-	Yes[+]
Nobel & Shiffrin (2001) Expt 1	-	-	*	**26s**	Yes

✔ Controls for this confound were implemented.

- No controls for this confound were implemented.

* No controls for this confound were implemented. However, the control was unnecessary in this case.

+ Results were reported collapsing across the retroactive and proactive designs.

Sternberg (1966) looked at recognition performance for sequences of symbols ranging in length from one to six. He found that the mean response latency increased as a function of the number of items on the list. This finding has come to be termed the Sternberg effect (Reed, 1976). The method used by Sternberg was also very different to that of most of the other list length studies in that the stimuli were few in number, participants were given the stimuli to learn prior to arriving for the experimental session and there was only one test probe per sequence. It is possible that the task in this experiment is solved in a different way to that of the other list length studies discussed previously. Nevertheless, significant effects

of list length have been identified using this paradigm and it could be included in the same model frameworks as the previous experiments. Contextual reinstatement is arguably the only potential confound of Dennis and Humphreys (2001) that is relevant in this type of design given the short duration of all lists.

Reed (1976) carried out a similar study to Sternberg (1966) involving a series of one, two or four confusable English consonants. At test, participants were instructed to respond as quickly as possible after a response signal was given. The onset of the response signal was manipulated between seven lags ranging from seven milliseconds to 4131 milliseconds, in an experimental design developed by Reed (1973). Response time was found to increase with list length in accordance with the Sternberg effect. Again, the majority of controls for potential confounds were not relevant in this case and no controls were implemented.

In the early 1970s, Atkinson, Juola and colleagues (see Atkinson and Juola, 1973; 1974; Juola, Fischler, Wood & Atkinson, 1971) conducted two similar experiments. Juola et al.'s Experiment 1 involved a between subjects manipulation of list length which was varied between 10, 18 or 26 words, while the study of Atkinson and Juola (1973; 1974) involved lists of 16, 24 and 32 words. No controls for the potential confounds of Dennis and Humphreys (2001) were implemented. In all experiments it was found that response latency increased as the length of the study list increased. The accuracy data was not analysed other than to state that the percentage of errors was similar across list lengths. There were several major differences between these experiments and other list length studies reported here. First, the study list was given to participants to memorise 18-24 hours prior to the experimental session, meaning that the test items were better memorised on the whole than in other list length studies. Further, participants were asked to recall the list in order prior to the recognition test and some recognition test items were presented more than once.

Ratcliff & Murdock (1976) conducted a series of experiments comparing recognition accuracy and latency. The third experiment in the series included a list length manipulation. Participants were presented with multiple lists of length 4, 8, 16, 32, or 64 words. No controls for the potential confounds were implemented. Ratcliff and Murdock analysed the data from the 16, 32 and 64 word lists and found that the mean response latency for items from longer lists was longer than for items from shorter lists. Significance tests for the accuracy data were not reported; however there were fewer hits and correct rejections in longer lists. That is, poorer recognition performance and a list length effect.

Monsell (1978) carried out two experiments which used similar methodology to Sternberg (1966). Participants studied lists of individual letters with the set size (list length) varied between one and four in Experiment 1 and between two and five in Experiment 2. Consistent with Sternberg, Monsell found that reaction time increased as a function of the length of the list. Controls for potential confounds were not necessary in these experiments as a consequence of the short duration of each list.

2.3 A Null List Length Effect?

However, as previously noted, there have been a number of studies that have not identified a significant list length effect. Schulman (1974) investigated recognition performance on lists of 25, 50 and 100 words using a two alternative forced choice test. The retroactive design was used with a filler following all study lists to control for retention interval, displaced rehearsal and contextual reinstatement. There was no significant difference in performance for the first 25 words of each list. Similarly, there was no difference in performance on words 26-50 between the 50 and 100 word lists. Thus,

Schulman (1974) concluded that recognition performance for a particular word was unaffected by the number of words that followed it at study.

Buratto and Lamberts (2008) conducted an experiment which involved both list strength and list length manipulations. The within subjects design involved participants studying a weak short list (60 words all presented once), a weak long list (120 words each presented once) and a strong list (30 targets presented once and 30 distractors presented three times, that is, 120 items). The retroactive design was used as a control for retention interval and targets were always presented before any interference items were presented for the second time. The encoding task, a size judgement task or pleasantness ratings task, was manipulated between subjects. Displaced rehearsal was controlled by means of a stimulating video game task following the weak short list. A self-paced study period was also included as an additional control for rehearsal which would prevent participants from using the encoding time to rehearse items that appeared earlier on the study list. A maximum encoding time of 30 milliseconds was set for each item. No control was in place for contextual reinstatement. No significant effect of list length was identified.

Jang and Huber (2008) conducted a series of experiments designed to investigate the way in which context changes throughout lists in a free recall task. To do this, they used the list-before-the-last paradigm of Shiffrin (1970). As its name suggests, this task involves testing participants on the list before the last list viewed, rather than the one immediately preceding the test list as is commonly done. The last list is referred to as the intervening list and the list before the last is the target list. Jang and Huber (2008) presented a series of lists in this manner with testing between lists. In their second experiment, they investigated the effect of using a free recall versus a forced choice recognition task at test. The free recall task involved participants recounting as many items as they were able from the study lists in any

order. In a within subjects design, participants viewed 34 study lists and were tested on all but the first and last lists. Half of the lists were short (6 items) and half were long (24 items). There were 16 free recall test lists and 16 forced choice recognition test lists. The nature of the task effectively controlled for both displaced rehearsal (continuous presentation of lists which would be tested) and contextual reinstatement (the list before the last paradigm necessarily requires reinstatement of the study list in question). Jang and Huber (2008) found no significant effect of the length of the target lists on recognition performance.

Dennis and Humphreys (2001) conducted two experiments to investigate whether a list length effect was evident when controls for all four potential confounds were implemented. In their first experiment, a between subjects design, participants studied either a short list (24 items) or a long list comprised of three blocks of 24 items (72 items in total). Controls for all four potential confounds were implemented with both retroactive and proactive designs used to control for retention interval. Attention was controlled using the pleasantness rating task as well as dividing the long list into separated blocks with puzzle filler in between each. Displaced rehearsal was controlled by means of an incidental test and an eight minute period of filler activity was used as a control for contextual reinstatement. No significant effect of list length was identified. The second experiment involved list length as a within subjects variable and also included a manipulation of list strength. Thus, participants were presented with three lists to study; a short list made up of two 20 item blocks, a long list made up of four 20 item blocks and a mixed list involving four 20 item blocks with the second block repeated as the third and fourth blocks. All potential confounds were controlled – retention interval was controlled using the retroactive design, a pleasantness rating task was used at study as a control for attention, displaced rehearsal was controlled by means of a stimulating filler task and there was a four minute period of filler as

a control for contextual reinstatement. There was no significant effect of list length on either the hit rates or false alarm rates between the short and long lists.

Criss and Shiffrin (2004c) conducted a series of associative recognition experiments. In Experiment 1, participants studied pairs of faces, pairs of words, and pairs made up of both faces and words. They were asked to rate how associated the two items were and this could act as a control for attention. No other controls were implemented. The test list included intact and rearranged pairs and responses were in the form of confidence ratings. Criss and Shiffrin found no effect of list length on recognition performance for word pairs in list lengths of 20, 30 and 40 pairs. However, significant list length effects were identified for face pairs and pairings of words and faces together.

Dennis et al. (2008) conducted a study designed to investigate the effect of controlling contextual reinstatement on the detection of the list length effect. In this within subjects design, participants studied lists of 20 and 80 words. In one condition, controls for all potential confounds were implemented. The retroactive design was used to control for retention interval and a pleasantness rating task was used to control for attention. Displaced rehearsal was controlled using a stimulating puzzle filler task and a further eight minutes of this puzzle was used before each test list as a control for contextual reinstatement. Results revealed a nonsignificant effect of list length on recognition performance. In the other condition, the control for contextual reinstatement was removed while all other controls remained in place. In this condition, a significant list length effect was identified. However, a Bayesian analysis of the data (see Chapter 3) did not find an effect of list length in either condition suggesting that the effect may have been generated by a small subset of participants.

The studies described above have reported nonsignificant effects of list length on

recognition performance dating back more than 30 years. Taken together, these studies challenge the dominant belief in the field that recognition performance decreases as the length of the study list increases. The results suggest that the difference between these studies and those that did identify a significant list length effect may be the way in which potential confounds were controlled. In particular, the retroactive / proactive distinction may also be critical in this respect.

2.4 Thesis Aims: Where to From Here?

This review of the literature has revealed that the method used to investigate the list length effect has varied significantly between studies. Different combinations of potential confounds have been controlled and the same confounds have been controlled in different ways between studies. More importantly, contradictory results regarding list length have been reported. One explanation for why the list length effect is identified in some cases and not in others may be a lack of experimental power in these experiments. Insufficient power in an experiment may lead to a failure to detect a significant effect when it does exist. Variations in power are likely between experiments in that each employed a different list length ratio and number of participants.

However, a review of Table 1 suggests that the potential confounds controlled in each of these experiments may explain the discrepant findings in line with Dennis and Humphreys' (2001) claim that it was the influence of these confounds that resulted in previous list length effect findings. They controlled for all confounds in two experiments and did not identify a significant list length effect. This was also true of the experiment of Dennis et al. (2008). However, it is interesting to note that Cary and Reder's (2003) Experiment 3 also controlled

for all four potential confounds but did identify a significant list length effect. Thus, it appears that there is no simple way to control for these potential confounds. The variation in the literature is indicative of a delicate balance between the potential confounds, perhaps with the retroactive/proactive distinction at its heart.

It is first important to consider the effect of controlling for these confounds. Cary and Reder's (2003) published results allow for comparison between an experiment in which all confounds were controlled (Experiment 3), one in which only contextual reinstatement was controlled (Experiment 2) and one in which no potential list length effect confounds were controlled (Experiment 1). The magnitude of the list length effect in each of Cary and Reder's three experiments was compared using a t-test for unequal samples to analyse the differences between the d' scores for the short and long lists. In Experiments 1 and 2, the present analysis involved only the 16 (short) and 64 (long) word lists in order to match the 1:4 list length ratio of the short (20 word) and long (80 word) lists in Experiment 3.

Analysis revealed that the difference between short and long list d' in Experiment 1 was not statistically significantly different to the same comparison in Experiment 2 ($t(46) = .40$, $p > .05$ (two-tailed)). This result suggests that controlling for contextual reinstatement alone may not prevent confounding of the list length effect. However, the difference between short and long list d' in each of these experiments was significantly different to 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 Figure 8. Each bar represents the difference in short and long list d' for each experiment, that is, each bar represents the magnitude of the list length effect in 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 smaller than the significant

effects identified in Experiments 1 and 2. It is clear in the present analysis that employing controls reduces the magnitude of the list length effect substantially from the original experiments. The fact that a list length effect is still identified may highlight the difficulty in adequately controlling the possible confounds. It appears that is not sufficient to simply control for the potential confounds, but rather, the way in which these controls are implemented is influential.

One difference between Cary and Reder's Experiment 3 and the studies of Dennis and colleagues was the duration of the control used to encourage contextual reinstatement following both the short and long lists. Dennis and Humphreys (2001) included an eight minute period of puzzle filler before each test list in their second experiment as did Dennis et al. (2008) while Cary and Reder's (2003) control for contextual reinstatement was a two minute period of algebra problem solving before each test list. The results of Dennis et al. suggest that two minutes of filler activity may not be sufficient to encourage participants to reinstate the study context at test rather than rely on the end of list context. However, Dennis et al. noted that while controlling for contextual reinstatement was important, it did not appear to be the most influential of the potential confounds. Thus, the contextual reinstatement control may partially explain the difference in Cary and Reder's results from those of Dennis et al, but there must be another explanation.

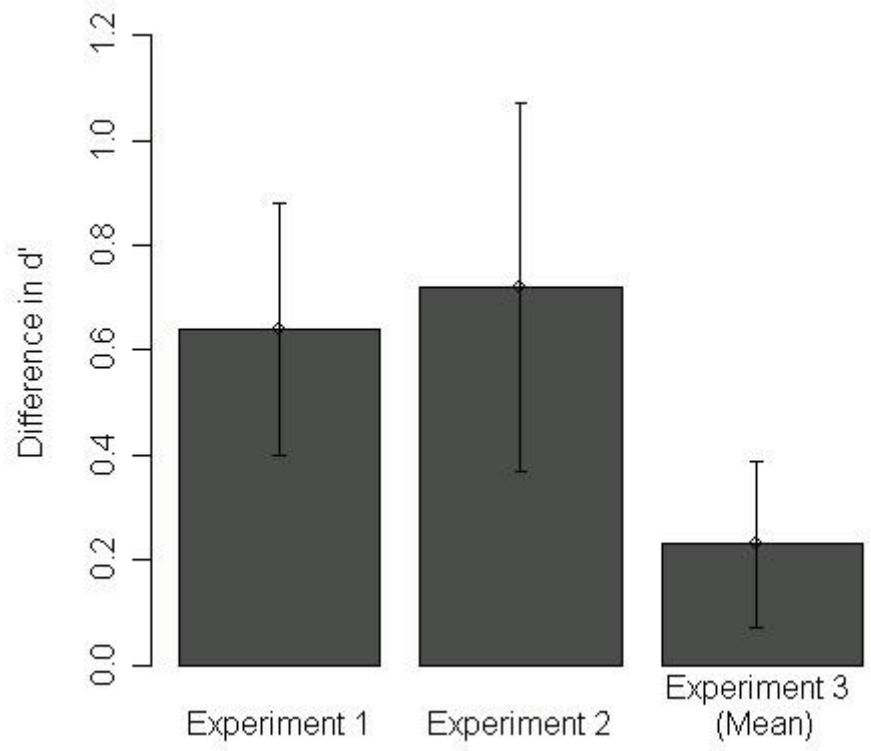

Figure 8. Differences in d' scores between the short and long lists in each of Cary and Reder's (2003) three experiments. Bars represent 95% confidence intervals of the differences between the means.

The way in which Cary and Reder analysed the results of Experiment 3 also differed to that of Dennis and Humphreys (2001). The results of the retroactive and proactive conditions were combined by Cary and Reder in the analysis of Experiment 3 data while these conditions were kept separate by Dennis and Humphreys (2001). Both the retroactive and proactive designs are used to control for retention interval differences between lists but these controls operate in different ways. For example, the retroactive design controls for attention but is subject to differences in contextual reinstatement, displaced rehearsal and study-test lag, while the proactive design controls for all of these issues except attention. Thus, if contextual reinstatement, displaced rehearsal or study-test lag are the most influential

confounds of the list length effect, then the effect is more likely when the retroactive design is used rather than the proactive design. Conversely, if attention is an important confound of the list length effect, then the proactive design is more likely to elicit the effect. It is unclear whether the effect of list length is more likely in one of these conditions and whether the effect in one and not the other can change the result when they are collapsed in the analysis.

Despite the uncertainty regarding the interactions of the retroactive and proactive designs with the other potential confounds of the list length effect, it is well established that retention interval can influence recognition performance. It is consistently found that recognition performance is better when the time period between study and test is smaller, particularly when there is no gap between the two (e.g. Schulman, 1974). Thus not controlling for retention interval would favour a short list where the retention interval is also shorter. Jang and Huber's (2008) experiment demonstrated the effect of retention interval. They found a significant effect of the length of the intervening list on recognition performance. The length of the intervening list was the retention interval and performance was better when the retention interval was small as would be the case following a short list. Thus, it is important to control for retention interval.

The experiments of Murnane and Shiffrin (1991) varied the retention interval control between experiments. In their first two experiments, a significant effect of list length was identified in the overall comparison (no retention interval control) and in a proactive design comparison. In Experiment 3 the retroactive design was introduced and no significant effect of list length was identified when short list performance was compared to performance on the equivalent portion of the long list. However, there was a significant effect when short list performance was compared to the entire long list. This difference in outcome depending on the comparison made stresses the importance of the retention interval control. However,

contrary to this result, in Experiment 4, a significant list length effect was identified when retention interval was controlled using both the retroactive and proactive conditions in a between subjects design regardless of the comparison.

The explanation for this contradictory result may lie with the power of the experimental design, with the effect significant in one experiment and not in the other, despite the same testing conditions. Alternatively, the explanation may rest with the controls for the other potential confounds. There was no control for attention in any of these experiments and the contextual reinstatement control in each was just 30 seconds which may not be sufficient to encourage contextual reinstatement according to the findings of Dennis et al. (2008). Thus it may be that the controls for the other confounds are important to maintain in addition to the control for retention interval.

Displaced rehearsal is another possible confound that, should it not be controlled, would favour performance on a shorter list. Rehearsal improves recognition performance, thus providing participants with an opportunity to rehearse after the short list and not the long list, as is the case in the retroactive design, would favour performance on the former. The possible confounding effect of displaced rehearsal appears to be widely accepted given that all of the previous studies which controlled for retention interval using the retroactive design also included a control for displaced rehearsal. Thus, we will assume that this control is essential when the retroactive design is used and no further investigation is required.

Finally, as Underwood (1978) suggested, the role of differential lapses in attention in confounding the list length effect finding may be significant. Underwood noted that this was an issue when the proactive design was used, however, the extent of the possible confounding effects of attention remains unclear. This will be the focus of Experiment 1 (Chapter 4). The use of an encoding task that requires a response at study can be used in an attempt to maintain

attention in the long list at the same level as that paid to the items on the short list. The influence of this task on maintaining attention and the resulting impact on the list length effect finding will be analysed. The control for retention interval will also be manipulated in this experiment, with the results from both the retroactive and proactive designs compared. The aim of this comparison is twofold; to assess the extent of differential lapses in attention in the proactive design as compared with the retroactive design and, in doing so, assess the relative benefit of each as a control for retention interval.

Another difference between Cary and Reder's (2003) Experiment 3 and the studies of Dennis and Humphreys was the use of the RK task at test in the former. The use of the RK task may prove to be an additional confound of the list length effect. It is possible that the inclusion of this task brought recall-like properties to the recognition experiment and could itself have confounded the results, given that the list length effect is widely accepted to occur in recall. Thus, the present thesis also aims to investigate the effect that the RK task has on the list length effect finding (Chapter 5).

Only the first list length experiment of Strong (1912) and the associative recognition experiment of Criss and Shiffrin (2004c) have used anything other than words as the stimuli. Words are overlearned and highly unitised stimuli and it may be these properties which influence the list length effect outcome when controls for the potential confounds are implemented. As a consequence, the role that other, less unitised and familiar stimuli have on the list length effect finding will be investigated in an attempt to define the boundary conditions of the effect (Chapter 6). This investigation will begin with an associative recognition list length experiment using word pairs as the stimuli and controlling for the four potential confounds of Dennis and Humphreys (2001). Faces, fractals and photographs will then be used as the stimuli in three separate list length experiments in which controls are

implemented for the four confounds. It may be that the list length outcomes for these stimuli differ from those of words. Alternatively, if the results are consistent with those of words, the use of the unfamiliar stimuli can be used in conjunction with the data on words to distinguish between the item noise and context noise models.

Finally, the impact that the results of this entire series of list length experiments has upon the item and context noise approaches to interference in recognition memory will be discussed. The existence of the list length effect is consistent with item noise models of recognition memory, while no significant effect of list length is predicted by context noise models. If the effect exists when a particular stimulus is used in the experiment and disappears when another stimulus is used, this would pose problems for both the item and context noise approaches.

Specifically, the aims of the present thesis are:

- To attempt to resolve the conflict in the literature regarding the list length effect and the influence of controls for potentially confounding variables, in particular the relative effectiveness of the retroactive and proactive designs.
- To explore the list length effect in stimulus sets other than words
- To use this investigation of the list length effect to differentiate between the item and context noise approaches to interference in recognition memory.

Chapter 3

Analysis Methods

The present thesis will use a variety of statistical analyses to investigate the list length effect in recognition memory. Based on the fact that several previously published experiments, which used designs not dissimilar to those that will be used in the present thesis, have failed to identify a significant effect of list length on recognition performance, it is possible that this will also be the case in the present experiments. The failure to find a statistically significant effect can on the one hand mean that the experiment is lacking in statistical power and has literally failed to find the effect. On the other hand, a nonsignificant result can be and, in the case of the list length effect, is theoretically meaningful. Particularly if nonsignificant effects of list length are identified in the present thesis, it is important to demonstrate that this outcome did not hinge on the method by which the data were analysed.

First, a standard analysis of accuracy data focussing on hit and false alarm rates will be carried out using traditional null hypothesis significance testing and analysis of variance (ANOVA). Accuracy data will also be analysed in terms of d' for sensitivity (using both equal and unequal variance signal detection theory). As mentioned in Chapter 2, it is also important to examine the effect of list length on response latency data in addition to the accuracy data. Thus, median response latencies for correct, incorrect and all responses will be analysed. Finally, a new Bayesian signal detection analysis of recognition memory designed by Dennis et al., (2008) will be applied to the data. The major benefit of this method is it allows for the accumulation of evidence in favour of the null hypothesis where standard null hypothesis significance testing (NHST) methods do not. A brief introduction to

this method will be included here but the reader is directed to this paper for further information.

3.1 Standard Accuracy Analyses

There are four response categories in a recognition memory experiment. A hit is recorded when a participant correctly identifies a target item as having appeared on the study list and a false alarm is registered when a participant incorrectly responds "yes" to a word that did not appear. A miss is recorded when a target is dismissed as new and a correct rejection occurs when the participant correctly responds "no" to a distractor (Snodgrass & Corwin, 1988). Hit rates and false alarm rates are commonly reported as measures of recognition performance. In the present thesis, ANOVAs will be performed on the raw hit and false alarm rates for each condition.

It is also important to analyse hit and false alarm rate data simultaneously to take sensitivity into account, thus an analysis of d' values will be undertaken. The calculation of d' is best understood by considering signal detection theory (SDT). SDT is a framework that can be used to describe the recognition process (Banks, 1970; Lockhart & Murdock, 1970). It is concerned with target and distractor distributions. These distributions lie on an axis indexing the strength of the memory, or familiarity, and are assumed to be normally distributed. The target distribution, therefore, lies further along this axis, as target items, having been presented at study, are assumed to be more familiar. The target distribution overlaps with the distractor distribution such that some new items are more familiar than target items (Wixted, 2007, see Figure 9). The mean of the distractor distribution is assumed to be fixed across conditions (Criss, 2009).

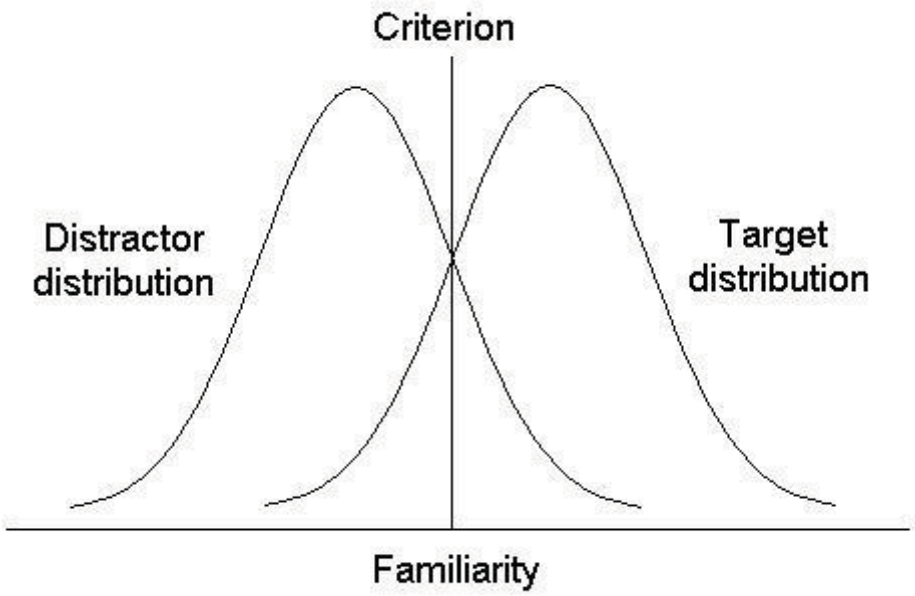

Figure 9. Signal detection model showing distractor and target distributions. A familiarity signal greater than criterion will elicit a "yes" response and familiarity signals below criterion will result in a "no" response.

In making the recognition decision, the participant compares the memory strength (familiarity) of the test probe to the response criterion that they have set. When the memory strength is greater than criterion, the participant will respond "yes" indicating that they believe the item to be a target. Similarly, when the memory strength is lower than criterion, a "no" response will be given. The portion of the target distribution that is greater than the criterion corresponds to the hit rate and the section of the distractor distribution greater than the criterion corresponds to the false alarm rate (Wixted, 2007).

The signal detection model allows for the calculation of a measure of discriminability. This measure, d', is the distance between the mean of the target distribution and the mean of the distractor distribution and represents the ability of the participant to distinguish between targets and distractors. d' is calculated using the following equation:

$$d' = s.z_H - s.z_{FA}$$

where z_H is the standardised false alarm rate, z_{FA} is the standardised hit rate and s is the standard deviation of the target and distractor distributions, standardised to be equal to one.

However, before calculating d', it is first necessary to perform corrections to the hit and false alarm rates to avoid infinite values of d' which occur with perfect recognition performance. As recommended by Snodgrass and Corwin (1988), the corrections in the present thesis were made by adding a value of 0.5 to the hit and false alarm counts and adding 1 to the number of target and distractor items. This correction was applied to all hit and false alarm rates for use in d' calculations.

Because d' takes both hit rates and false alarm rates into account it is a measure of recognition performance that is theoretically free and in practice relatively free from bias (see Snodgrass & Corwin, 1988). It is the difference in the means of the two distributions that is of importance and this is unaffected by where the individual sets criterion.

The signal detection model described thus far has assumed that the target and distractor distributions have equal variance and this is assumed to be equal to one. However, there has been research to suggest that this is not the case and that the variance of the target distribution is greater than that of the distractor distribution. If the two distributions have the same variance, the slope of the standardised ROC curve should be equal to one (Wixted, 2007). The slope of this curve is generally found to be less than one and Ratcliff, Sheu and Gronlund (1992) have found that there is generally a slope of 0.80. This value means that the variance of the target distribution is 1.25 (1/0.80) times greater than that of the distractor distribution (Mickes, Wixted & Wais, 2007; Wixted, 2007, see Figure 10). Thus, the d' equation becomes:

$$d' = (1.25 \times z_H) - z_{FA}$$

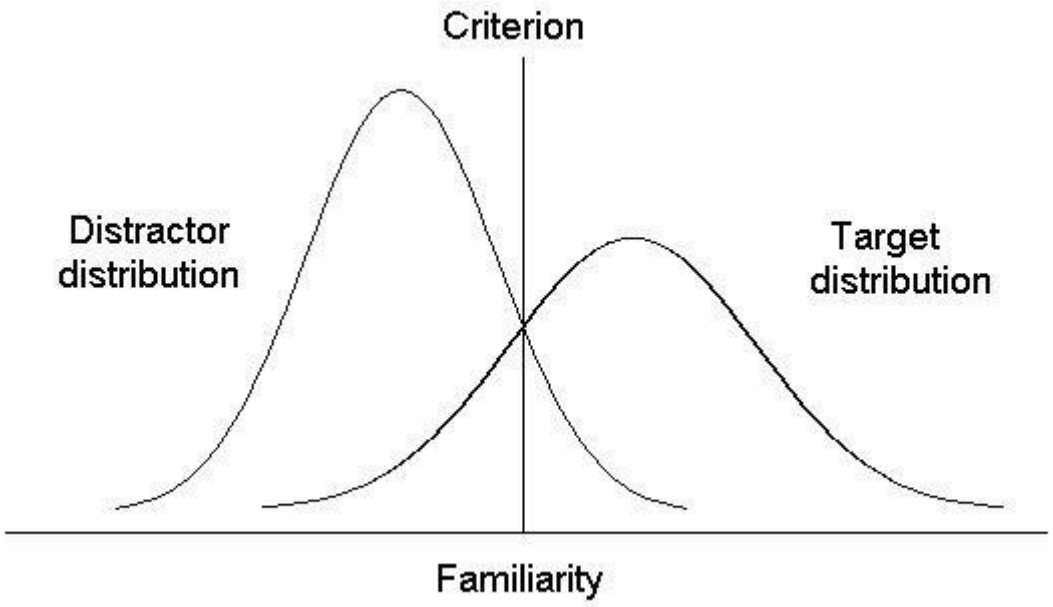

Figure 10. Unequal variance signal detection model showing distractor and target distributions.

Both the equal variance and unequal variance SDT models make assumptions about the variance of the target and distractor distributions. Both analyses will be carried out on the data in the present thesis to ensure that the type of SDT model applied to the data does not alter the substantive outcomes of the experiments. . However, to avoid repetition, only the equal variance d' analyses will be reported, unless the outcome of the unequal variance analysis is in opposition.

3.2 Response Latency Analysis

Response latency data will be collected for each experiment in this thesis. As noted in Chapter 3, the speed accuracy trade-off means that the presence or absence of a significant effect in the accuracy data will not necessarily be reflected in the response latency data. However, the relationship between results in the accuracy data with those in the response latency data is unclear and the two are not necessarily indicative of the same effect.

Several studies which have investigated the list length effect have analysed response latency data (e.g. Atkinson & Juola, 1973, 1974; Bowles & Glanzer, 1983; Juola et al., 1971; Nobel & Shiffrin, 2001; Ratcliff & Murdock, 1976; Reed, 1976; Sternberg, 1966) however the majority have not. Only the studies of Bowles and Glanzer (1983) and Nobel and Shiffrin (2001) reported the effect of list length on both accuracy and response latency data, with a significant effect identified in each case. All of these experiments which examined response latency identified a significant effect of list length, however they generally did not implement controls for any of the potential confounds of the list length effect.

In the present thesis, median response latencies for the correct and incorrect responses will be calculated for each participant, these will then be averaged across participants with the mean value used in the analysis. Using median values should limit the effect of outliers on the analysis.

3.3 The Word Frequency Effect

In order to investigate the list length effect in recognition memory, a design of sufficient power is required to detect the effect if it exists. An a priori power analysis is

complicated by the fact that the majority of studies do not report effect sizes. In addition, each experiment implemented a different combination of controls meaning that the magnitude of the effect will be different depending on the experimental design. To address this issue, another effect, potentially of similar magnitude to the list length effect, can be incorporated into the experimental design. In the present thesis, a manipulation of word frequency will be included in all experiments in which the stimuli are words. The ability to detect a significant word frequency effect, a well replicated finding, in the experiments would indicate that any nonsignificant list length effect finding would not be because the power of the experiment was too poor to detect any effects and demonstrate that the design at least had minimal power.

Both item and context noise models predict the word frequency effect, although it is arguably more naturally accounted for by the latter in that the context noise process involves the retrieval of all previous contexts in which an item has been seen.

3.4 A Bayesian Analysis of Recognition Memory

The Bayesian signal detection method of Dennis et al. (2008) was designed to analyse recognition memory data and overcome a number of issues with null hypothesis significance testing (NHST). In the present experiments, the use of this Bayesian method was motivated by the possibility of failing to identify significant effects of list length (as in Dennis & Humphreys, 2001 and Dennis et al., 2008) despite having experimental power comparable to previous experiments which had identified an effect (e.g. Cary & Reder, 2003). In the experiments of the present thesis, the absence of a significant effect should be taken as a theoretically interesting outcome and not simply a failure to identify an effect. However,

NHST does not allow for acceptance of the null hypothesis, that there is no effect of length on recognition performance.

The main benefit of the Bayesian analysis is that it allows for the accumulation of evidence in favour of several competing theoretical positions, in this case, both for the existence of the list length effect and the absence of the effect. The standard NHST approach assumes that the null hypothesis is true until data proves otherwise and it is not possible to accumulate evidence for a null effect or any other theoretical position (Dennis et al., 2008; Rouder, Speckman, Sun, Morey & Iverson, 2009). Using the standard NHST analysis allows only for evidence to be accumulated in favour of the list length effect, but the alternative is that the null hypothesis, that there is no significant effect of list length cannot be accepted, we can only fail to reject it. As Wagenmakers and Grunwald (2006, p. 641) have noted, it is important to consider "the plausibility of both the null hypothesis and the alternative hypothesis" and the Bayesian analysis allows for this to happen.

Using the standard NHST analysis it is possible that a minority of participants displaying the effect are driving the overall finding. The inclusion of more and more participants in the experiment will eventually lead to a significant finding despite the possibility that it is a minority of them who are displaying the effect. This is problematic when the findings from these experiments are then generalised to the entire population (Dennis et al., 2008). The Bayesian method makes its inference from the majority of participants. It is concerned with what is true for 90% of the participants 90% of the time. This also eliminates the problem of excluding participants whose results are outliers. While these participants may have a large effect on the result of the standard NHST analysis they will not affect what is true of 90% of participants in the Bayesian method.

It is standard in recognition memory to perform edge corrections on hit and false

alarm data to prevent obtaining infinite d' values (c.f. Snodgrass & Corwin, 1988). Performing these edge corrections can greatly affect the results, in particular, because there are a number of different ways that they can be calculated. Results should not be determined by the type of edge correction performed. The Bayesian analysis is carried out on the raw hit and false alarm counts with no edge correction required.

Finally, the standard NHST analysis does not capture sampling variability in that it is insensitive to the number of observations within each cell. The more observations from which the data are drawn, the more certain one can be about the results obtained. The Bayesian method automatically takes into account this uncertainty (Dennis et al., 2008).

3.4.1 How the Model Works and the Implementation of the Model in this Thesis

The Bayesian analysis allows for several competing models to be compared against each other based on a given set of data. With regard to the list length effect, the comparison of interest is whether there is a systematic difference in discriminability in the short list compared with the long list. In this case there are two models being compared. The first is known as the 'error-only' model and assumes that any difference in discriminability between the two groups of interest, in this case, short and long lists, is not systematic and is the result of random noise. This noise is assumed to come from a Gaussian distribution with a mean of zero and unknown variance. The second model is known as the 'error-plus-effect' model in which there is both a systematic positive difference in discriminability between the two groups and also random noise. The difference in discriminability is assumed to come from the sum of a Gamma distribution and a Gaussian distribution with mean of zero.

Each participant has a certain difference in discriminability between the two conditions which in the case of the list length effect would refer to short and long lists. Based on this difference in discriminability, inferences are made regarding whether each participant is best modelled by the error-only model or the error-plus-effect model. Both of these models involve an error component, while only the error-plus-effect model involves an effect component.

The results are reported as a pair of values indicating the posterior probability that at least 90% of participants conform to the error-only or the error-plus-effect model, respectively. Note it is possible that neither of these results occur if, for instance, half of the participants follow an error-only model and half follow an error-plus-effect model. As an example, the results of the Bayesian analysis may be reported as (.74, .15). In this case, the .74 indicates that there is a 74% probability that at least 90% of the participants are best represented by the error-only model. The probability that at least 90% of the participants are represented by the error-plus-effect model is 15%. Here, the probability that 90% of participants are best explained by a combination of the error-only and the error-plus-effect model is the remaining 11% probability. In this example, it is clear that the greater probability lies with the error-only model and has allowed evidence to be accumulated in favour of the null hypothesis. As a general rule of thumb, an effect is considered small when the probability is .50, moderate when .70 and strong when .90.

The implementation of the Bayesian method in this thesis will use the model of Dennis et al. (2008). The results will be obtained by implementing the unequal variance signal detection model in WinBUGS (Lunn, Thomas, Best & Spiegelhalter, 2000). The unequal variance version of the model was employed with the ratio of standard deviations of target and distractor distributions set to 1.25. A burn-in period of 10^3 samples was run in

order to ensure that the Markov-Chain Monte-Carlo (MCMC; see Chen, Shau & Ibrahim, 2000) was sampling from the posterior distribution. This was followed by the collection of 10^4 samples which was used to calculate the proportion of participants best modelled by the error-only model, the error-plus-effect model, or both were calculated.

It is hoped that using a variety of analysis methods will allow for thorough investigation of the list length effect in recognition memory, particularly in the event that the effect is not significant in one or more of these analyses. With each additional analysis run, however, there is an increased risk of obtaining a type I error, that is, of rejecting the null hypothesis when it is actually true (Gravetter & Wallnau, 1985). However, since the motivation for running these additional analyses is to ensure that if a nonsignificant effect of list length is obtained, that this (non) effect is consistent across all analyses. Obtaining a nonsignificant effect in this case is also proof that no type I error has resulted.

Another way to deal with the possibility of type I error is to reduce the alpha level in the analyses which would make statistical significance more difficult to achieve. This will not be done in the present thesis and any nonsignificant effect will not be the consequence of a reduction in the alpha level.

Chapter 4

Experiment 1 - The Effect of Attention on the Detection of the List Length Effect in Recognition Memory

The aim of Experiment 1 was to investigate the influence of attention as a potential confound of the list length effect in recognition memory. Lapses in attention are more likely in the long list than in the short list. It may be this comparative lack of attention paid to the long list items which leads to poorer recognition performance and a spurious list length effect finding. Differential lapses in attention can be controlled with the inclusion of an encoding task which requires a response at study, for example, pleasantness ratings (Cary & Reder, 2003; Dennis & Humphreys, 2001; Dennis et al., 2008). The effectiveness of this task in eliminating the potentially confounding effects of differential lapses in attention has never been investigated. Thus, the present experiment will manipulate the inclusion of the pleasantness rating task between subjects, with a significant list length effect more likely when the task is not used.

Further, one of the major differences between the retroactive and proactive designs, which are implemented to control for differences in retention interval between the long and short lists, is the way that they interact with differential lapses in attention. Underwood (1978) noted that differences in attention are more problematic in the proactive design where the comparison of interest is the entire short list with the last items of the long lists. These are precisely the long list items that are likely to have been paid the least amount of attention. Thus the use of the proactive design and its accentuation of differences in attention may also confound the list length effect. The retroactive design effectively controls for attention in that

short list items are compared with the items from the beginning of the long list, before lapses in attention become substantial. In Experiment 1, the use of the retroactive and proactive designs will also be manipulated between subjects. As noted in Chapter 2, the proactive design, while contributing to the potential confounding effect of attention provides better protection against the other potential confounds than does the retroactive design. Thus, the present analysis will allow assessment of the relative benefits of the retroactive and proactive designs as a control for retention interval.

4.1 Method

4.1.1 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.

4.1.2 Design

This experiment had a 2 x 2 x 2 x 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 while attention task and design were between subjects manipulations.

4.1.3 Materials

The stimuli for this study were 140 five and six letter words from the Sydney Morning Herald Word Database (Dennis, 1995; see Appendix A for experimental stimuli). Half of the words were of high frequency (100-200 occurrences per million) and half were low frequency (1-4 occurrences per million). All lists had the same number of five and six letter, and high and low frequency words. All words were randomly assigned to lists with no participant seeing the same word twice, except for targets.

4.1.4 Procedure

The general procedure of this experiment followed that of Dennis et al. (2008) which, in turn, was based on the methods of Cary and Reder's (2003) Experiment 3 and Dennis and Humphreys' (2001) Experiment 1.

Participants were first given an overview of the study and introduced to the filler activity that would be used throughout the experiment. A computerised sliding tile puzzle was used as the filler task. An image of a fractal, a complex geometric image, 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 in Cary and Reder's (2003) study. Each study word appeared for 3000ms. Test lists were made up of 20 targets and 20 distractors. All lists had half high frequency and half low frequency words. All words were presented in lower-case letters in the centre 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 six 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 3000ms 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 three minute 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 three minutes 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 seconds 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 recognised the word from the study list and to respond “no” if they did not recognise 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 either 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).

Contextual reinstatement was encouraged following both short and long lists with an eight minute period of sliding tile puzzle filler activity before each test list.

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.

4.2 Results

4.2.1 List Length

A 2 x 2 x 2 x 2 (length x frequency x task x 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, η_p^2= .07).

For comparison with the results of Cary and Reder (2003) a 2 x 2 x 2 (length x frequency x 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 (F(1,78) = 3.70, p = .06) and the false alarm rate (F(1,78) = .02, p = .89). Cary and Reder (2003) 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 will be seen, 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. For consistency with the previous analyses and results, oneway ANOVAs were used in place of t-tests. Both these analyses and the Bayesian analyses examined the effect of list length on both accuracy and response latency data, collapsed across word frequency.

4.2.1.1 Retroactive Pleasantness Condition. In the Retroactive Pleasantness condition, repeated measures ANOVAs revealed a nonsignificant effect of list length on d' ($F(1,39) = .38$, $p = .54$, see Figure 11 for short and long list d' in each condition), the hit rate ($F(1,39) = 1.55$, $p = .22$) and the false alarm rate ($F(1,39) = 3.95$, $p = .054$, see Table 4 for hit and false alarm rate data in each condition). Similarly, the Bayesian analysis found in favour of the error-only model (.81, .01), suggesting no effect of list length.

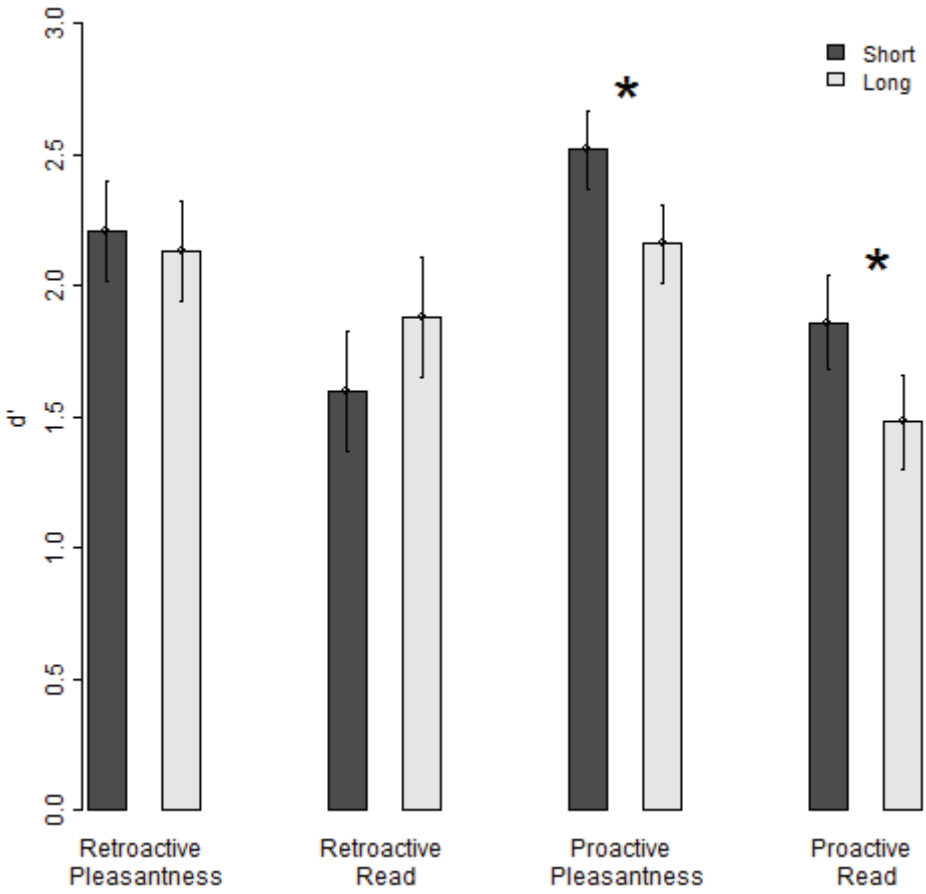

Figure 11. d' values for each of the four attention conditions. There was a nonsignificant effect of list length when the retroactive design was used and a positive list length effect when the proactive design was used. Bars represent 95% within subjects confidence intervals.

Table 4

Mean hit and false alarm rates for each of the four attention conditions (standard deviations in parentheses)

	Hit Rate		False Alarm Rate	
	Short List	Long List	Short List	Long List
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)*

Repeated measures ANOVAs on the response latency data also yielded nonsignificant effects of list length on the median time taken to make correct responses ($F(1,39) = .35$, $p = .56$), incorrect responses ($F(1,37) = .05$, $p = .82$) and both correct and incorrect responses combined ($F(1,39) = .002$, $p = .97$, see Figure *12*).

4.2.1.2 Retroactive Read Condition. Repeated measures ANOVAs in the Retroactive Read condition yielded nonsignificant effects of list length on d' ($F(1,39) = 3.06$, $p = .09$, see Figure 11) and the false alarm rate ($F(1,39) = .60$, $p = .44$). However, there was a statistically significant effect of list length on the hit rate ($F(1,39) = 9.95$, $p = .003$, $\eta_p^2 = .20$, see Table 4). In addition, in this condition, the effect of list length on the unequal variance d' was statistically significant ($F(1,39) = 4.22$, $p = .047$, $\eta_p^2 = .10$). It should be noted, however, that in this condition, performance on the long list was superior to that of the short list, meaning that this result is significant in the opposite direction to that previously identified in the literature. The Bayesian analysis again found in favour of the error-only model (.78, .06).

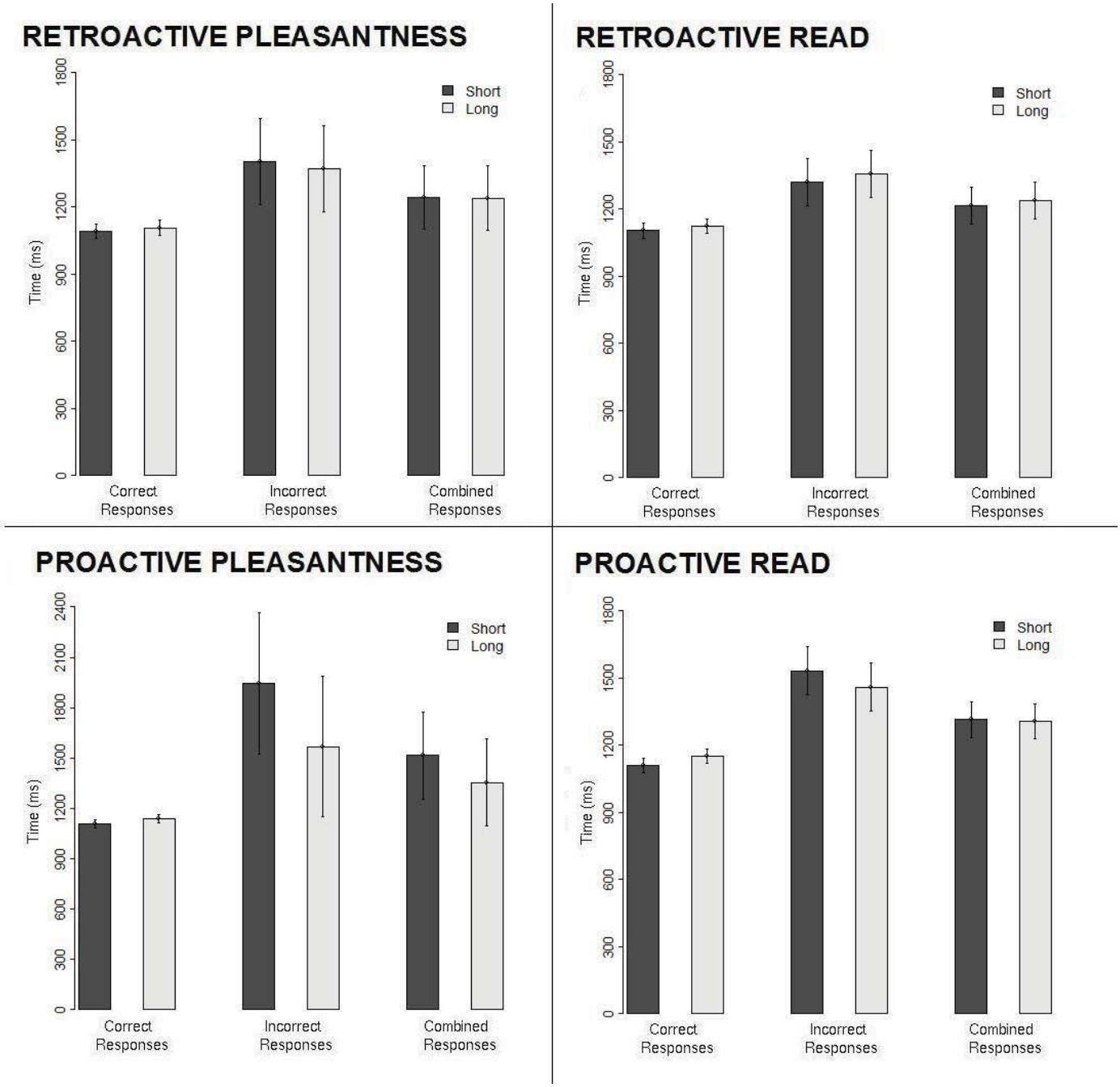

Figure 12. Median response latency for correct, incorrect and combined responses in each condition (bars represent 95% within subjects confidence intervals). The graph for each condition was drawn on the same scale with the exception of the Proactive Pleasantness condition, where the effect of outliers influenced the size of the confidence intervals.

Similarly, repeated measures ANOVAs yielded nonsignificant effects of list length on median response latencies for correct responses ($F(1,39) = .62$, $p = .44$), incorrect responses ($F(1,38) = .23$, $p = .64$) and correct and incorrect responses combined ($F(1,39) = .33$, $p = .57$, see Figure 12).

4.2.1.3 Proactive Pleasantness Condition. In the Proactive Pleasantness condition, repeated measures ANOVAs yielded statistically significant effects of list length on both d' ($F(1,39) = 11.55$, $p = .002$, $\eta_p^2 = .23$, see Figure 11) and false alarm rate ($F(1,39) = 6.72$, $p = .013$, $\eta_p^2 = .15$). There was no significant effect on the hit rate ($F(1,39) = 2.42$, $p = .13$, see Table 4. In contrast to the ANOVA d' results, the Bayesian analysis found in favour of the error-only model (.68, .13).

Oneway repeated measures ANOVAs on the response latency data suggested a marginally significant effect of list length on the median response time for correct responses ($F(1, 39) = 3.76$, $p = .06$). There was a nonsignificant effect of list length on both the median time taken to make incorrect responses ($F(1, 37) = 1.56$, $p = .22$) and both correct and incorrect responses combined ($F(1,39) = 1.26$, $p = .27$, see Figure 12). Note that the mean response latency for the incorrect responses on the short list in this condition was driven up by one outlier. Removal of the outlier did not alter the substantive conclusions drawn in this case and was therefore left in the analysis.

4.2.1.4 Proactive Read Condition. A repeated measures ANOVA in the Proactive Read condition yielded a significant effect of list length on d' ($F(1,39) = 8.26$, $p < .001$, $\eta_p^2 = .17$, see Figure 11). 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 ($F(1,39) = 3.65$, $p = .06$, see Table 4),

although the false alarm rate was close to significance. The Bayesian analysis was ambiguous for this condition (0.46, 0.27) but favoured the error-only model.

Oneway repeated measures ANOVAs for the response latency data yielded nonsignificant effects of list length on the median time taken to record correct responses (F(1, 39) = 3.52, p = .07), incorrect responses (F(1,37) = .90, p = .35) and both correct and incorrect responses combined (F(1, 39) = .02, p = .88, see Figure 12).

4.2.2 Word Frequency

A 2 x 2 x 2 x 2 (length x frequency x task x design) repeated measures ANOVA yielded a significant effect of word frequency on d' (F(1,156) = 232.79, p < .001, η_p^2= .60) in the overall data. Further, planned comparisons were carried out on the word frequency data in each of the four conditions, collapsing across list length.

4.2.2.1 Retroactive Pleasantness Condition. Three repeated measures ANOVAs yielded significant effects of word frequency on d' (F(1,39) = 52.4, p < .001, η_p^2= .57), hit rate (F(1,39) = 14.36, p < .001, η_p^2= .27) and the false alarm rate (F(1,39) = 47.88, p < .001, η_p^2= .55, see Figure 13 for hit and false alarm rates in each condition) in the Retroactive Pleasantness condition. Similarly, the Bayesian analysis of the d' values found in favor of the error-plus-effect model (0, .83).

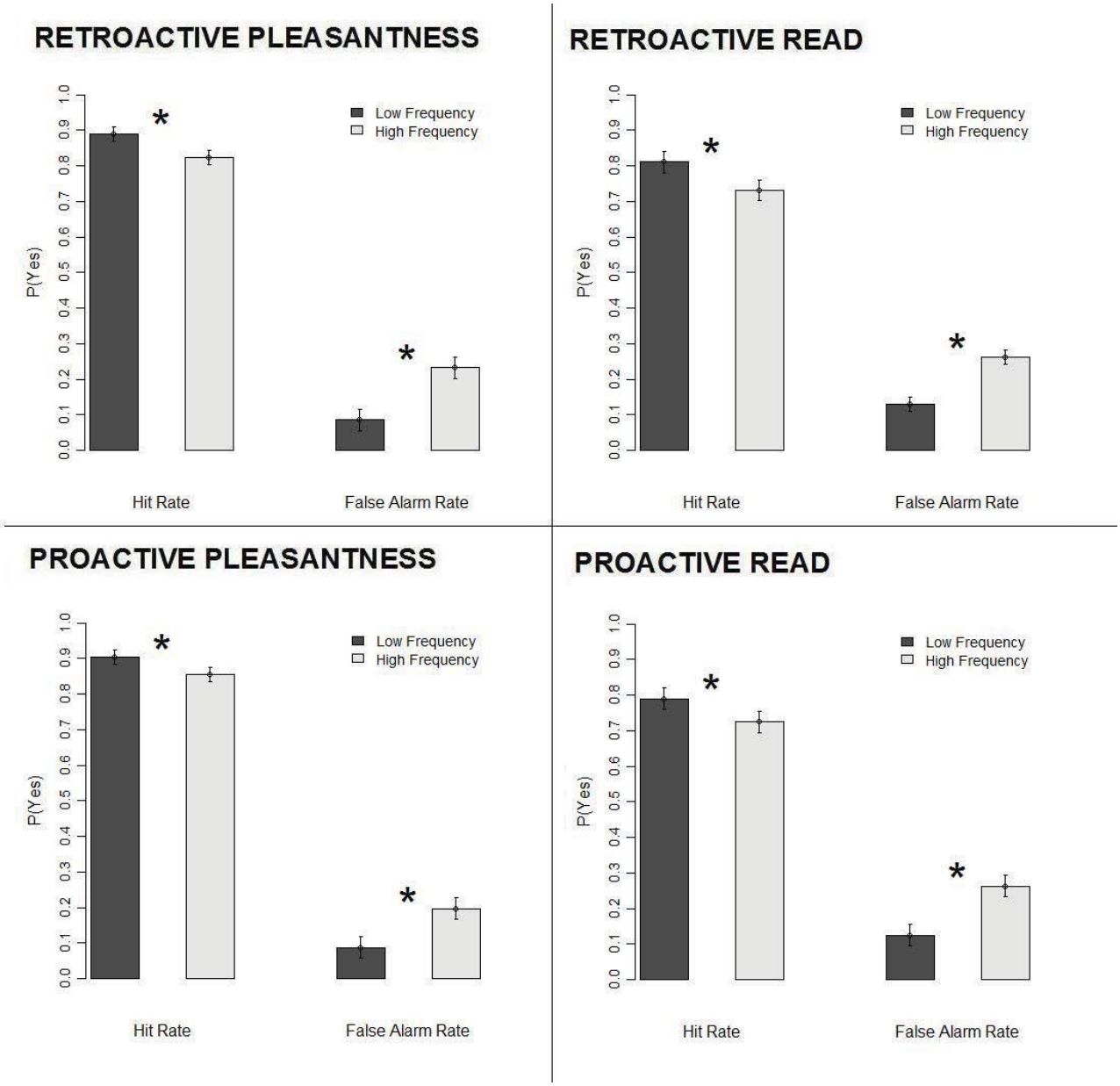

Figure 13. A significant word frequency effect was identified in both hit and false alarm rates in each condition. Bars represent 95% within subjects confidence intervals.

4.2.2.2 Retroactive Read Condition. In the Retroactive Read condition, repeated measures ANOVAs yielded significant effects of word frequency on d' ($F(1,39) = 66.12$, $p < .001$, $\eta_p^2 = .63$), the hit rate ($F(1,39) = 11.83$, $p = .001$, $\eta_p^2 = .23$) and the false alarm rate ($F(1,39) = 59.19$, $p < .001$, $\eta_p^2 = .60$, see Figure 13). The Bayesian analysis of the d' values

again supported the error-plus-effect model (0, .89).

4.2.2.3 Proactive Pleasantness Condition. Repeated measures ANOVAs in the Proactive Pleasantness condition yielded significant effects of word frequency on d' ($F(1,39) = 49.12$, $p < .001$, $\eta_p^2 = .56$), the hit rate ($F(1,39) = 11.32$, $p = .002$, $\eta_p^2 = .22$) and the false alarm rate ($F(1,39) = 26.76$, $p < .001$, $\eta_p^2 = .41$, see Figure 13). The Bayesian analysis of the d' values found in favour of the error-plus-effect model (.05, .64).

4.2.2.4 Proactive Read Condition. In the Proactive Read condition a significant effect of word frequency was found on d' ($F(1,39) = 57.82$, $p < .001$, $\eta_p^2 = .60$), the hit rate ($F(1,39) = 9.21$, $p = .004$, $\eta_p^2 = .19$) and the false alarm rate ($F(1,39) = 29.24$, $p < .001$, $\eta_p^2 = .43$, see Figure 13). The Bayesian analysis of d' values supported the error-plus-effect model (0, .89).

A strong word frequency effect was identified under all conditions and using both the standard ANOVA analysis and the Bayesian analysis. These findings suggest that the power of this experiment was not so poor that no effects of any kind were identified.

4.3 Discussion

The aim of Experiment 1 was to investigate the effect of attention on the detection of the list length effect in recognition memory. Two aspects of attention were examined. First, the use of the pleasantness rating task as a means of controlling for differential lapses in attention was explored. This was done by comparing performance on short and long lists

which used the pleasantness rating task at study with short and long study lists that did not involve any encoding task. It is thought that the inclusion of an encoding task that requires a response at study can help to ensure that all study items are processed to an equivalent level in each list length. Secondly, differential lapses in attention may be exacerbated when the proactive design is used in the experiment rather than the retroactive design (Underwood, 1978). If this is the case, then the magnitude of the list length effect would be expected to be greater in the proactive design condition. Experiment 1 included the design as a between subjects factor.

The experiment involved four between subjects conditions each with a different level of control for differential lapses in attention. The Retroactive Pleasantness condition involved controls for all four confounds outlined by Dennis and Humphreys (2001) including a pleasantness rating encoding task at study and the use of the retroactive design. In this condition, no significant effect of list length was identified. At the other end of the scale, the Proactive Read condition involved no control for attention under circumstances (the proactive design) in which inattention was likely. A significant list length effect was identified in this condition. In addition, it is interesting to note that in the Retroactive Read condition long list performance was actually superior to short list performance. The Bayesian analysis favoured the error-only model in each condition, suggesting no significant effect of list length.

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 where the retroactive and proactive design conditions were collapsed for analysis. When the present results were collapsed in the same manner, a significant list length effect was identified. However, the effect of list length was nonsignificant when the

results of only those participants who completed the retroactive design were analysed. It appears that the design of the experiment is important and that it is the proactive condition which drives the effect.

The observed list length effects evident in the d' values in the proactive conditions are attributable to differences in the false alarm rates between lists. These differences occur with no significant change in the hit rates based on list length. The difference in false alarm rates but not hit rates could come about as a result of an increase in accuracy and an increase in the decision criterion when greater attention is paid to short list items. In terms of optimal performance, increasing the decision criterion with increased sensitivity would result in a minimisation of errors overall.

Taken together, the results of Experiment 1 suggest that differential lapses in attention are the most influential confound of the list length effect finding. Further, the use of a ratings task at study does little to control for this potential confound. The best way to control for differential lapses in attention is to use the retroactive design. In doing so, however, controls for displaced rehearsal and contextual reinstatement must also be implemented. The results of the present experiment also suggest that study-test lag does not confound the list length effect finding. No controls were implemented for this confound in the retroactive design and no significant effect of list length was identified (note that the use of the proactive design is a natural control for differences in study-test lag).

No significant effect of list length on the response latency data was identified in any of the conditions, however, the difference between the means was greatest for the two conditions in which the proactive design was used rather than the retroactive design, a finding consistent with the results of the accuracy data.

Chapter 5

Experiment 2 - The List Length Effect and the Remember-Know Task

Cary and Reder's Experiment 3 was a partial replication of Dennis and Humphreys' (2001) experiments. In both cases, controls were implemented for the four potential confounds of the list length effect but Cary and Reder identified a significant effect while Dennis and Humphreys did not. The results of Experiment 1 have suggested that this difference may be attributable to the collapsing of the retroactive and proactive designs in the analysis of Cary and Reder's data. Further, Dennis et al. (2008) contended that the different results were attributable to Cary and Reder's use of a shorter period of filler activity with which to control for contextual reinstatement. Another possible explanation for the contradictory findings is Cary and Reder's use of the Remember-Know (RK) task in their experiments. It may be that the use of the RK task at test is an additional confound of the list length effect. This will be the focus of Experiment 2.

Originally developed by Tulving (1985), the RK paradigm has been used as a means of investigating an individual's conscious experience and awareness of the recognition task (Dunn, 2004). The RK task is easily incorporated into the standard yes/no paradigm with the most common method being a two-step procedure. After making a "yes" response, indicating that they have recognised 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" judgement. 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 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).

With regard to the list length effect, the inclusion of the RK task in the study design could be another potential confound. Remember responses, in that they are based on recollection of the study experience, have been said to involve a recall-like process (Clark, 1999; Diana, Reder, Arndt & Park, 2006). Participants are no longer just asked whether or not they recognise a particular item from the study list, but rather, they are asked how they recognise that item. That is, they are asked to recall elements of the study event. As Diana et al. (2006) note, 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 in the yes/no paradigm. Thus, the use of the RK task may induce some recall-like components to the recognition task. Since the list length effect is widely accepted to occur in recall, this may confound the list length effect finding in recognition. The present experiment used the retroactive design only and manipulated the inclusion of the RK task between subjects.

5.1 Method

5.1.1 Participants

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

5.1.2 Design

A 2 x 2 x 2 factorial design was used in this study. The factors were list length (short or long), word frequency (low or high) and test task (RK task or yes/no 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 of the power of the experiment.

5.1.3 Materials

One hundred and forty words chosen from the Sydney Morning Herald Word Database (Dennis, 1995) were used as stimuli in this experiment (see Appendix A). As with Experiment 1, there were an equal number of five and six letter, and high (100-200 occurrences per million) and low frequency (1-4 occurrences per million) words. All words were randomly assigned to lists for each participant.

5.1.4 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. 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 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. Lists were counterbalanced for order.

5.2 Results

5.2.1 List Length

A 2 x 2 x 2 (length x frequency x task) repeated measures ANOVA did not identify a significant interaction between list length and task on d' ($F(1,78) = 1.32$, $p = .25$). A similar pattern was identified in both the hit and false alarm rates ($F(1,78) = .02$, $p = .90$ and $F(1,78) = .08$, $p = .78$, respectively).

In addition, two planned comparisons were carried out on each of the 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.

5.2.1.1 Yes/No Task Condition. In the Yes/No Task condition, repeated measures ANOVAs yielded nonsignificant effects of list length on d' ($F(1,39) = .43$, $p = .51$, see Figure 14 for short and long list d' in each condition), the hit rate ($F(1,39) = 4.30e^{-30}$, $p = 1$) and the false alarm rate ($F(1,39) = .15$, $p = .70$, see Table 5 for hit and false alarm rate data). Similarly, the Bayesian analysis found strongly in favour of the error-only model (.84, .03).

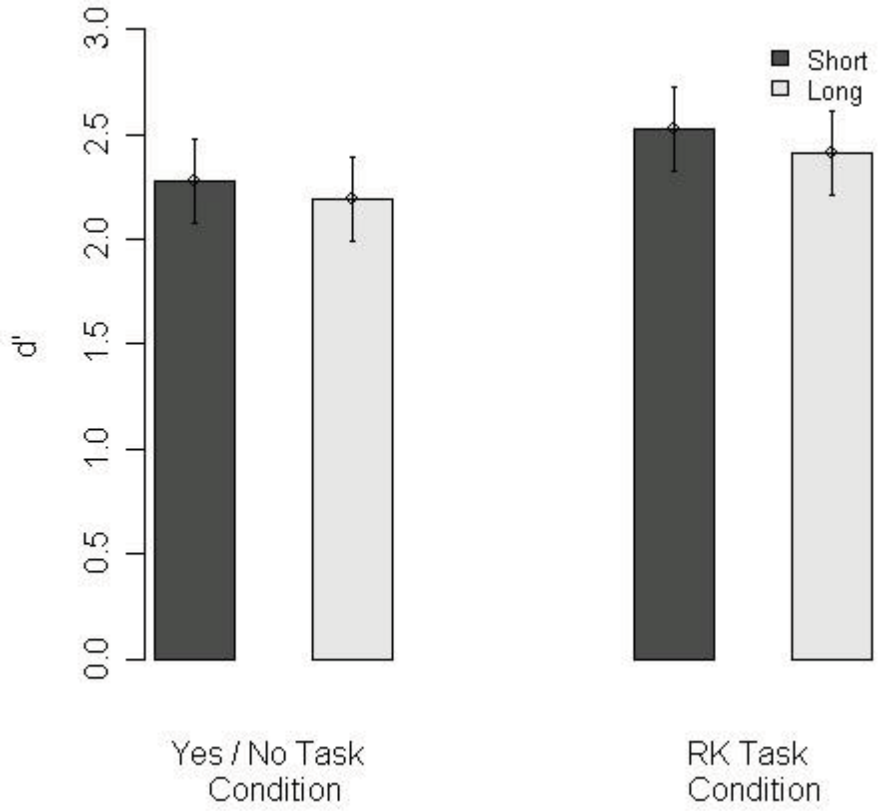

Figure 14. d' values for the Yes/No Task and RK Task conditions. There was no significant effect of list length in either condition. Bars represent 95% within subjects confidence intervals.

Finally, three repeated measures ANOVAs on the response latency data in the yes/no task condition did not identify a significant effect of list length on the median time taken to record correct responses ($F(1,39) = .14$, $p = .71$), incorrect responses ($F(1, 36) = .37$, $p = .55$) and both correct and incorrect responses combined ($F(1, 39) = .29$, $p = .59$, see Figure 15).

Table 5

Mean hit and false alarm rates for the Yes/No Task and RK Task conditions in Experiment 2 (standard deviations in parentheses)

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

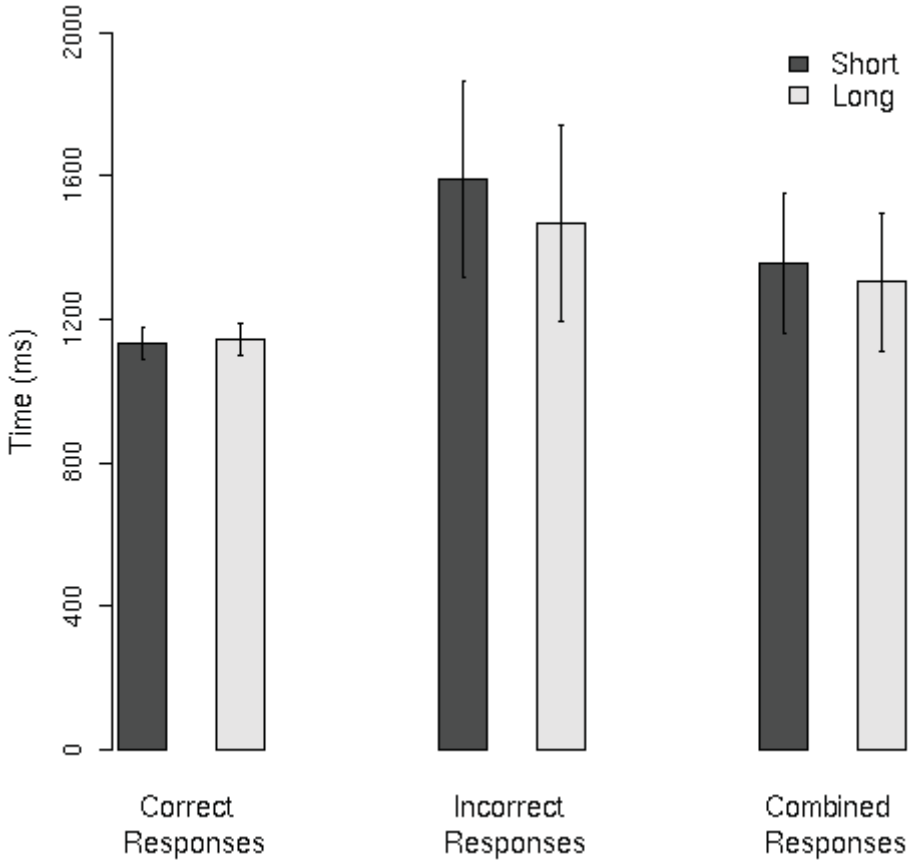

Figure 15. Median response latency for correct, incorrect and correct and incorrect responses combined for the yes/no task condition. There was no significant effect of list length on response latency. Bars represent 95% within subjects confidence intervals.

5.2.1.2 Remember-Know Task Condition. Similarly, in the RK Task condition, three repeated measures ANOVAs yielded nonsignificant effects of list length on d' (F(1,39) = 1.21, p = .28, see the bars on the right side of Figure 14), hit rate (F(1,39) = .03, p = .86) and false alarm rate (F(1,39) = 1.24, p = .27, see Table 5). Again, the Bayesian analysis favoured the error-only model (.69, .22).

Analysis of the response time data via oneway repeated measures ANOVAs yielded a nonsignificant effect of list length on the median time taken to record incorrect responses (F(1,35) = 1.97, p = .17) and correct and incorrect responses combined (F(1,39) = 2.10, p = .16). However, there was a statistically significant effect of list length on the median time taken to record correct responses (F(1,39) = 4.71, p = .04, η_p^2 = .11, see Figure 16).

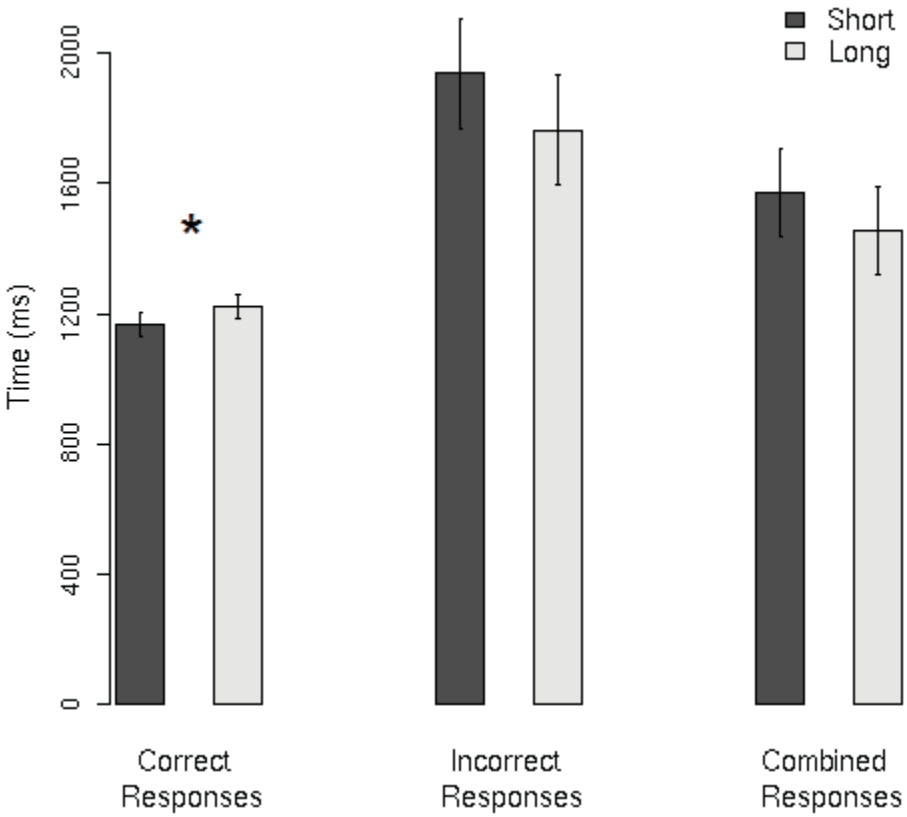

Figure 16. Median response latency for correct, incorrect and correct and incorrect responses combined in the RK Task condition. There was a significant effect of list length on response latency for correct responses. Bars represent 95% within subjects confidence intervals

5.2.2 Analysis of Remember-Know Data

A 2 x 2 (list length x whether word studied) repeated measures ANOVA was conducted on all remember (R) responses given by participants in the RK task condition. There was no significant effect of list length on the number of R responses given ($F(1,39) = .001$, $p = .97$, see Table 6 for all Remember-Know data). There was an anticipated effect of whether or not a particular word was studied on the number of R responses ($F(1,39) = 520.75$, $p < .001$, $\eta_p^2 = .93$). Given that R responses should only be given if the participant has a recollection of the word appearing in the study list, the strength of this effect was expected. There was also a nonsignificant interaction between list length and whether or not a particular word was studied on the number of R responses given ($F(1,39) = .38$, $p = .55$). It should be noted that this interaction was significant in the study of Cary and Reder (2003).

Table 6

Remember and Know responses as proportions of the total number of hits and false alarms

	Hits		False Alarms	
	Short List	Long List	Short List	Long List
Remember responses	.81	.80	.42	.45
Know responses	.19	.20	.58	.55

The results for the know (K) responses followed a similar pattern to the R responses. Again, the effect of list length on the number of K responses was nonsignificant ($F(1,39) = .41$, $p = .52$). Whether or not a word was studied had a significant effect on the number of K 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 ($F(1,39) = .01$, $p = .90$).

5.2.3 Word Frequency

A 2 x 2 x 2 (length x frequency x 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.

5.2.3.1 Yes / No Task Condition. In the Yes/No Task condition, repeated measures ANOVAs yielded statistically significant effects of word frequency on d' ($F(1,39) = 18.74$, $p < .001$, $\eta_p^2 = .32$) and the false alarm rate ($F(1,39) = 15.24$, $p < .001$, $\eta_p^2 = .28$). The effect of word frequency on the hit rate was not significant ($F(1,39) = 3.29$, $p = .08$, see Figure 17). The Bayesian analysis of d' values found in favour of the error-only model (.52, .20).

5.2.3.2 Remember-Know Task Condition. Three repeated measures ANOVAs in the RK Task condition yielded statistically significant effects of word frequency on d' ($F(1,39) = 30.49$, $p < .001$, $\eta_p^2 = .44$), the hit rate ($F(1,39) = 11.57$, $p = .002$, $\eta_p^2 = .23$) and the false alarm rate ($F(1,39) = 21.86$, $p < .001$, $\eta_p^2 = .36$, see Figure 17). Similarly, the Bayesian analysis of d' values found in favour of the error-plus-effect model (.03, .59).

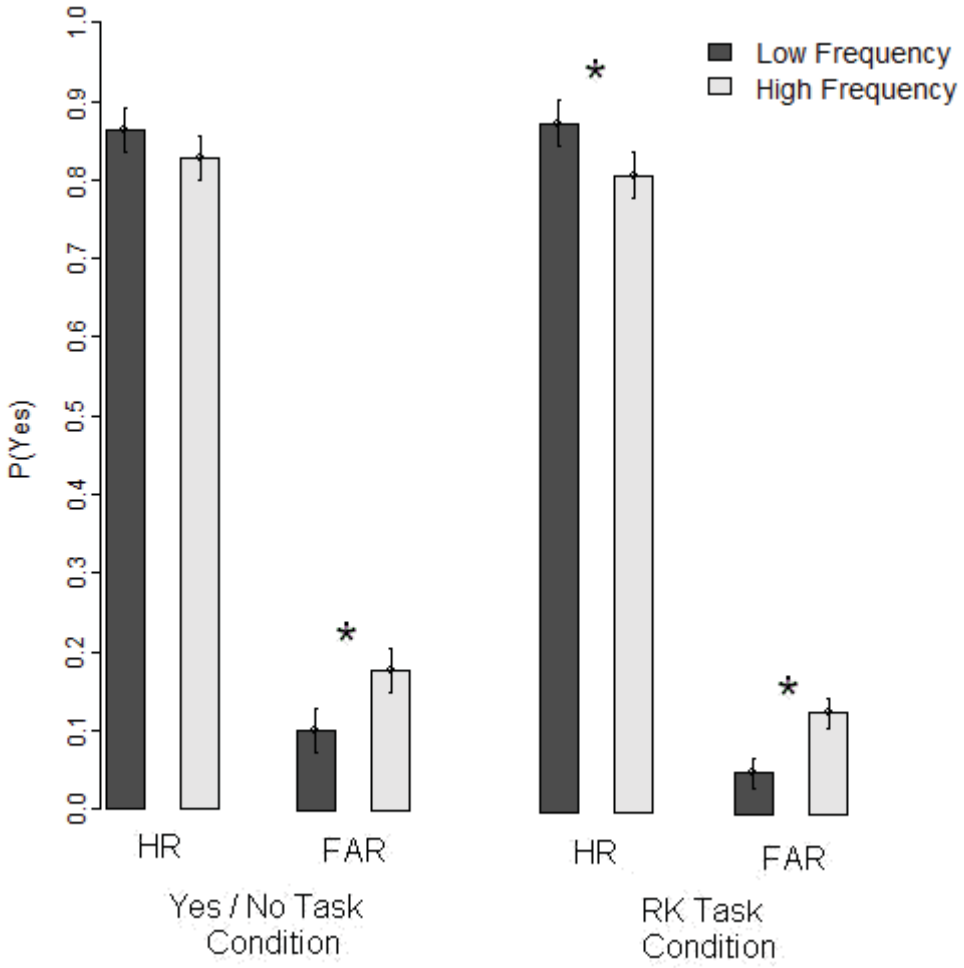

Figure 17. A significant word frequency effect was identified in all comparisons but the hit rate in the Yes/No Task condition. Bars represent 95% within subjects confidence intervals.

5.3 Discussion

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. This experiment was motivated by the possibility that the inclusion of the RK task in the experimental design may be an additional confound of the list length effect in that it may have induced recall-like strategies in participants. The list length effect is a

hallmark finding in recall. The results indicated no significant effect of list length on the accuracy data under either the Yes/No or the RK Task conditions using either the standard ANOVA analysis or the Bayesian analysis. However, there was a significant effect of list length on the median response latency data for correct responses in the RK Task condition. The response latency for correct responses on a short list was 57 milliseconds faster than the response latency for correct responses on the long list.

The nonsignificant effect of list length in the accuracy data of both conditions is consistent with the results of Dennis and Humphreys (2001), Dennis et al. (2008) and the results of Experiment 1 of this thesis, all of which suggest that when controls for the four potential confounds of the list length effect are controlled the effect is not significant. There was a significant effect of list length in the response latency data of the RK task condition which suggests that this task may, as hypothesised, induce a recall-like component into the recognition task (c.f. Diana et al., 2006) and result in a spurious list length effect finding. However, it is unclear exactly how the accuracy and response time data interact and whether a statistically significant effect of list length in the response time data is analogous to a statistically significant effect in the accuracy data. However, the claim that there is a recall-like process involved in the RK task must be mediated by the absence of a significant effect of list length on the number of remember responses. This effect should be present if remember responses are taken to be generated by recall-like processes. It seems safe to conclude that while there may be a recall-like process introduced to the recognition experiment with the inclusion of the RK task, this effect is minimal.

There was a significant word frequency effect under most conditions, however there was a nonsignificant effect on the hit rate component in the Yes/No Task condition as evidenced in both the ANOVA and Bayesian analysis. This result was consistent with several

other studies which identified disruptions to the hit rate component under a variety of encoding conditions (e.g. Criss & Malmberg, 2008; Criss & Shiffrin, 2004b, Glanc & Greene, 2007; Hirshman & Arndt, 1997). As in Experiment 1, the significant word frequency effects identified in this experiment suggest that the power of the experimental design was sufficient to identify a significant effect.

Chapter 6

The List Length Effect with Stimuli other than Words

With the exception of the very first list length study, that of Strong in 1912 which used newspaper advertisements, and the associative recognition experiment of Criss and Shiffrin (2004c) which included faces, all other published list length studies have used words as the stimuli. Strong did not control for any of the four potential confounds of Dennis and Humphreys (2001) in his study, while Criss and Shiffrin's experiment only included a control for attention. Thus there has not been a fully controlled list length experiment using stimuli other than words. An investigation using different stimulus sets may also help to identify the boundary conditions of the list length effect: under what circumstances does it exist and when does it not.

On the basis of the results of the previous two experiments of this thesis, it seems that the source of interference in the recognition of single words is from the previous contexts in which they have been encountered. The role of item noise in this process in negligible at best. However it remains unclear if the nonsignificant effect of list length is specific to words or if item noise plays a role when different stimuli are used. It is possible that there is something special about words. They are unitised stimuli and, despite the fact that they are very commonly encountered, they have distinctive representations and are not easily confused with each other. As Greene (2004, p. 261) noted "[A]ll words contain strong semantic features that make each one unique and distinctive". It may be that other, less unitised, stimuli behave differently. Thus, one aim of the present thesis was to examine whether the list length effect is evident when stimuli other than words are used. The final

four experiments of this thesis were designed to address this issue.

6.1 Experiment 3: Word Pairs

One variation on using single words as the stimulus is to pair two words together as a pair as in the associative recognition paradigm. Three experiments have manipulated list length in an associative recognition design with divided results. Clark and Hori (1995) and Nobel and Shiffrin (2001) identified significant list length effects while Criss and Shiffrin (2004c) did not. None of these experiments controlled for all of retention interval, attention, displaced rehearsal and contextual reinstatement. The aim of Experiment 3 was to investigate the list length effect in associative recognition when controls for the four potential confounds of Dennis and Humphreys (2001) are implemented.

In the experiment of Clark and Hori (1995), list length was manipulated between 34 and 100 pairs (list length ratio of 1:3). Contextual reinstatement was the only potential confound for which a control was implemented by means of a 45 second period of mental arithmetic. Nobel and Shiffrin (2001) manipulated list length between 10 and 40 pairs in their first experiment (list length ratio of 1:4), while in their third experiment, list length was either 10 or 20 pairs (list length ratio of 1:2). They also controlled for contextual reinstatement by including a period of mental arithmetic for 26 seconds prior to each test list, however this may not be of sufficient length to encourage contextual reinstatement following the long list (Dennis et al., 2008). Criss and Shiffrin's (2004c) experiment involved participants being presented with three types of lists; pairs of faces (FF), pairs of words (WW) or one word and one face paired together (WF). For lists of word pairs, list length was varied between 20, 30 and 40 pairs (list length ratio of 1:1.5:2). Participants were asked to

rate the degree of association between the two items that made up each pair, which could act as a control for attention.

Criss and Shiffrin's (2004c) experiment had the weakest list length ratio making it the least likely to identify a significant effect of list length, as was the case. However it is also interesting to note that this study included a control for attention while the other three experiments all controlled (weakly) for contextual reinstatement only. Based on the results of Experiment 1 it is feasible that Criss and Shiffrin did not identify the list length effect as a result of this attention control.

There may also be something special about using the associative recognition design rather than a single item recognition experiment. Hockley (1991; 1992) carried out a series of experiments comparing forgetting rates in the recognition of single words and of word pairs. The consistent finding was that the rate of forgetting was greater for single words than it was for word pairs. He concluded that memory for the word pairs is "more resistant to the effects of decay, interference from intervening events, or both" than is memory for single items (Hockley, 1992, p. 1328). However, Weeks, Humphreys & Hockley (2007) argue that Hockley's (1991; 1992) finding of flat forgetting curves does not show a complete absence of forgetting, but suggests that both targets and distracters are forgotten at the same rate. Nevertheless, it may be that there is no significant effect of list length when the stimuli are word pairs simply because they are more resistant to the influence of interference from other study pairs. If the associative recognition cue is formed by combining the items of the pair using a matrix outer product, as is the case in the Matrix model (Pike, 1984) or convolution, as in TODAM (Gronlund & Elam, 1994; Murdock, 1982), then it may be that pairs of words are less similar to each other than are the words from which they are composed.

Another difference between the associative and single item recognition paradigms

involves the relative influence of recall in the task. Clark et al. (1993) and Nobel and Huber (1993) have argued that the associative recognition task involves a combination of global matching and recall. If there is a recall component in the associative recognition process, a significant effect of list length would be the expected result. Clark and Hori (1995) argue that as the length of the list increases, the role of recall is minimised, suggesting that short list performance may be aided by recall strategies while long list performance may not.

A partial replication of the studies of Criss and Shiffrin (2004c), Clark and Hori (1995) and Nobel and Shiffrin (2001) is the basis of the present experiment using word pairs and the associative recognition design. Controls for all potential confounds will be introduced and for consistency with the previous experiments in this thesis, the list length ratio set at 1:4. The primary aim of Experiment 3 was to determine whether word pairs behave differently to single words and whether there is a significant list length effect when the former are used as the stimuli.

6.1.1 Method

6.1.1.1 Participants. Participants were 40 first year Psychology students at the Ohio State University who participated in exchange for course credit.

6.1.1.2 Design. A 2 x 2 factorial design was used in this experiment. The factors were list length (short or long) and word frequency (low or high, note that in this case, word frequency refers to the frequency of both items in the pair, so two high frequency words make up a high frequency pair and two low frequency items make up a low frequency pair). The word frequency manipulation was included in this experiment for consistency with the

previous experiments. However, the nature of the word frequency effect in associative recognition is not as clear as it is in single item recognition. In single item recognition, performance for low frequency words is superior to that of high frequency words, while in associative recognition, the reverse is generally true (Clark, 1992; Clark et al., 1993, Experiments 1 and 2; Clark & Shiffrin, 1992; Clark & Burchett, 1994), however some studies (e.g. Clark et al., 1993, Experiment 3; Hockley, 1992) have not identified a significant effect of word frequency at all. For this reason, the word frequency effect cannot be used as a check of the experimental power in the present experiment.

6.1.1.3 Materials. The stimuli for this experiment were 240 five and six letter words from the Sydney Morning Herald Word Database (Dennis, 1995; see Appendix A). Half of the words were of high frequency (100-200 occurrences per million) and half were low frequency (1-4 occurrences per million). Words were randomly paired with another word of the same frequency. All word pairs were randomly assigned to lists for each participant with no participant seeing the same pair of words twice, with the exception of targets.

6.1.1.4 Procedure. The procedure for Experiment 3 was similar to the overall procedure of Experiments 1 and 2, but there were some key differences. Upon arrival, participants were given an overview of the experiment and completed a practice session of the sliding tile puzzle activity that would be used throughout the experiment. Participants completed two study lists, one short (24 word pairs) and one long (96 word pairs) with the presentation order counterbalanced for participants. Each pair appeared on screen for 3000ms. High frequency pairs comprised half of the list and the other half were low

frequency pairs. The two words of each pair were presented side by side on a computer screen.

During study, participants were asked to rate how related the two words of each pair were to each other on a six point Likert scale (1: unrelated, 6: related) by clicking on the appropriate number displayed below the word pair. They were instructed to make this rating while the word pair was on screen and to move on to and rate the next pair should they miss giving a rating.

The experiment had a retroactive design as a control for retention interval. This meant that the short list was followed by a three minute and 36 second period of sliding tile puzzle filler and that it was the first eight word pairs from both the short and long list that were included as targets at test. One word was taken from each of the next 16 word pairs at study to create rearranged pairs at test. Each of the words in a rearranged pair was presented in the same screen position (left or right) as it had been at study as part of the original pair.

Participants were given a 20 second warning before the start of the test list. They were instructed to respond "yes" if they recognised a word pair from the study list and to respond "no" if they did not. Responses were recorded when the participant clicked on the appropriate button which was displayed on screen below the word pair. Test lists were comprised of eight target pairs and eight rearranged pairs. There was no time limit for responding and there were no missing data.

Contextual reinstatement was encouraged following both lists using the sliding tile puzzle filler task. Eight minutes was spent completing the puzzle before the onset of each test list, in addition to the puzzle as a control for retention interval.

6.1.2 Results

A 2 (length) x 2 (frequency) repeated measures ANOVA did not yield a statistically significant effect of list length on d' ($F(1,39) = .20$, $p = .66$, see Figure 18), the hit rate ($F(1,39) = 3.03$, $p = .09$) or the false alarm rate ($F(1,39) = .51$, $p = .48$, see Table 7 for hit and false alarm rate data). The Bayesian analysis favoured the error-only model (.54, .11) suggesting no effect of list length.

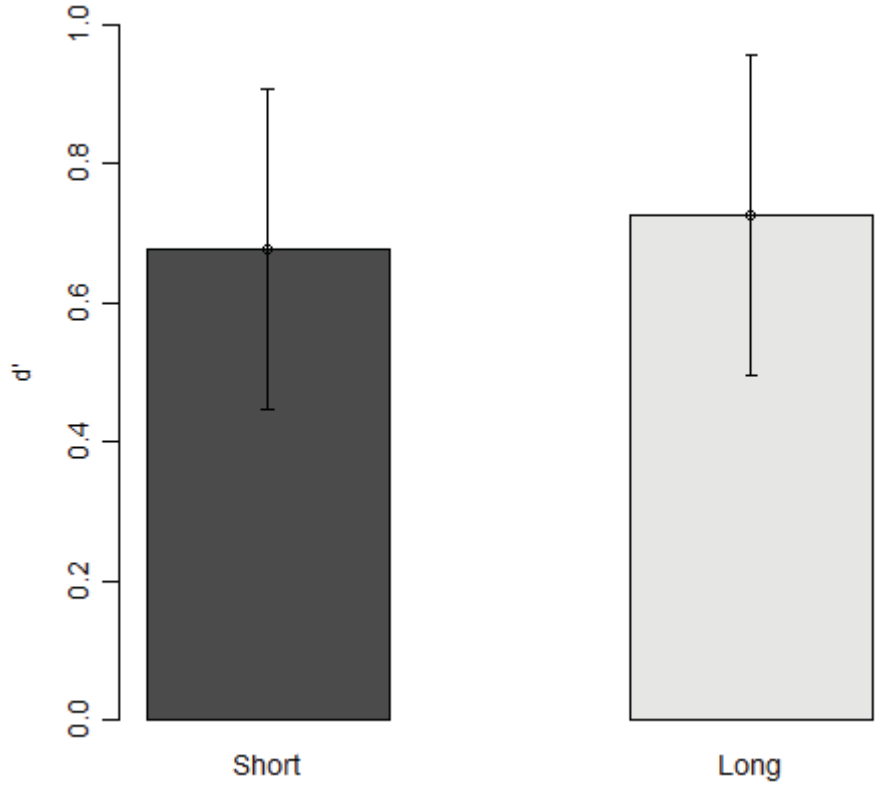

Figure 18. d' values for short and long lists in Experiment 3. Bars represent 95% within subjects confidence intervals.

Table 7

Mean hit and false alarm rates for word pair data in Experiment 3 (standard deviations in parentheses)

	Hit Rate		False Alarm Rate	
	Short List	Long List	Short List	Long List
High Frequency	.72 *(.22)*	.79 *(.22)*	.28 *(.27)*	.34 *(.27)*
Low Frequency	.73 *(.30)*	.78 *(.23)*	.46 *(.33)*	.45 *(.31)*

There was also a significant main effect of word frequency on d' ($F(1,39) = 10.17$, $p = .003$, $\eta_p^2 = .21$), and the false alarm rate ($F(1,39) = 17.04$, $p < .001$, $\eta_p^2 = .30$). However, there was no significant effect of word frequency on the hit rate ($F(1,39) = .01$, $p = .94$, see Figure 19 for hit and false alarm rate data).

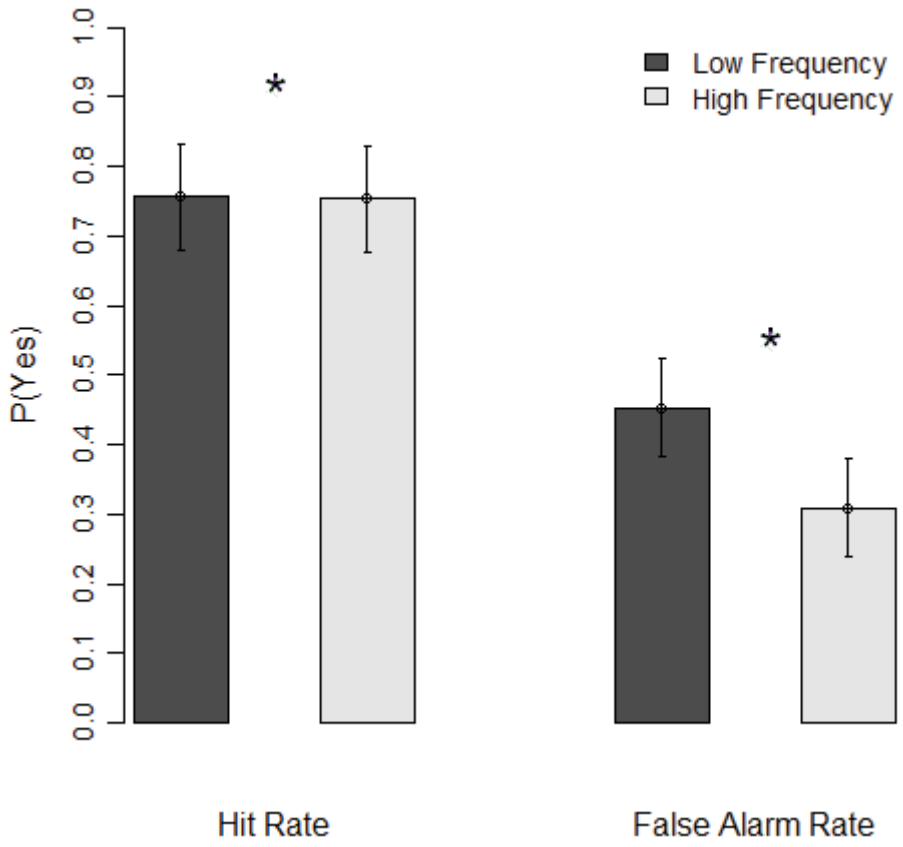

Figure 19. A significant effect of word frequency was identified only for the false alarm rate in the word pair data. There was no significant effect on the hit rate. Bars represent 95% within subjects confidence intervals.

Oneway ANOVAs also yielded nonsignificant effects of list length on the median response latency for correct responses ($F(1, 39) = .68$, $p = .42$), incorrect responses ($F(1, 26) = .37$, $p = .55$) and correct and incorrect responses combined ($F(1, 38) = .92$, $p = .34$, see Figure 20).

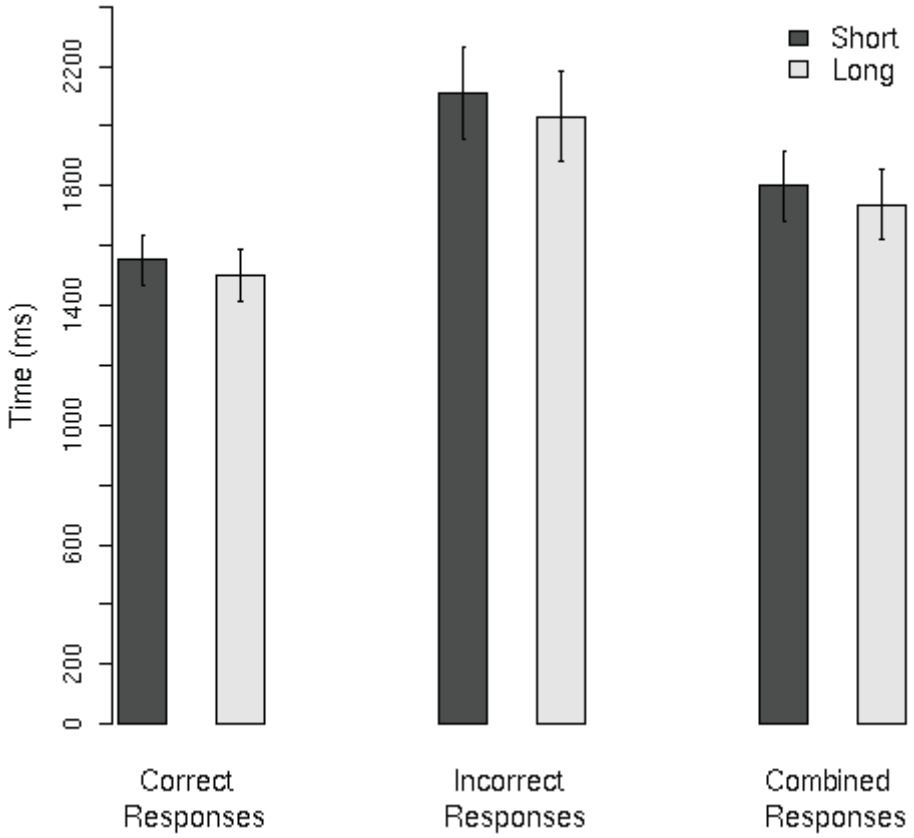

Figure 20. The mean of the median response latency for correct, incorrect and combined (correct and incorrect) responses for Experiment 3. Bars represent 95% within subjects confidence intervals.

6.1.3 Discussion

Results of Experiment 3 using word pairs as the stimuli were consistent with the findings for word pairs of Criss and Shiffrin (2004c), however they were in contrast to those of Clark and Hori (1995) and Nobel and Shiffrin (2001). No significant list length effect was identified in either the accuracy or response latency data. In addition, the Bayesian analysis

favoured the error-only model. The list length ratio in the present experiment was stronger than in the studies of Criss and Shiffrin, Clark and Hori, and Nobel and Shiffrin's Experiment 1. The larger list length ratio would make it more likely that a significant effect of list length would be identified but this was not the case. However, the increase in the list length ratio was perhaps offset by the introduction of controls for the four potential confounds.

It should also be acknowledged that using the retroactive design to control for retention interval may have had an influence on the results of this experiment. While the first eight pairs of each study list were included as targets at test, the distractor pairs were created using one word from each of the next 16 pairs meaning that they were more recently encountered than were the targets. This could perhaps have affected recognition performance. However, the situation was the same for both the short and long lists and would not have an impact on the list length finding.

The absence of a significant effect of list length suggests that recall-like processes cannot have a dominant role in the associative recognition paradigm contrary to the suggestions of Clark et al. (1993) and Nobel and Huber (1993).

The word frequency effect was identified in the d' and false alarm rate data with better performance for high frequency pairs, consistent with some previous research (Clark, 1992; Clark et al., 1993, Experiments 1 and 2; Clark & Shiffrin, 1992; Clark & Burchett, 1994). However, the effect was nonsignificant in the hit rate data consistent with other previous findings (Clark et al., 1993, Experiment 3; Hockley, 1992). The present word frequency effect results do little to resolve the nature of the effect in associative recognition.

The results of Experiment 3 suggest that, in terms of list length, word pairs behave in the same way as do words, although with poorer recognition performance. The effect of list length on performance is not significant for either of these stimuli.

6.2 Experiment 4: Faces

Experiment 4 used images of novel faces as the stimuli. While the structure of faces is overlearned in a similar way to words, these particular examples have never been encountered before. There is also no real meaning attached to images of novel faces and they are not unitised stimuli. Criss and Shiffrin's (2004c) associative recognition experiment also revealed that words behave differently to faces in that the list length effect was identified for the latter but was not significant for the former class of stimuli.

There has been much research devoted to identifying differences in the processing, encoding and retrieval of words and faces. It has been suggested that words and faces are processed in the same way, as a series of parts (Martelli, Majaj & Pelli, 2005). However, others have argued that faces are processed holistically while words are processed as a series of parts that together make up the whole (Farah, Wilson, Drain & Tanaka, 1998). Thus, a nonsignificant effect of list length when the stimuli are words does not necessarily mean that there will also be a nonsignificant effect when faces are used as the stimuli.

Chalmers (2005) noted that faces have nameable features (e.g. eyes, nose, lips) but that these features are common to every face making the stimuli difficult to describe in a unique way. This too is different to words which, even with different combinations of features (letters), combine to form a unique word which is easy to describe. When it comes to a recognition memory test, this may mean that there is more overlap in the encoding of faces than there is for words. As Jacoby and Dallas (1981) noted, the processing of verbal stimuli like words is more fluent than that of nonverbal stimuli, such as images of novel faces. Thus adding extra faces to a study list may result in greater interference than adding extra words and make it more likely that a significant list length effect will result for face

stimuli.

Alternatively, it has been argued that words and faces behave in a similar manner. Xu and Malmberg (2007) conducted an associative recognition study in which the type of stimuli used for the pairs was manipulated between subjects. The lists were either made up of word pairs, face pairs, pseudoword pairs or Chinese character pairs. Xu and Malmberg found that the pattern of results for words resembled that of faces, with pseudowords and Chinese characters behaving differently to both. They proposed that words and faces behave in the same way because they are more commonly encountered in everyday life than are pseudowords and Chinese characters (by non-Chinese speakers).

The aim of Experiment 4 was to investigate whether there was a significant effect of list length on recognition performance when novel faces were used as the stimuli.

6.2.1 Method

6.2.1.1 Participants. Forty first year Psychology students from the Ohio State University participated in this experiment. They received course credit for their participation.

6.2.1.2 Design. Length (short or long) was the only factor manipulated in this experiment and was a within subjects manipulation.

6.2.1.3 Materials. The stimuli in this experiment were 140 colour images of faces taken from the AR Face Database (Martinez & Benavente, 1998, see Figure 21 for examples and Appendix B for all stimuli). Half of the images were of males and half were females. All images were 460 x 460 pixels in size and were randomly assigned to lists with no image

appearing twice.

Figure 21. Examples of face stimuli from the AR Face Database used in Experiment 4. Half of the images were of females and half were of males.

6.2.1.4 Procedure. The procedure of Experiment 4 closely resembled those of the previous experiments presented in this thesis. The experiment was first described to participants who then completed a practice session of the sliding tile puzzle activity that would be used throughout the experiment. Participants completed two study lists, one short (20 items) and one long (80 items) with order counterbalanced across participants. Each face appeared on screen for 4000ms. Male faces made up half of each list and the other half were female faces. Test lists were comprised of 20 targets and 20 distractors. All faces were presented in the middle of the computer screen.

During study, participants were asked to rate the pleasantness of each image on a six point Likert scale (1: least pleasant, 6: most pleasant) by clicking on the appropriate number displayed below the image. They were instructed to make this rating while the image was being displayed on screen and to move on to and rate the next image should they miss giving

a rating in the allotted time.

The experiment had a retroactive design as a control for retention interval. The short list was followed by a four minute period of sliding tile puzzle filler and it was the first 20 face images from the long list that were included as targets at test.

Participants were given a 15 second warning before the start of the test list. Using the yes/no recognition paradigm, participants were instructed to respond "yes" if they recognised a face image from the study list and to respond "no" if they did not. Responses were recorded when the participant clicked on the appropriate button displayed on screen. There was no time limit for responding and there were no missing data.

Contextual reinstatement was encouraged following both lists by including an eight minute period of sliding tile puzzle before the onset of each test list, in addition to the puzzle as a control for retention interval.

6.2.2 Results

A one way repeated measures ANOVA yielded a statistically significant effect of list length on d' ($F(1,39) = 6.53$, $p = .01$, $\eta_p^2 = .14$; see Figure 22). This was driven by a significant effect of list length on the false alarm rate ($F(1, 39) = 12.16$, $p = .001$, $\eta_p^2 = .24$). However, the effect of list length on the hit rate was not statistically significant ($F(1,39) = .06$, $p = .81$, see Table 8 for hit and false alarm rate data). The Bayesian analysis again favoured the error-only model (.53, .14).

Oneway repeated measures ANOVAs yielded nonsignificant effects of list length on the median response latency for correct responses ($F(1, 39) = 2.45$, $p = .13$), incorrect responses ($F(1, 38) = .05$, $p = .83$) and correct and incorrect responses combined ($F(1, 39) =$

.30, p = .59, see Figure 23).

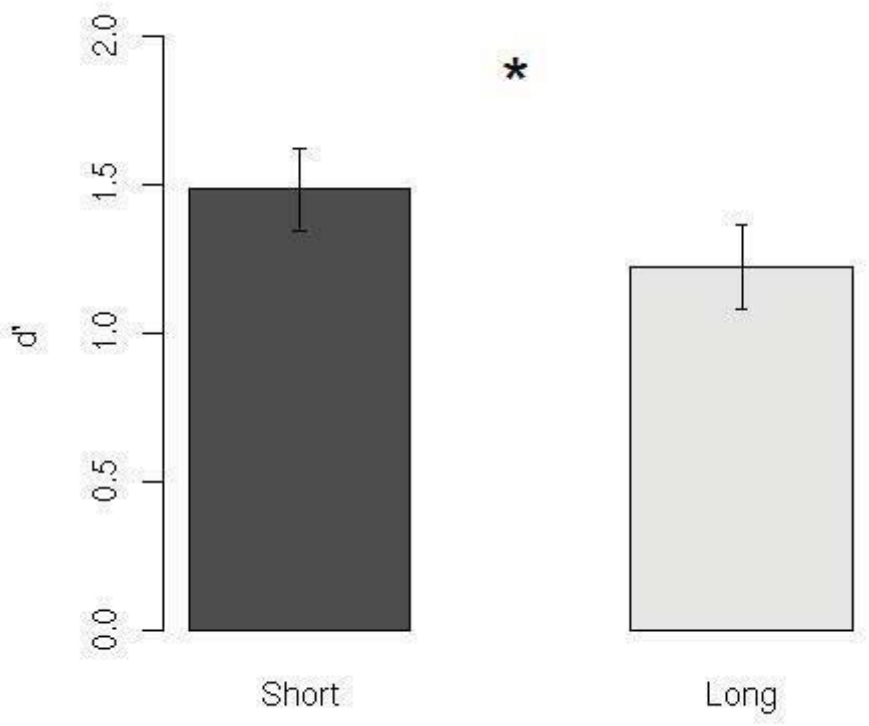

Figure 22. d' values for short and long lists in Experiment 4. Bars represent 95% within subjects confidence intervals.

Table 8

Mean hit and false alarm rates for the short and long lists in Experiment 4 (standard deviations in parentheses)

	Hit Rate	False Alarm Rate
Short List	.79 *(.15)*	.25 *(.16)*
Long List	.80 *(.14)*	.34 *(.17)*

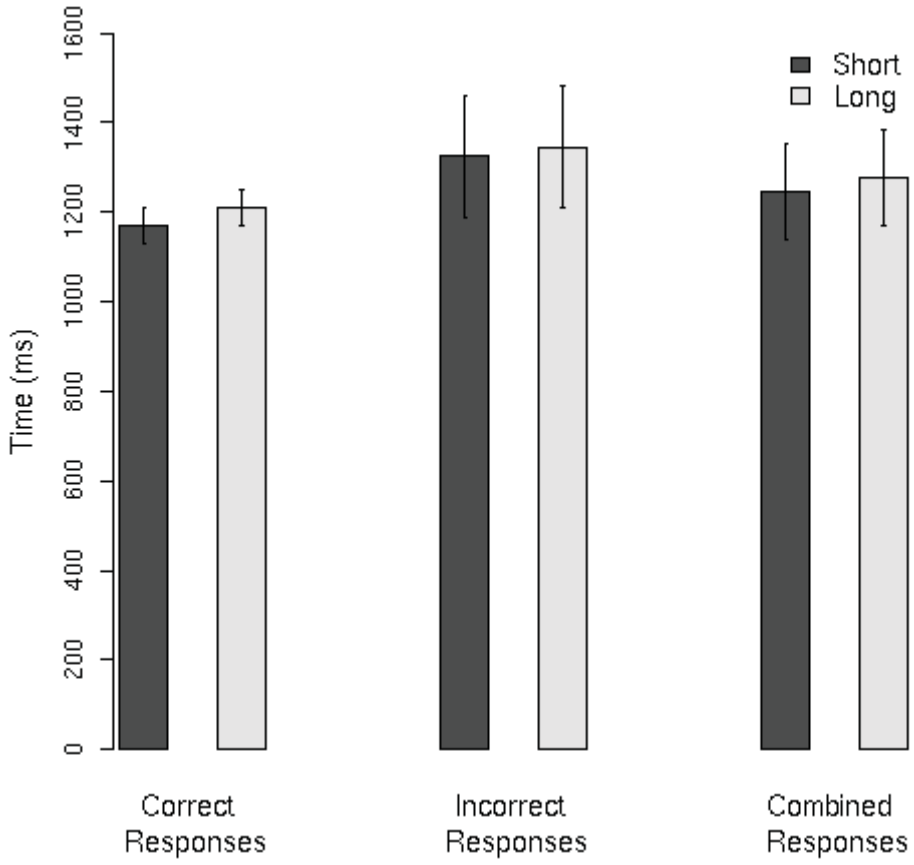

Figure 23. The mean of the median response latency for correct, incorrect and combined (correct and incorrect) responses for Experiment 4. Bars represent 95% within subjects confidence intervals.

6.2.3 Discussion

Consistent with the results of Criss and Shiffrin (2004c) a statistically significant effect of list length was identified on recognition performance when unfamiliar faces were used as the stimuli (but the Bayesian analysis favoured the error-only model). However, this contrasts with the nonsignificant effects of list length reported by Dennis and Humphreys (2001), Dennis et al. (2008) and the previous experiments in this thesis. It seems that

contrary to the work of Xu and Malmberg (2007) there is something different about words and faces. The different findings may be attributable to the possible differences in processing of faces and words, that is, holistically or as a combination of parts respectively (see Farah et al., 1998).

Alternatively, the hypothesis of Chalmers (2005), that there is a lack of unique ways in which one can describe and encode faces may explain the different results. This difficulty in encoding face stimuli may result in greater overlap in the representations of the faces that appeared at study versus a similar list comprised of words. Thus, the addition of more faces to a study list would lead to greater interference than would the addition of other words to the study list, and hence the significant list length effect finding in the present case. This greater interference may also be reflected in the false alarm rates which were high compared with those in the previous experiments with single words as the stimuli, consistent with the idea that "yes" responses are more probable when the stimuli are nonverbal (e.g. Greene, 2004; Whittlesea &Williams, 2000; but see Xu & Malmberg, 2007).

While the face stimuli in this experiment were unfamiliar, in that it is highly unlikely that a participant would have ever seen the faces in any other context, participants are certainly familiar with looking at faces in everyday life. Thus, there are many previous instances of witnessing faces of varying degrees of similarity to those presented at study. Therefore, while the majority of interference in this experiment was from other list items it is still possible that there was an influence of context noise, depending on the similarity of the experimental stimuli with those of real faces seen in everyday contexts. Consequently, the impact on the list length effect of a stimulus that is perhaps more unfamiliar and at least less often encountered than are faces and words, for example, images of fractals, which will be investigated next.

6.3 Experiment 5: Fractals

A fractal is "a geometrical figure in which an identical motif repeats itself on an ever diminishing scale" (Lauwerier, 1991, xi). Like faces, images of fractals contain several nameable features (e.g. certain colours, shapes, patterns) but these are also common to many other fractals and cannot be used to distinguish between different images of fractals with certainty. Further, as a consequence of being both a novel difficult to verbalise stimulus, the encoding of the fractals may be negatively affected by the lack of readily available appropriate labels with which to tag and encode the stimulus (Curran, Schacter, Norman & Galluccio, 1997; Gardiner & Java, 1990). With faces, there are common features which can be described in different ways, for example, big nose, blonde hair and blue eyes. With fractals, it is not as apparent what features there are to describe, for example, a certain pattern or shape may be present in some but not all fractal images, while a nose and mouth are present on every face. This may lead to a longer description of each image being employed and lead to greater overlap in encoding than is the case with words.

This difficulty in encoding, together with being less commonly encountered than faces, means that the fractal images may be subject to more interference in that they are more confusable with each other, than are images of faces, especially since recognition performance for faces is considered to be quite high (e.g. Bahrick, Bahrick & Wittlinger, 1975). Experiment 5 aimed to investigate whether the list length effect was evident when fractals were used as the stimuli.

6.3.1 Method

6.3.1.1 Participants. Participants in this experiment were 40 first year Psychology students at the Ohio State University. They each received course credit in exchange for their participation and gave informed consent.

6.3.1.2 Design. This experiment had a 2 x 2 factorial design with the factors being list length (short or long) and fractal type (circle or leaf). Both were within subjects manipulations.

6.3.1.3 Materials. The stimuli used in this experiment were 140 colour images of fractals, each 600 x 400 pixels in size. Half of the fractals were classed as circle fractals (see Figure 24 for examples) involving a circular shape as the centrepiece and the remainder were termed leaf fractals (see Figure 25 for examples and Appendix C for all stimuli) involving leaf-like shapes scattered across the image. All images were randomly assigned to lists with no image appearing twice, targets being the exception.

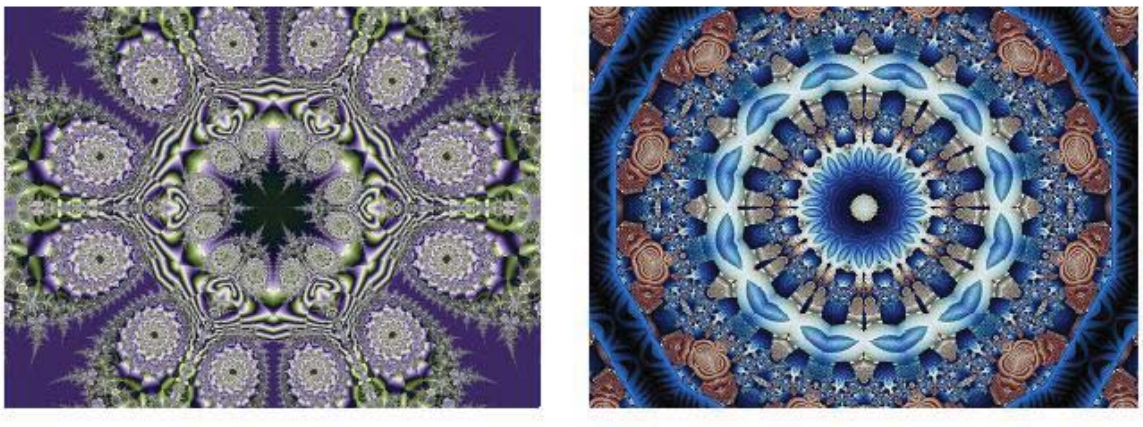

Figure 24. Two examples of ‘circle' fractals.

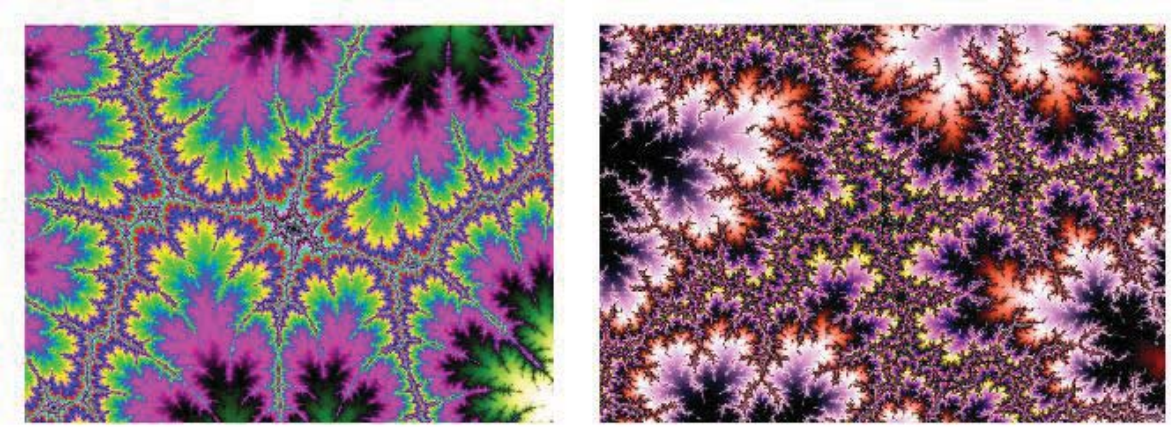

Figure 25. Two examples of 'leaf' fractals.

6.3.1.4 Procedure. The procedure for this experiment was identical to that of the previous experiment, with the only difference being the stimuli used. Each list was made up of an equal number of both circle and leaf fractals.

6.3.2 Results

A 2 x 2 (length x fractal type) repeated measures ANOVA did not yield a statistically significant effect of list length on d' ($F(1, 39) = 1.66$, $p = .21$, Figure 26) or the hit rate ($F(1,39) = 2.61$, $p = .11$), however there was a significant effect of list length on the false alarm rate ($F(1,39) = 10.86$, $p = .002$, $\eta_p^2 = .22$, see Table 9 for hit and false alarm rate data). The Bayesian analysis of list length data favoured the error-only model (.75, .08).

The 2 x 2 ANOVA analysis also yielded a statistically significant effect of fractal type (leaf vs. circle) on d' ($F(1,39) = 13.39$, $p < .001$, $\eta_p^2 = .26$) and the false alarm rate ($F(1,39) = 12.30$, $p = .001$, $\eta_p^2 = .24$). The effect of fractal type on the hit rate was nonsignificant ($F(1,39) = .34$, $p = .56$, see Table 9 for hit and false alarm data).

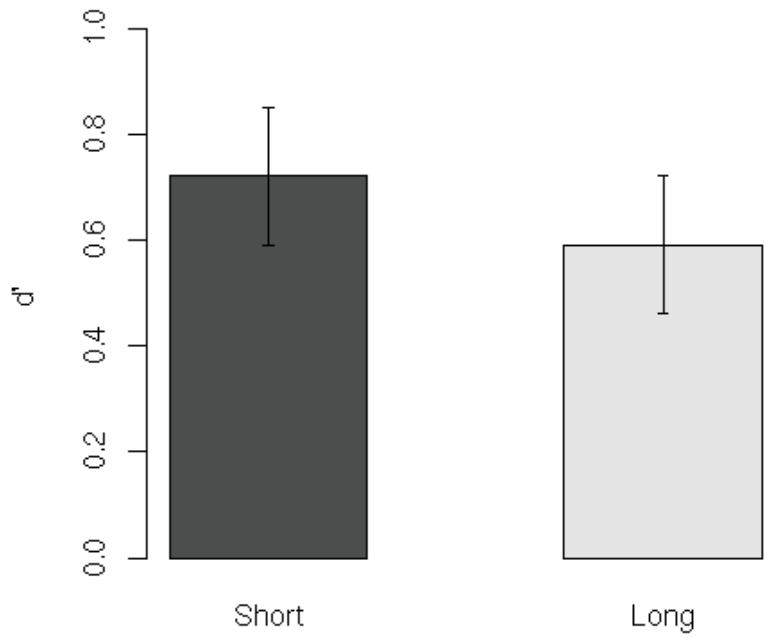

Figure 26. d' values for short and long lists in Experiment 5. Bars represent 95% within subjects confidence intervals.

Table 9

Mean hit and false alarm rates for fractal data in Experiment 5 (standard deviations in parentheses)

	Hit Rate		False Alarm Rate	
	Short List	Long List	Short List	Long List
Circle Fractals	.66 *(.21)*	.69 *(.22)*	.30 *(.21)*	.37 *(.25)*
Leaf Fractals	.62 *(.26)*	.69 *(.19)*	.40 *(.21)*	.54 *(.25)*

Analysis of the response latency data, however, produced different results. Oneway repeated measures ANOVAs yielded statistically significant effects of list length on the median response latency for correct responses ($F(1,39) = 17.85$, $p = .0001$, $\eta_p^2 = .31$), incorrect responses ($F(1, 39) = 24.29$, $p = 1.57e^{-5}$, $\eta_p^2 = .38$) and correct and incorrect responses combined ($F(1,39) = 25.57$, $p = 1.05e^{-5}$, $\eta_p^2 = .40$, see Figure 27).

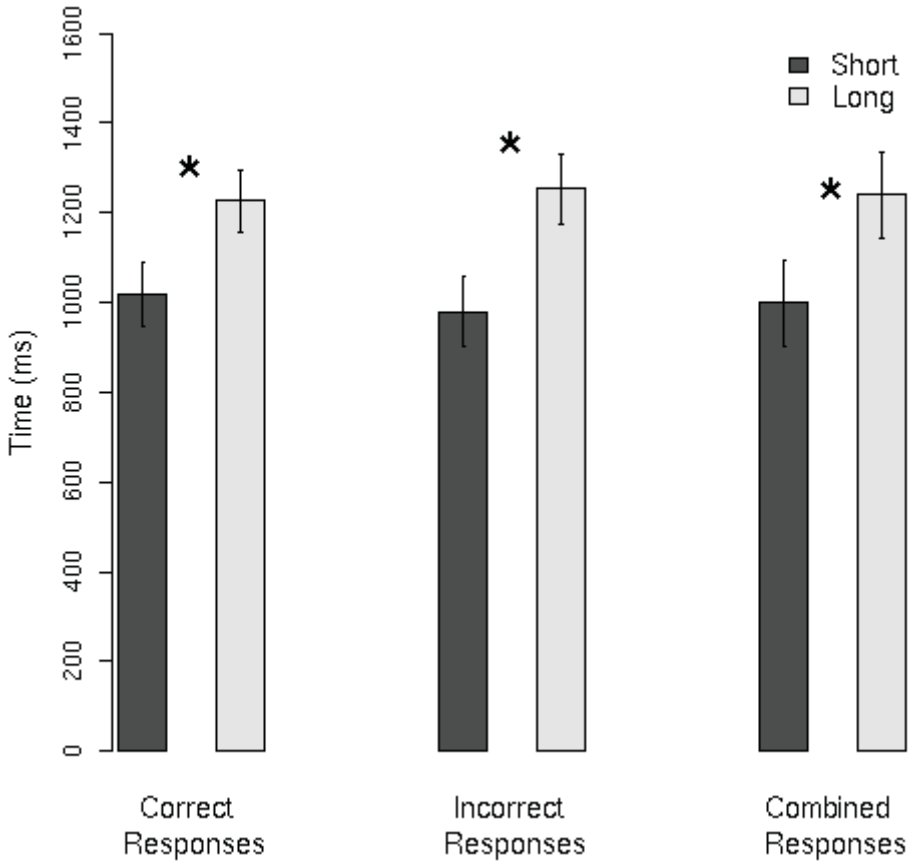

Figure 27. The mean of the median response latency for correct, incorrect and combined (correct and incorrect) responses for Experiment 5. Bars represent 95% within subjects confidence intervals.

6.3.3 Discussion

The results of the present experiment indicated a statistically significant effect of list length on the response latency data for correct, incorrect and all responses combined but this was not reflected in the accuracy data (with the exception of the false alarm rate). This discrepancy highlights the importance of analysing and reporting both sets of data. The Bayesian analysis was consistent with the results of the accuracy analysis and favoured the error-only model, suggesting no significant list length effect.

The findings of this experiment are consistent with those of Experiment 4 which identified a significant effect of list length when faces were used as the stimuli. In both experiments the stimuli were novel items meaning that the majority of the observed interference is item noise given that the specific stimulus items cannot have been seen in prior contexts.

Overall recognition performance in this experiment was low. This decreased performance was the likely result of overlapping representations of the fractal images. As Chalmers (2005) noted with faces, there is a shortage of unique ways in which complex stimuli of this type can be described and uniquely encoded into memory. When this is the case, the addition of extra items to the list has a more detrimental effect on performance, compared with word lists. As was the case with the faces in the previous experiment, the false alarm rates for fractals were higher than those in previous experiments of this thesis. This result is another example of a tendency for participants to provide more "yes" responses for nonverbal stimuli (Greene, 2004; Whittlesea & Williams, 2000; but see Xu & Malmberg, 2007).

6.4 Experiment 6: Photographs

Experimental results to this point suggest that, when controls for potential confounds are in place, no significant effect of list length is identified for single words or word pairs in an associative recognition paradigm. There is, however, a significant list length effect when novel faces and fractal images are used as the stimuli. The final experiment of this thesis was designed in an attempt to fall somewhere in between these two types of stimuli and to help establish the boundary conditions between the two.

Shepard (1967) looked at recognition performance on lists of words (540 items), sentences (612 items) and pictures (612 items). In a forced choice recognition paradigm involving 68 pairs, Shepard found that performance for all three types of stimuli was high and noted that discriminability was at its best when the "stimuli were meaningful, colored pictures" (1967, p. 159). This issue of meaningfulness of the stimuli was re-visited by Chalmers (2005) who noted that this may, in part, explain differences in performance on novel faces, very low frequency (novel) words, and pictures of complex scenes. Thus, colour photographs of complex scenes will be used as the stimuli in Experiment 6. These should be meaningful to the participants, visual in nature, rather than word-based, and elicit high recognition performance.

Photographs of this kind are also more nameable than the images of either faces or fractals and participants should be better able to encode them uniquely (Chalmers, 2005). This should also improve performance and allow us to ascertain whether it is the visual nature of faces and fractals or the difficulty with unique encoding that has resulted in the contradictory list length results for these stimuli with those of words.

6.4.1 Method

6.4.1.1 Participants. Forty first year Psychology students from the Ohio State University participated in this experiment in exchange for course credit.

6.4.1.2 Design. List length (short or long) was manipulated within subjects in this experiment.

6.4.1.3 Materials. Stimuli for this experiment were 140 different colour photographs of everyday scenes, for example images of a library interior, a beach and a classroom (see Figure 28 for examples and Appendix D for all stimuli). Each photograph was 800 x 600 pixels in size.

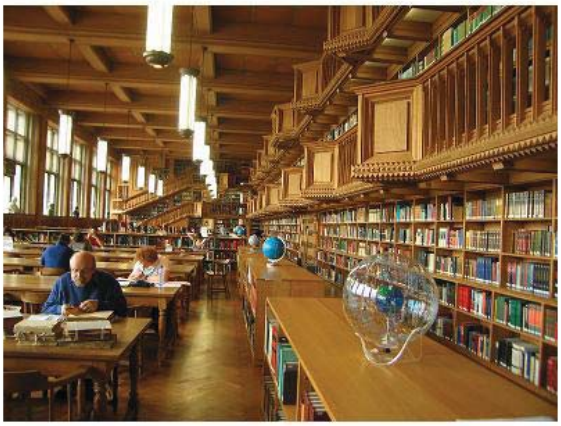

Figure 28. Examples of photographs used as stimuli in Experiment 6.

6.4.1.4 Procedure. The procedure for this experiment largely followed that of the previous two experiments. The main difference was in the duration of presentation of the

items at study. A pilot study was carried out in which photographs were presented for 3000ms at study as in the previous experiments, however performance was at ceiling. Therefore, in an effort to reduce performance, the rate of presentation of the stimuli at study was cut to 500ms with a 250ms inter-stimulus interval during which time the screen was blank. As a consequence of the shortened rate of presentation, it was not possible to request a response to an encoding task while the stimuli were on screen, as had been the case in all previous experiments, as a control for attention. Given the short overall duration of the present experiment, and the results of Experiment 1 which revealed that the encoding task does not alter the list length effect finding, this does not seem problematic in terms of the controls. This experiment used the retroactive experimental design with the first 20 items of the long list included as targets at test. There was an additional 45 seconds of filler activity following the short list to equate the retention interval with that of the long list. All other details were as in Experiments 4 and 5.

6.4.2 Results

A oneway ANOVA did not yield a significant effect of list length on recognition performance as measured by d' ($F(1,39) = .77$, $p = .39$, see Figure 29). The effect of list length on the hit rate ($F(1,39) = 2.09$, $p = .16$) and false alarm rate was also nonsignificant ($F(1,39) = .01$, $p = .91$, see Table 10 for hit and false alarm rates). Note that the long list d' and hit rate were superior to that of the short list. The nonsignificant effects of list length were also reflected in the Bayesian analysis, which found in favour of the error-only model (.78, .06).

Table 10

Mean hit and false alarm rates for the short and long lists in Experiment 6 (standard deviations in parentheses)

	Hit Rate	False Alarm Rate
Short List	.80 *(.16)*	.10 *(.12)*
Long List	.84 *(.14)*	.10 *(.12)*

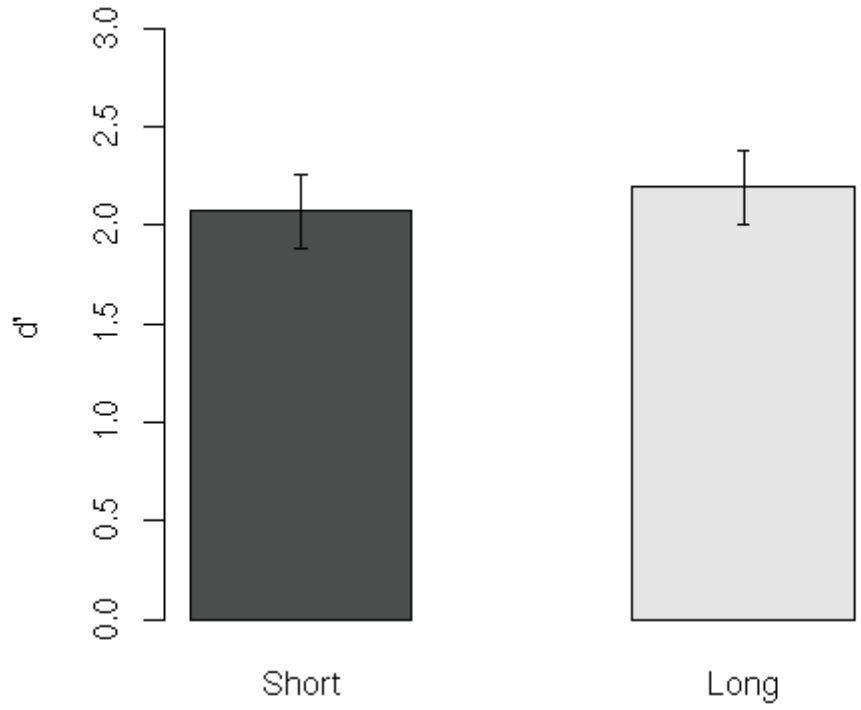

Figure 29. d' values for short and long lists in Experiment 6. Bars represent 95% within subjects confidence intervals.

There was also no significant effect of list length on the median response latency for correct responses ($F(1, 39) = 2.20$, $p = .15$), incorrect responses ($F(1, 35) = 2.56$, $p = .12$) and

both correct and incorrect responses combined ($F(1,39) = 2.71$, $p = .11$, see Figure 30) as revealed by oneway repeated measures ANOVAs.

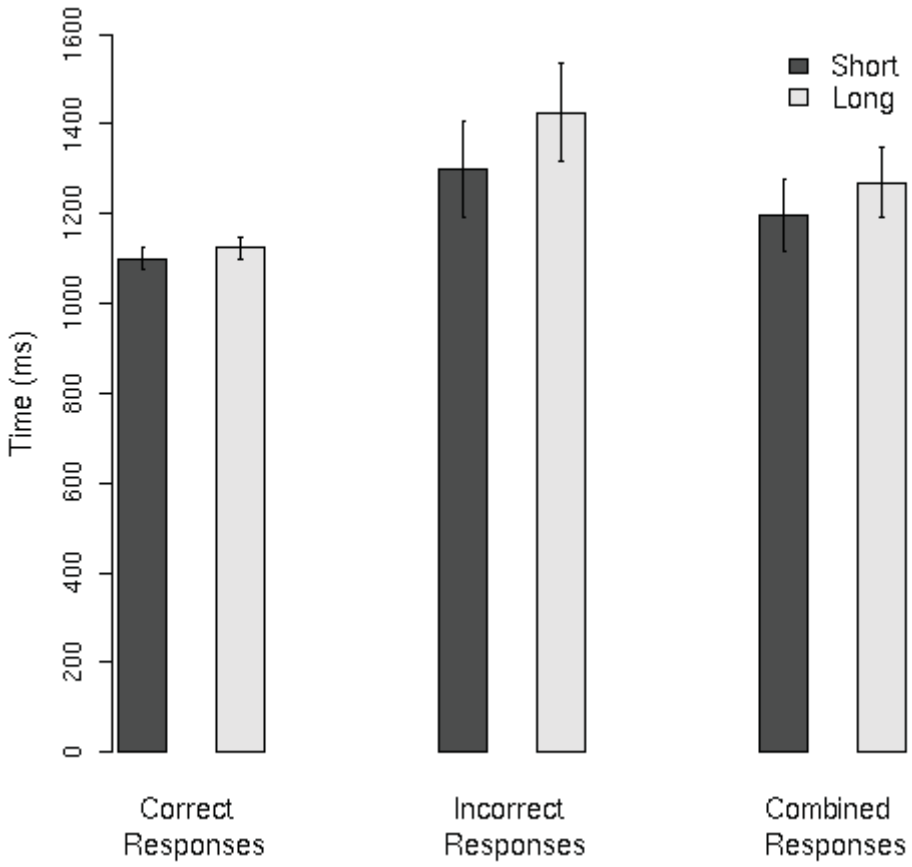

Figure 30. The mean of the median response latency for correct, incorrect and combined (correct and incorrect) responses for Experiment 6. Bars represent 95% within subjects confidence intervals.

6.4.3 Discussion

Following the pilot study in which performance was at ceiling, the reduced presentation time of the photographic stimuli in Experiment 6 successfully brought about

deterioration in recognition performance. No significant effect of list length was identified in this experiment in either the accuracy or response latency data and the Bayesian analysis also favoured the error-only model.

Contrary to the images of faces and fractals that were the stimuli in the previous two experiments, the photographic images in the present experiment were straightforward to describe and to name. There was no overlap in the scenes depicted in the photographs and these could be described in most cases using just one word, for example, “library”, “beach” and “classroom”. In this way, the photographic images were similar to words, and this may explain why the list length findings for the two classes of stimuli follow the same pattern. In addition, the false alarm rate in the present experiment was similar to those for the word stimuli in Experiments 1 to 4 and lower than the false alarm rate when faces and fractals were the stimuli. This finding suggests that despite the photographs technically being a nonverbal stimulus like faces and fractals, they are behaving as if they were verbal, suggesting that participants may have been basing their decisions on the verbal labels given to the photographs (Greene, 2004; Whittlesea & Williams, 2000; but see Xu & Malmberg, 2007).

6.5 General Discussion/Conclusions

In an attempt to identify and define the boundary conditions of the list length effect, a series of four experiments was conducted, each involving recognition testing of a different stimulus; word pairs, faces, fractals and photographs. Controls for the potentially confounding effects of retention interval, attention, displaced rehearsal and contextual reinstatement were implemented in each case with the general format of the method following that of Experiments 1 and 2. The results varied across experiments. No significant

effect of list length was identified on recognition performance for word pairs or photographs in either the accuracy or response latency data. There was a significant list length effect on recognition performance for faces based on the accuracy data, while the response latency data for fractals showed a statistically significant effect of list length that was absent in the accuracy analysis. The Bayesian analysis suggested no significant effect of list length under any conditions.

One explanation for the different results depending on the stimuli used is the differences in the ease with which the stimuli can be named or labelled, which may also be related to the similarity of the stimulus items to others of the same class. Words are themselves a unique label for the stimulus, and the photographic images, while complex, can be easily and uniquely described with a single word, making them a verbal stimulus. This ease of labelling reduces the similarity within the types of stimuli. Labelling was more difficult for the nonverbal face and fractal stimuli in which there was necessarily more overlap in feature descriptions, especially with greater similarity within the stimuli class. Thus, the more of these items that are on the study list, the more problematic it is for the participant to create unique labels for the items. This was reflected in the false alarm rates which were higher for faces and fractals than any other stimulus, consistent with the suggestions of Greene (2004) and Whittlesea and Williams (2000) (but see Xu and Malmberg, 1997). It may be that the different types of stimuli lie on a continuum, from the nonverbal, difficult to label and more similar faces and fractals, through photographs to words which are verbal stimuli.

The suggestion of Xu and Malmberg (2007), that it may be that the experience one has in encountering the particular type of stimuli which influences the results, does not explain the results of the experiments presented here. If that were the case, one would expect

the same effect of list length on the often encountered words, word pairs, faces and photographs, and for these results to contrast with those obtained when fractals were the stimuli. The results showed otherwise.

The results of the four experiments presented in this chapter have serious consequences for mathematical models of recognition memory and both the item and context noise approaches to interference. The findings indicate that despite controls for potential confounds being implemented in all cases, there is a significant list length effect in recognition memory for certain stimuli and a nonsignificant effect when other stimuli are used. REM, Minerva II and other item noise models are able to predict the significant list length effect that is evident for faces and fractals, however, they cannot produce the nonsignificant effect for words, word pairs and photographs. Conversely, context noise models such as BCDMEM can produce the nonsignificant effect of list length for words, word pairs and photographs but are unable to predict the significant effect for faces and fractals. The solution may lie with combined interference models such as SAC, or the modification to REM (Criss & Shiffrin, 2004a).

In terms of identifying the boundary conditions of the list length effect, the four experiments presented in this chapter have been a good starting point for this investigation. The results revealed that the stimulus used in a particular experiment was influential to the eventual outcome regarding list length, with a significant effect identified in some cases and not in others. The explanation for this seems to lie in the intra-stimulus similarity and the ease with which the stimuli can be uniquely named and described.

Chapter 7

Another potential confound of the list length effect?

All experiments reported in the present thesis have used a within subjects design. This was done to ensure that the experiments had the greatest experimental power possible so that any failure to find a list length effect was not a consequence of lack of power. In addition, the within subjects design allowed for consistency with the studies of Dennis and Humphreys (2001) and Cary and Reder (2003) who had also employed this design.

The problem with the within subjects design stems from the fact that by the end of the experimental session, each participant will have studied and been tested on both a short and a long list. All participants will have studied 100 items which may result in an elimination of the list length manipulation by the end of the experimental session despite the counterbalancing of list order across participants. In part, this relates to the concept of a list. While the experimenter deems there to be two discrete lists as part of the within subjects design, the experience of studying lists at all could be so novel to the participants that they fail to adequately isolate the list contexts from each other.

The inclusion of list order as a factor in the within subjects analysis revealed that it was indeed an issue. There was a significant interaction between list length and list order on d' in the four conditions of Experiment 1 (Retroactive Pleasantness - $F(1,38) = 5.63$, $p = .02$, $\eta_p^2 = .13$, Retroactive Read – $F(1,38) = 10.78$, $p = .002$, $\eta_p^2 = .22$, Proactive Pleasantness – $F(1,38) = 7.93$, $p = .01$, $\eta_p^2 = .17$, Proactive Read – $F(1,38) = 9.06$, $p = .01$, $\eta_p^2 = .19$) . The interaction was also significant for the face ($F(1,38) = 15.29$, $p = .0004$, $\eta_p^2 = .29$), fractal ($F(1,38) = 4.63$, $p = .04$, $\eta_p^2 = .11$) and photograph stimuli ($F(1,38) = 20.83$, $p = 5.13e^{-5}$, η_p^2

= .35). The interaction was nonsignificant for both conditions of Experiment 2 (Yes / No Task – $F(1,38) = 2.05$, $p = .16$, RK Task – $F(1,38) = 1.82$, $p = .19$) and the word pairs in Experiment 3 ($F(1,38) = 1.94$, $p = .17$).

Despite the fact that participants were informed of the study test procedure before beginning the experiment, the order effects in the within subjects design may be explained by the fact that the data for the first list studied came from a naive participant while the data for the second list studied came from a participant who had already experienced the experimental design. When list length was manipulated within subjects, a participant who studied the long list first may then have had an expectation that the second list would be of equal length. This expectation may have influenced the study strategy employed by that participant in the second list, for example, they may have prepared to spread their attention over a longer series of items and not pay as much attention to the short list items as a participant who studied the short list as their first list.

Similarly, a participant who viewed a short list to begin with may have had an expectation that the second list would be equally as short and attend to the items in the same manner. In the retroactive condition in particular, this attention at the start of the long list would have been focussed on precisely the items that would be tested which may also have positively influence performance on the long list when it was viewed second. Thus, collapsing long and short lists regardless of their order may have averaged out performance on the lists such that no significant effect of list length was identified overall.

Further, a participant who viewed the short list second would have already seen 80 long study items as well as 40 test items prior to the start of the second (short) list. Consequently, this participant may have been bored prior to the start of the second list and therefore have been less likely to pay attention to the short study list than if it had been

viewed first. This would favour performance on the long list when it was viewed second.

7.1 A Between Subjects Re-analysis of the Data

7.1.1 The Effect of Re-analysis on the Statistical Significance of the Results

One way to check if the within subjects design confounded the results of the experiments in the present thesis and led to the nonsignificant findings identified is to use only the data from the first list studied by each participant in the analysis and determine whether the conclusions hold. In this re-analysis, the data would essentially come from a between subjects design.

Table 11 shows both the original within subjects ANOVA results (reported in the earlier chapters) and the between subjects re-analysis of the effect of list length on d' for each of the major conditions in the six experiments of this thesis (hit rate and false alarm rate analyses and re-analyses are available in Appendix E). What is immediately evident is that the ANOVA results fluctuate with the re-analysis due to the reduction in the number of data points per cell and the consequential loss of experimental power, which will be discussed in greater detail below. The conclusions that would be drawn based on the ANOVA results of the accuracy data, however, remain relatively unchanged, with the exception of Experiments 4 and 5 which involved faces and fractals as the stimuli and the Retroactive Pleasantness condition of Experiment 1. The conclusions drawn from the word frequency and response latency data do not change as a result of the between subjects re-analysis (see Appendix E for this data).

Table 11

ANOVA results revealing the effect of list length on d' for all experiments using both the within subjects and between subjects (first list only) analyses. Results that are statistically significant are marked with an asterisk(*). Grey shading is used to indicate the results about which the conclusions drawn changed depending on the analysis.

	Within Subjects Analysis	Between Subjects Analysis
Experiment 1 – Attention		
Retroactive Pleasantness	$F(1, 39) = .38$, $p = .54$	$F(1,38) = 3.43$, $p = .07$
Retroactive Read	$F(1,39) = 3.06$, $p = .09$	$F(1,38) = .01$, $p = .93$
Proactive Pleasantness	$F(1,39) = 11.55$, $p = .002$*	$F(1,38) = 17.13$, $p = .0002$*
Proactive Read	$F(1,39) = 8.26$, $p = .007$*	$F(1,38) = 21.58$, $p = .00004$*
Experiment 2 – The Remember Know Task		
Yes/No Instructions	$F(1,39) = .43$, $p = .51$	$F(1,38) = .05$, $p = .82$
RK Instructions	$F(1,39) = 12.21$, $p = .28$	$F(1,38) = 1.29$, $p = .26$
Experiment 3 – Word Pairs		
	$F(1,39) = 2.15$, $p = .15$	$F(1,38) = 3.39$, $p = .07$
Experiment 4 – Faces		
	$F(1,39) = 6.53$, $p = .01$*	$F(1,38) = 3.53$, $p = .07$
Experiment 5 – Fractals		
	$F(1,39) = 1.66$, $p = .21$	$F(1,38) = 8.25$, $p = .007$*
Experiment 6 – Photographs		
	$F(1,39) = .77$, $p = .39$	$F(1,38) = .25$, $p = .62$

In Experiment 4, the effect of list length on recognition performance (d') for faces went from strongly significant in the within subjects analysis to nonsignificant but arguably marginally significant when just the first list studied was analysed. This change means that the results support the claim of Xu and Malmberg (2007) who suggested that words and faces

behave in the same way by virtue of the fact that they are both commonly encountered in everyday life. Using the between subjects re-analysis, there is no significant effect of list length identified when either words or faces are used as the stimuli.

The results of Experiment 5 changed in the opposite direction, with the effect of list length on recognition performance for fractals changing from nonsignificant to strongly significant in the between subjects analysis. However, note that in Experiment 5 there was a significant effect of list length on median response time when fractals were the stimuli. This effect was maintained in the between subjects re-analysis. Therefore, the conclusions drawn from Experiment 5 regarding the list length effect have not changed.

The other change of note was in the Retroactive Pleasantness condition of Experiment 1. Bearing in mind that this is a condition in which the four potential confounds of Dennis and Humphreys (2001) were controlled, the change in significance level from nonsignificant to $p = .07$, and arguably marginally significant, is of considerable importance. The effect on d' in the between subjects re-analysis was driven by a strongly significant effect of list length on the false alarm rate in this condition, however this was offset by an increase in the hit rate for the long lists, resulting in the nonsignificant d'.

One could argue that since the between subjects analysis involves a decrease in experimental power, running more participants in the retroactive condition may have led to further reduction of the *p* value and statistical significance in this condition. However, Dennis et al. (2008), in their summary of problems with traditional methods of null hypothesis significance testing, raised the issue that these methods, including ANOVA, are sensitive to the effect of outliers. This is heightened in the present case where there is a small sample size. When the results of the Retroactive Pleasantness condition were re-analysed using data from the first list studied only and after the removal of the most extreme outlier,

the effect of list length on d' was nonsignificant ($F(1,37) = 2.46$, $p = .13$), however the effect on the false alarm rate was maintained. The Bayesian analysis of Dennis et al. (2008) may be useful in this case if expanded to allow for between subjects designs as the method should be less affected by the removal of outliers.

7.1.2 Differences in Effect Size Between the Analyses

While the switch to the between subjects analysis affected the conclusions of only a minority of conditions, the effect sizes for the differences in list length did change, and were larger in the between subjects analysis. Figure 31 illustrates the results for each of the conditions across the six experiments presented in this thesis based on the within subjects analysis (Figure 32 shows the same results based on the between subjects analysis). Each bar represents the difference in d' for each condition as measured by short list performance minus long list performance. Thus, positive values indicate that performance on the short list was superior to that of the long list and negative values indicate the opposite. Numbers above each bar are the partial eta squared effect sizes for each effect.

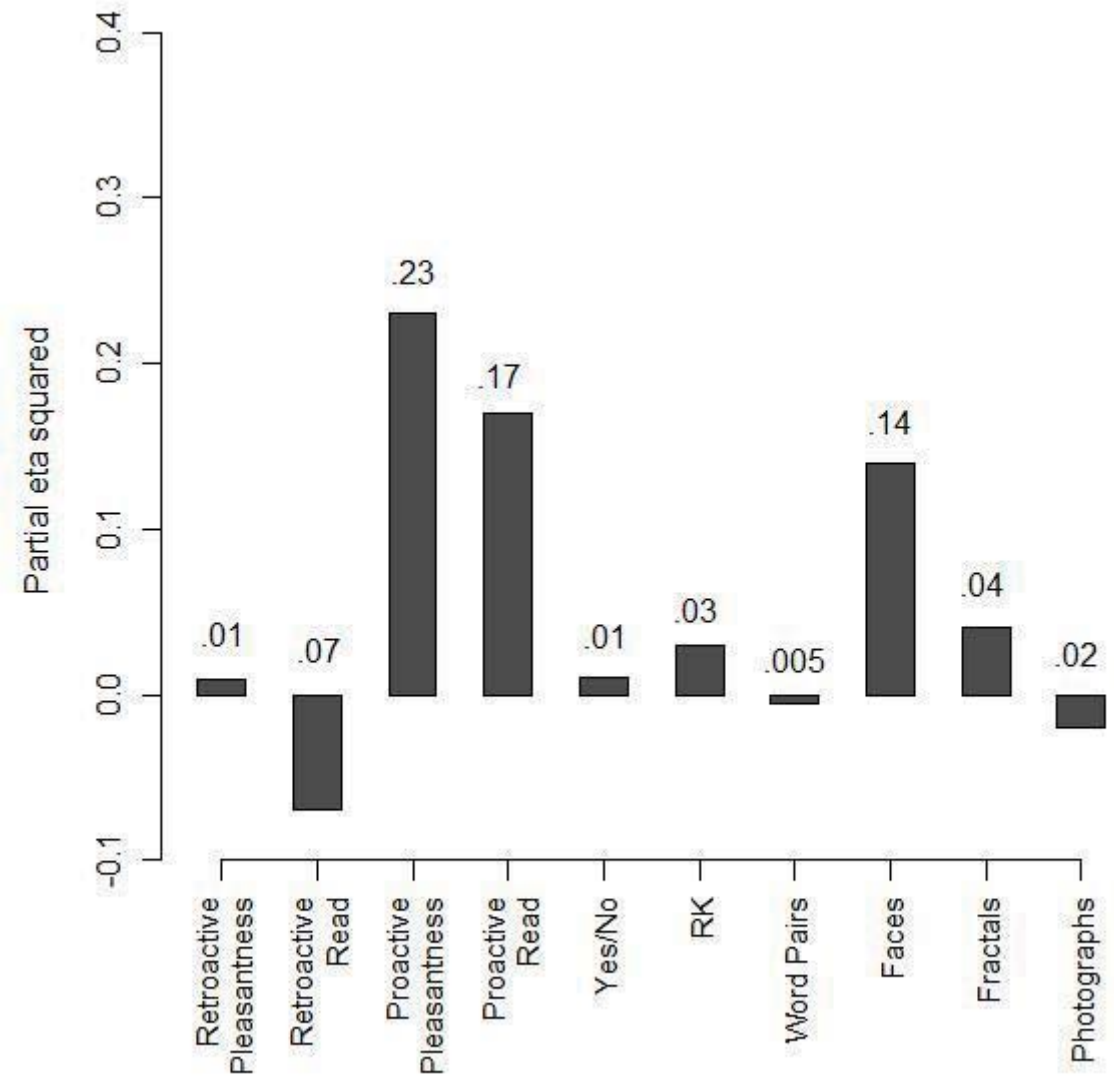

Figure 31. The partial eta squared effect sizes for each of the conditions of all experiments in the present thesis. The results in this figure are based on the within subjects analysis of the data.

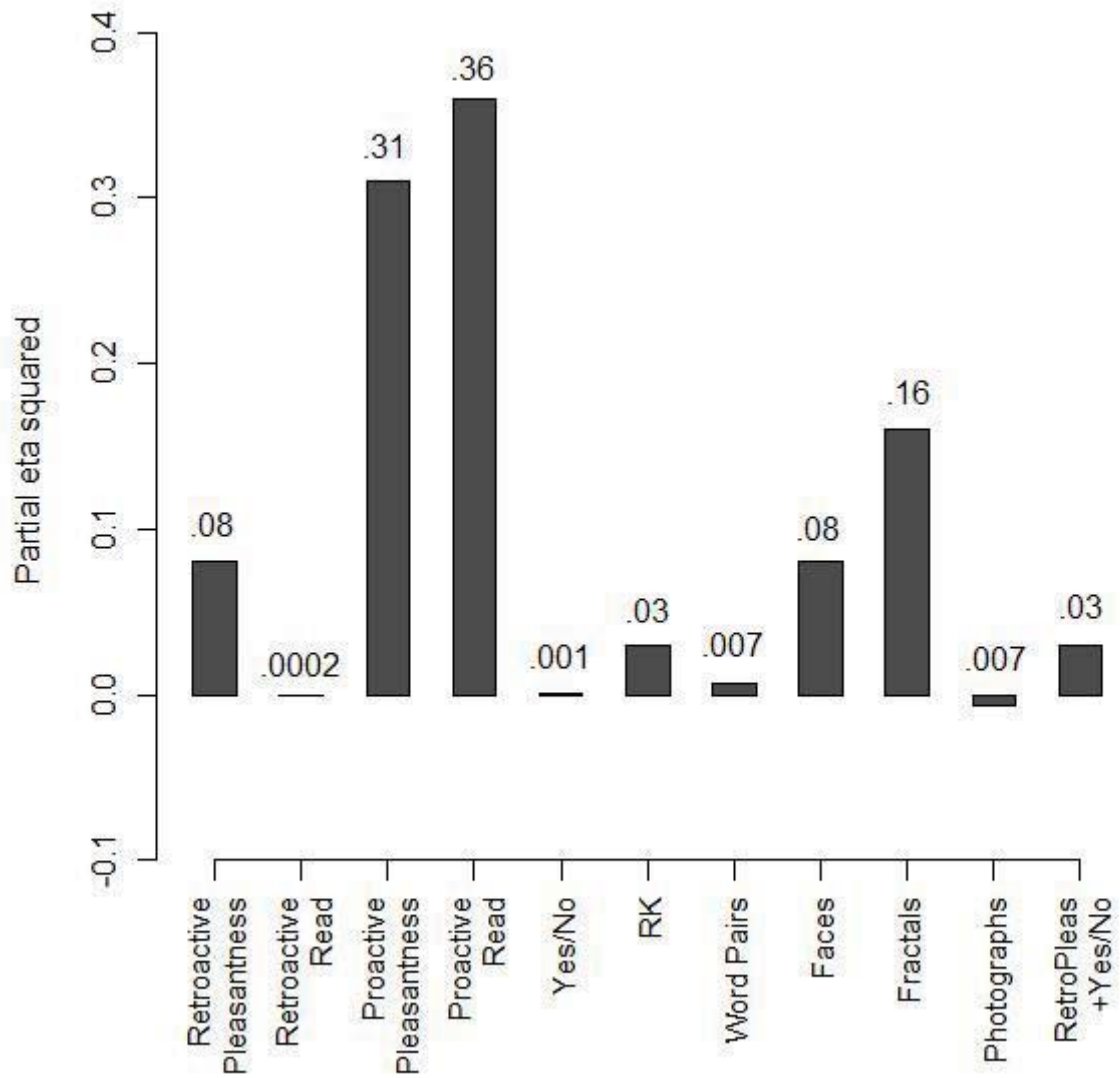

Figure 32. The partial eta squared effect sizes for each of the conditions of all experiments in the present thesis. The results in this figure are based on the between subjects analysis of the data.

Figures 31 and 32 illustrate the change in the magnitude of the list length effect in each condition with the change in the analysis, with the effect more pronounced in the between subjects analysis. The effect sizes are generally larger in the between subjects

analysis than those in the within subjects design. Cohen (1988), with reference to the magnitude of correlations (*r*), recommended that an effect size of .10 should be considered small, .30 medium and .50 large. Using this as a frame of reference, the effect size in the proactive conditions could be considered medium in both analyses (mean $\eta_p^2 = .20$ in the within subjects analysis and mean $\eta_p^2 = .34$ in the between subjects analysis). For the retroactive conditions in which words are the stimuli (the Retroactive Pleasantness condition, the Retroactive Read condition, the Yes/No Task condition and the RK task condition), the mean effect size failed to reach Cohen's description of a small effect and can be considered negligible in both the within subjects and between subjects analyses (mean $\eta_p^2 = .03$ in each). The effect sizes for word pairs ($\eta_p^2 = .01$ within subjects and $\eta_p^2 = .07$ for between subjects) and photographs ($\eta_p^2 = .02$ within subjects and $\eta_p^2 = .01$ for between subjects) were similarly very small. The effect sizes for the Faces (mean $\eta_p^2 = .14$ within subjects and $\eta_p^2 = .08$ between subjects) and Fractals (mean $\eta_p^2 = .04$ within subjects and $\eta_p^2 = .16$ between subjects) stimuli fell in between these two groups but may still be considered small effects. Note that all of these effect sizes, with the exception of the proactive conditions, are small, especially when we consider that they are based upon a fourfold increase in list length. If item noise plays a substantial role in recognition memory, a larger effect would be expected given the four times increase in interference. For comparison, the mean word frequency effect size for each of the conditions in Experiments 1 and 2 was large (mean $\eta_p^2 = .52$ within subjects) or medium-large (mean $\eta_p^2 = .37$ between subjects) depending on the analysis.

However, regardless of the analysis performed, there are several conditions in which performance on a long list is better than that on a short list; the Retroactive Read condition of Experiment 1 and for the photograph stimuli in Experiment 6. For the word pair stimuli in

Experiment 4, long list performance was superior to short list performance only in the within subjects analysis. It is also clear from the graphs that in the rest of the conditions, the mean of the short list is greater than the mean of the long list, however, this difference was only statistically significant for both of the proactive conditions in Experiment 1, face stimuli in Experiment 4 (for the within subjects analysis only, although bear in mind that there was a significant effect of list length on the median response latency in this condition) and for fractals in Experiment 5 (for the between subjects analysis only).

7.1.3 Analysis of the Second Lists Studied Only

The previous analysis of the first lists studied provides the purest list length manipulation available. When only the first list studied is analysed as a between subjects design, short list performance is generally better than long list performance. Conversely, inspection of performance on the second list studied revealed superior performance on the long list than the short list (however, none of these effects were statistically significant) which can explain the averaged results in the within subjects design. Figure 33 shows the difference in d' for each of the conditions based on the second list viewed by each participant. In this case it is clear that long list performance was superior to that of the short list in the majority of conditions, including the proactive conditions.

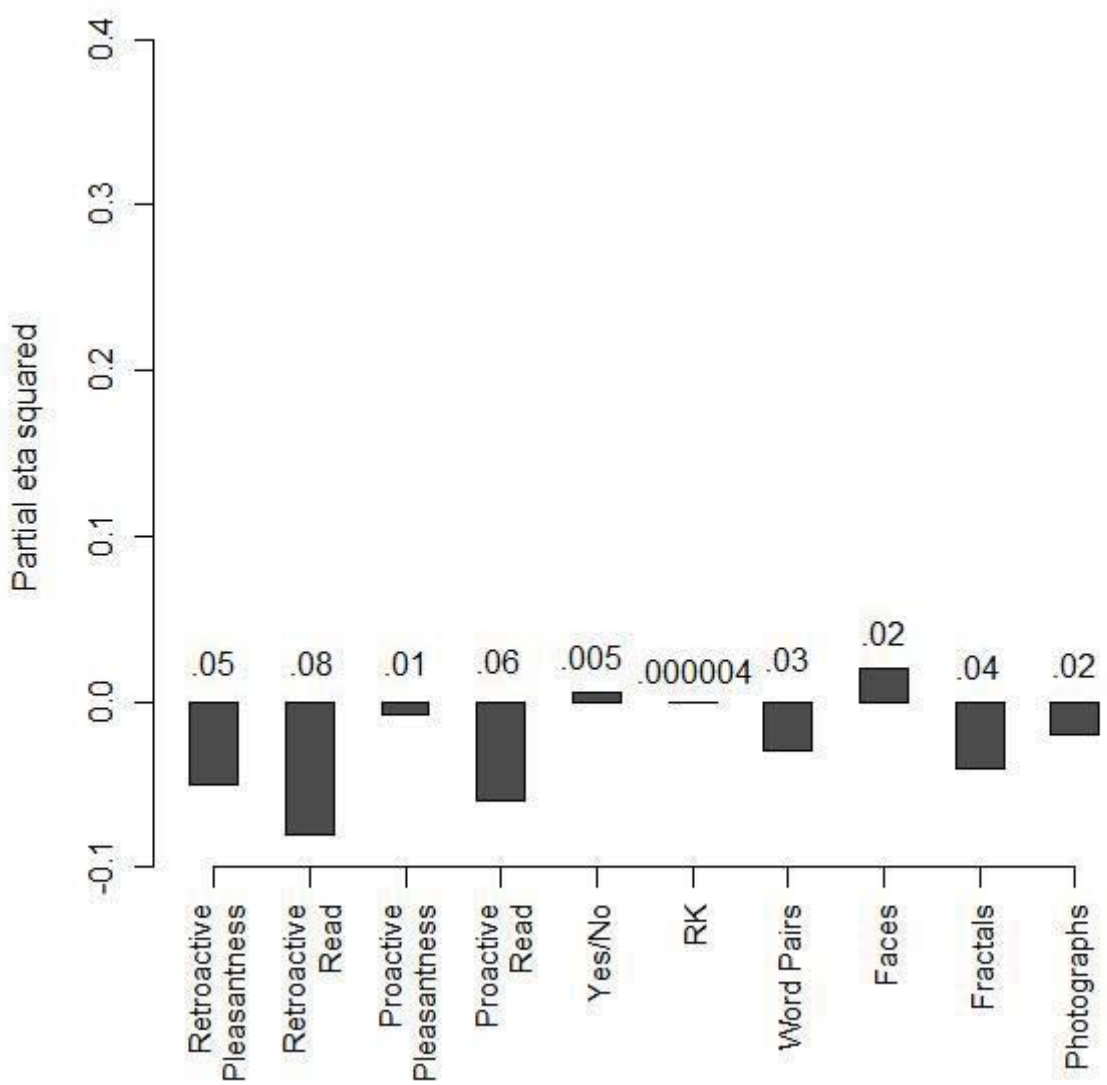

Figure 33. The partial eta squared effect sizes for each of the conditions of all experiments in the present thesis. The results in this figure are based on the second list that was studied by each participant in a between subjects analysis of the data.

One explanation for this reversal in performance based on list order, which would explain the poorer performance for short lists when viewed second, is attention. Prior to the second list, the participant would have already paid attention to an 80 item long list and

completed a test list. They may already have been bored before the onset of the short list and consequently have paid less attention to the short list items than another participant who began with this list. This reversal which sees better recognition performance for long list items than short list items when viewed second, in the within subjects design, cannot be accounted for by greater interference. At the start of the second test list, regardless of the order in which they viewed the study lists, all participants will have seen an equal number of items and been subject to the same amount of interference regardless of list length. When the two lists are combined as in the within subjects analysis, the result is an averaging of the order effect and minimal difference between the list lengths.

In addition, regardless of list length, performance on the first list studied was superior to performance on the second list studied. This finding suggests that it may not be the list length manipulation that is affecting performance but, rather, may be an issue of list identification or discrimination and appropriate reinstatement of the study list context. The second list is not isolated in memory, instead, performance is based on a blend of the two study lists. While the experimenter has deemed there to be two separate lists that have been studied, the experience of studying a list of items for a memory test is likely to be a unique event in a participant's life and they may be basing decisions on the entire experimental context rather than the specific study list in question.

7.1.4 Combining the Results From Previous Experiments as a Between Subjects Analysis

The re-analysis of the data from the first list of the within subjects experiments as a between subjects analysis is not ideal. As has already been noted, the switch to a between

subjects design necessarily comes at the cost of a loss of experimental power. When nonsignificant effects are a matter of interest this is of great consequence. However the results of the re-analysis revealed some strongly significant effects, suggesting that the analysis did have sufficient power to detect effects. However these were generally in conditions in which the controls for the confounds were relaxed and a significant list length effect was predicted, for example in the proactive conditions of Experiment 1. In the fully controlled conditions, such as the Retroactive Pleasantness condition of Experiment 1 and the Yes/No Task condition of Experiment 2, the effect size would be expected to be much smaller with the controls implemented (as was the case when comparing Cary and Reder's three experiments, Figure 8) and it is not possible to say for certain that the present experiments, when using the between subjects analysis, had sufficient power to detect a significant effect should one exist.

Further, analysing the experiments as between subjects experiments meant halving the sample size for long and short list data from 40 participants to 20 participants. With such a small sample size, the effect of outliers on the data is accentuated (Dennis et al., 2008) and could influence the overall outcome, as was the case in the Retroactive Pleasantness condition. However, determining the required number of participants is also problematic. Conducting an a priori power analysis to calculate the appropriate sample size required is made difficult by the question of what is an appropriate effect size to use in the calculations. Part of this problem stems from the lack of effect size reporting in the literature in general (the exceptions being Cary & Reder, 2003; Clark & Hori, 1995; Criss & Shiffrin, 2004c; Murnane & Shiffrin, 1991; Nobel & Shiffrin, 2001; Ohrt & Gronlund, 1999) and part of it also lies in the fact that many of the previous studies which have investigated the list length effect have used a within subjects design as the experiments presented here had done. When

a power analysis was conducted using the effect size obtained from the within subjects analysis of the Retroactive Pleasantness condition in Experiment 1, with an alpha level of .05 and beta value of .95, the suggested sample size was in excess of 1200 participants. There is also a logical issue with attempting to calculate the power required to detect an effect that some (e.g. Dennis & Humphreys, 2001, Dennis et al., 2008) argue does not exist. The smaller the effect is expected to be, the more extreme the required sample size becomes as a consequence.

One solution to this problem, save for running a new between subjects experiment with a seemingly unfeasible number of participants, is to increase the sample size using data already obtained. The control condition in each of the six experiments of this thesis involved the same controls being implemented for the four potential confounds of Dennis and Humphreys (2001). In effect, this means that the Retroactive Pleasantness condition of Experiment 1 and the Yes/No Task condition of Experiment 2 were identical to one another. Both of these conditions used words as the stimuli, followed the retroactive design, required pleasantness ratings to be given for each item at study, used the same sliding tile puzzle as a filler task, had an eight minute period of puzzle activity to encourage contextual reinstatement and used the yes/no recognition paradigm. It was therefore feasible to combine the results of these two conditions and double the sample size for the between subjects analysis and, consequently, increase the experimental power. In both cases, the data were taken from conditions in which all potential list length confounds were controlled, including the use of the between subjects design and provides the best example of what happens when all potential confounds are controlled. When the data from these two control conditions were analysed, there was a nonsignificant effect of list length on d' ($F(1,78) = 2.21$, $p = .14$, $\eta_p^2 = .03$, see Figure 34) and the hit rate ($F(1,78) = 1.25$, $p = .27$, see Table 12 for hit and false

alarm rate data). The effect of list length on the false alarm rate was significant ($F(1,78) = 9.94$, $p = .002$, $\eta_p^2 = .11$). The effect in the false alarm rate may be due to a change in the decision criterion between the lists with participants more likely to respond "yes" to items from a long list.

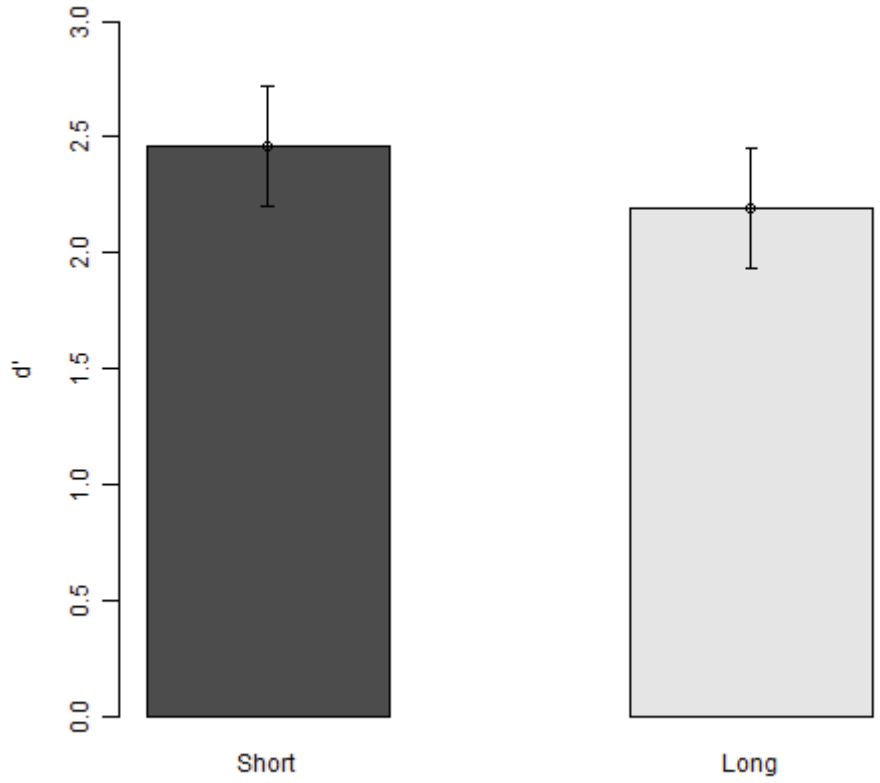

Figure 34. d' values for short and long lists when the data from the Retroactive Pleasantness condition of Experiment 1 and the Yes/No Instructions condition of Experiment 2 were combined in the analysis. Bars represent the 95% between subjects confidence intervals.

Table 12

Hit and false alarm rates for the short and long lists when the data from the first lists studied in the Retroactive Pleasantness condition of Experiment 1 and the Yes/No Task condition of Experiment 2 are combined in the analysis (standard deviations in parentheses)

	Hit Rate	False Alarm Rate
Short List	.82 *(.12)*	.10 *(.10)*
Long List	.85 *(.12)*	.19 *(.16)*

7.2 Conclusions

In addition to the four potential confounds of the list length effect that were raised by Dennis and Humphreys (2001), it appears that the use of a within subjects design to investigate the list length effect may also confound the findings. Where the four original confounds, retention interval, attention, displaced rehearsal and contextual reinstatement inhibit performance on the long list, the use of the within subjects design appears to negatively affect short list performance across multiple lists and lead to an averaging of the effect. Re-analysis of the data from the first list studied by each participant in a between subjects analysis did influence the results obtained. The qualitative conclusions drawn regarding the effect of list length on d' changed in one of the conditions (Experiment 4) across all experiments with some minor variations in the conclusions drawn from two other conditions. In addition the effect size of the list length comparison was greater in the between subjects analysis than it was in the within subjects analysis. It is important to note, however, that these effects were still small in magnitude in the between subjects analysis,

particularly when words were used as the stimuli in a retroactive design with additional controls for potential confounds implemented. Both the within subjects and between subjects analyses suggest that the magnitude of the list length effect is small at best.

In summary, the use of the within subjects design does appear to have an impact upon the list length effect finding and a between subjects design should be favoured in experiments designed to investigate the effect. Calculating the required sample size for these experiments may prove problematic, however, if we are to assume that the magnitude of the effect may be very small (e.g. Dennis & Humphreys, 2001; Dennis et al., 2008). The collapsing of the Retroactive Pleasantness condition data from Experiment 1 and the Yes/No Task condition of Experiment 2 increases the sample size in an analysis in which all potential confounds are controlled. No significant effect of list length was identified, even when the between subjects design was used.

Chapter 8

General Discussion

The primary aim of this thesis was to determine whether, and under what conditions, there is a list length effect in recognition memory. This effect critically distinguishes models of memory based on item and context noise. In item noise models, any interference comes from the other items that make up the study list, thus these models predict a list length effect. In context noise models, interference is assumed to come from all of the previous contexts in which an item has been seen before. These models do not predict a list length effect. The purest forms of these models are at opposite ends of an interference continuum and it is also possible that there is a combination of each at work in the recognition process.

It has been well accepted in the literature that recognition performance for items from a short list is better than performance for long list items. Many studies have documented and replicated this effect (e.g. Cary & Reder, 2003; Gronlund & Elam, 1994; Murnane & Shiffrin, 1991; Ohrt & Gronlund, 1999; Ratcliff & Murdock, 1976; Strong, 1912). However, several nonsignificant effects of list length have also been reported (Buratto & Lamberts, 2008; Criss & Shiffrin, 2004c; Dennis & Humphreys, 2001; Dennis et al., 2008; Jang & Huber, 2008; Murnane & Shiffrin, 1991; Schulman, 1974). Dennis and Humphreys (2001) have proposed that these contradictory results may be due to four potential confounds of list length which could have produced a spurious list length effect when no such effect exists. These potential confounds are retention interval, attention, displaced rehearsal and contextual reinstatement.

A survey of the literature revealed large variation both in terms of which potential confounds were controlled in various list length studies, as well as the method used to control

for them. It seems plausible that these differences can explain the contradictory findings regarding the list length effect and that Dennis and Humphreys' (2001) claim, that there is a nonsignificant effect of list length when appropriate controls for all confounds are implemented, is valid. The review of the literature also suggested that the use of the retroactive or proactive designs as a control for retention interval could be influential. This was the starting point for the present research.

The experiments conducted within this thesis investigated the effect of attention, including the retroactive / proactive distinction, and the use of the Remember-Know (RK) task at test on the detection of the list length effect. The influence of different stimuli was also investigated in an attempt to identify the boundary conditions of the list length effect.

The principle result of the present thesis was that the list length effect in recognition memory appears to be dependent upon a number of factors. There was a negligible and nonsignificant effect of list length on recognition performance for single words when the retroactive design was used. However, the magnitude of this effect was greater when the proactive design was used with words as the stimuli. Finally, there was evidence that the magnitude of the list length effect may be larger when stimuli that are both more similar to each other and more confusable than words are used in the experiment. In the following sections, the results of the six experiments will first be reviewed before the overall implications of the findings are discussed.

8.1 Experiment 1 – The Effect of Attention

The aim of Experiment 1 was to examine differences in the retroactive and proactive designs in terms of attention and also to investigate the extent to which a failure to control for

attention, in this case using a ratings task, may account for the list length effect finding. In the retroactive design, the short list was followed by a period of filler activity such that the duration of these combined was equal to the duration of the long list. Only those items that appeared at the beginning of the long list were tested. In the proactive design, the filler activity preceded the short list. The duration of the short list and the filler activity combined was again equal to the length of the long list. Only the items that appeared at the end of the long list were tested. Differential lapses in attention are more likely to occur in the proactive design.

The results of Experiment 1 revealed a significant list length effect only when the proactive design was used at study (in both within and between subjects analyses). This was the case both when the proactive design was used in conjunction with a pleasantness rating task and when it was not. Conversely, no significant effect of list length was identified when the retroactive design was used, again, regardless of whether the pleasantness rating task was used at study. The magnitude of the list length effect was medium in the proactive conditions but negligible in the retroactive conditions. The inclusion of a pleasantness rating task at study did not affect recognition performance when compared with simply reading the study list items.

8.2 Experiment 2 – The Remember-Know Task

Experiment 2 aimed to investigate whether the inclusion of the RK task at study, as in Cary and Reder's (2003) experiments, had an influence on the list length effect finding. This effect is a standard finding in recall and to the extent that a remember response, and the RK task more generally, makes use of a recall-like process, the task could result in a spurious list

length effect finding in recognition. This experiment manipulated the test task between subjects with the RK task in one condition and the standard yes/no task in the other.

The accuracy data in Experiment 2 did not reveal a significant effect of list length in either condition (this was true of both the within and between subjects analyses). This finding is consistent with the results from the retroactive designs of the previous experiment and the nonsignificant effects of list length previously reported in the literature (e.g. Dennis & Humphreys, 2001; Dennis et al., 2008). However, there was a significant effect of list length on the median response latency for correct responses in the RK task condition (within subjects analysis), although this was small in magnitude (57ms and $\eta_p^2 = .11$).

8.3 Experiments 3 to 6 – Stimuli Other than Words

Considering Experiments 1 and 2, as well as the two experiments of Dennis and Humphreys (2001) and Dennis et al.'s (2008) experiment, there have now been five experiments which have controlled for the four potential confounds and have failed to identify a significant effect of list length on recognition performance. On the basis of these experiments it could be concluded that it is possible to institute controls for retention interval, attention, displaced rehearsal and contextual reinstatement and, in doing so, substantially minimise the size of the list length effect to the extent that the effect is negligible, nonsignificant and could potentially be completely eliminated.

However, all five of these fully controlled experiments used only words as the stimuli. Very little is known about whether this result can be generalised to other classes of stimuli. The final four experiments of this thesis were designed to address this issue and potentially identify the boundary conditions of the list length effect.

Experiments 3 to 6 shared the same experimental design with only the stimulus type varied between the experiments. Word pairs were used in Experiment 3, novel faces in Experiment 4, images of fractals in Experiment 5 and photographs in Experiment 6. Results varied depending on the stimuli with very small and nonsignificant effects of list length for word pairs and photographs but slightly larger and significant effects of list length when faces (in the within subjects accuracy data only) and fractals (in the between subjects accuracy data and the response latency data from both the within and between analyses) were used as the stimuli. This pattern of results suggests that the similarity of the test item to other list items influenced the interference observed. When items that were similar to one another, such as the faces and fractals, were the stimuli there was comparatively more interference in long lists than short lists, and the largest list length effects were observed. When the stimuli were not similar to each other and were randomly chosen items, such as the photograph stimuli and arguably the word pairs (pairs of randomly chosen words) the effect of list length was smaller and nonsignificant.

8.4 Summary of Main Findings

8.4.1 The List Length Effect in Recognition Memory

The general trend in the results from the experiments of this thesis was that the magnitude of the list length effect is substantially minimised with the introduction of controls for potential confounds to the point that it is negligible, nonsignificant and all but eliminated. Despite the fourfold increase in list length, the general finding was a nonsignificant effect of list length, especially when words were used as the stimuli in the retroactive design with

additional confounds controlled.

However, in the majority of experimental conditions presented in this thesis, the means fell in the direction of the list length effect (see Figures 35 and 36), however, most of these differences were not statistically significant. What then should be concluded about the status of the list length effect in recognition memory? This issue is problematic and the question should perhaps be re-phrased in terms of the magnitude of the list length effect. The effect was greatest in the proactive conditions of Experiment 1 where the controls for the potential confounds were arguably at their weakest (particularly in the Proactive Read condition). The introduction of the controls for the four potential confounds was met with a reduction in the difference in recognition performance across list lengths, that is, a reduction in the magnitude of the list length effect. This was true regardless of whether the within or between subjects analysis was used. Note also that there were small effects of list length at various levels of overall performance. Therefore, negligible effects of list length were not a consequence of either floor or ceiling performances.

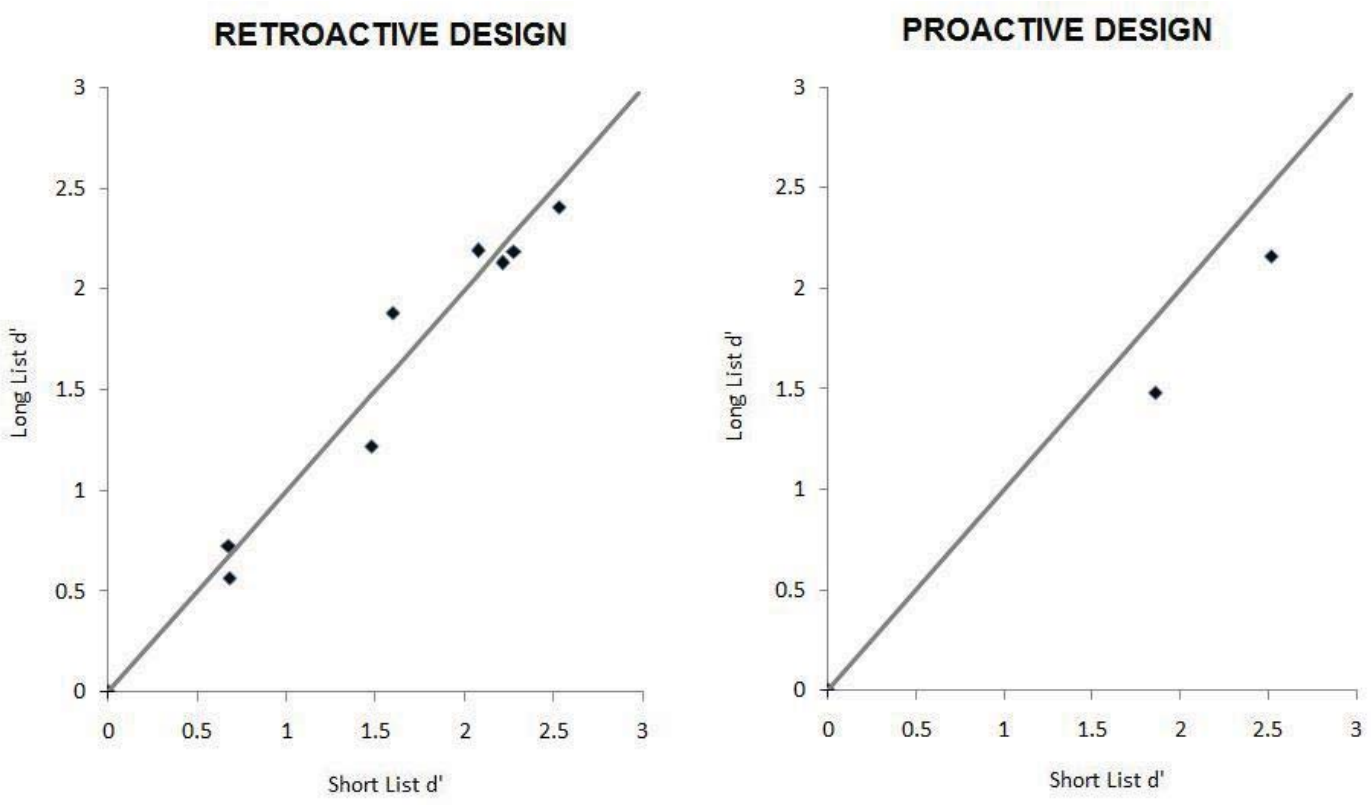

Figure 35. Long list d' plotted against short list d' for all retroactive and proactive design conditions in the within subjects analysis. Data points falling along the middle line would indicate no difference in performance based on list length. Points below the line are instances where short list performance is superior.

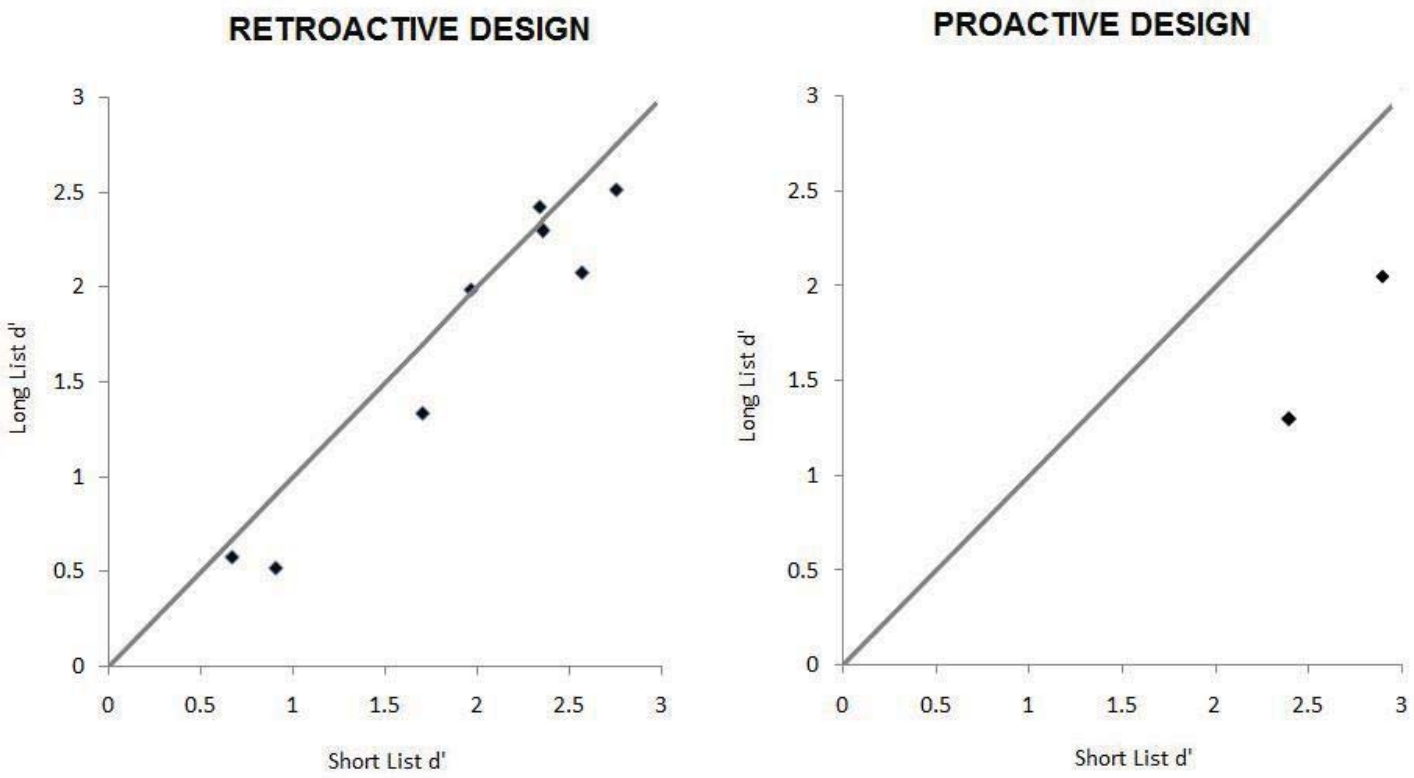

Figure 36. Long list d' plotted against short list d' for all retroactive and proactive design conditions in the between subjects analysis. Data points falling along the middle line would indicate no difference in performance based on list length. Points below the line are instances where short list performance is superior.

There are a number of possible explanations for mean recognition performance being higher for short lists. The first is that the method used to control for the four potential confounds may not have been ideal. For example, participants may still have attempted to rehearse short list words in the filler period despite the presence of the puzzle. Thus, they may still have tired and paid differential amounts of attention to various points of the long list, and they may have required explicit encouragement to reinstate the study context at test rather than relying on an enforced delay to encourage them to do so. Thus, it is possible that

while the inclusion of the controls for the confounds minimised the magnitude of the effect, a tightening of the way in which this was done may prove beneficial and further reduce the size of the effect. In addition, it may be that we have not yet established all of the possible confounds of the list length effect and others remain uncontrolled.

Despite being unable to accept the null hypothesis, that there is no list length effect in recognition memory, the fact that the effect is not statistically significant and the effect size is negligible, suggests that the magnitude of the effect is smaller than one would wish for a "touchstone of recognition memory" (Gronlund & Elam, 1994, p.1355). When controls for the four confounds established to date are implemented, the consistent outcome is of a nonsignificant effect of list length in recognition memory when words are used as the stimuli, in both within and between subjects analyses (Dennis & Humphreys, 2001; Dennis et al., 2008). The exception is another within subjects design, Cary and Reder's (2003) Experiment 3. However, in this study the data from the retroactive and proactive designs were collapsed together in the analysis based on a nonsignificant interaction between list length and design. The results of Experiment 1 of the present thesis also identified this nonsignificant interaction; however when the data from the retroactive and proactive conditions were analysed separately, there were significant effects of list length in the latter conditions only.

The results of the present thesis suggest that the magnitude of the list length effect is negligible when words are used as the stimuli in the retroactive deign with controls for potential confounds implemented. Indeed, the item noise models would predict an effect much larger than that obtained given that the driving force of interference in these models is other list items. Perhaps the use of the Bayesian SDT method of Dennis et al. (2008) will help to shed light on this issue, in that evidence can be accumulated for the null list length effect hypothesis. In fact, the Bayesian analysis favoured the error-only model in every

experimental condition of the present thesis (based on within subjects data), suggesting no significant effect of list length. This method must first be expanded to include analysis of between subjects data.

However, the above claims regarding the negligible and nonsignificant list length effect findings must be mediated by the fact that, despite the controls for the potential confounds being in place, the present thesis has identified slightly larger and significant effects of list length when there is greater similarity between the stimuli (in both the within and between subjects designs).

8.4.2 The Retroactive vs. Proactive Designs

One of the major findings of the present thesis was that when words were used as the stimuli, a retroactive design implemented and additional confounds in place for displaced rehearsal and contextual reinstatement, a negligible and nonsignificant effect of list length resulted (Experiments 1 and 2). However, the results of Experiment 1 also revealed a larger and statistically significant list length effect when the proactive design was used. It is the retroactive design which, with the implementation of controls for displaced rehearsal and contextual reinstatement, provides the better protection against a spurious list length effect finding than does the proactive design. This finding is despite the proactive design providing natural controls for displaced rehearsal and contextual reinstatement as well as study-test lag.

Attention is a point of difference between the retroactive and proactive designs and is the only potential confound not naturally controlled with the proactive design. The results of Experiment 1 suggest that attention may be the most telling of the potential confounds of the list length effect. However, the manipulation of the study task (pleasantness ratings versus

reading) did not affect the list length outcome. It is possible that the ratings task is not sufficient to maintain a participant's attention in the long list to the extent of the attention paid in the short list and it did not fully engage processing of the items. Thus, it appears that the use of the retroactive design, with additional controls for displaced rehearsal and contextual reinstatement, provides sufficient protection against the potentially confounding effects of attention.

An alternative explanation to lapses in attention is that participants are better at reinstating the start of list context than they are the end of list context, which would result in a smaller effect of list length in the retroactive design. However, this is suggestive of a primacy effect, superior performance for items at the beginning of a study list, an effect generally found in recall but not recognition. Further research could experiment with the inclusion of several different controls for attention in order to determine whether it is attentional lapses or the ease of reinstatement that results in more equivalent performance across list lengths when the retroactive design is used.

Whatever the cause of differences between the retroactive and proactive designs, these results highlight the importance of reporting the results of the retroactive and proactive conditions separately; the data should never be collapsed in an analysis.

8.4.3 Stimuli Other Than Words

The results of Experiments 3 to 6 provide some evidence that the magnitude of the list length effect is greater when confusable items are used as the stimuli, such as faces and fractals. The effect sizes for faces (in within subjects data) and fractals (in between subjects data) were larger than those for words under similarly controlled conditions. These findings

suggest that the nature of the stimuli used in the experiment matters in terms of the effect of list length. It is likely that the similarity within a stimulus type may play a role in the differing results, although this was not explicitly tested in the experiments of this thesis. When the stimuli are more similar to one another, as with faces and fractals, there is likely to be greater overlap in the encoding of these items resulting in greater interference when more items are added to the list and a significant list length effect finding. This idea is consistent with the predictions of models, such as REM, which can produce small effects when the stimuli are words chosen at random. However, larger effects are produced when the items are chosen specifically because of their similarity to each other (e.g. Criss, 2006; Criss & Shiffrin, 2004a; Shiffrin, Huber & Marinelli, 1995).

The significant list length effect finding for faces is consistent with the results of Criss and Shiffrin (2004c) who identified a significant list length effect for face but not word pairs. In addition, there is also a dissociation between face and word stimuli in terms of the list strength effect which is consistent with the present list length results. There is a null effect in the recognition of words, but the effect is significant for faces (Norman et al., 2008; Ratcliff et al., 1990).

8.4.4 Order Effects in the Within Subjects Design

Another major finding of the present thesis was that the use of a within subjects design to investigate list length introduces order effects which influence the results. All of the experiments of this thesis involved the manipulation of list length within subjects. This possible confounding was addressed in Chapter 7 with all results re-analysed using a between subjects analysis, that is, analysing only the data from the first list studied by each participant.

The majority of the conclusions drawn from the results remained unchanged but the effect size was generally greater when the between subjects analysis was used.

There are advantages and disadvantages of both the within and between subjects designs. The main advantage of the within subjects design is experimental power which is important when considering the possibility of nonsignificant effects. However, using this design appears to lead to an averaging of the list length effect as a consequence of presentation order. Performance on a short list is better than a long list when viewed first, but the reverse appears to be true for the second list viewed (although not significantly so). Thus, the main benefit of a between subjects design is that it allows for a pure measure of list length at the cost of a lack of experimental power and uncertainty regarding the number of participants required to detect the effect should it exist. Nevertheless, the results of the between subjects re-analysis of the data in the present thesis suggest that the between subjects design should be favoured over the within subjects design for investigation of the list length effect.

8.5 Implications for Mathematical Models of Recognition Memory

The general finding of a negligible and nonsignificant effect of list length when controls for potential confounds were in place is consistent with context noise models of recognition memory but is not accommodated within the item noise framework. To the extent that the magnitude of the list length effect is larger, potentially in the case of face and fractal stimuli, context noise models will struggle to predict the effect. Thus, it is also possible that, as suggested by Criss and Shiffrin (2004a), both item and context noise play a role in recognition memory performance. However, it is important to note that it must be

possible for these combined interference models to completely eliminate any influence from item noise under certain conditions. Nonsignificant and negligible in magnitude list length effect findings may be the impetus for another change to the existing mathematical models just as the null list strength effect was to the GMMs in the 1990s.

The first result to be addressed is the larger list length effect in the proactive design. Models must be able to produce effects of greater magnitude when this design is used. However, it is likely that all models, both item and context noise, can produce this differential result by manipulating their attention parameters.

The second problematic result concerns the fact that there appears to be a larger effect of list length when the stimuli are faces and fractals than for words in the retroactive design. While the size of the effects for these stimuli can still be considered small, the effect is notably larger than the effect when words are the stimuli. The specific implications of the results of the present thesis on the existing mathematical models of memory will now be discussed.

8.5.1 Item Noise Models

8.5.1.1 Minerva II and the GMMs. 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. While these models differ from each other, they all predict the effect by virtue of a global matching process. This process necessarily involves all retrieved study list items in the decision process meaning that a list length effect is predicted under all conditions and that interference from other items is the

driving dynamic of noise in recognition memory. The GMMs, including Minerva II, can accommodate the significant list length effect results in the present thesis. However, the nonsignificant effects of list length identified in this thesis and their generally negligible magnitude are problematic for the GMMs. Furthermore, it seems doubtful that these models can be easily modified to capture this finding.

Minerva II can achieve varying performance between the retroactive and proactive designs by manipulating the learning parameter for each list. In the long list of the proactive design, items would receive less attention and this would be reflected in a less fully encoded vector containing more zeroes. A similar manipulation of the learning or attention parameters can be applied to the other GMMs to achieve this result.

8.5.1.2 REM. The REM model (Shiffrin & Steyvers, 1997) also predicts a list length effect in recognition memory. With increasing list length there is an additional chance that the extra items will spuriously match the test probe. The odds ratio calculation involves dividing by the number of study list items, meaning that they are an integral part of the recognition process and a list length effect is predicted.

REM's *u* and *c* parameters can be manipulated so as to produce the differences in the list length effect depending on whether the retroactive or proactive design is used. In the proactive condition, lapses in attention in the long list would result in fewer features of the study item being stored in memory (*u*) and a reduced likelihood of these features being copied correctly (*c*). The lower values for these parameters in the long lists of the proactive design would lead to differentially lower performance for the long list in this design than the long list in the retroactive design. This would enable REM to account for the larger list length effect in the proactive design.

8.5.2 Context Noise Models

8.5.2.1 BCDMEM. BCDMEM (Dennis and Humphreys, 2001) is a context noise model of recognition memory. The test item is the cue to retrieve all previous contexts in which that item has been encountered from memory. 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 as other list items are not considered during retrieval. Thus, BCDMEM does not predict a list length effect and is consistent with the negligible and nonsignificant effects of list length presented in this thesis for words in the retroactive design with controls implemented. It is unable to accommodate larger and significant effects of list length as may be the case when the stimuli are fractals and faces.

Manipulating the learning parameter, *r*, in BCDMEM can help account for the different effects of list length between the retroactive and proactive designs. The long list in the proactive design would receive less attention than any other list and thus the value of *r* would be small, reducing the weight of the links between the input and output layers. This would reduce the number of nodes in the output layer that are activated at test leading to more errors in this condition.

Another way in which BCDMEM could achieve the differential effect of list length depending on whether the retroactive or proactive design is used in the way that the study context is reinstated. It may be that participants reinstate the start of list context rather than that of the entire list. In this case, reinstatement of the start of list context would favour a more equivalent performance across lists in the retroactive design and disadvantage the long

list in the proactive design.

8.5.3 Combined Interference Models

With respect to which of the item or context noise approaches best accounts for the results of the present thesis, the question is best re-phrased in terms of what is the driving force of interference in recognition memory? It is along this line that the item and context noise models are divided. Item noise models posit that the driving force of interference is the other items present on the study list. If this is the case, a significant effect of list length would be the anticipated result. However, that was not the outcome of the experiments reported here where the pattern of results shows that there is a small effect across the entire set (excluding the proactive designs) and the effect is negligible under some conditions (words in the retroactive design with additional confounds controlled). Interference from other items cannot be the driving force when there is such a small effect of list length despite the four times increase in the number of items on the study list given the performance is not at floor.

For context noise models, the driving force of interference is the previous contexts in which a particular item has been seen. With the manipulation of list length, no additional interference is expected given that it is the number of list items that is being altered, not the number of previous contexts. Therefore, no significant effect of list length is predicted. However, to the extent that the controls implemented to counter potential confounds are not perfect, a negligible effect of list length may be anticipated. The small effect sizes when words are the stimuli, and with potential confounds controlled, fits more naturally with the account of context noise models than it does with item noise models. This result suggests that the concept of context noise must be taken more seriously than it has been to date.

Particularly since context noise models can also naturally account for the word frequency effect with interference from previous contexts an item has been seen in the driving force of interference. However, if we are to accept that the magnitude of the list length effect is greater when stimuli other than words are used in the experiment, particularly faces and fractals, then context noise models would encounter problems accounting for the larger effects.

Thus, while context noise models can better account for the results of the present thesis, there are issues with both these and item noise models in accounting for the entire pattern of results from the experiments of this thesis. It may be that a model which involves both types of interference may be best equipped to deal with the present findings. If interference from other items is not the driving dynamic in recognition memory, then these combined models must be able to all but eliminate the influence of item noise. The challenge for item noise models is to account for both forgetting and a negligible effect of list length under certain conditions.

It may be possible to make changes to the existing item and context noise models of recognition memory to incorporate interference from the other source under certain conditions. For example, as Criss and Shiffrin (2004) noted, BCDMEM (Dennis and Humphreys, 2001) could be altered to take the similarity of the test probe to other list words into account. However, this would have to be implemented only when the stimuli were highly confusable, like faces and fractals and not when words are the stimulus.

It is not immediately clear how item noise models could be altered to radically reduce item noise under certain conditions. Criss and Shiffrin's (2004a) modifications to REM which involved both item and context noise features in the decision process is a step in the right direction, however, it should again be noted that the role of item noise must be

eliminated in order to account for the nonsignificant list length effect findings.

The SAC model (Reder et al., 2000) also involves interference from both other items and previous contexts. At test the concept and context nodes are activated. More items on the study list means that there are a greater number of links between the context node and the event nodes which results in lower activation of the event node at test, fewer "yes" responses and a lower recollection based hit rate for longer lists. As a consequence, participants then rely more on familiarity and adopt a more lenient concept threshold resulting in more false alarms for long lists.

In principle, SAC incorporates both item and context noise. The contribution of item noise could be made arbitrarily small in the same way that the contribution of item noise can be reduced for SAC to account for the null list strength effect in recognition. SAC can account for the null list strength effect by reducing the role of the recollection component which then leads to an increased role of familiarity (Diana & Reder, 2005). The same assumption could be made for list length. Thus, SAC may be able to account for the results of the present studies when the effect of item noise is reduced significantly.

8.6 Problems to Address

The major problem facing the results of the present thesis is the use of the within subjects analysis. However, steps were made to address this issue in Chapter 7. It is apparent that a between subjects design should be favoured over a within subjects design despite the former lacking in experimental power. It is the between subjects design which provides a pure manipulation of list length.

Further, the investigation of the list length effect in the present thesis has revealed that

the effect itself is a difficult phenomenon to explore. It may be impossible to conduct an experiment which completely eliminates all potential confounds of the effect and, while this remains the case, it will be difficult to ascertain whether any observed effect is attributable to differences in list length or the result of confounding. Ideally, to investigate the list length effect, a between subjects design would be used, though the number of participants may be great, the retroactive design would be used with words as the stimuli and additional controls implemented for displaced rehearsal and contextual reinstatement. However, even the addition of these controls into the experimental design is problematic. It may be that there are better ways to minimise the potential effects of confounding variables. Alternatively, as previously noted, it may be that we have not yet identified all of the potential confounds of the list length effect.

8.7 Future Research Directions

The experiments reported in the present thesis have provided a good foundation for investigating the source of interference in recognition memory and the list length effect more generally. These experiments have laid the foundations for future research directions that could include a similarly controlled investigation exploring different list lengths and list length ratios, exploring the concept of a list and continuing the manipulation of similarity in stimuli other than words.

The starting point for further analyses should be the replication of several of the studies of the present thesis using a between subjects design and a greater number of participants. This replication would allow for stronger conclusions to be made regarding the list length effect and its potential confounds as well as the influence of this result on item and

context noise models of recognition memory. In addition, further manipulation of the displaced rehearsal, attention and contextual reinstatement controls in the between subjects design would be beneficial in terms of eliciting the best method for controlling for their differential influence.

The list length ratio in each experiment presented in this thesis was 1:4 with 20 items for short lists and 80 items for long lists in all but Experiment 3 where the short list was 24 pairs long and the long list was made up of 96 pairs. It may be beneficial to increase this ratio and examine the influence of this on the list length effect finding. It would be important to maintain the controls for the four potential confounds of Dennis and Humphreys (2001) with longer list lengths and establish whether a significant list length effect results as the difference between the two list lengths increases. It may be that the relative influence of item and context noise changes with the list length ratio.

Another similar manipulation could involve smaller list lengths using a procedure resembling that used by Sternberg (1966). That study involved recognition testing of symbols that were part of a series ranging in length from one to six items with one test probe per series. Sternberg identified a significant list length effect on the response latency data.

Another avenue for future research could involve an investigation of the concept of a list. Particularly when a within subjects design is used, the experimenter is aware that there is a short and long list and views them as two separate entities. However, for participants the idea of studying a list or two in this experimental setting is likely to be quite a unique experience in their lives. When they have studied two different lists, they may encode this as one long study experience hence the small effect of changing the design to a between subjects analysis. In item noise models, the concept of a list is taken somewhat for granted. At test, these models posit that participants use the test probe as a cue to retrieve all items present on

the study list. It is not clear how they are specifically able to do this and whether they call to mind just the list in question, all lists from the study session, all recent lists, or any other type of similar experience.

Potential future experiments could involve manipulation of the way in which lists are presented to participants at study. For example, participants could be told that they will be studying a series of lists. After the first 20 items, participants are told that they have viewed the first list. Then 80 more items are presented, this is the second list. They would then just be tested on one of the lists. Another experiment could investigate the list length effect in a design in which participants are aware that they will only be tested on the first (or last) items of the long list. In this case, however, controls would need to be in place to ensure that participants continued to attend to all items on the long study list, otherwise the comparison would just be between the two short lists, although in practice, this may be difficult to guarantee. Experiments of this kind would allow us to ascertain whether the nature of the presentation of study lists plays a role in the list length effect outcome and influences the source of interference in recognition memory.

The results of the six experiments of the present thesis have shown that the stimulus used in the experiment may matter in terms of the list length effect finding and the source of interference in recognition memory. Though still considered small, the magnitude of the effect when faces and fractals were used in the experiment was larger than the size of the effect for words in the retroactive design. A logical extension of the work presented here would be to further manipulate the intra-stimulus similarity which may also provide a conclusive result in terms of whether the stimulus used in the experiment mediates the effect of list length. Photographs provide the best opportunity for manipulation, rather than having each photograph depict a separate identifiable scene, they could show variations upon the

same theme, for example, images of sunsets. A major problem with manipulating the similarity of the stimuli is finding the right balance in terms of recognition performance. For example, in Experiment 6, performance with photographs was at ceiling until the rate of presentation at study was significantly shortened. Conversely, a pilot experiment involving random patterns of 20 dots was met with floor performance, with the stimuli too similar to one another to be discriminated better than chance. Further investigation would enable better definition of the boundary conditions of the list length effect.

The experiments in the present thesis have paved the way for several different avenues of future research to further investigate the list length effect finding and the source of interference in recognition memory.

8.8 Summary and Conclusions

The main finding of the present thesis was that the well-documented existence of the list length effect in the recognition memory literature is not as well established as it first appears. It is not possible to accumulate evidence for a null hypothesis using ANOVA, however the results of the present experiments suggest that at the very least, the magnitude of the list length effect is significantly smaller than that which should be predicted by a model in which the major source of interference is the other items on the study list. This is particularly so in the present case where the list length manipulation involved a fourfold increase in the number of study list items. The explanation for this seems to be that there are a number of potential confounds, retention interval, attention, displaced rehearsal and contextual reinstatement, which result in a spurious or at least magnified list length effect (Dennis &

Humphreys, 2001). The results of Experiments 1 and 2 suggest that when single words are the stimuli and these confounds are controlled, including the use of the retroactive design, there is a negligible and nonsignificant effect of list length in recognition memory, consistent with the work of Dennis and Humphreys (2001) and Dennis et al., (2008).

The primary finding in this regard was that the magnitude of the list length effect was substantially smaller in the retroactive design than in the proactive design. This finding suggests that the proactive design is subject to confounding from another source, most likely attention. The retroactive and proactive designs are not interchangeable and should be considered and analysed separately at all times.

The second issue addressed in the present thesis was the attempt to identify and define the boundary conditions of the list length effect. The results of Experiments 3 to 6 suggest that the stimuli used in the recognition memory task influences whether a significant effect of list length is identified. The effect of list length on recognition performance for word pairs and photographs was nonsignificant, however, there was a significant and slightly larger list length effect when faces and fractals were used as the stimuli.

In addition, the results and re-analysis of the data from this thesis suggest that a within subjects manipulation of list length introduces additional confounds and an averaging of performance across list order. These order effects are eliminated when a between subjects design is used and as a result, this type of analysis should be favoured.

Finally, the pattern of results from the series of experiments in the present thesis, in both within and between subjects analyses, pose significant problems for existing mathematical models of recognition memory. In particular, item noise models in which the source of interference is exclusively the other items that were present on the study list, are called into question when one considers that the negligible size of the list length effect

finding when words are the stimuli and potential confounds are controlled is not consistent with this assertion. Other items cannot be the dominant source of interference when a four times increase in the number of items (four times more interference in terms of items noise models) results in only a negligible effect of list length. Context noise models are more naturally able to account for these results, although they are challenged to the extent that the effect of list length when faces and fractals are the stimuli is significant. It may be that neither item noise nor context noise alone can account for the observed interference in recognition memory. Based on these results it may be that we must now look to a new generation of mathematical models of memory which primarily involve context noise with the possibility for some influence from other items to explain the present results.