Exploring the Utility of Recollected Content in a List-Specific Recognition Task.

Krista Louise Murray
BA (Hons).

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in 2013
School of Psychology
ABSTRACT

There has been considerable debate within the literature between those who view recognition as the result of two processes (recollection and familiarity) and those who consider that familiarity is all that is required. Within those models that propose a role for recollection, the dominant view is that recollection involves recalling qualitative content from the study episode and thus is thought to be responsible for tasks that require episodic memory such as list discrimination tasks. Familiarity is not generally thought to be helpful in identifying source although it may be correlated with it. Eluding the debate between a single and dual process interpretation, I instead propose that a role for content is likely for some tasks, but not necessarily for all episodic tasks.

In this thesis, the necessity of recalling qualitative information in order to adequately perform an episodic recognition task was examined across seven experiments. In order to properly address this question it was necessary that the methodology inhibited the use of content in distinguishing between episodes. All experiments contained the same basic design where participants receive multiple study and test lists constructed of the same items. In this list-specific design, the same words are tested on each list, however the targets and distracters are randomized on every list. In consequence all items become more familiar and recollected content becomes less useful with each successive list. Despite the difficulty of the task and the reduction in the availability of discriminating content, across all experiments it was found that most participants were still able to perform the task.

The Remember-Know (RK) procedure has often been applied in order to ascertain the involvement of recollection and familiarity in recognition tasks, however there has been a fair amount of controversy regarding the validity of the RK procedure. Despite the controversy the procedure is still frequently used. Including RK judgments allowed some assessment of whether content could still be useful in a list specific task. In addition, I was able to evaluate how the addition of the RK procedure might affect recognition performance.

All three experiments incorporated in the second chapter demonstrated that participants could adequately perform the list specific task. Some large differences in recognition performance were observed when RK judgments were included for the first time on the final list compared to conditions where participants were well
practiced at the task or were yet to make RK decisions. However it was difficult to form strong conclusions from the data due to limitations of the design. It was further difficult to make strong conclusions about whether or not content was still useful to participants, because despite some very poor performance with the inclusion of the RK task on the final list, participants were still more likely to attribute R judgments to hits than false alarms.

In the third chapter the methodology of the initial experiments was refined. From the results of the experiments in this chapter I was able to more categorically conclude that the inclusion of the RK paradigm can alter recognition and sometimes R judgments. I was also able to more accurately measure the contribution of recollection to the task. In the last experiment in this chapter I included a justification condition where participants were asked to write down the content they had recollected. This manipulation greatly reduced the likelihood that participants would produce a R judgment and also adversely affected recognition performance. In addition, while R judgments were still frequently accurate (more often made to hits than false alarms), a chi square analysis which tests for an association between R judgments and hits, failed to indicate that content could have been useful for the majority of participants. However, there was some indication that participants were more capable of associating hits and R judgments when they were more practiced at the RK procedure.

The final experiment, included in chapter 4, supported previous findings and also provided further clarification as to how RK judgments might be affecting recognition. It was possible that knowledge about a role for familiarity and recollection in recognition memory might be enough to alter performance on the recognition task. It was also possible that actively making RK judgments affected the task. The results of the final experiment indicated more strongly that actively making RK judgments as participants perform the recognition task, could adversely affect recognition performance. Implications for the RK paradigm are discussed, as are implications for the role of content recollection in recognition memory and episodic memory.
DECLARATION BY AUTHOR

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the General Award Rules of The University of Queensland, immediately made available for research and study in accordance with the Copyright Act 1968.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.
PUBLICATIONS DURING CANDIDATURE

Peer reviewed journals


Conference Posters


PUBLICATIONS INCLUDED IN THIS THESIS

No publications included.
CONTRIBUTION BY OTHERS TO THE THESIS

Em. Prof. Michael Humphreys, my PhD supervisor, supervised and contributed to the design of the experiments and the execution and analysis of the experiments. Em. Prof. Michael Humphreys also made editorial contributions to the thesis so as to contribute to interpretation. Associate Professor Jennifer Burt also provided editorial assistance to written material.

STATEMENT OF PARTS OF THE THESIS SUBMITTED TO QUALIFY FOR THE AWARD OF ANOTHER DEGREE

None.
ACKNOWLEDGEMENTS

I would like to thank my principal advisor Em. Prof. Michael Humphreys for his unwavering loyalty in the face of what must have been a great deal of uncertainty. This surely is the result of a very kind and patient teacher. I am also extremely grateful to my advisor for throwing as many work opportunities my way as possible. This was not only essential for my continued survival in the world but also allowed me to explore other disciplines and meet new and interesting people. It also helped me to understand that multi-tasking requires much more than the capacity to text and watch T.V at the same time.

There have been a number of colleagues who offered sage advice and support along the way, a special thanks to Emerald Quinn, Dr. Tania Xiao, Dr. Angela Maguire and Assoc. Prof. Jennifer Burt.

Thanks must also go to my family and friends for their love and support, especially to my partner Daniel. Thanks also to my dog Naboo whose incessant need for a walk helped to keep me fit and healthy throughout my candidature.
KEYWORDS
Recognition, recollection, familiarity, remember-know, episodic memory.

AUSTRALIAN AND NEW ZEALAND STANDARD RESEARCH CLASSIFICATIONS (ANZSRC)
ANZSRC code: 170201, Computer Perception, Memory and Attention, 100%

FIELDS OF RESEARCH (FoR) CLASSIFICATION
FoR code: 1702, Cognitive Sciences, 100%
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>DECLARATION BY AUTHOR</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>CHAPTER 1</td>
<td>1</td>
</tr>
<tr>
<td>Literature Review</td>
<td>1</td>
</tr>
<tr>
<td>Overview of Chapters</td>
<td>22</td>
</tr>
<tr>
<td>CHAPTER 2</td>
<td>23</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>23</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>24</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>25</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>27</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>28</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>28</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>28</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>31</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>32</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>32</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>32</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>34</td>
</tr>
<tr>
<td>CHAPTER 3</td>
<td>34</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>36</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>36</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>36</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>41</td>
</tr>
<tr>
<td>Experiment 5</td>
<td>43</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>43</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>44</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>48</td>
</tr>
<tr>
<td>Experiment 6</td>
<td>50</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>50</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>50</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>58</td>
</tr>
<tr>
<td>CHAPTER 4</td>
<td>60</td>
</tr>
<tr>
<td>Experiment 7</td>
<td>61</td>
</tr>
<tr>
<td><em>Method</em></td>
<td>63</td>
</tr>
<tr>
<td><em>Results</em></td>
<td>63</td>
</tr>
<tr>
<td><em>Discussion</em></td>
<td>71</td>
</tr>
<tr>
<td>CHAPTER 5</td>
<td>73</td>
</tr>
<tr>
<td>General Discussion</td>
<td>73</td>
</tr>
<tr>
<td>Summary of Results</td>
<td>73</td>
</tr>
<tr>
<td>Future Research</td>
<td>87</td>
</tr>
<tr>
<td>Conclusion</td>
<td>90</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>92</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Experiment 1. Hits and False Alarms Broken Down by List Number and RK Response Interval ................................................................. 25
Table 2. Experiment 1. RK Judgments as a Function of Recognition Accuracy and RK Response Interval ......................................................... 27
Table 3. Experiment 2. Hits and False Alarms as a Function of List and RK Frequency ...................................................................................... 29
Table 4. Experiment 2. RK Judgments as a Function of List, Accuracy and RK Frequency .............................................................................. 30
Table 5. Experiment 3. Proportion of Hits, False Alarms and Dprime Values Broken Down by List and RK Group ............................................. 32
Table 6. Experiment 3. RK Judgments as a Function of Accuracy in the Final Condition .............................................................................. 33
Table 7. Experiment 4. D-prime values, hits and false alarms as a function of list number and RK Frequency condition .................................. 36
Table 8. Experiment 4. Mean proportion of RK judgments as a function of RK frequency, list number and accuracy ................................... 38
Table 9. Experiment 5. Hits, false alarms and d-prime values broken down by list number and RK frequency condition .............................. 44
Table 10. Experiment 5. Mean proportion of RK judgments as a function of list number, recognition accuracy and RK frequency condition .... 45
Table 11. Experiment 6. Recognition results presented as hits, false alarms and d-prime values broken down by list, RK frequency and justification conditions ................................................... 50
Table 12. Experiment 6. Mean proportion of RK judgments as a function of list number, recognition accuracy, RK frequency and justification condition ................................................... 52
Table 13. Experiment 7. Proportion of hits, false alarms and d-prime values broken down by list and RK response interval ................................ 63
Table 14. Experiment 7. RK judgments as a function of accuracy, list, RK response interval and RK frequency ........................................ 65
Table 15. Experiment 7. Comparing old decisions on test 1 (recognition) and test 2 (RK) in the delayed conditions ......................................... 71

LIST OF FIGURES

Figure 1. Experiment 1. False Alarms as a Function of List and RK Response Interval .................................................................................. 26
Figure 2. Experiment 2. Proportion of False Alarms Across List and RK Frequency Condition ................................................................. 30
Figure 3. Experiment 4. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency ......... 39
Figure 4. Experiment 4. Individual correlation scores between d-prime and chi square values for list 4 in the continuous condition (r = .083) ................................................... 40
Figure 5. Experiment 4. The relationship between individual d-prime and chi square values for list 4 of the final condition (r = .090) ...................... 41
Figure 6. Experiment 5. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency .......... 46
Figure 7. Experiment 5. Individual correlation scores between d-prime and chi square values for list 4 in the continuous condition (r = .391) ........... 47
Figure 8. Experiment 5. Individual correlation scores between d-prime and chi square values for list 4 in the final condition ($r = .344$). ............................................... 47

Figure 9. Experiment 6. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency in the justification condition. ........................................................................................................... 54

Figure 10. Experiment 6. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency in the no justification condition. ........................................................................................................... 55

Figure 11. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the continuous no justification condition ($r = .759$). . 56

Figure 12. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the continuous justification condition ($r = .177$). ...... 56

Figure 13. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the final no justification condition ($r = .184$). ........... 57

Figure 14. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the final justification condition ($r = .003$). .................. 57

Figure 15. Experiment 7. Individual chi square values as a function of RK frequency in the delayed condition................................................................. 68

Figure 16. Experiment 7. Individual chi square values as a function of RK frequency in the immediate condition. ................................................................. 68

Figure 17. Experiment 7. Individual correlation scores between d-prime and chi square values for list 4 in the continuous immediate condition ($r = .355$). ....... 69

Figure 18. Experiment 7. Individual correlation scores between d-prime and chi square values for list 4 in the interrupted immediate condition ($r = .467$)........... 69

Figure 19. Experiment 7. Individual correlation scores between d-prime and chi square values for list 4 in the continuous delayed condition ($r = .337$)............. 70

Figure 20. Experiment 7. Individual correlation scores between d-prime and chi square values for list 4 in the interrupted delayed condition ($r = -.030$).............. 70
CHAPTER 1

Literature Review

*Episodic Memory and Autonoesis*

Episodic memory has accumulated a large number of defining features since its conception. It has been defined broadly as the memory for past experience and more specifically it is considered to be more recent to evolve, late to develop and easy to deteriorate (Tulving and Szpunar, 2009). The most distinguishing feature remains Tulving’s (1983) assertion that a special type of phenomenological experience termed ‘autonoetic consciousness’ or ‘self-knowing’ accompanied episodic memory. It is thought that the experience of recalling information about the time and place of the event results in autonoetic awareness. This type of conscious awareness has been contrasted to the type of awareness experienced during semantic retrieval. Recalling from semantic memory, for example recalling that Paris is the capital of France, does not elicit autonoetic awareness. Instead, the type of awareness thought to accompany semantic retrieval has been described as ‘noetic’ or simply ‘knowing’ (Tulving, 1983).

As a key characteristic of his theory it was unavoidable that consciousness would have to be in some way quantified. To this end, Tulving (1985) tested his theory by collecting subjective reports from participants on whether their state of awareness at retrieval was best characterized by a remember (R) or a know (K) judgment. In order to allow participants to distinguish between different states of awareness, Tulving (1985) indicated that a R judgment should be made if qualitative detail of the study event was recalled (i.e., such as a thought or feeling one experienced at the time of study). In comparison, a K judgment should be made when memory of the event was not accompanied by any details of the episode and rather should be made when the participant simply felt that they knew they had studied the item. Therefore, a R judgment was only to be made when episodic content was recalled and consequently should reflect episodic memory, while a K judgment should reflect an absence of such detail and therefore should reflect semantic memory and noetic consciousness.

Tulving (1983) introduced the synergistic ecphory model of recall and recognition. Ecphoric information was defined as an amalgam of episodic and semantic information, which could then determine the nature of the recollective
experience. Within this model different memories could be formed depending on whether there was more episodic information or more semantic information available. Tulving (1985) postulated that if there was more episodic trace information then the recollective experience will mostly be characterised by autonoetic consciousness (eliciting a R response), whereas when there was more semantic cue information then the recollective experience would mostly be characterised by noetic consciousness (eliciting a K response).

In his first experiment, Tulving (1985) varied ecphoric information across three successive tests after participants studied a list of category names and single category instances (e.g., musical instrument – viola). The first test was a free recall task. The second test provided category cues for items that had not been recalled on the first test. Similarly, any remaining to-be-remembered items from the second test were then tested on the third test, which provided the category cue and also the initial letter of the target item. The first test was described as having relatively rich episodic trace information combined with relatively impoverished (semantic) retrieval information. Whereas by the final test, there was abundant retrieval information combined with relatively poor trace information. Tulving (1985) found that the proportion of R judgments reflected the changes in episodic (trace) information across tasks.

In the second experiment of Tulving’s (1985) paper RK judgments were collected from participants in two recognition tests. The retention interval varied between groups such that one group studied and were tested on a list of 36 words on the same day and another group studied the same list on day 1 however their memory for the items was not tested until 7 days later on day 8. The hit rates decreased and the false alarm rate increased over the 7-day retention interval. The proportion of R judgments for correctly recognized words was lower on day 8 than on day 1. It was concluded that autonoetic awareness was more clearly present in the recollection of recently encountered events than in that of events encountered a week earlier.

The distinction Tulving (1983) made in terms of consciousness was intuitive as most people can discriminate between the experiences of recalling a self-experienced event and recalling from general knowledge. However in response to Tulving’s (1984) paper, Klatzky (1984) commented that there were some difficulties with providing experimental evidence for separate systems, especially when based on differing phenomenological experience. Klatzky (1984) indicated that a distinction
could be made between semantic and episodic knowledge. That is a simple content distinction can be made between information relating to ‘what, when and where’ (www) and other information. Klatzky (1984) pointed out that different phenomenological experience was not necessarily an indication of separate systems but rather a difference in awareness occurring due to the retrieval of two different types of information. While not explicitly acknowledged in the literature, the content distinction pointed out by Klatzky (1984) has often been relied upon in order to differentiate between episodic and semantic memory.

A simple content distinction as outlined by Klatzky (1984) can be applied to the thoroughly documented case of the amnesic K.C who could recall general self-knowledge however he could not recall a single self experienced episode. While K.C was aware that his family owned a holiday house, he could not remember one single instance of actually being there, despite many visits in the past (Tulving, 2005). Thus K.C could not recall information relating to www, however he could report other information from general knowledge. On the basis of this it was concluded that K.C did not possess autonoetic awareness and therefore that he did not possess episodic memory. The assumption implied in this conclusion was that recollection of content was indicative of the memory for a particular episode, and that a memory for a particular episode can only occur if content is recollected.

A distinction was eventually made between episodic knowledge and autonoetic awareness (Suddendorf and Corballis, 2007). Specifically, a distinction is made between knowing www and remembering www. Thus in order to possess episodic memory one had to mentally relive the event, not just possess episodic knowledge. Due to this distinction, scrub-jays, a food-caching bird, that can provide evidence for knowledge of www can only be deemed as having episodic-like memory (Clayton, Yu and Dickenson, 2001). That is, their behaviour may indicate they possess knowledge of an episode, but it can not be concluded that they have experienced autonoesis. In addition, while children under the age of four can retain knowledge they seem to be without the capacity to recall the event that provided them with that knowledge (Wimmer, Hogrefe and Perner, 1988). That is, three year olds can tell what they know but rarely can indicate the source of that knowledge, or why they know what they know (Wimmer et al., 1988; Perner & Ruffman, 1995; Perner, Kloo, & Gornik, 2007). The proposed distinction between episodic knowledge and episodic memory proposes that contents recalled may not necessarily relate to a
particular episode, however to the best of my knowledge, this premise has not been explicitly tested within traditional laboratory based memory experiments.

Recollection and Content in Recognition Memory

The initial content distinction Tulving (1985) made between episodic and semantic memory had a lasting influence on the way in which recollection is more often conceptualized, particularly within the recognition literature. Some theories propose that recognition involves the use of two underlying processes, familiarity and recollection. As the following section will illustrate, where a role for recollection is proposed, it is often assumed to involve the recollection of qualitative content from the study episode. Theories differ on how recollection is defined and some are more heavily influenced by Tulving’s theory of episodic memory than others. The next section reviews these different opinions with the aim of illustrating that the content distinction Tulving (1985) initially made can be applied to some current theories of recognition memory.

The idea that recognition might require some form of recall/recollection and also some sort of familiarity has been in the literature for quite some time (i.e.. Atkinson and Juola, 1974; Mandler, 1968; 1980, Humphreys 1976; 1978, Rabinowitz, Mandler and Barsalou, 1977). Across these earlier theories familiarity and recall/recollection have had varied definitions, for example, Rabinowitz et al., (1977) proposed the idea of a perceptual code and conceptual code, roughly synonymous with ideas of familiarity and recall respectively. Humphreys (1976, 1978) drew a distinction between item and relational information, also roughly synonymous with ideas of familiarity and recall. Earlier models made reference to simply recalling the item, however more current theories often consider recollection as the retrieval of qualitative aspects of the study episode, which is more congruent with Tulving’s (1983) theory of episodic memory.

At this point, a distinction can be made between recall and recollection. That is between simply recalling the item or recollecting aspects of the study experience. The distinction I make between recall and recollection has often been either obfuscated or overlooked in the literature. For example, while Mandler (1980) focused on a role for recall and familiarity in recognition, as more popularly evidenced by his ‘butcher on the bus’ example, he remained ambiguous about whether recall referred to recollected content, the recall of the to-be-recognized item as had been implied in earlier models, or the act of recalling something. In order to
distinguish from earlier ideas about recall or any other definition, when recollection is viewed solely through the type of content retrieved, I will use the term content recollection.

There are a number of current approaches that consider that recollection and familiarity are required for recognition decisions (i.e., Jacoby, 1991; Yonelinas, 1994; The source of activation confusion (SAC) model: Reder, Nhouyvanisvong, Schunn, Ayers, Angstadt & Hiraki, 2000). Wixted and Stretch (2004) argue that recollection and familiarity combine to produce a single strength signal on which recognition decisions are based. The STREAK model (Rotello, Macmillan and Reeder, 2004) also views the recognition decision as a reflection of different combinations of recollection and familiarity. Some of these theories have been heavily influenced by Tulving’s (1983, 2005) theory of episodic memory. For example, Jacoby (1991) and Yonelinas (1994) view recollection as the effortful and accurate retrieval of episodic information. However the role of familiarity is far more vague. It is not considered to be source specifying, nor be accompanied by any qualitative information as is thought to occur during recollection. In sum, differences are outlined in terms of speed, accuracy, effort and content. That qualitative information is present for one process and not the other has become a fundamental distinguishing feature. Furthermore, these models, similar to Tulving’s (1983) theory of episodic memory, assume that recollected content is informing the participant about whether or not the test item appeared at a certain time and place.

A clear example of how these latter dual process models view recollection is embodied within the process dissociation procedure (PDP) devised by Jacoby (1991). The broader purpose of the PDP was to distinguish intentional from automatic processes. Jacoby (1991) also used it to affirm assumptions regarding content recollection (intentional) and familiarity (automatic) in a list discrimination task. This procedure involves participants studying two lists and then receiving a recognition task. In the ‘inclusion’ condition, they are required to respond, “yes” if the item was in either of the two lists. In the ‘exclusion’ condition, they are required to respond “yes” only if the item appeared in the target list (either list 1 or list 2) and not in the non-target list. Jacoby (1991) assumes that, in the inclusion condition, the subject will respond yes if the item is familiar or if it is recollected (or both). In the exclusion condition, when the subject should be responding no, it is assumed that they will mistakenly respond yes if the item is familiar but they fail to recollect which list it
occurred in. The PDP relies on a number of assumptions; however the critical assumption made with this model that I wish to examine, is that recollected contents are informing participants both about the list membership of each item and occurrence in the experiment.

In contrast to ideas about recollection in the PDP and other dual process models there are some models that propose that list discrimination tasks, such as those employed in the PDP can be accomplished via a process that does not bring content to mind. Content may or may not accompany this process, and may or may not be useful, but it is not considered to play the important role given to it within the PDP. So called single-process models view recognition as a reflection of one source of information that varies in strength and can be described by signal detection theory. Signal detection theory applied to decision processes in old/new recognition has been well established (Green and Swets, 1966). For example, global matching models, while differing in specifics, propose a match between a cue or target item and other relevant information in memory. The strength of the match is reflected in the recognition judgment (e.g., SAM (Gillund and Shiffrin, 1984), REM (Shiffrin and Steyvers, 1997), The Matrix Model (Humphreys, Bain and Pike, 1989) and BCDMEM (Dennis and Humphreys, 2001). These models provide a possible alternative to a dual process understanding of recognition performance.

There has been substantial debate within the literature regarding how many processes underlie recognition memory and as of yet there is no clear consensus. It is not my aim to argue for a single or dual process interpretation, but rather to ascertain how useful content recollection is in a task that requires the participant to consider information pertaining to www. In going about this I consider the common methodologies used to measure the presence or absence of content recollection. In particular, the following section discusses the use of the RK paradigm in recognition memory and in doing so, expands on different conceptualizations of content recollection and also how these differences might be reflected in RK judgments. This is particularly relevant considering the extensive use of the paradigm to provide converging evidence for recollection and familiarity in recognition performance.

*Application of RK Judgments to Recognition Memory*

It was not only Tulving’s (1983) theory of declarative memory that influenced ideas about recollection but also his 1985 use of the RK paradigm. Typical instructions on a RK task in recognition memory ask participants to indicate their
recognition decision as R if they recall qualitative aspects about the study episode (such as thoughts or reactions they had at the time of study) or K if the item is familiar but they cannot recall any content from the study episode. Thus a R response is thought to index conscious recollection, while a K response is thought to index the feeling of familiarity. The way in which the RK task is interpreted differs depending on how recollection and familiarity are conceptualized within a recognition task.

Those dual process theories that consider a recognition decision to be based on a combination of recollection and familiarity, for example Wixted and Stretch (2004) and Rotello et al., (2004) also view RK judgments as a reflection of different combination of recollection and familiarity. That is, they do not consider that RK judgments will reflect a process-pure measure of recollection and familiarity. However others, particularly those who consider that recollected content is responsible for list discrimination or source identification consider a R judgment at least to be a reasonably pure measure of recollection.

Tulving (2001) notes that the application of the RK paradigm to recognition studies have indicated that, as well as episodic recollection, recognition can be achieved through familiarity (i.e. Gardiner, 1988; Yonelinas and Jacoby, 1995). In general single list recognition tests have been preferred in the literature and Tulving (2001) points out that such a test may not necessarily rely on episodic memory, as an old decision does not necessitate that the participant know exactly what, when and where they had viewed the item. This is due to the fact that more recently viewed items will tend to be more familiar and therefore a general feeling of familiarity may suffice. While there has been a preference for such single list designs, there are list and source discrimination studies that have also utilised the RK paradigm.

In Yonelinas and Jacoby (1995) the RK paradigm was applied to the PDP with the result that those judgments converged nicely with estimates of recollection and familiarity gained form the procedure. The authors do not propose a process pure interpretation of RK judgments as K responses reflect familiarity in the absence of recollection (F(1 − R)) (Yonelinas and Jacoby, 1995). However they do expect R to be a reliable indicator that recollection has occurred, provided participants are responding as the experimenter expects them to. In essence, the assumption made by the PDP is that R judgments should reflect the process of recollection even if familiarity is also contributing to the R judgment. Therefore, according to the PDP
perspective, R judgments are usually only made when content has been recalled and therefore should be a reasonable indicator of content recollection.

The PDP view of content recollection and RK judgments can be contrasted to how these processes are viewed in the source monitoring framework (SMF). Source memory, such as deciding that the test item had been heard in a male or female voice at the time of study, is thought to require the retrieval of source specific information which may be considered synonymous to recollected content (i.e., Johnson, Hashtroudi and Lindsay, 1993). However the functionality of content recollection as it is embodied within the SMF differs somewhat to how it is viewed in episodic recognition or dual process models of recognition. Source memory recollection does not necessarily need to be accurate or effortful, nor is it seen as independent of any strong feeling of familiarity (Qin, Raye, Johnson & Mitchell, 2001).

An example of how the two theories differ can be observed by comparing interpretations of data present in Yonelinas (1999). In his fourth experiment participants discriminated between words from two lists that had been presented five days apart. The results supported the idea that a continuous process was responsible for completion of the memory task, thus, Yonelinas (1999) argues that it was familiarity (in the form of recency) that provided a strong cue to source. Qin et al (2001) noted however, that such data often results even when familiarity could not be used to indicate source. That is, content recollection and familiarity are thought to occur continuously within the SMF therefore making a clear distinction between familiarity and recollection does not fit with a source monitoring perspective. As such, Qin et al (2001) argued that the results Yonelinas (1999) observed were due to differences in the qualitative characteristics of memory as memory for source should be continuous. While the SMF consider that R judgments should exist on a continuum, and therefore is at odds with some dual process views on recollection, both consider that content recollection, as reflected by R decisions, should be largely responsible for source or list decisions.

As previously mentioned, there are models commonly referred to as single process theories that do not focus on a familiarity/recollection distinction in recognition memory. These models also predict that RK judgments like recognition judgments are also made on a strength based continuum. As such, these theories propose that task difficulty may be all that is required to produce a dissociation in RK judgments. For example in Donaldson’s (1996) account of a recognition task,
participants must establish a criterion at which point an item can be labeled ‘old’ and all memories that fall beneath this criterion are labeled ‘new’. When making RK judgments another criterion is established for R responses and everything that is ‘old’ but below this second criterion is identified as a K response. Using Tulving’s (1985) recognition experiment as an example this means that after a greater interval between study and test, R responses would be affected simply because task difficulty is confounded across conditions. Thus as the memory weakens over the longer retention interval more items would fall below the R criterion but above the new criterion relative to the number that fall above the R criterion.

It may be that two processes underlie recognition memory and it may be that sometimes RK judgments adequately reflect those processes. However there has been considerable disagreement in the literature about both of these proposals. The following section reviews some attempts to validate a role for content recollection in recognition memory and, in doing this, I further explore attempts to validate the RK procedure.

Can RK Data Adequately Reflect Two Distinct Processes?

There has been some evidence that estimates of recollection and familiarity obtained from the RK paradigm, the processes dissociation procedure (PDP) and from receiver-operating-characteristic (ROC) curves converge (i.e., Yonelinas and Jacoby, 1995; Yonelinas, Kroll, Dobbins, Lazzara & Knight, 1998, Arndt and Reder, 2002). However Malmberg (2008) has reviewed this literature and has concluded that the estimates are not sufficiently similar to support the conclusion that these procedures are producing estimates of the same constructs. This does not necessarily indicate that content recollection is not involved, but rather that there is not enough convincing evidence to categorically conclude that the different procedures are measuring what they are meant to measure.

ROC data in particular has been used along with RK data in order to provide converging evidence for either a single or dual process theory of recognition memory. To explain, the ROC is based on confidence ratings given by participants in regards to a comparative old/new decision. The different levels of confidence are interpreted as different criterion settings, so that a high confidence old response corresponds to a high, conservative criterion and a high confidence new response to a low liberal one. When graphed, the ROC curve is a plot of the proportions of the ratings to hits as a function of the proportions of ratings to false alarms. From a dual process point of
view of recognition memory, familiarity is thought to produce a curvilinear plot. Whereas recollection is thought to represent a threshold retrieval process so that R responses are only given to the highest confidence responses (i.e., Yonelinas, 2001). However, the single process view assumes that R and old responses are merely high and low confidence recognition decisions, thus a curvilinear ROC in line with signal detection theory is predicted for all responses.

Yonelinas (2001) found that, across three different measurement procedures, familiarity was well described by signal detection theory and recollection by a threshold process (since R responses were clearly identified with only high confidence responses). However, Rotello Macmillan and Reeder (2004) provide alternative interpretations for some of this data, their experiments indicated that RK responses tended to reflect more graded information. However they did not find evidence that this occurred along one continuum as proposed by Donaldson (1996), instead according to Rotello et al., (2004), an evaluation of the data led to their conclusion that neither the dual process nor single process theories provided an adequate fit for the data, instead they came up with an entirely new model (the STREAK model). In their review however, Wixted and Stretch (2004) determined that when procedures across experiments were controlled ROC and RK data could produce similar ROC slopes in line with a signal detection theory account for RK judgments. It remains unclear as to which interpretation is superior, however it is clear that RK data can produce evidence both for and against each model. While it may be that recollection and familiarity were involved with such tasks, it does not necessarily follow that the RK task was adequately reflecting those processes.

Neuroimaging data has further been explored in order to validate RK judgments. Different brain regions are thought to be involved, generally such that the hippocampus supports recollection, while familiarity is supported through other regions, such as the perirhinal cortex (i.e., Yonelinas et al., 1998; Yonelinas, Otten, Shaw and Rugg, 2005; Rugg and Yonelinas, 2003; Yonelinas, Kroll, Quamme, Lazzara, Suave, Widaman and Knight 2002). Some evidence, using a combination of ROC data and RK data has been used to verify this observation. For example, Yonelinas et al., (1998) provided convincing evidence as he found that amnesics with hippocampal lesions have a selective detriment to their recollection process, but not the familiarity process. However Wais, Wixted, Hopkins and Squire (2006) found that amnesics produce similar ROC’s to controls when their memory was strengthened.
(here amnesics studied ten words for a recognition test whereas controls studied 50 words). Therefore an entirely conceivable account is that the increased brain activity associated with a R response in Yonelinas et al., (1998) reflects a corresponding higher level of activity associated with the retrieval of strong memories. In support of the idea that the hippocampus may be involved in stronger memories, not necessarily as the result of recollection, Wixted and Squire (2010) have shown that high confidence K responses activate the hippocampus just as do R responses.

**RK Judgments Vary Depending on Task Attributes**

A further method commonly utilised in order to determine that RK judgments reflect different processes is to demonstrate that independent variables have a different but interpretable pattern of results on R and K judgments. Dissociations have been observed in the literature across a number of variable manipulations. For example Reder et al., (2000) found that low frequency words were more likely to be accompanied by recollection than familiarity and the opposite was observed for high frequency words. Also, Gardiner, Gawlik and Richardson-Klavehn (1994) found that recollection supported elaborative rehearsal after a short cue delay in a directed forgetting task, while knowing supported maintenance rehearsal after a long cue delay. As well as many other examples (i.e., Dobbins, Kroll & Yonelinas, 2004, Gardiner and Java, 1990, Gardiner, Gawlik & Richardson-Klavehn, 1994, Rajaram, 1993, Rajaram 1996, Rajaram 1998, Rajaram and Geraci 2000). However, despite all of this evidence, Dunn (2008) reviewed 37 such studies and found that the data was also consistent with unequal signal detection theory. Showing that when variation in decision criteria is controlled, the two memory components appear to show little tendency to vary independently. Therefore it becomes difficult to know if one is really dissociating recollection from familiarity or just varying the criterion for RK decisions.

In accordance with dual process logic R responses should not be affected by a change in criterion placement, as the conscious state of recollection should be independent of other decision processes. However K responses may vary depending on decision manipulations. In contrast, a signal detection explanation would predict movement for both R and K responses. There was some evidence that a criterion shift failed to affect R judgments from Strack and Forster (1995). In their study recognition instructions were manipulated to indicate that either 30% of the test list items had been studied or 50% (while it was always the case that half the items were old and
half were new as in a standard recognition task). In agreement with dual process predictions, they found that R responses did not vary with the manipulation however K responses did. However, Hirshman and Henzler (1998) found that when the manipulation was strengthened such that participants were told that either 30% or 70% of the test items would be old then R responses did vary. Specifically more R and K judgments were made with the less conservative criterion.

In general manipulating instructions has often led to changes in the way in which RK judgments are made and this has further sometimes led to changes in recognition performance. For example, Mather, Henkel and Johnson (1997) found that instructing participants to give a more detailed examination of recollective detail at test resulted in higher recognition accuracy. Specifically RK judgments were compared with MCQ (Memory Characteristics Questionnaire; Johnson, Foley, Suengas & Raye, 1988) data, which provided a more stringent evaluation of recollected content (i.e., questions probed for information across a wide array of attributes such as perceptual, semantic and confidence) than the more global RK instructions. Mather et al., (1997) noted that this may be because the MCQ condition might increase the number of characteristics recalled and therefore more information would be available for the participant in order to make an old/new decision.

Rotello, Macmillan, Reeder and Wong (2005) found that when more specific instructions were given as in Yonelinas (2002) (subjects were told that they might need to explain their R judgments to the experimenter) very few R responses to words recognized with lower confidence were given. A neutral condition (with instructions modeled off Rajaram, 1993) resulted with subjects providing R responses to words recognized over a range of confidence levels. Here the instructions produced data consistent with either a dual process or single process result. Rotello et al., (2005) commented that the Rajaram (1993) instructions are standard and often the subsequent data produced from these instructions have actually been consistent with a dual process view. In addition, adding a third guess category has been shown to change the pattern of RK results (i.e., Gardiner, Richardson-Klavehn & Ramponi, 1997). That instructions and response measures can change the number of R responses is particularly problematic for dual process theories as their conceptualization of recollection is that it should not be affected by such manipulations.
In response to the difficulties noted for a dual process interpretation of RK data, Parks and Yonelinas (2007) and Yonelinas, Aly, Wang and Koen (2010) have indicated that greater care must be taken with the wording of RK instructions. That is, only strict instructions regarding recollection are likely to yield accurate estimates of recollection gained through R judgments (in accordance with the Yonelinas (2002) results). While this may be true, this may also be yet another way of moving the RK criterion without really gaining an accurate measure of recollection and familiarity in the process.

**RK and Accuracy**

As outlined previously, it is generally assumed by dual process models that R judgments provide an indication that qualitative aspects of the study episode have been recalled (Gardiner, 1988; Rajaram, 1993; Yonelinas and Jacoby, 1995). The theory initially did not take account of R judgments made to false alarms and they were rarely considered within the literature. Yonelinas (2002) treated R-false alarms as error and proposed that true recollection may be estimated by substracting R-false alarms from R-hits. However, others have viewed R-false alarms as more interesting than error that should be removed. For example, Higham and Vokey (2004), had participants make R and K judgments without first making a recognition response. Under these conditions participants made a large number of R responses to new words. Some have even used the RK paradigm in an effort to further understand false memory (i.e., Rodieger and McDermott, 1995; Whittlesea, 2002). A dual process view has some trouble accounting for the fact that this research has often found that R responses are made to false memories, however other models such as the SMF previously mentioned can accommodate false remembering.

In support of the idea that RK judgments are made along a continuum, K judgments have been found to support source accuracy. According to the dual process account of recognition memory, K responses should not reflect recollection and therefore should not be useful in source identification. Although a R response tends to always elicit significantly more source accuracy at test than a K response, K judgments still produced source accuracy above chance levels (i.e., Conway and Dewhurst, 1995). In addition, Hicks, Marsh and Ritschel (2002) found that a high percentage of know judgments were made after an accurate source decision. Therefore, while R responses may contain richer source information, a feeling of familiarity may still suffice when making source decisions. It should be noted that
within these experiments K responses could not be made due to an increased familiarity for more recently viewed items as these were not list discrimination studies. It is possible and in general, always a concern, that participants may not have a firm grasp of what a K response is meant to engender, as such participants may be recollecting partial detail and incorrectly labeling their experience as K. It is also possible however that K judgments can reflect a type of familiarity that is source specifying, (e.g., a memory that includes some kind of contextual information without the addition of recollected content) an explanation at odds with a dual process conceptualization of familiarity.

The only firm conclusion that can be drawn from studies utilising the RK task is that all possible expectations for the task appear possible. The critical implication is that content recollection may or may not occur, however the degree to which it is necessary for source memory or a list discrimination task is uncertain given the variable nature of the data. There may be issues with the task, however it may also be that content is sometimes important and sometimes not. The following section explores the idea of content recollection that is not diagnostic or useful in a list or source discrimination task.

**Criterial and Noncriterial Recollection: Implications for Episodic Recognition.**

The idea that the content that comes to mind is the information used to specify source is intuitive. However there is also some intuitive understanding that perhaps there are situations where content is more likely to be helpful than others. As an example, consider the design of an experiment Reder et al., (2000) conducted in order to ascertain if R judgments were more likely to reflect contextual information about the encoding event. Using a list discrimination task they made sure that the font type and background colour were different for every list. Reder et al., (2000) found that correct list discrimination occurred more often after R judgments were made than K judgments. The distinctive features that accompanied list membership (font type and background colour) increase the available content in memory indicative of list membership. In such a design it appears entirely likely that participants would be more likely to make list discrimination judgments due to content recollection.

As well as indicating that the most stringent of instructions should be used when utilising the RK paradigm, Yonelinas et al., (2010) notes that procedures should be used that increase the likelihood that content recollected will be discriminating.
Naturally the more available recollected content is to the memory system then the more likely it will be helpful on a list discrimination task. Many studies have found that R judgments are easily manipulated by varying the level of processing at study (i.e., Gardiner 1988; Rajaram 1993; Gardiner and Parkin, 1990). However, apart from those who have been interested in false recollection (i.e., Roediger and McDermott, 1995) little consideration has been given to the idea that in some tasks in which the RK paradigm is applied, participants may have very little to base a R response on, and in such a case, what might a R response represent? This may be the reason why sometimes R judgments are made to false alarms. To comment that content is helpful only under certain conditions does not explain how necessary content is in an episodic task, it only explains how useful it is in an episodic task that can easily enable the recollection of source specifying content.

Others have often considered the real life examples that support a role for recollected content. For example, the idea of mentally reliving a previous moment, filled with detail that can involve many content rich experiences, such as greeting your father at the airport after a long separation, your first overseas trip or going on a skiing holiday with a group of friends. However, if you always have the same breakfast, at the same time in the morning and in the same spot then there is very little content available to distinguish yesterday’s breakfast from eating breakfast last week. Yet despite this, common experience also allows that we can still have some ability to distinguish such episodes. This possibly has implications when we consider that an episodic memory task is though to be accomplished via the retrieval of source specifying content. That is, content may frequently be available to the memory system however the content recalled may not always be helpful in discriminating between different episodes.

In consideration of issues raised regarding the RK procedure, Yonelinas and Jacoby (1996) went on to explore a possible role for the kind of recollection that does not discriminate between sources in a source-monitoring task. Yonelinas and Jacoby (1996) had participants study words that were either presented on the left or right side of the screen, in small or large font. Participants were explicitly instructed to remember the words and their position on the screen (and were encouraged to associate one side of the screen with a particular person or place), but were not told to focus on the change in font size. There were four test conditions, one tested source monitoring for the word position (left or right) which was considered to be the easy
discrimination condition, while the other tested source monitoring for font size (difficult discrimination condition). Participants responded under speeded or non-speeded response conditions. Participants responded, “yes” if the word was on the appropriate side/font, or if they could not remember how the word was presented at study but thought the word had been studied, or “no” if the word was new or if it was presented on the inappropriate side/font size. Inclusion responses were counted as the participant responding “yes” if they accept the word from the left side under left side instructions and if they accept the word from the right side under right side instructions. The same was true for the difficult discrimination (font size) condition. Exclusion responses were counted as participants responding “yes” to accepting a word from the left side under right side instructions and vice versa and the same for the font size condition. According to process dissociation logic, recollection (R) was estimated as the difference between inclusion (I) and exclusion (E) scores ($P(R) = P(I) - P(E)$). Familiarity (F) was estimated as the probability of accepting an item under exclusion conditions, divided by 1 minus the probability of recollection ($P(F) = P(E)/(1-P(I)+P(E))$).

Yonelinas and Jacoby (1996) found that estimates of recollection were greater when participants had more time to respond and in the easy discrimination condition, thus supporting the dual process assumption that recollection is effortful. In addition estimated familiarity did not vary between the response speed conditions which is congruent with the assumption made that familiarity is automatic (therefore always fast and unaffected by a speeded response). However, estimated familiarity did increase in the hard discrimination condition compared to the easy condition. This is contrary to the assumption that familiarity is acquired in an automatic fashion. Yonelinas and Jacoby (1996) attributed the change in estimated familiarity to a type of recollection that did not discriminate between sources. They postulated that this type of recollection would have been more likely to occur in the difficult discrimination condition. That is, during the difficult source task, recollecting which side the word had occurred on would be irrelevant to the size decision but would still have been useful in deciding whether the word had occurred in the experiment. They termed this form of recollection ‘noncriterial’. Yonelinas and Jacoby (1996) argued that if noncriterial recollection were acting like ‘criterial’ (source specifying) recollection than it should have been strategic and therefore differed under speeded and non-speeded conditions. In the difficult discrimination condition estimates of
familiarity were not changed by speeded responding, therefore the authors proposed that sometimes content recollection can occur that does not inform of source and it can occur quickly or automatically and thus act more like familiarity, rather than being the result of an effortful process (however see Parks, 2007).

There are a number of implications of noncriterial recollection that Yonelinas and Jacoby (1996) fail to explore. Yonelinas and Jacoby (1996) acknowledge that noncriterial recollection is problematic for the RK paradigm, considering that R responses do not distinguish between criterial and noncriterial information. This could account for the results I mentioned earlier that indicate that R responses are sometimes reported based on very partial information and can support inaccurate memory judgments. However, this must also be considered a problem for the inclusion condition of the PDP. Successful responses to inclusion questions might be accompanied by both criterial and noncriterial information, however this is not considered by Yonelinas and Jacoby (1996). The authors further do not consider the role that retrieval intention plays in their results. That is, they state that recollected content that participants intend to retrieve is effortful, whereas recollected content that they don’t intend to retrieve is not effortful, without specifically addressing this change in retrieval intention. It may simply be that a certain amount of memorable information comes to mind quickly, though this may not necessarily be criterial in nature. In the case that participants need to look for discriminatory information then this process may involve something more. This could require the use of contextual cues or some kind of directed search. Regardless of how this process may occur, this issue has implications for estimates of recollection and familiarity gained from the PDP.

*Is Recollected Content, as Reflected by R Judgments, Useful in List Discrimination or Simply More Available to the Memory System Under Certain Conditions?*

As noted, Yonelinas and Jacoby (1996) found that estimated familiarity was higher in the difficult discrimination condition and they assumed that this was due to the use of noncriterial information. However when they found that the use of this information did not vary with the speed of responding they assumed that it was acting like familiarity, which violated the assumption of content recollection. As an alternative, Humphreys et al., (2003) proposed that information that participants were not intending to recall was experienced as familiarity and not as recollection.
Therefore, the alternative was that assumed equality of familiarity for both the inclusion and exclusion conditions could be violated as participants may rely on familiarity more in the inclusion condition than the exclusion condition. This possibility would undermine estimates of recollection and familiarity gained from the PDP.

In their second experiment Humphreys et al., (2003) asked participants to first make recognition judgments to heard or read words along with RK judgments and then gave participants a source monitoring task presented in the exclusion format. Half the participants were asked to respond “yes” if the words were heard and to respond “no” if new or not heard, the others were asked to respond the same way for read words. These instructions were referred to as congruent if the process of studying the items was congruent with the test question (i.e., studied read words (studied visual), asked to say “yes” if the word was read (tested on visual)). They compared the probability of saying “yes” to incongruent or congruent questions on test 2 conditional on RK judgments from test 1. They found that for both congruent and incongruent questions, the probability of a “yes” response conditional on a R response was greater than the probability of a “yes” response conditional on a K response. According to dual process assumptions the occurrence of recollection on test 1 should have made it more likely that recollection would occur on test 2. However, recollection on test 2 should have supported a “yes” response only in the congruent condition. In the incongruent condition recollection should have supported a “no” response. Furthermore, source accuracy was greater for congruent questions given a K response than incongruent questions given a K response. Thus K responses could support correct source decisions and R responses could support incorrect source decisions. While it may be that sometimes content may be useful in order to perform the exclusion condition, the Humphreys et al., (2003) results support the idea that content recollection is not always necessary for a source decision.

Humphreys et al., (2003) suggested that by asking two different types of questions they were changing the retrieval intention of participants between conditions. That is, participants were looking for different types of evidence depending on the question asked (heard or visual). For example, if the question asked is about the visual content of a test item when that item had been heard at study then visual content would not come to mind, as there is none. In addition, if auditory content comes to mind the participant will respond, “no” not “yes”. Thus it seems
feasible that information that participants were not intending to recall was experienced as familiarity. That is, they may have been responding, “yes” to stronger memories that may feel like a strong sense of familiarity. This poses a problem for estimates of recollection and familiarity in the PDP, as it may be that different information supports the answers to inclusion and exclusion questions simply because those conditions ask for different information, not necessarily because one type of information is necessary for one condition and not another.

Results from Bodner and Lindsay (2003) indicate that certain contextual features of the test list may influence the likelihood that participants will willingly report noncriterial recollection. Employing a level of processing (LOP) task Bodner and Lindsay (2003) found that reported content recollection, as indicated by R judgments, could be affected by the test list context. In their first experiment they had participants study two lists with strength of LOP manipulated between groups. Both groups studied the same list of medium items, however one group also studied a list of shallow items (medium-shallow) and the other group studied a list of deep items (medium-deep). Participants were then tested on those items along with new items. They found that although recognition was not greatly affected by the strength manipulation, that R judgments were much more likely to be made to medium items in the medium-shallow than the medium-deep test context condition. Their second experiment confirmed that this was an effect of test context rather than study context. In their third experiment they ran a source-monitoring task instead of a RK task and found that source accuracy did not differ for the medium group between test context conditions. Therefore, both recognition performance and source identification for the medium items did not differ even if R responses did. One possibility, offered by Bodner and Lindsay, is that content was useful for identifying medium items, however different recollective experiences dictated by the test list context led to different definitions of R and K for those items. A related interpretation to consider is that at times content was not judged to be useful (at least for medium items) and was simply more likely to be reported in one condition for medium items than the other.

The fact that performance on medium items did not change while R responses did, suggests that the two dependent variables were not strongly related. This conclusion is further supported by the fact that the same information was reported in one condition and not the other. One would expect that if recollected content were helpful in making a memory decision then it would be just as likely to be reported in
both conditions considering that recognition performance did not differ. Therefore it seems entirely possible that the information reported to medium items in the medium-shallow test condition was largely noncriterial, that is, it was not helpful in completing the task. Furthermore, the fact that participants only reported it in one condition indicates that participants may have been aware that the information retrieved was not useful. At the very least the Bodner and Lindsay (2003) results have implications for the RK paradigm, specifically that contextual elements can change the way that information is reported. It is also possible that when this information is objectively unhelpful or is not perceived to be helpful, participants may still label their subjective experience as R. The incentive to report noncriterial recollection is most likely due to demand characteristics of the RK task.

Typical RK instructions start by first introducing to participants the idea that recognition memory can be accomplished via two processes, recollection or familiarity. This is instead of giving participants the option of responding R or K when they feel like either process is helping them to make their recognition decision. It is of little doubt then that participants might then expect to use either recollection and familiarity and therefore may report R simply because the instructions have informed them that recollection should occur. In Bodner and Lindsay (2003) this may have resulted in R responses to noncriterial information. Instructions do not usually indicate that not all recollected information might be helpful, however the instructions for participants to focus on criterial information used in Yonelinas (2002) might help to prevent the reporting of R to noncriterial information.

While it is clear that participants can sometimes retrieve content, it remains unclear as to how useful this content can be. The PDP assumptions regarding the role of content recollection may not hold under certain conditions (Humphreys et al., 2003) and there is some evidence to suggest that content recollection is not always useful, though still may be reported (Bodner and Lindsay, 2003). In partial support of Humphreys et al., (2003), Jacoby, Shimuzu, Daniels and Rhodes (2005) found that directing participants to a particular source at test produced a qualitative change in the type of information used for memory judgments. This led Jacoby et al. (2005) to propose that under some conditions, source information (context) is used to constrain what comes to mind during retrieval, and that this in turn would affect what information is used to make a recognition decision. It may be that this occurred in Jacoby et al. (2005); however in Bodner and Lindsay (2003) reported information
changed for the same items when memory performance didn’t, and in Humphreys et al., (2003) if participants were relying on recollection at all, then it supported both correct and incorrect source judgments depending on the question asked. I cannot rule out the possibility that content recollection is sometimes useful during source monitoring or episodic recognition tasks. However, I cannot decipher whether it is content that is directing the memory decision (as assumed by Jacoby and colleagues) or if recollected content occurs after the memory decision or not at all, as assumed by Humphreys et al., (2003).

**Research Questions and Proposed Methodology**

I do not deny that typical recognition testing with old/new instructions will sometimes be accompanied by the subjective experience of remembering thoughts and memories of other words and that these recollections will be generally attributable to what happened at study as indexed by a R judgment. However I would like to know which of the following interpretations is correct:

1) Recognition decisions in a list specific task are not based on content recollection.

2) Content recollection can occur but is not pivotal for list discrimination.

3) Content recollection, as revealed in subjective reports, plays a crucial role in list discrimination or source monitoring.

Put simply, I wish to ascertain how necessary content recollection is in a list specific recognition task. Ascertaining the usefulness of content recollection in recognition tasks has been hindered by a general preference for single study and test list recognition paradigms. In such single list designs, successful recognition can be achieved through a number of different paths. It can occur via acontextual familiarity, contextual familiarity, recalling a list item using other list items or a category label as a cue or through content recollection. However, by using a task that will limit the use of content recollection I can test the possibility that content recollection is not the only way to perform episodic recognition. My proposal is to have participants study and be tested on multiple lists constructed of the same items. In this list-specific design, the same list of words will be tested on each list, however the targets and distracters will be randomized on every list. In this situation the thoughts and reactions one has while studying a word will become less unique in each successive list. Thus content should become more and more irrelevant over the repeated presentations in different lists. Because the same words will be presented on every list
all of the words will also become increasingly familiar. Therefore participants will not be able to easily rely on acontextual familiarity or content recollection in order to perform the task. Thus the task will be episodic, however participants should not find content recollection useful.

Despite the increased difficulty of the task, it should be possible for participants to discriminate old from new items even by the last list. I state this mostly because similar methodology has been applied elsewhere for a similar purpose, (i.e., Weeks, Humphreys and Hockley 2007, Heathcote, Raymond and Dunn, 2006). While performance was negatively affected in these studies, participants were still able to perform the task. It is possible for participants to complete this kind of task if participants have the opportunity to reinstate the study context (i.e., Humphreys, Bain and Pike, 1989). This might be helped by the addition of instructions indicating to participants to recognize items from the previous list only. Participants may also spontaneously reinstate aspects of the study context (Smith and Vela, 2001). In the event that both a dual process interpretation is ruled out and participants are able to still perform the task, then further consideration will be given to how successful recognition might be achieved in the final discussion.

Further to examining the role of content recollection, I can also examine a question related to my content recollection question. I would further like to understand what the relationship is between accuracy of R judgments and accuracy of recognition by including a RK task. After making an old recognition judgment, participants will be asked to describe their recognition as R or K. I am not sure what will happen to R responses as recollection becomes increasingly noncriterial in nature. As the Bodner and Lindsay (2003) results indicate, participants might realize that their subjective experience is noncriterial and stop making R responses. Alternatively they may continue to make R responses. The methodology I employ should make criterial recollection quite difficult to achieve, the inclusion of the RK paradigm will either support this or may reflect noncriterial recollection. If R judgments support hits rather than false alarms however then I may need to consider the possibility that content is still useful.

Overview of Chapters

Chapter 2 includes three experiments. All three experiments aim to understand how useful content can be in an episodic recognition task using the basic design described previously. In addition these experiments tested how making RK decisions
may affect the recognition task. This sometimes involved asking these decisions at a delay or asking these decisions on every list or only on the final list. Attempts were made to reduce the likelihood that participants would be affected by the knowledge of recollection and familiarity, so that it could be understood as to how the act of making judgments might affect the task rather than how participants might approach the task with prior knowledge of RK. Results from these studies indicated that participants could perform the task, however it was difficult to interpret whether or not content was useful to participants.

Chapter 3 includes three experiments that replicate and extend on the experiments included in chapter 2. In particular, a new method of assessing a relationship between RK judgments and hits was introduced in this chapter which indicated that content was not helpful for the majority of participants. The addition of a condition where participants were asked to justify their R responses further supported this conclusion.

The final experiment included in chapter 4 clarifies a number of prior effects. In particular it was possible to examine speculation that the introduction of the RK procedure may encourage a reliance on recollection without familiarity. The inclusion of a longer study-test retention intervals (as had been applied in experiments 1 to 4) across all conditions and lists allowed me to examine how increasing retention interval potentiated a decrease in recognition performance on the final list with the inclusion of the RK task.

CHAPTER 2

Experiment 1

It is possible that the act of asking participants to make R or K responses will change the way in which participants perform the task (Humphreys et al., 2003, Dennis and Humphreys, 2001). In particular Humphreys et al., (2003) point out that whenever two judgments occur together, we might expect the first judgment to possibly be affected by the knowledge that another judgment will follow. Furthermore, typical RK instructions begin by informing participants that their recognition of words will involve recollection and familiarity. This strongly suggests to participants that they should use the two processes, thus participants may actively try to recollect details when they otherwise would not in order to satisfy the RK task
demands. If my interpretation of Bodner and Lindsay (2003) is correct then it is also possible that these instructions induce participants to make an R response even if they are aware that their recollection is noncriterial.

In experiment 1, participants received four study and test lists. RK judgments were made only on the final list under two between subject conditions. In the delayed condition, participants made RK judgments after making all of their recognition judgments for that list. In the immediate condition RK judgments were made as they usually are, that is after each recognition decision. I introduced the RK decisions after the study session on the final trial to rule out the possibility that participants would change the way they study the items due to their anticipation of having to make a RK decision. The delayed condition was an attempt to try and rule out the possibility that participants may change the way they perform the recognition task while making RK judgments.

Method

Participants

40 students from The University of Queensland participated in the study. Participants were randomly assigned to one of two conditions, delayed or immediate. The data from 5 participants was lost due to computer error, which left 18 participants in the delayed condition and 17 participants in the immediate condition. The variables to be measured were both RK decisions and recognition accuracy.

Procedure and Materials

The experiment was computer administered. The Nelson, McEvoy, and Schreiber (1998) norms were used to find 40 medium frequency unrelated words for the study. Every participant received four study and test lists. The same pool of 40 words was randomised to produce different targets and distracters on every list without any restrictions regarding how often a word was studied. Participants studied a list of 20 words; each word appeared one at a time and remained on the screen for two seconds before the next word appeared. At test one word was presented to the center of the screen and participants were asked to indicate if that word was “OLD” or “NEW” by clicking on the appropriate on screen button.

Participants were warned when they were about to enter a study or test phase. From the second list onwards, participants received test instructions before each test phase that explicitly stated that they were only to respond old to items that they had
studied on the previous study list, and not to words that may have appeared on prior lists.

On the final list participants made RK decisions either immediately or at a delay. In the immediate condition after making an old recognition decision participants were presented with two more buttons; “Remember” or “Know”. Participants were asked to describe their recognition experience as R or K by clicking on the appropriate button. RK instructions were modeled off Rajaram (1996) and were presented to participants before the test phase of the final list.

In the delayed condition participants completed the recognition task on the final list as they had on previous lists. After this, participants were presented with the items they had just been tested on. All items remained in the same list position from the recognition test in order to make the RK task easier to complete. Participants received RK instructions before the RK test but after the recognition test on the final list. During the RK task, participants were presented with one word at a time along with three response buttons ‘Remember’, ‘Know’ and ‘New’. They were clearly instructed that they were to describe their prior old recognition judgments as R or K and click on the ‘new’ button if they had previously indicated that the item was new in the recognition task.

Results

Recognition

Table 1. Experiment 1. Hits and false alarms broken down by list number and RK response interval

<table>
<thead>
<tr>
<th>List</th>
<th>Hit</th>
<th>False Alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td>List 1</td>
<td>.847</td>
<td>.064</td>
</tr>
<tr>
<td></td>
<td>.809</td>
<td>.047</td>
</tr>
<tr>
<td>List 2</td>
<td>.817</td>
<td>.272</td>
</tr>
<tr>
<td></td>
<td>.827</td>
<td>.327</td>
</tr>
<tr>
<td>List 3</td>
<td>.800</td>
<td>.244</td>
</tr>
<tr>
<td></td>
<td>.829</td>
<td>.315</td>
</tr>
<tr>
<td>List 4</td>
<td>.792</td>
<td>.261</td>
</tr>
<tr>
<td></td>
<td>.832</td>
<td>.606</td>
</tr>
</tbody>
</table>
The results of a 2 between RK response interval (delayed vs immediate) by 4 within (list: 1, 2, 3, 4) ANOVA on hit rates (see Table 1) failed to produce a significant main effect of RK response interval, $F(1, 33) = .082, p = .776, MSE = .046, \eta^2_p = .002$; or list, $F(3,99) = .114, p = .951, MSE = .016, \eta^2_p = .003$ and no significant interaction between list and RK response interval was observed, $F(3,99) = .674, p = .570, MSE = .016, p = .570, \eta^2_p = .020$.

The same analysis conducted on false alarms revealed a main effect of RK response interval, $F(1, 33)=5.558, p = .024, MSE = .080, \eta^2_p = .144$ such that the immediate group were more likely to incorrectly respond ‘old’ ($M = .32$) than the delayed group ($M = .21$); a main effect of list was also noted, $F(3,99) = 35.615, p < .001, MSE = .024, \eta^2_p = .519$. Follow up comparisons revealed that, all lists significantly differed ($p < .05$) except list 2 and 3 ($p = .516$). In addition, the analysis revealed a significant interaction between list and RK response interval, $F(3,99) = 9.168, \ p < .001, MSE = .024, \eta^2_p = .217$. The interaction between list and RK response interval can be viewed in Figure 1. When followed up by examining simple effects it was found that more false alarms were produced on list 4 than all prior lists in the immediate condition ($p < .05$), however this was not observed in the delayed condition.

![Figure 1](image-url)  
*Figure 1. Experiment 1. False Alarms as a function of list and RK response interval*
Table 2. Experiment 1. RK judgments as a function of recognition accuracy and RK response interval

<table>
<thead>
<tr>
<th></th>
<th>Delayed</th>
<th></th>
<th>Immediate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>K</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>Hit</td>
<td>.567</td>
<td>.297</td>
<td>.477</td>
<td>.371</td>
</tr>
<tr>
<td>FA</td>
<td>.439</td>
<td>.294</td>
<td>.344</td>
<td>.279</td>
</tr>
</tbody>
</table>

My interest was primarily on how R judgments were affected by the task and therefore only R judgments are analysed (see Table 2 for a full summary of RK data). A 2 between RK response interval (delayed vs immediate) by 2 within accuracy (hit vs false alarm) mixed ANOVA was conducted on R responses. The results of this analysis revealed that R judgments were significantly more likely to be made to hits ($M = 0.52$) than false alarms ($M = 0.39$), $F(1, 33) = 11.280, p = .002, MSE = .026, \eta_p^2 = .255$. The effect of accuracy did not vary by RK response interval ($F(1, 33) = 0.003, p = .953, \text{ns}$), nor did the delayed and immediate groups significantly differ ($F(1, 33) = 1.548, p = .222, \text{ns}$).

Discussion

The most interesting result from experiment 1 was that the introduction of the RK paradigm in the immediate condition significantly increased the likelihood that participants would respond old to test items on the final list. Despite this participants could otherwise perform the recognition task quite well. Participants in the immediate condition were also adept at making R responses more to hits rather than false alarms. Although participants were more likely to make R responses to hits than false alarms in the delayed condition, participants in the delayed condition were clearly not performing the RK task very well. Summing across R and K judgments made to hits in table 2, the number of hits on the delayed test exceeds the mean number of hits made on the recognition test (see Table 1) by .072. While this is not all that troubling, the mean number of false alarms on the delayed test exceeds the mean number of false alarms on the recognition test by .472. Taken together, it is clear that participants were likely to overestimate how many old decisions they had made on the recognition test. It may be that participants are not remembering their recognition judgment and are on the second test making that judgment again. If this is the case then by the time
participants perform the delayed test then the extra familiarity from the recognition test makes most items appear old.

Experiment 2

Experiment 2 was devised to replicate the immediate condition of experiment 1 and also to explore how RK judgments might change from list to list. It was queried as to whether or not the poor performance on the final list would still be present if participants were well practiced at the task. In this case, it was considered that participants may become more adept at evaluating their R judgment (e.g., they may have become aware that the content recollected was noncriterial). It was also possible that a gradual decline in recognition performance may occur in line with the results from the final condition. Like experiment 1, experiment 2 will again have 2 between subject conditions. The final condition will replicate the immediate condition in experiment 1. In the continuous condition, participants will make RK decisions on all four trials. Judgments in both conditions will always be made immediately.

Method

Participants

40 participants were randomly assigned to one of two conditions, final or continuous. The variables to be measured will be both the nature of the RK decisions and recognition accuracy. Due to computer error the data from two participants was lost leaving 19 participants in each group.

Procedure and Materials

The procedure and materials were largely identical to that in experiment 1 with two exceptions. The first was that all RK judgments were made immediately. The second was that a continuous condition was added, such that participants made RK judgments on every list. Participants in the continuous condition received RK instructions before the test phase of the first list, while participants in the final condition received RK instructions before the test phase of the final list.

Results

Recognition
Table 3. Experiment 2. Hits and false alarms as a function of list and RK frequency

<table>
<thead>
<tr>
<th>List</th>
<th>Hit</th>
<th>False Alarm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Final</td>
<td>Continuous</td>
</tr>
<tr>
<td>List 1</td>
<td>.868</td>
<td>.837</td>
</tr>
<tr>
<td></td>
<td>.053</td>
<td>.068</td>
</tr>
<tr>
<td>List 2</td>
<td>.811</td>
<td>.768</td>
</tr>
<tr>
<td></td>
<td>.247</td>
<td>.255</td>
</tr>
<tr>
<td>List 3</td>
<td>.813</td>
<td>.763</td>
</tr>
<tr>
<td></td>
<td>.284</td>
<td>.224</td>
</tr>
<tr>
<td>List 4</td>
<td>.832</td>
<td>.737</td>
</tr>
<tr>
<td></td>
<td>.447</td>
<td>.168</td>
</tr>
</tbody>
</table>

A 2 between RK frequency (continuous vs final) by 4 within (list) mixed ANOVA conducted on hit rates (see Table 3) failed to find a significant main effect of list, $F(3, 108) = 2.323, \ p = .079, \ MSE = .018, \ \eta^2 = .061$; nor of RK frequency, $F(1,36) = 2.027, \ p = .163, \ MSE = .056, \ \eta^2 = .053$. RK frequency also did not vary depending on list, $F(3,108) = .418, \ p = .740, \ MSE = 0.018, \ \eta^2 = .011$.

The same analysis on false alarms (see Table 3) produced a significant main effect of list, $F(3, 108) = 23.268, \ p < .001, \ MSE = .019, \ \eta^2 = .393$. Follow up comparisons revealed that list 1 produced fewer false alarms to all other lists ($p < .001$) however no other lists significantly differed (list 2 and list 3: $p = .920$, list 2 and list 4: $p = .109$, list 3 and list 4: $p = .106$). There was no significant main effect of RK frequency, $F(1, 36) = 2.079, \ p=.158, \ MSE = .114, \ \eta^2 = .055$. There was, however, a significant interaction between list and RK frequency, $F (3, 108) = 9.375, \ p < .001, \ MSE = .019, \ \eta^2 = .207$. The significant interaction can be viewed in Figure 2. Follow up simple effects revealed that significantly more false alarms were made on list 4 than all other lists in the final condition, ($p < .05$). However this was not observed in the continuous condition.
Figure 2. Experiment 2. Proportion of false alarms across list and RK frequency condition

**Table 4. Experiment 2. RK judgments as a function of list, accuracy and RK frequency**

<table>
<thead>
<tr>
<th></th>
<th>Final</th>
<th></th>
<th>Continuous</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
<td>K</td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>List 1</td>
<td>Hit</td>
<td>-</td>
<td>.482</td>
<td>.355</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>-</td>
<td>.021</td>
<td>.047</td>
</tr>
<tr>
<td>List 2</td>
<td>Hit</td>
<td>-</td>
<td>.516</td>
<td>.253</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>-</td>
<td>.124</td>
<td>.132</td>
</tr>
<tr>
<td>List 3</td>
<td>Hit</td>
<td>-</td>
<td>.526</td>
<td>.253</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>-</td>
<td>.089</td>
<td>.134</td>
</tr>
<tr>
<td>List 4</td>
<td>Hit</td>
<td>.540</td>
<td>.292</td>
<td>.518</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>.271</td>
<td>.176</td>
<td>.076</td>
</tr>
</tbody>
</table>

A full breakdown of RK results can be viewed in Table 4. A 2 between RK frequency (continuous vs final) by 2 within (accuracy: hit vs false alarm) mixed ANOVA was conducted on R responses. As in experiment 1 I again found that R judgments were more likely to accompany hits ($M = .53$) than false alarms ($M = .17$), $F(1, 36) = 51.046, p < .001, MSE = .047, \eta^2_p = .586$. Accuracy did not vary significantly across the continuous and final conditions, $F(1, 36) = 3.050, p = .089,$
ns. Nor did the final or continuous groups differ significantly, $F (1, 36) = 3.177, p = .083$, ns.

Discussion

As in experiment 1, the results of experiment 2 clearly showed that participants could perform the task adequately. Also, again the introduction of RK judgments on the final trial greatly decreased the ability of participants to distinguish between items that were studied on the previous list and items studied on all previous lists. However, when participants were practiced at the RK task then recognition performance was not adversely affected. Also like experiment 1, R responses were more likely to be made to hits than false alarms.

Although the results from experiments 1 and 2 indicated that the surprise inclusion of the RK paradigm on a demanding task may negatively effect performance, this may have been due to the fact that retention interval varied between conditions. It was necessary to present the RK instructions before the test phase as I did not wish participants to change the way they approached the task. All instructions were self-paced and the RK instructions presented for participants in the final condition would take on average about 40 seconds longer to read through than the same recognition test instructions presented on the continuous condition. The same retention interval difference was also present on the first list of the continuous condition. This did not produce a significant difference between groups, however the task was much less demanding on list 1.

A longer retention interval may also have produced the observed differences between the immediate and delayed conditions in experiment 1. In the immediate condition participants received RK instructions before the recognition test (since both judgments were made at the same time), however RK instructions were presented to participants in the delayed condition after the recognition test and therefore again, the retention interval was larger in the immediate than the delayed condition. Furthermore the presentation of the instructions in the delayed condition (after the recognition test and before making RK decisions) would have made the RK task a lot more difficult. A simple experiment was run in order to directly test if the poorer performance on the final list with the addition of the RK task was still observable when retention interval was controlled.
Experiment 3

In experiment 3 I essentially replicated experiment 1 excluding the delayed judgment condition. Thus one group simply performed the recognition task without RK judgments. The focus from experiment 1 and 2 was placed ultimately on the abrupt change on the final list with the introduction of the RK task and thus the third experiment was focused on clarifying that effect. The experiment was run with a control for the retention interval for the recognition only condition.

Method

Participants
A total of 40 participants recruited from the same source as the previous two experiments participated in the study. Twenty participants were randomly assigned to one of two between subject test conditions, final or recognition only. The variables to be measured were both RK decisions and recognition accuracy.

Procedure and Materials
The procedure and materials were largely identical to that in experiment 1 with the exception that participants in the recognition only condition did not receive a RK task. Also while participants in the final condition received RK instructions after studying list 4, participants in the recognition only condition received a paragraph about confirmation bias, which was tested by friends of the experimenter to take approximately 50 seconds to read through (the same amount of time to read through the RK instructions). This particular paragraph was chosen because it was interesting and novel to participants and therefore likely to engage attention to a similar level to those participants in the final condition reading the RK instructions.

Results

Recognition

Table 5. Experiment 3. Proportion of hits, false alarms and d-prime values broken down by list and RK group

<table>
<thead>
<tr>
<th></th>
<th>Recognition Only</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>List 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>.890</td>
<td>.900</td>
</tr>
<tr>
<td>False Alarm</td>
<td>.118</td>
<td>.108</td>
</tr>
</tbody>
</table>
Instead of simply examining hits and false alarms, I also considered comparing d-prime values in order to examine overall accuracy. D-prime values were calculated with the standard correction (see Table 5) and a 2 between (group: recognition only vs final) by 4 within (list) mixed ANOVA was conducted on those values. There was a main effect of list $F(3,144) = 47.797, p < .001, \text{MSE} = .417, \eta^2_p = .557$. Follow up main comparisons revealed that accuracy declined significantly with each list ($p < .05$). Recognition accuracy did not differ significantly between the groups ($F(1, 38) = 1.120, p = .297, \text{ns}$). The interaction was also not significant, $F(3, 144) = 1.126, p = .342, \text{ns}$).

In order to explore this further I examined hits and false alarms separately (see Table 5). The same analysis conducted on the false alarms and hits reflected the same pattern of results of the d-prime analysis and are therefore not mentioned.

**Remember-Know**

Table 6. *Experiment 3. RK judgments as a function of accuracy in the final condition*

<table>
<thead>
<tr>
<th></th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R</td>
</tr>
<tr>
<td>Hit</td>
<td>.390</td>
</tr>
<tr>
<td>FA</td>
<td>.198</td>
</tr>
</tbody>
</table>

A one way repeated measures ANOVA on list 4 data for standard RK judgments produced a significant difference between R responses made to hits and to false alarms, such that R judgments were significantly more likely to be made to hits
than false alarms \((F (1, 19) = 13.522, p = .002, MSE = .027, \eta_p^2 = .416)\). See Table 6 to view the mean proportion of RK judgments made to hits and false alarms.

Discussion

The results of experiment 3 did not replicate the dramatic change in the false alarm rate on the final list in the RK condition compared to the recognition condition. However a perusal of the hits and false alarms does indicate a trend in the same direction. It should be noted that the increase in retention interval for both conditions in experiment 3 did not affect performance all that badly, or at least, the effect was not as dramatic as what had been observed in experiments 1 and 2. There is some ambiguity surrounding the results of the experiments in chapter 1 as a result of this. While it appears that the difference observed in recognition performance between the two groups might be due to the increased retention interval, I can not conclude this absolutely. It is possible that the effect of introducing the RK task on the final list was present but small and that the longer retention interval potentiated the effect noted in the first two experiments. In addition experiment 3 did not include a continuous group with a control for retention interval, adding a continuous condition might help to clarify the effect of the retention interval compared to the effect of making RK judgments for the first time on the final list.

Another consideration is that the results of the RK paradigm indicate that participants are quite accurate with R judgments and we did not note any group differences in the first two experiments. This seems odd considering that recognition performance was so poor in the final condition in these experiments, yet there were no concurrent group differences in R accuracy, if recollection were helpful then surely the false alarm rate should not have been so high. The traditional way of measuring R responses, may tell us something about accuracy but cannot tell us what this response represents and whether or not content recollection was responsible for the recognition decision. In order to establish how useful content is, I reconsider the typical way of measuring RK judgments and instead offer an alternative in the next chapter.

CHAPTER 3

In order to establish how useful content is, I will have to reconsider the typical way of measuring RK judgments. RK judgments are usually scored by taking the proportion of R responses made to hits and then dividing that number by the total
number of old items. In addition R responses made to false alarms are then divided by the total number of new items. This provides a good measure of how accurate R responses are, however it does not inform us about whether or not recollection was responsible for the recognition decision. The alternative is that the participant first makes a recognition decision and then assigns a R response to some proportion of the items that were identified as old. Whenever participants are more likely to correctly identify an old item as old than to incorrectly identify a new item as old then the procedure of first recognizing and then assigning a R response will result in a larger number of R responses to old words than to new words. That is, the traditional way of presenting RK results will indicate that R responses are accurate but will not tell us that recollection was responsible for the accuracy. Given this issue, my proposal is to calculate a chi square statistic for each participant that will test for a relationship between RK judgments and recognition decisions. If we reject the null hypothesis and if R responses are more likely for hits than for false alarms then such a result would be consistent with the idea that recollection is accurate and is driving the recognition decision. However, if we accept the null hypothesis then we cannot reject the hypothesis that R responses are made to recognized items independently of whether the recognized item constitutes a hit or a false alarm. Note that there is some existing support for the idea that the recognition decision is at least partially responsible for the R response as opposed to the idea that content recollection is responsible for the recognition decision. That is, Higham and Vokey (2004) found that R responses, especially to new words, were more likely in the case that there was no prior recognition decision.

In addition to the above we can examine the relationship between recognition accuracy and the relationship between R judgments and recognition judgments (as measured by the chi square statistic). If participants who achieve a high chi square value and therefore are more likely to respond R to hits than false alarms, also perform well on the recognition task (as indicated by d-prime values) then this would be consistent with the theory that those participants are using content in order to perform the recognition task. However, I may also find a lack of correlation between the two. A dual process explanation would not be supported for example if I found that the majority of participants who performed well on the recognition task failed to produce significant chi square values. Examining the correlation between these two measures should help to clarify this relationship. It should be noted however, that
these techniques can not prove definitively that participants are or are not using content effectively, rather the results of these measures will either be congruent or incongruent with the content hypothesis.

Experiment 4

Method

Participants

40 students from the University of Queensland were randomly assigned evenly to one of two between subject test conditions, final or continuous. The variables to be measured were both RK decisions and recognition accuracy.

Procedure and Materials

The procedure and materials were similar to experiment 2 except that in experiment 4 a control was introduced for the retention interval difference. For participants in the continuous condition, RK instructions were presented before they viewed the first study list, while participants in the final condition received RK instructions before being tested on the final list of words. Five friends of the experimenter were asked to read through the instructions and on average, it took them 50 seconds. Therefore to equate the retention interval across conditions, before receiving the test phase in list 4, participants in the continuous condition read through a paragraph explaining confirmation bias (which was measured to take approximately as long to read through as the RK instructions). Participants read through a paragraph simply to give them an intervening task that was similar to reading through instructions for participants in the final condition. This seemed preferable to a blank screen, where they may have been able to rehearse items.

Results

Recognition

Table 7. Experiment 4. D-prime values, hits and false alarms as a function of list number and RK frequency condition

<table>
<thead>
<tr>
<th></th>
<th>RK Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Continuous</td>
</tr>
<tr>
<td>List 1</td>
<td>Hit</td>
</tr>
<tr>
<td></td>
<td>False Alarm</td>
</tr>
<tr>
<td></td>
<td>d'</td>
</tr>
</tbody>
</table>
List 2

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>False Alarm</th>
<th>d’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.80</td>
<td>0.24</td>
<td>1.53</td>
</tr>
</tbody>
</table>

List 3

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>False Alarm</th>
<th>d’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.77</td>
<td>0.30</td>
<td>1.21</td>
</tr>
</tbody>
</table>

List 4

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>False Alarm</th>
<th>d’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.73</td>
<td>0.29</td>
<td>1.25</td>
</tr>
</tbody>
</table>

In order to examine recognition performance d-prime values were calculated with the standard correction (see Table 7) and a 2 between RK Frequency (continuous vs final) by 4 within list ANOVA was conducted on those values. I found a significant difference between the two groups such that making RK judgments on every list (continuous) resulted in more sensitive decisions than the final group who did not make RK judgments until list four, \( F(1, 38) = 7.218, p = .011, \text{MSE} = .959, \eta_p^2 = .160 \). Follow up simple effects of a significant interaction between list and RK Frequency \( F(3, 114) = 8.19, p < .001, \text{MSE} = .316, \eta_p^2 = .177 \) revealed that the groups differed on every list \((p < .05)\) except the first \((p = .186)\). Unsurprisingly we found a significant main effect of list such that earlier lists produced more accurate responses than later lists \((p < .001)\) except lists 3 and 4 \((p = .193)\), \( F(3, 114) = 68.883, p < .001, \text{MSE} = .316, \eta_p^2 = .644 \).

The between group differences in d-prime scores indicate that participants that made RK judgments on every list seemed to be able to adapt to the recognition task more efficiently by the second list than those participants that only made RK judgments on the final list. In order to explore this further I examined hits and false alarms separately by employing the same analysis that was applied on d-prime values. Only results that add to the interpretation of the d-prime analysis are mentioned. I found that significantly more false alarms were made in the final condition compared to the continuous condition on list 2 and list 4 \((p < .05)\), congruent with the interaction observed between RK frequency and list in the d-prime analysis. In addition, hits differed between groups on list 1, with significantly more hits made in the final condition than the continuous \((p = .011)\). This difference was not observed in the d-
prime analysis due to the similar false alarm rate on list 1 across groups, this is not surprising as false alarms were rare on list 1 and therefore unlikely to differ between groups.

It would therefore appear that having practice at making RK judgments resulted in a decrease in incorrectly labeling a new word as old. The difference in hit rates on the first list may also further indicate that initially making RK judgments is a little more difficult than just making old/new judgments. When taken together these findings suggest that basing old/new decisions on separate decisions about recollection and familiarity may not be the normal way that the participants approach single item recognition.

**Remember-Know**

Table 8. *Experiment 4. Mean proportion of RK judgments as a function of RK frequency, list number and accuracy.*

<table>
<thead>
<tr>
<th>List</th>
<th>Continuous R</th>
<th>Continuous K</th>
<th>Final R</th>
<th>Final K</th>
</tr>
</thead>
<tbody>
<tr>
<td>List 1</td>
<td>0.47</td>
<td>0.37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.06</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>List 2</td>
<td>0.46</td>
<td>0.34</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>0.13</td>
<td>0.12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>List 3</td>
<td>0.48</td>
<td>0.29</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>0.15</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>List 4</td>
<td>0.51</td>
<td>0.22</td>
<td>0.37</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>0.12</td>
<td>0.26</td>
<td>0.25</td>
</tr>
</tbody>
</table>

As discussed previously, I will provide analysis results of the standard way of measuring RK judgments (see Table 8) and then I will also conduct a chi square analysis of this data for the final list across conditions and for all lists of the continuous condition (refer to Figure 3). I conducted a 2 between RK frequency (final vs continuous) by 2 within accuracy (hits vs false alarms) on R responses calculated in the standard way, made on the final list. In regards to accuracy, results indicated that overall R responses were more likely to be made to old than new items, $F(1, 38) = 39.691, p < .001, MSE = .026, \eta_p^2 = .511$. Collapsing across accuracy the groups did not significantly differ ($F(1,38) = 0.197, p = .659$, ns.), however there was a significant interaction between group and accuracy, $F(1, 38) = 10.251, p = .003$, .
MSE = .026, $\eta_p^2 = .212$, revealing that R responses were much more likely to be accurate (a greater difference between hits and false alarms) in list 4 of the continuous condition than they were in the final condition.

Thus not only did the continuous group perform more accurately than the final group on the recognition task, but they were also more likely to be correct when they indicated that their decision was based on recollection. Taken together, these results indicate a possible learning component to the RK paradigm. While the overall analysis indicates that R responses were more likely to be made to hits than false alarms, this does not tell us how useful recollected content is. That is, it does not tell us whether recollection is responsible for the recognition decision or whether the recognition decision is responsible for the R response.

![Figure 3](image.png)

**Figure 3.** Experiment 4. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency.

This analysis involved creating a 2x2 table (with 1 degree of freedom) for each participant. Raw scores produced by each participant were placed in the contingency table, with hits and false alarms placed on the vertical axis of the table and R and K judgments placed on the horizontal axis. I calculated chi square values across all lists. Participants could only be included if they produced both hits and false alarms, since false alarms were rare on list 1, the number of participants able to be included in the analysis were smaller on this list. However the main focus was on
performance in list 4 and it was rare for a participant to not produce a false alarm on that list.

A significant chi square value for 1 degree of freedom would have to be larger than 3.84. In the 2–4 value range (see Figure 3), only three participants had values that exceeded 3.84, two of these participants were in the continuous condition and one from the final condition. Therefore, as Figure 3 illustrates, the large majority of values were non-significant. We summed the individual chi square values for each condition in order to gain some idea of differences between conditions. List 4 of the final condition just produced a significant chi square ($\chi^2(20) = 32.164, p < .05$). However, the total summed chi square value on list 4 in the continuous condition failed to reach significance ($\chi^2(18) = 28.869, \text{ns}$). (However it should be noted that the difference between the two conditions was really due to the inclusion of one participant in the final condition who produced a chi square value over 6 – see Figure 3). In the continuous condition we are able to examine if chi square values change across lists, the first three lists produced a significant summed chi square value (list 1: $\chi^2(14) = 35.999, p < .05$; list 2: $\chi^2(19) = 52.938, p < .05$; list 3: $\chi^2(17) = 41.038, p < .05$). Note that a large proportion of these values were provided by a small number of participants who produced large chi square values in each condition.

**Correlation between chi square and recognition accuracy**

![Figure 4](image4.png)

*Figure 4*. Experiment 4. Individual correlation scores between d-prime and chi square values for list 4 in the continuous condition ($r = .083$).
Figure 5. Experiment 4. The relationship between individual d-prime and chi square values for list 4 of the final condition ($r = .090$).

A pearson product-moment correlation coefficient was computed in order to examine the relationship between recognition accuracy (d-prime values) and chi-square values for performance on the fourth list of both the final and continuous conditions. This was only performed on data from the fourth list as, as stated previously, performance on this list was likely to produce better data and was theoretically of more interest. While positive, the correlation between d-prime and chi-square values on the list 4 of the continuous condition (see Figure 4) was very small and non-significant ($r = .083, n = 16, p = .770$). Similarly the correlation between chi-square and d-prime values on list 4 of the final condition (see Figure 5) was small and non-significant ($r = .087, n = 19, p = .723$). Overall indicating that in general no strong relationship between recognition accuracy and chi square values occurred.

Discussion

The results from experiment 1 indicate that participants could reasonably perform the task, but that this became more difficult as they proceeded. The pattern of RK judgments appeared to support the idea that recollected contents were more likely to accompany correct old decisions rather than false alarms, however with the
inclusion of the chi square analysis I was not able to reject the hypothesis that the majority of participants made recognition decisions and then assigned a R response to a proportion of those decisions independently of whether the decision had been a hit or a false alarm. That is I do not have evidence that recollection is driving the recognition decision for most of our participants. However there are limitations to the chi square analysis that should be considered when interpreting these results. First, it was necessary to eliminate any participant who did not make a false alarm, although by list 4 this was less of an issue as it was rare for a participant to avoid making a false alarm by list 4. In addition I often observed small cell values, which may affect the reliability of the analysis. However, if I had applied Yates correction for small cell values then the obtained chi square values would have been smaller. I chose not to use the Yates correction in order to present a conservative case.

In addition to the chi square results, I further examined a possible relationship between recognition performance as measured by d-prime values and chi square values. By the final list it is clear in both RK frequency conditions that there is no strong evidence of a relationship between chi square and d-prime values, rather a range of relationships appear possible. Importantly there are participants who were able to perform well on the recognition task and not produce high chi square values. In addition the individuals responsible for the highest chi square values in both conditions failed to produce large d-prime values. It does appear as though some participants can produce both reasonable d-prime and chi square values, however this does not describe the majority of participants. This further supports the idea that, while recollection may effectively be utilised by some participants, this does not appear to be the method by which the majority of participants completed the task.

The results revealed a possible learning component to the RK task, such that participants who made RK judgments on every list were better able to perform the recognition task and were more likely to make accurate R judgments than participants who made RK judgments on the final list only. However on the first list, the analysis of hits revealed that the final group had an advantage as they were more likely to correctly respond old than participants who were making RK judgments for the first time in the continuous condition. Thus the results indicate that the extra task of rating subjective awareness is having an effect on the recognition task. It may be that simply informing participants that recollection and familiarity are involved in recognition memory affected their performance on list 1. In addition, it is possible that the process
of making judgments on each list made participants more stringent judges of their recognition decision.

In experiment 5 all participants will be given RK instructions before list 1, this was done for a couple of reasons. First it enables me to separate the effects due to participants’ knowledge about familiarity and recollection from the effects due to the act of making RK judgments. Second it will make the larger retention interval between the study and test phases of list 4 unnecessary. Discrimination was generally much poorer on list 4 than the previous lists and, since we compare RK performance on this list only, it would be helpful to be able to have a more reliable comparison between groups and to prior lists. Furthermore there was no opportunity for participants to practice making RK decisions. Elsewhere there has been some concern that participants do not make RK decision in the manner in which the experimenter expects. That is, the judgments may reflect confidence rather than content (i.e., Donaldson, 1996). Thus in experiment 5, in both conditions, participants will perform a practice trial in order to make sure they have understood the instructions. The continuous group will again make RK judgments on all four lists, and participants in the final condition will not make them until the final list. Both groups will be told to follow test instructions carefully so that they know when they are required to make RK decisions.

Experiment 5

Method

Participants
A total of 40 students from the University of Queensland participated in the study. 20 participants were randomly assigned to one of two conditions (final or continuous). The variables to be measured were both RK decisions and recognition accuracy.

Procedure and Materials
The procedure and materials were the same as that used in experiment 4, except that participants first had a practice trial before the four lists to get to know the remember know paradigm. During the trial, five words were studied (a mix of three letter and ten letter words that were unrelated to any of the words used in the four lists) and then they were presented with ten words at test and were asked to make Old/New and R/K judgments as in experiment 4. When the practice trial was finished the
The experimenter asked the participants as to what constituted a R or K response. When
the experimenter was satisfied that participants understood the instructions then they
continued on with the task. In the continuous condition participants were told that
they would be asked to make RK judgments for all lists. Participants in the final
condition were instructed that they would be informed in the instructions before each
test list as to whether or not they would have to make RK judgments.

Results

Table 9. Experiment 5. Hits, false alarms and d-prime values broken down by list
number and RK frequency condition.

<table>
<thead>
<tr>
<th>List</th>
<th>RK Frequency</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Continuous</td>
<td>Final</td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>0.82</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>False Alarm</td>
<td>0.06</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>d'</td>
<td>2.42</td>
<td>2.55</td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>0.77</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>False Alarm</td>
<td>0.23</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>d'</td>
<td>1.51</td>
<td>1.56</td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>0.73</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>False Alarm</td>
<td>0.19</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>d'</td>
<td>1.49</td>
<td>1.21</td>
<td></td>
</tr>
<tr>
<td>Hit</td>
<td>0.75</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>False Alarm</td>
<td>0.26</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>d'</td>
<td>1.34</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

The analysis performed on d-prime values in experiment 4 was repeated on
those values in experiment 5 (see Table 9). I found that the two groups did not differ
significantly in accuracy as they did in experiment 1 \(F(1, 33) = .254, p = .617, \text{ns}\),
however accuracy did again reduce significantly between lists \(F(3, 99) = 35.163, p < .001, MSE = .344, \eta^2_p = .516\), follow up comparisons revealed that list 1 produced
significantly more accurate recognition responses to all other lists and list 2 resulted
in more accurate performance than to list 4 \(p < .05\), there was no significant
difference in performance between list 3 and 4 \(p = .214\). Unlike experiment 4, I
failed to observe a significant interaction between RK frequency and list \((F (3, 99) = 1.446, p = .234, \text{ns})\).

I analysed the hits and false alarms separately as I did previously in experiment 4 (see Table 9). I found hits decreased with every list \((p < .05\), except between list 3 and 4 \((p = .186\)). False alarms significantly increased from list 1 to list 2 \((p < .05\) but did not differ between lists after that. No other significant effects were observed.

In summary, recognition performance was quite similar to experiment 4 as participants could still differentiate between old and new items, but with less accuracy as they continued through the task. In particular the declining hit rate across lists observed in experiment 4 was again observed in experiment 5. While there was some evidence that false alarms increased with each list, this effect was less dramatic in experiment 5. Unlike experiment 4, group differences were not significant.

Remember-Know

Table 10. Experiment 5. Mean proportion of RK judgments as a function of list number, recognition accuracy and RK frequency condition.

<table>
<thead>
<tr>
<th>List</th>
<th>Hit</th>
<th>FA</th>
<th>Hit</th>
<th>FA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>0.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>0.55</td>
<td>0.11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>0.47</td>
<td>0.09</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>0.52</td>
<td>0.12</td>
<td>0.38</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>0.23</td>
<td>0.14</td>
<td>0.26</td>
<td>0.11</td>
</tr>
</tbody>
</table>

R responses (see Table 10) were scored and analysed as they were in experiment 4. I again found a significant main effect of accuracy such that R responses were more likely to be made to hits \((M = .45)\) than false alarms \((M = .14)\), \(F (1, 33) = 59.06, p < .001, MSE = .029, \eta_{p}^2 = .642\). I also again found that groups did not differ overall \((F (1, 33) = 1.04, p = .315, \text{ns})\). However, as in experiment 4 a significant interaction revealed that that R responses were more accurate in the
continuous condition than in the final condition $F(1, 33) = 4.338, p = .045, MSE = .029, \eta^2_p = .116$.

Figure 6. Experiment 5. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency.

See Figure 6 for a summary of chi square values produced by participants and note that again most chi square values were not significant. This time no participants managed a significant chi square value in the 2 – 4 value range. Due to the inclusion of some participants who produced very high chi square values, all summed chi square values reached significance. List 4 of the final condition produced a significant summed value, $\chi^2(17) = 34.328$, $p < .05$. Similarly the summed value of the final list of the continuous condition reached significance, $\chi^2(14) = 60.931$, $p < .05$. All other lists in the continuous condition again produced significant summed values (List 1: $\chi^2(9) = 27.342$, $p < .05$; list 2 $\chi^2(14) = 39.660$, $p < .05$; list 3 $\chi^2(15) = 34.973$, $p < .05$). Again, these large values were driven by a minority of participants.
Correlation between chi square and recognition accuracy

![Graph](image)

**Figure 7.** Experiment 5. Individual correlation scores between d-prime and chi square values for list 4 in the continuous condition ($r = .391$).

Both RK accuracy conditions produced small positive pearson product-moment correlations between d-prime and chi-square values on list 4. While larger than experiment 4, both correlations again failed to reach significance $r = .391$, $n = 14$, $p = .167$ (see Figure 7) in the continuous condition and $r = .344$, $n = 17$, $p = .176$.
(see Figure 8) in the final condition. As illustrated in the scatterplots there are a small number of participants who produce high chi square and high d-prime values, while the majority of participants data falls to the bottom left quadrant of the figures.

Discussion

The RK data accurately replicates what was found in experiment 4. While in this experiment it was found that making RK judgments either on every list or on the final list did not affect recognition performance, the general way in which the recognition task was performed was quite similar to experiment 4. That is, there was a decrease in hit rate and a steady or increasing false alarm rate. Also similar to prior experiments participants accurately made R judgments so that they were more likely to occur to hits than false alarms. However, again our results did not support the idea that R judgments were associated with hits and thus there was no indication that R judgments were driving the recognition decision. The fact that R responses again differed between groups and this time without a corresponding difference in recognition accuracy indicates more strongly that some aspect of the RK task is learnt and that it does not necessarily have anything to do with the quality of content recalled, or the recognition decision.

The individual chi square values again indicated that recollection was not driving the recognition response. While all conditions produced a significant summed chi square value this was usually due to the inclusion of a few participants who were able to produce large individual values. This makes a comparison between conditions quite difficult to interpret. However, it is clear that while the majority of participants did not produce a significant chi square value, a minority of participants can and sometimes these values can be quite large. Thus it is entirely possible that the recognition task can be achieved through content recollection, however this is not necessary for participants to complete the recognition task.

The conclusion drawn from observing the chi square data is also supported by the result of examining the correlation between chi square and d-prime values. As in experiment 4, I again observed that not all participants are producing a relationship in line with a dual process perspective. Some participants produced low chi square and high d-prime values and some participants produced high chi square and low d-prime values (both indicating a negative relationship between the two). There are a minority of participants in both conditions who produced a very high score on both measures,
thus it is possible that a few participants were using content recollection in order to complete the recognition task. However, these small and non-significant correlations do not indicate a strong relationship overall between the two performance measures.

The design of experiment 5 allowed an easier interpretation of the results. That is, the initial practice trial enabled a closer examination the effect of awareness (regarding recollection and familiarity) and the effect of making RK judgments on recognition and R judgments. The inclusion of a practice trial further helped to rule out the possibility that participants did not understand how to correctly apply the RK task. Contrary to the results from experiment 4, group differences were not observed in recognition performance. However it was noted that R judgments were again more accurate in list 4 of the continuous condition compared to the final condition. The results of experiment 5 therefore raise the possibility that awareness of the two processes may have affected recognition performance, while the act of making judgments may affect accuracy of R judgments. However considering the variability of results usually afforded by the RK paradigm and the lack of replication, it would be premature to make any firm conclusions.

While the majority of participants do not produce significant chi square values, there are some participants, notably in the continuous conditions who are able to produce large values. At this point it might be wondered what could be driving R responses. One way to ascertain what is driving R decisions is to simply ask participants to justify their R judgments. It has previously been noted that asking for a justification can affect the way in which participants make R judgments. Yonelinas (2002) informed participants that they may have to justify their R judgments after the task and found that participants were more likely to make more conservative R judgments. Similarly Mather, Henkel and Johnson (1997) found that participants who were asked to provide information about their R responses tended to make fewer R judgments. However, as we noted previously, Bodner and Lindsay (2003) did not note a difference in the way in which RK judgments were made when participants were asked to justify their decisions. A reasonable explanation of these differences is that in some tasks, like LOP tasks, the content that is recollected (a memory for how the item was processed) clearly ties the contents of recollection to the list. However, in tasks without a difference in processing the contents of what is recollected may not be closely tied to occurrence in the list. If participants are aware that the content of their recollections are not closely tied to occurrence in the list then a requirement to
disclose the content of their recollections will encourage them to become more conservative with their R responses. It remains to be seen as to whether or not this will affect how accurate those decisions are or how tightly bound they are to the recognition decisions.

Experiment 6

Method

Participants

A total of 80 first year students participated in the study for course credit. 20 participants were randomly assigned to one of four between subject conditions (final, final justification, continuous or continuous justification). The variables to be measured were both RK decisions and recognition accuracy. One participant did not understand the instructions properly and was excluded from the analysis, which resulted in 19 participants in the continuous justification condition.

Procedure and Materials

The procedure and materials were similar as that used in experiment 5, however participants were now asked to justify their R responses in the justify conditions. Participants in the justify conditions were also asked to justify their R responses during the initial trial phase. A sheet of paper was provided for participants next to their keyboards with a space for them to write the word in question and next to that to write down the details that accompanied their memories. Once they had made a R judgment they were given a prompt on screen to write down the details that accompanied that judgment.

Results

Recognition

Table 11. Experiment 6. Recognition results presented as hits, false alarms and d-prime values broken down by list, RK frequency and justification conditions.

<table>
<thead>
<tr>
<th>List</th>
<th>Continuous Justification</th>
<th>Continuous No Justification</th>
<th>Final Justification</th>
<th>Final No Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hit</td>
<td>0.85</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>0.09</td>
<td>0.04</td>
<td>0.06</td>
</tr>
</tbody>
</table>
A 2 between RK frequency (continuous vs final) by 2 between justification (justification vs no justification) by 4 within (list) mixed ANOVA was conducted on d-prime scores (see Table 11). This analysis revealed a main effect of list such that accuracy declined across list ($F(3, 225) = 121.598, p < .001, MSE = .363, \eta_p^2 = .619$), follow up comparisons revealed that all lists differed significantly ($p < .05$) except between lists 2 and 3 ($p = .084$). When I examined between group differences, I failed to find a significant difference in accuracy between the final and continuous conditions ($F(1, 75) = .007, p = .933, \text{ns}$) however, collapsing across RK frequency and list, participants in the no justification condition ($M = 1.85$) were more accurate in their recognition responses than participants in the justification condition ($M = 1.58$), $F(1, 75) = 5.02, p = .028, MSE = 1.144, \eta_p^2 = .063$. Despite not finding an effect overall for RK frequency, I did find a significant interaction between list and RK frequency $F(3, 225) = 4.333, p = .005, MSE = .363, \eta_p^2 = .055$. In particular, follow up simple effects revealed that sensitivity was greater in the final conditions compared to the continuous conditions on list 1, though this difference was marginal, ($p = .054$). The groups did not differ significantly on any other list ($p > .05$).

As in the previous two experiments I examined the pattern of hits and false alarms. Performing the same analysis as that applied to the d-prime data, I found that false alarms increased with each list ($p < .05$) and an analysis of the hit rates showed that hits decreased across lists as they did in the previous two experiments ($p < .05$).

<table>
<thead>
<tr>
<th>List 2</th>
<th>Hit</th>
<th>0.78</th>
<th>0.75</th>
<th>0.82</th>
<th>0.83</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>0.23</td>
<td>0.24</td>
<td>0.23</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>d’</td>
<td>1.58</td>
<td>1.45</td>
<td>1.60</td>
<td>1.80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>List 3</th>
<th>Hit</th>
<th>0.71</th>
<th>0.80</th>
<th>0.75</th>
<th>0.79</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>0.18</td>
<td>0.19</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>d’</td>
<td>1.49</td>
<td>1.71</td>
<td>1.14</td>
<td>1.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>List 4</th>
<th>Hit</th>
<th>0.67</th>
<th>0.71</th>
<th>0.61</th>
<th>0.71</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>0.35</td>
<td>0.22</td>
<td>0.30</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>d’</td>
<td>0.87</td>
<td>1.36</td>
<td>0.75</td>
<td>1.18</td>
</tr>
</tbody>
</table>
continuous conditions produced fewer false alarms ($M = .19$) than the final group ($M = .22$) on the 3\textsuperscript{rd} list ($p = .004$). No other significant effects were observed.

In order to make the results of experiment 6 more comparable to previous experiments I examined the no justification conditions separately, however this analysis did not reveal anything new and is therefore not reported here. Similar to the results from experiment 4, a higher false alarm rate in the final condition compared to the continuous condition was noted, however this time this was only significant for list 3.

While not evident in our analysis of hits, the analysis conducted on d-prime values indicates that the first list of the final condition produced marginally more sensitive recognition judgments than the same list of the continuous condition. This result was also similar to what I had found in experiment 4. Thus providing some evidence that recognition accuracy may be hurt by the inclusion of the RK paradigm even when recollected content is more likely to be criterial.

I was also able to examine group differences that may have occurred between the two continuous conditions. A 2 justification (justification vs no justification) by 4 within list mixed ANOVA on d-prime data and hits failed to reveal anything that might add to my previous analysis. However I did find a marginally significant interaction between list and justification when this analysis was conducted on false alarms, $F(3, 111) = 2.655, p = .052, MSE = .015, \eta^2_p = .067$. The trend in follow up comparisons of justify at each level of list, found that group differences approached significance on list 1 ($p = .063$) and list 4 ($p = .068$). Basically reflecting the trend for fewer false alarms in the no justify condition to the justify condition on list 1 and list 4, which did not occur on list 2 or 3.

**Remember Know**

Table 12. *Experiment 6. Mean proportion of RK judgments as a function of list number, recognition accuracy, RK frequency and justification condition.*

<table>
<thead>
<tr>
<th></th>
<th>Continuous</th>
<th>Final</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Justification</td>
<td>No Justification</td>
</tr>
<tr>
<td>R</td>
<td>K</td>
<td>R</td>
</tr>
<tr>
<td>List 1</td>
<td>Hit</td>
<td>0.27</td>
</tr>
<tr>
<td>FA</td>
<td>0.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>
A 2 between RK frequency (continuous vs final) by 2 between justification (justify vs no justify) by 2 within accuracy (hit vs false alarm) mixed ANOVA on R responses (see Table 12) made in list 4, revealed a significant main effect of accuracy such that R responses were more likely to occur for hits ($M = .33$) than false alarms ($M = .10$), $F(1, 75) = 118.502, p < .001, MSE = .018, \eta_p^2 = .612$, thus replicating previous results from experiments 4 and 5. Accuracy interacted with justification such that better discrimination was observed in the no justify condition ($M(R\text{-}hit) = .48$, $M(R\text{-false alarm}) = .14$) than the justify condition ($M(R\text{-hit}) = .19$, $M(R\text{-false alarm}) = .07$), $F(1, 75) = 28.110, p < .001, MSE = .018, \eta_p^2 = .273$. In addition, significantly more R responses were made in the no justify ($M(R\text{-hit}) = .31$, $M(R\text{-false alarm}) = .07$) than the justify conditions ($M(R\text{-hit}) = .19$, $M(R\text{-false alarm}) = .07$), $F(1, 75) = 30.161, p < .001, MSE = .042, \eta_p^2 = .287$. Therefore the results indicate that R responses were much more likely to be made when no justification of content recalled was required. As such, asking participants to produce proof of their recollected content severely limited their willingness to do so. It should be noted that the superior discrimination in the no justify than in the justify conditions may simply reflect a floor effect as R responses were so rarely made in the justify condition.

The pattern of R responses made in the continuous condition did not differ significantly overall to R responses made in the final condition, $F(1, 75) = 0, p = .982$, ns. Accuracy did not interact with RK frequency, $F(1, 75) = 1.214, p = .274$, ns. Nor was there a significant interaction between the RK frequency conditions and justification conditions; $F(1, 75) = 1.945, p = .167$, ns, or between accuracy, RK frequency and justification $F(1, 75) = 2.455, p = .121$, ns. Separate analysis of the no justification conditions again did not reveal anything of interest and is therefore not reported.

In order to examine differences that might occur with practice between justifying a R response and not justifying that response, I compared the pattern of R
responses made in the two continuous conditions. A 2 between justification (no justify vs. justify) by 4 within (list) by 2 within accuracy (hits vs. false alarms) mixed ANOVA was applied to R responses. The analysis reflected results of the main analysis, which are not reported here. In addition to the expected results, it was further found that the justify conditions varied across block $F(3, 111) = 4.370$, $p = .006$, $MSE = .013$, $\eta_p^2 = .106$. This significant interaction indicated that while R responses increased from list 1 to list 2 in the no justification condition, R responses decreased from list 1 to list 2 in the justification condition, thus possibly exhibiting that by the second list participants in the justification condition were aware that content was not helpful, while participants in the no justification condition were not.

Figure 9. Experiment 6. Proportion of individual chi square values occurring in each value range broken down by list and RK frequency in the justification condition.
As illustrated in Figures 9 and 10, the pattern of chi square values found in the previous two experiments was again replicated in experiment 6. That is, the majority of values fell between 0 and 2 and, therefore, the majority of values failed to reach significance. In Figure 9, one participant in the final condition produced a significant value in the 2–4 range, no other value included in that value range reached significance. In Figure 10 none of the participants managed a significant value in this range.

Summing across participants I found fairly low chi square values in list 4 of the final conditions (final no justification ($\chi^2(15) = 15.346, \text{ns}$) and final justification ($\chi^2(18) = 16.344, \text{ns}$)). In the continuous justification condition the first two lists resulted in significant summed chi square values, while the final two lists failed to reach significance (list 1: $\chi^2(9) = 17.491, p < .05$; list 2: $\chi^2(13) = 22.362, p < .05$; list 3: $\chi^2(13) = 13.040, \text{ns}$; list 4: $\chi^2(13) = 13.308, \text{ns}$). In the continuous no justification condition all lists reached significance, (list 1: $\chi^2(7) = 35.243, p < .05$), list 2: ($\chi^2(16) = 65.877, p < .05$, list 3: $\chi^2(16) = 50.775, p < .05$, list 4: $\chi^2(14) = 33.024, p < .05$).

Again the significant chi square values were helped by the inclusion of a small number of large values with the majority of the individual chi square values being far
from significant. Note also that there are very few R judgments in the justification conditions so the chi square values in those conditions are likely to be less accurate than the chi square values in the other conditions.

**Correlation between chi square and recognition accuracy**

*Figure 11*. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the continuous no justification condition \((r = .759)\).

*Figure 12*. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the continuous justification condition \((r = .177)\).
A significant positive correlation between recognition accuracy and chi square performance was observed in the continuous no justification condition (see Figure 11), $r = .759$, $n = 14$, $p = .002$, indicating that participants who performed well on the recognition task were also more likely to have large chi square values. All other conditions produced non-significant correlations: continuous justification (see Figure 12), $r = .177$, $n = 13$, $p = .555$; final no justification condition (see Figure 13), $r = .184$, $n = 15$, $p = .511$ and final justification, (see Figure 14) $r = .003$, $n = 18$, $p = .990$.

Figure 13. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the final no justification condition ($r = .184$).

Figure 14. Experiment 6. Individual correlation scores between d-prime and chi square values for list 4 in the final justification condition ($r = .003$).
Reported Content

A perusal of the type of recollections participants reported provides some clear evidence that recollected content was largely not useful in discriminating between lists on our task. There were some exceptions, for example a participant in the final justification condition utilised a strategy that may have been helpful. They chose one word from the study list and linked it to other words via a series of sentences. Others reported details about the recency (‘it was the last word in the list’) or primacy (‘it was the first word shown’) of the test item, which may have been helpful. A participant also noted that two words studied in succession started with the same letter, this may have been helpful in identifying one or both of these words correctly at test. However, the large majority of participants who wrote down their recollection reported basic observations at the time that have very little to do with distinguishing between lists (i.e., ‘Patrol’ – ‘I thought about soldiers’). Others stated that they recalled visualizing the items and others reported noting something orthographic about the item like the inclusion of a double letter. Sometimes participants made the observation that the repeated presentation of a word had reminded them of studying that word in a prior list. Sometimes this took the form of explaining that they had seen it a couple of times before or it is more specific and they say that they had seen it in the previous list. They also note down words that they feel they have not studied before until they studied that word in the prior list. In addition to this, in the continuous condition a few participants repeat the initial thought that they had about a particular word on subsequent tests. That is, to use a previous example ‘patrol’ – ‘I thought about soldiers’ would be repeated anytime a participant remembered studying the word patrol. This would not have been useful to them in identifying list membership for the item.

Discussion

The standard way of scoring RK data again indicated that R responses were made accurately, however again the chi square analysis showed that for the vast majority of participants there was no significant association between R responses and hits. As in the earlier experiments, it was found that recognition performance in the continuous condition was more accurate than the final condition. However, this was only significant on list 3. I also found that participants in the final conditions were
marginally more accurate on the recognition task in list 1 than participants in the continuous conditions. This same effect was observed in experiment 1 and there had been a slight trend towards this effect in experiment 2.

Unlike the previous experiments a significant positive correlation occurred between d-prime and chi square values in the continuous no justification condition. A look at the scatterplot (Figure 11) of this data shows that four participants who performed very well on both measures largely drove the correlation. Given previous results such a strong correlation was surprising. A look at the correlations on previous lists in that condition showed that this was not a consistent pattern: list 3; $r = .066$; list 2; $r = -.228$; list 1; $r = .009$. It is possible that the four participants who performed well on list 4 became progressively better at the task and failed to impact largely on the correlation prior to list 4. As in the prior experiments there was no clear pattern to the data, this is also evident by examining the other scatterplots, where again as in previous experiments, all relationships were possible between the two performance measures.

Participants were more likely to make old/new decisions accurately and more likely to report R when they did not need to justify their R response. The decrease in responding R in the justification condition (particularly from list 1 to list 2), indicates that participants may have become aware that content was not helpful. It is interesting that recognition performance was adversely affected in this condition. While it may be tempting to attribute this to greater retention intervals due to participants having to write down a justification for their R responses, when we consider how rarely participants made a R response in this condition then this explanation becomes extremely unlikely. Instead it appears more likely that this detriment to performance was due to some aspect of having to justify a R response or possibly through examining the information employed to make a R response.

It was noted in experiment 5 that making participants aware of the RK task by incorporating an initial practice task may have been responsible for the failure to find a difference in recognition performance between the final and continuous conditions. Only a small difference was observed between these conditions in experiment 6 on one list, rather than a main effect. However a difference was observed between the justification conditions, such that the no justification condition resulted in more accurate recognition. Thus again, the manipulation of the RK task affected recognition performance, and in that case it must have been due to actually making
CHAPTER 4

There is some evidence across all three experiments included in chapter 3 that when the RK task is introduced for the first time it can adversely affect performance. It is possible that the RK instructions are too suggestive. This not only poses a problem for the interpretation of RK data, but may also affect performance on the RK task and/or the recognition task. While there is some variation in these effects, the general trend is that recognition performance is negatively affected on list 1 in the continuous condition, while recognition and/or R judgments are more likely to be hurt by list 3 or list 4 of the final condition. The smaller effect on the first list is most likely due to the fact that content recollection is more likely to be criterial, whereas it is more likely to be noncriterial by the final list. Furthermore in experiment 6, being made to justify R responses negatively affected recognition performance compared to the standard RK response condition, which strongly supports the notion that the RK task can affect memory performance adversely, especially when content is not useful.

There are a number of potential explanations as to why performance differences were observed between the continuous and final conditions. First, these differences may indicate a possible learning aspect to the RK task. That is, recollected content may be assessed differently if RK judgments are made on every list, as compared to just the final list. As I mentioned in the introduction, the instructions for the RK paradigm do not specify that discriminatory information must be retrieved to satisfy a R response, rather anything that pops into mind will do. It may be that on first using the RK task participants assume that any recollected content that comes to mind is accurate and hence on the final list this could only damage recognition performance. That is, by the final list the content that comes to mind may refer to studying or being tested on that same word in prior lists. Participants in the continuous conditions however, may become more astute reviewers of their own recollections. That is, they may have learnt not to uncritically accept anything that comes to mind.

Differences between the continuous and final conditions may also be observed if participants are more likely to rely on recollection in the absence of familiarity.
when initially presented with the RK task. In accordance with the idea that the RK instructions are suggestive, participants may be focusing on recalling content without necessarily focusing on the recognition task. It may be that participants in the continuous condition initially rely on recollection in the absence of familiarity but then, with practice, come to report recollection only when it is accompanied by familiarity. Relying purely on recollection in the fourth list in particular would adversely affect performance, as explained previously and in accordance with the results. The process of recollection on list 1, though more helpful than in list 4, may not always be as accurate in the absence of familiarity.

Another alternative to consider is that making RK judgments may encourage participants to evaluate their recognition decisions, and this extra cognitive effort may be negatively affecting performance. Presumably, greater effort expended on evaluating the subjective experience of a recognition decision would be required on first encountering the RK task. On subsequent lists in the continuous condition the task may become less arduous. This possibility is supported by the fact that the increased effort required to justify a response also had deleterious affects on recognition performance. The concept of effort as described by Yonelinas and Jacoby (1996) may apply in source monitoring tasks or when two lists have been processed in two different ways. However, it is less clear as to how effort may be applied to an episodic recognition task.

Thus the results from experiments 4, 5 and 6 indicated that it was likely that a combination of knowledge about recollection and familiarity and the act of making RK decisions was affecting recognition and/or RK performance. Furthermore it may be that in the continuous conditions participants might have relied on recollection only when items were first familiar, while participants in the final conditions may have been relying on recollection in the absence of familiarity. A final experiment was conducted in order to further examine these possibilities.

Experiment 7

Experiment 7 includes all four conditions previously utilised in experiment 1 and 2 (RK frequency and RK response interval conditions). However there were some additional changes. Instead of a practice RK trial initially as had been employed in experiments 5 and 6, all participants made judgments on list 1. This allows a tighter
comparison of the effect of practice while allowing all participants to have the opportunity to understand the RK procedure. While participants in the continuous conditions continue to make RK judgments on all subsequent lists, participants in the interrupted condition will only make them again on list 4. Thus participants in the continuous condition will be practiced at making RK judgments on a list discrimination task, while participants in the interrupted condition will be practiced on the task, but not while also completing a list discrimination task. A longer study-test retention interval is further re-introduced in experiment 7. The effect on the final list when participants were not practiced at making RK judgments was a lot smaller across experiments 5 and 6, however it may be that a longer retention interval may recover the effect, as it may be more noticeable under more demanding conditions.

The inclusion of the delayed condition may help to clarify previous speculation. The addition of the RK task may encourage participants to rely on recollection, however participants in the delayed condition have already completed the recognition task by the time they attempt the RK task. It is also possible to separate the effects of RK knowledge from the act of making RK judgments by incorporating the delayed condition. If a decrement in performance is observed on the last list of the interrupted condition compared to the continuous condition, when judgments are made immediately and not at a delay, then the interpretation that the act of making judgments can affect R and/or recognition performance is supported. This would also support the idea that the effect on the final list is due to participants relying on recollection in the absence of familiarity. Furthermore, in the case that participants are only relying on recollection in the presence of familiarity in the continuous condition, then a difference in performance between the delayed and immediate continuous conditions would not be expected.

The results of experiment 1 indicated that participants were unable to make RK judgments at a delay without overestimating prior old decisions on test 1. However, it is possible that the inclusion of a larger retention interval between the recognition test and the RK test may have been the reason for this poor performance. Or it could be that, as participants had no opportunity to practice the task, they became confused about when they should be responding R or K. It is further possible that such a task may be more achievable if participants are well practiced at making RK judgments at a delay. In particular practice at making RK responses when there is
a list discrimination task (lists 2, 3, and 4) may be important in allowing participants to calibrate their R responses.

Method

Participants

A total of 92 participants from the University of Queensland, were included in the study. 23 participants were randomly assigned to one of four between subject test conditions (delayed: interrupted or continuous; and immediate: interrupted or continuous). The data from two participants was lost due to computer error leaving 46 participants in the delayed conditions and 44 participants in the immediate conditions. For each RK response interval condition (delayed and immediate), an equal number of participants were included in each RK frequency condition (interrupted and continuous).

Procedure and Materials

All participants were instructed about the RK task after the first study list but before the first test list. In order to keep study and test interval constant across lists, test instructions remained on the screen for 50 seconds before each recognition test list across all conditions. The procedure and instructions for the delayed and immediate conditions were similar to experiment 1 however there were some changes made regarding when participants received instructions. As indicated previously all participants received RK instructions before the recognition test on list1. Further to this participants in the delayed continuous condition received a reminder about the RK task between the recognition test and the RK test, they also received clear instructions indicating that they were to describe their subjective experience only for the items they had just been tested on. Participants in the interrupted conditions were instructed before the recognition test phase of list 2 and list 3 that they would not be required to make RK judgments on the subsequent test phase. They were then instructed before the test phase of list 4 that they would be required to make RK judgments.

Results

Recognition

Table 13. Experiment 7. Proportion of hits, false alarms and d-prime values broken down by list and RK response interval.
### Table 13

<table>
<thead>
<tr>
<th></th>
<th>Delayed</th>
<th></th>
<th>Immediate</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Continuous</td>
<td>Interrupted</td>
<td>Continuous</td>
<td>Interrupted</td>
</tr>
<tr>
<td>List 1</td>
<td>Hit</td>
<td>.861</td>
<td>.791</td>
<td>.833</td>
</tr>
<tr>
<td></td>
<td>False Alarm</td>
<td>.074</td>
<td>.077</td>
<td>.137</td>
</tr>
<tr>
<td></td>
<td>D-prime</td>
<td>2.546</td>
<td>2.216</td>
<td>2.101</td>
</tr>
<tr>
<td>List 2</td>
<td>Hit</td>
<td>.804</td>
<td>.775</td>
<td>.789</td>
</tr>
<tr>
<td></td>
<td>False Alarm</td>
<td>.315</td>
<td>.248</td>
<td>.339</td>
</tr>
<tr>
<td></td>
<td>D-prime</td>
<td>1.440</td>
<td>1.423</td>
<td>1.233</td>
</tr>
<tr>
<td>List 3</td>
<td>Hit</td>
<td>.794</td>
<td>.727</td>
<td>.707</td>
</tr>
<tr>
<td></td>
<td>False Alarm</td>
<td>.398</td>
<td>.291</td>
<td>.417</td>
</tr>
<tr>
<td></td>
<td>D-prime</td>
<td>1.172</td>
<td>1.255</td>
<td>0.706</td>
</tr>
<tr>
<td>List 4</td>
<td>Hit</td>
<td>.752</td>
<td>.736</td>
<td>.741</td>
</tr>
<tr>
<td></td>
<td>False Alarm</td>
<td>.363</td>
<td>.213</td>
<td>.367</td>
</tr>
<tr>
<td></td>
<td>D-prime</td>
<td>1.134</td>
<td>1.552</td>
<td>0.977</td>
</tr>
</tbody>
</table>

In order to examine recognition performance d-prime values were calculated with the standard correction (see Table 13) and a 2 between RK frequency (interrupted vs continuous) by 2 between RK response interval (delayed vs immediate) by 4 (list; 1, 2, 3, 4) mixed ANOVA was conducted on those values. As in previous experiments a main effect of list was observed ($F(3, 258) = 59.060, p < .001, \text{MSE} = 0.477, \eta^2_p = .407$). Accuracy significantly ($p < .05$) decreased with each list except between list 3 and 4 ($p = .606$). Recognition performance across list did not differ significantly depending on RK response interval ($F(3, 258) = 1.061, p = .366, \text{ns}$) nor did performance differ across list depending on RK frequency ($F(3, 258) = .726, p = .537, \text{ns}$). Collapsing across list and RK frequency, the delayed and immediate groups did not significantly differ ($F(1, 86) = .048, p = .828, \text{ns}$).

However overall the continuous group ($M = 1.59$) was more accurate on the recognition task than the interrupted group ($M = 1.27$), ($F(1, 86) = 4.455, p = .038, \text{MSE} = 2.120, \eta^2_p = .049$). Recognition performance in the RK frequency conditions did not vary depending on RK interval condition, $F(1, 86) = 0.001, p = .973, \text{ns}$.

However there was a significant three way interaction between list, RK frequency and RK interval, $F(3, 258) = 2.733, p = .044, \text{MSE} = 0.477, \eta^2_p = .031$. 
A perusal of the data indicated that this was most likely to be the result of a difference in the RK frequency by list interactions at each level of the RK response interval conditions. Thus, the three way interaction was followed up by examining the interaction between list and RK frequency separately for the delayed and immediate group. The obtained values were adjusted so as to incorporate the correct error term and degrees of freedom from the initial model. As expected, a significant interaction in the delayed condition was not observed ($F_{adj} (3, 261) = .615, \text{ns}$). However the interaction between RK frequency and list was significant in the immediate condition ($F_{adj} (3, 261) = 2.953, p = .033$). Follow up simple effects of this interaction revealed that the continuous and interrupted conditions only differed significantly on list 4, specifically participants in the interrupted condition were less accurate than in the continuous condition ($F_{adj} (3, 261) = 14.376, p < .001$). The interrupted and continuous groups did not differ significantly on any other list, ($p > .05$).

The analysis applied to the d-prime values was applied to the hits and false alarms. The large majority of each analysis reflected what was found in the d-prime analysis, with the exception of a marginally significant main effect of RK frequency found when examining the false alarms, $F (1, 86) = 3.808, p = .054, MSE = .112, \eta^2_p = .042$. This reflected the fact that more false alarms were made in the interrupted conditions ($M = .32$) on average than in the continuous conditions ($M = .25$). Neither the analysis of the hits or false alarms revealed a significant three way interaction as had been observed in the d-prime analysis.

**Remember-Know**

Table 14. *Experiment 7. RK judgments as a function of accuracy, list, RK response interval and RK frequency.*

<table>
<thead>
<tr>
<th></th>
<th>Delayed</th>
<th>Immediate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Continuous</td>
<td>Interrupted</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>K</td>
</tr>
<tr>
<td>List 1</td>
<td>Hit</td>
<td>.450</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>.020</td>
</tr>
<tr>
<td>List 2</td>
<td>Hit</td>
<td>.496</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>.183</td>
</tr>
<tr>
<td>List 3</td>
<td>Hit</td>
<td>.483</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>.191</td>
</tr>
</tbody>
</table>
A 2 between RK response interval (delayed vs immediate) by between 2 RK frequency (continuous vs interrupted) by 2 accuracy (hits vs false alarms) was conducted on R responses made on the final list only. A significant main effect of accuracy illustrated that participants made more R responses to hits ($M = .43$) overall than false alarms ($M = .16$), $F(1, 86) = 87.809, p < .001, MSE = .036, \eta^2_p = .505$. This effect of accuracy did not differ between the delayed and immediate conditions ($F(1, 86) = .316, p = .575, \text{ns}$), however the effect of accuracy did differ between the continuous and interrupted conditions ($F(1, 86) = .316, p = .575, \text{ns}$). Follow up simple effects revealed that the interaction between accuracy and RK frequency resulted because participants were more accurate with their R responses in the continuous condition (mean difference between R-hit and R-fa: $M$-diff = .32) than participants in the interrupted condition ($M$-diff = .21) on list 4.

Overall, delayed and immediate conditions did not differ significantly, ($F(1, 86) = 2.059, p = .155, \text{ns}$; nor did the continuous or interrupted conditions, $F(1, 86) = .007, p = .933, \text{ns}$). However RK response interval and RK frequency conditions did produce a marginally significant interaction, $F(1, 86) = 3.802, p = .054, MSE = .054, \eta^2_p = .042$. This interaction indicates that more R responses occurred in the interrupted than the continuous condition when judgments were made immediately (see Figure 14). However, in the delayed condition the trend went the opposite way, so that R responses were more likely in the continuous than the interrupted condition. The three way interaction between accuracy, RK response interval and RK frequency was not significant ($F(1, 86) = .007, p = .933, \text{ns}$).

Chi-square

A chi square analysis was also applied to the data here as it was in the three experiments reported in chapter 3. Again the large majority of participants failed to produce a significant chi square value (see Figure 15 and Figure 16). In the delayed condition (see Figure 15), only one participant in the continuous condition produced a significant value (exceeding 3.84) on the first list in the 2 – 4 value range. A participant in the interrupted condition also produced a significant value in the 2 – 4

<table>
<thead>
<tr>
<th>List 4</th>
<th>Hit</th>
<th>.459</th>
<th>.220</th>
<th>.333</th>
<th>.287</th>
<th>.459</th>
<th>.277</th>
<th>.464</th>
<th>.211</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA</td>
<td>.154</td>
<td>.159</td>
<td>.139</td>
<td>.161</td>
<td>.118</td>
<td>.095</td>
<td>.243</td>
<td>.159</td>
<td></td>
</tr>
</tbody>
</table>
range on the first list, however no other values were significant in this range. In the immediate condition (see Figure 16), two participants produced significant values in this range (one on list 2 and one on list 4), no other significant values occurred in this range.

Despite the delay, the participants in the delayed conditions were still capable of occasionally producing large chi square values (see Figure 15) and on all but the third list, significant summed values were produced (list 1: $\chi^2(10) = 29.596, p < .05$; list 2: $\chi^2(18) = 57.234, p < .05$; list 3: $\chi^2(17) = 20.504, ns$; List 4: $\chi^2(17) = 36.107, p < .05$). While the first list in the interrupted condition was significant ($\chi^2(17) = 35.041, p < .05$), the summed chi square value on the final list of the interrupted condition was not significant, $\chi^2(22) = 14.671, ns$.

The immediate conditions produced similar summed chi square results to the delayed conditions. Only list 4 of the interrupted condition produced a non-significant summed chi square value, $\chi^2(19) = 25.342, ns$. All other summed values were significant; immediate interrupted list 1: $\chi^2(11) = 24.507, p < .05$; immediate continuous list 1: $\chi^2(14) = 55.272, p < .05$; list 2: $\chi^2(21) = 79.113, p < .05$ list 3: $\chi^2(20) = 73.729, p < .05$; list 4: $\chi^2(19) = 42.466, p < .05$).

The summed chi square values we observed generally indicated that larger values were more likely to occur in the immediate than delayed conditions, this was more noticeable across the continuous conditions. A perusal of the summed chi square values indicate this, also a greater number of significant values occurred in the immediate continuous condition than the delayed condition. The number of significant values across lists for the delayed continuous condition were: 3, 4, 2, 3 (for list 1, 2, 3, 4 respectively). While for the immediate continuous the values were: 5, 7, 8, 3. This is not unexpected considering that we might expect more of a relationship between recognition and RK judgments when made in quick succession rather than when a delay occurs between each decision.
Figure 15. Experiment 7. Individual chi square values as a function of RK frequency in the delayed condition.

Figure 16. Experiment 7. Individual chi square values as a function of RK frequency in the immediate condition.
Correlation between chi square and recognition accuracy

*Figure 17.* Experiment 7. Individual correlation scores between d-prime and chi square values for list 4 in the continuous immediate condition ($r = .355$).

*Figure 18.* Experiment 7. Individual correlation scores between d-prime and chi square values for list 4 in the interrupted immediate condition ($r = .467$).
A Pearson product-moment correlation coefficient was computed in order to examine the relationship between recognition accuracy (d-prime values) and chi-square values for performance on the final list of all four conditions. While positive, the correlation between d-prime and chi-square values on the list 4 of the continuous immediate condition (see Figure 17) was not significant ($r = .355$, $n = 19$, $p = .136$). The same correlation in the interrupted immediate condition (see Figure 18) produced
a marginally significant correlation between recognition accuracy and d-prime performance \((r = .467, n = 18, p = .051)\). There was no significant relationship between d-prime and chi square performance in the continuous delayed condition \((r = .337, n = 16, p = .202)\) or the interrupted delayed condition \((r = -.030, n = 22, p = .896)\).

**Discussion**

Differences were not noted between any of the conditions on the first list in terms of recognition accuracy or R judgments, and therefore all groups initially used the RK task in a similar manner. The pattern of results indicated that differences in recognition accuracy (between continuous and interrupted conditions) were noticed primarily when judgments were made immediately and not at a delay. Furthermore performance on the RK task was marginally more accurate in the continuous than the interrupted conditions (collapsing across RK response interval conditions). In addition participants tended to make a stronger association between hits and R judgments when RK judgments were made immediately rather than at a delay. However this was more evident in the continuous conditions as chi square values were generally quite low in the interrupted conditions (on list 4).

The larger chi square values in the immediate conditions (and more notably in the continuous immediate condition) might be expected considering that some forgetting may occur in the delayed condition between the first and second test. Table 9 compares how participants responded on test 1 (immediate recognition) and test 2 (delayed RK) in the delayed condition. R and K judgments are summed to produce a hit or false alarm probability across each list for test 2 responses. As the data illustrates participants were slightly more conservative with their RK judgments at a delay, even in the interrupted conditions. However performance on the delayed test was quite good. Indicating that participants were likely to slightly underestimate prior old recognition decisions but were still reasonably accurate.

Table 15. **Experiment 7. Comparing old decisions on test 1 (recognition) and test 2 (RK) in the delayed conditions.**

<table>
<thead>
<tr>
<th></th>
<th>Continuous</th>
<th></th>
<th>Interrupted</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R+K</td>
<td>Recog</td>
<td>R+K</td>
<td>Recog</td>
</tr>
<tr>
<td>List 1</td>
<td>Hit</td>
<td>0.783</td>
<td>0.861</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>0.057</td>
<td>0.074</td>
<td>0.097</td>
</tr>
<tr>
<td>List 2</td>
<td>Hit</td>
<td>0.726</td>
<td>0.804</td>
<td></td>
</tr>
</tbody>
</table>
Similar to previous experiments no overall strong relationship between chi square and d-prime values was evident in the data. A perusal of figures 17 – 20, indicates that in general the majority of participants fall close to the x-axis. Thus reflecting the chi square results, as most participants did not produce a large chi square value. The larger positive correlations often included one or two outstanding performances from individuals as occurred in experiment 6. Again there are examples of negative correlations, with some participants performing very well on the recognition task and not very well on the chi square task. In particular there are always a number of subjects with d-prime values greater than 1 with chi square values less than 2.

The main results from experiment 7 indicate that R responses are not as correlated with hits in the delayed conditions as in the immediate conditions. In the delayed conditions it looks like there is a greater tendency for the recognition decision to drive the R response as opposed to the R response driving the recognition decision. As making the judgments at a delay did not affect performance on the final list in the interrupted condition and also chi square values were not as large in the continuous delayed condition. However, in the immediate continuous condition participants are performing quite well on the RK task and associating hits with R judgments – possibly indicating that they are not responding R unless the word is familiar, or that they are specifically looking for criterial or more information as participants may have in Mather et al., (1997) upon receiving the MCQ task.

The results of experiment 7 are consistent with the idea that participants may be relying on recollection in the presence of familiarity in the continuous conditions but not in the last list of the interrupted conditions. The chi square analysis supports this, as does the significant three-way interaction and the failure to find a difference in recognition performance between delayed and immediate continuous conditions. The results were further consistent with the idea that a longer retention interval potentiates the effect of introducing the RK task for the first time on the final list. This effect was

<table>
<thead>
<tr>
<th>List 3</th>
<th>Hit</th>
<th>0.726</th>
<th>0.794</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FA</td>
<td>0.296</td>
<td>0.315</td>
</tr>
<tr>
<td>List 4</td>
<td>Hit</td>
<td>0.679</td>
<td>0.752</td>
</tr>
<tr>
<td></td>
<td>FA</td>
<td>0.313</td>
<td>0.363</td>
</tr>
</tbody>
</table>
CHAPTER 5

General Discussion

Summary of Results

In the experiments included in this thesis I examined the possibility that participants could adequately perform an episodic recognition task when content recollection and acontextual familiarity were no longer useful. In order to examine the role of recollected content, a list-specific task was used where words were randomly assigned as targets and distractors from the same pool of words for each study and test list. It was proposed that by the final list, the utility of acontextual familiarity and content recollection would be greatly reduced, thus if the task was still achievable then some other process must be responsible for performance. In order to provide a measure of content recollection I applied the RK paradigm. In addition, I was curious as to how noncriterial recollection might be reported, if at all, in such a design. The possibility that recognition performance or RK judgments might be affected by having participants well practiced at the task or not was also explored.

The first three experiments included in chapter 2 indicated that, while discrimination tended to decrease with each list, most participants could reasonably perform the list-specific recognition task. In experiments’ 1 and 2 it appeared as though introducing the RK task on the final list dramatically increased false alarms and that the inclusion of the RK task on every list avoided this affect. However, it was possible that these effects were only noted in the final condition because of a longer retention interval between study and test on the final list (due to the introduction of RK instructions). The third experiment controlled for retention interval and this reduced but may not have eliminated the effect noted in the first two experiments. Across all experiments R responses were more likely to occur to hits than false alarms on the final list. This seemed odd given the lack of available content on the final list and also that this remained the case even in the final condition of experiments 1 and 2 that also included a significant increase in false alarms.

The next chapter included three experiments which all refined on the methodology of the initial experiments. Experiment 4 replicated and extended on
experiment 3 by including a continuous condition. A decrement to recognition performance in the final condition compared to the continuous condition was observed. It was also observed that R responses were more likely to accompany hits than false alarms overall, however this time, the effect was greater in the continuous condition. It is possible that an effect was noted in experiment 4 and not experiment 3 because the continuous condition was more likely to result in more accurate recognition, thereby increasing the performance differences between groups on the final list.

When we introduced a practice trial on experiment 5 in order to eliminate the longer retention interval on the last list, we failed to note recognition differences between groups however R responses again were more likely to be accurate in the continuous than the final condition. Experiment 6 included a justification condition. When participants had to justify their responses they were less likely to respond R and their recognition performance was not as accurate as it was in the no justification conditions. In addition, small differences were noted between the continuous and final conditions however this effect was not as large as we had observed in experiment 4.

The final experiment included in chapter 4 helped to clarify previous results. Instead of a practice task, participants in all conditions made RK judgments on the initial list. In the continuous condition, participants are practiced on the RK task while performing a list discrimination task, whereas the interrupted condition only has practice on the task without list discrimination (on list 1). In addition, the longer retention interval of experiments 3 and 4 was included on every list. The objective was to see if the longer interval potentiated the effect on final list recognition performance when making RK judgments for the first time. Results were consistent with this. Also reduced performance on the final list in the interrupted condition was only evident when RK judgments were made immediately rather than at a delay. This result was consistent with the idea that participants may be relying on recollection in the absence of familiarity on that task. In addition, similar to previous experiments, it was again found that R responses were more accurate on the continuous than final conditions.

In the last four experiments of the thesis we applied a chi square analysis to RK judgments. The results of this analysis indicated that the vast majority of participants were not more likely to associate R to hits rather than false alarms. However a minority of participants were capable of producing a significant value,
thus it is possible that content recollection was useful for some participants. In particular this was noted in the continuous conditions. We further examined the correlation between individual d-prime and chi square scores for list 4 of the last four experiments. While most of these correlations were positive and therefore inline with a dual process interpretation, they were usually small and only positive due to a small number of participants who performed very well on both measures. In addition to this there were a number of participants who exhibited no relationship between d-prime and chi square values. It was also possible for participants to perform quite well on the d-prime task without producing a significant chi-square value. In general there was no convincing evidence that a dual process understanding of recollection was necessary to perform the task. Rather some participants produced data congruent with that interpretation, but the majority of participants did not.

There were some very clear trends across all seven experiments. Performance gradually declined with each list. The introduction of the RK paradigm on the first or final list tended to negatively affect recognition performance, although this was more pronounced on the final list and when RK judgments were made immediately. R judgments were more likely to be accurately made in the continuous than final groups and when no justification was required. Furthermore R judgments were more likely to be associated with hits when judgments were made on every list and when those judgments were made immediately rather than at a delay.

**Main Conclusions**

Across all experiments we have observed that R responses are made more often to hits than false alarms. However, our chi square analysis tends not to support the idea that recollection is more helpful for hits than false alarms for the majority of our participants. This is further supported by an examination of the contents of recollection in experiment 6. What is more likely is that the difference observed between hits and false alarms is driven by accuracy of recognition and not greater sensitivity (of applying R to hits rather than false alarms). For the few participants that produced significant chi square values and hence were more likely to associate R judgment with hits, there are a number of possible explanations. Some participants may have engaged in learning strategies, such as assigning a list number to the words in each list. In a similar way participants may have strung words in one list together with sentences or, as one participant reported in experiment 6, choosing one word per list by which to relate all other words in the list. All of these examples would count as
instances of content recollection. In a manner similar to the scenario outlined by Donaldson (1996), it is further possible that some participants were relying on strong feelings of familiarity and were reporting that as R as opposed to K. However it is important to note that only a few participants could have been relying on content recollection or the use of familiarity to discriminate between R and K responses. Thus there is no indication that this is the natural or the usual way of performing the task.

The expectation is that a R judgment will tend to reflect the presence of qualitative content, however it is an inference that this recollected content is driving the recognition response. The data does not tend to support the idea that content was useful for the majority of participants however I cannot absolutely rule out the involvement of content. Nor does the chi square analysis ultimately prove that the content reported by participants even with a high chi square value was ultimately what enabled them to perform the recognition task. The correlations help to support this idea however, again, can provide nothing definitive in this regard. Rather these measures can be either congruent or incongruent with a dual process understanding of content recollection. In this way these measures can assist in deciphering whether content was reported in a way that supports the idea that content was useful in the recognition task or not. The ultimate conclusion to be drawn from my data is that some participants perform in a manner that is congruent with a dual process understanding of recognition performance, however this did not describe the majority of the data. In addition to this, some participants produced data congruent with the idea that content recollection was not necessary in order to perform the task.

There are a number of indications present in the data that there are certain limitations to the RK task, especially when employed in a multi-list design. For example, it was clear that participants were labeling noncriterial recollection as a R experience. This can be concluded for a number of reasons. First, R judgments were made to false alarms, in which case any recollected content would had to have been noncriterial. In addition, by the final list recollected content would be largely unhelpful however participants still willingly reported R. In experiment 6 we further found that when participants did not have to justify their R judgments then they were more likely to say R on the second than the first test list (when information would be more noncriterial). However when they had to justify their response then they were less likely to say R on the second test list. However, even though participants were potentially more aware that content was unhelpful in the justification condition, they
still willingly reported content that was not criterial. It is therefore possible that participants willingly report noncriterial recollection, largely as a result of a motive to comply with the RK instructions. Participants may be more able to comply with the RK instructions I used as these instructions do not specifically indicate that only criterial recollection should be reported. This is consistent with previous findings that R responses are made to false alarms (Roediger and McDermott, 1995), and is also consistent with the interpretation of the Bodner and Lindsay (2003) results outlined in the introduction, that R judgments may be made to noncriterial recollection.

Another indication that the task was not working in the list-specific paradigm was that participants who were well practiced at the RK task (in the continuous conditions), were more accurate with their R judgments on list 4 compared to the final condition. In addition to this, it was often observed that recognition performance was improved in the continuous conditions compared to the final conditions. This effect was due to both diminished performance on the last list in the final condition and also to improved performance on the continuous condition, not just by the final list, but as a whole and sometimes on prior lists.

Some consideration was given to whether the difference observed between the continuous and final conditions occurred simply due to awareness of the RK task. Being aware of a possible role for recollection and familiarity may have led participants to approach the recognition task in a way that they ordinarily would not have. It was also considered that differences may have resulted from the more active component of making RK judgments (which must necessarily also include a role for awareness). In experiments 1 – 4, participants in the final condition were not made aware of the RK task until right before the fourth test list. When the RK practice task was introduced on experiment 5, the only group difference observed was that R responses were more accurate in the continuous than final condition. Thus when all participants were made aware of the RK task initially, we did not observe the effect of less accurate recognition performance on the last list of the final condition compared to the continuous. This indicated a role for awareness in the RK task, although it should also be noted that the retention interval for list 4 in experiment 5 was not as long as it was in the first four experiments and in experiment 7.

While it is likely that awareness of the task played a role, the changes we observed on later experiments provide a stronger indication that actively making RK judgments also affected the task. For example, more accurate recognition
performance was noted on list 3 in the continuous condition compared to the final condition in experiment 6 even though all participants were knowledgeable about recollection and familiarity. In addition, participants who were asked to justify their R responses performed more poorly on the recognition task than participants who were not asked to justify their responses in experiment 6. Both groups were knowledgeable about the proposed role of recollection and familiarity in recognition and therefore a change in the way judgments were made must have played a role in performance. The effect of making RK judgments was further observed in experiment 7 where judgments were made at a delay or immediately. The results indicated that performance was negatively impacted by RK decisions in the final list only when those judgments were made immediately and not when they were made at a delay.

A possible reason for the observed detriment to performance on the final list when the RK task is introduced for the first time could conceivably be explained if we consider that participants may be relying on recollection in the absence of a contextual or differential familiarity. That is, in the case that participants are focusing on recollected content to inform their recognition decisions and doing this without using contextual familiarity, then, especially by the final list, this would not be helpful. If participants are relying on a contextual familiarity then anything recollected that comes to mind is going to be more likely to have occurred within that list-specific context. If participants are not relying on contextual familiarity, than anything recollected to the item may have occurred at any prior study or test list. It is reasonable to consider that participants may have been focused on recollecting content rather than relying on any contextual familiarity due to the suggestibility of the RK instructions. Participants who make RK judgments on every list may, with practice, come to report recollection only when it is accompanied by familiarity. The stronger association present between R judgments and hits in the continuous than the final conditions supports this.

The possibility that participants in the continuous conditions learn to make R judgments only when recollection is accompanied by familiarity was supported by the results of experiment 7. Participants in the delayed condition made recognition judgments first and then made R judgments on a following test. As such those participants were not encouraged to rely on recollection when they made recognition judgments. Performance did not differ greatly between the continuous conditions (delayed and immediate) when it came to recognition accuracy thus supporting the
idea that participants in the continuous immediate condition were making R judgments only when familiarity was present. However, a difference in recognition performance was noted between the delayed and immediate conditions in the interrupted condition. Specifically, performance was negatively affected when RK judgments were made immediately on the final list of the interrupted condition but not when those judgments were made at a delay. Thus supporting the idea that participants in the immediate condition were relying on recollection in the absence of familiarity, while participants in the delayed condition were not.

There was some indication from the data that supported the idea that participants were more likely to use the recognition response to drive the R judgments in the delayed condition than they were in the immediate condition. For example the result of the significant three-way interaction is congruent with the idea that participants in the immediate conditions are using R judgments to inform their recognition opinion. Obviously when they are not practiced on that, then this is detrimental to performance. In the delayed condition participants have to make the recognition judgment first and therefore are not affected by the inclusion of the RK task. The greater incidence of forming a strong association between R judgments and hits in the continuous immediate condition than the continuous delayed condition further supports the idea that participants in the immediate conditions were using the R judgment to inform their recognition decision. In this case, as speculated previously, relying on recollection in the presence of familiarity may improve the usefulness of recollection. It could also be that participants in the continuous immediate condition are searching for more information before making a recognition decision. Both of these processes would result in a higher association between R judgments and hits.

The increased false alarm rate on list 4 of the final condition compared to the continuous condition in experiment 7 was the first time since introducing a practice task on experiment 5 that we noted that particular effect. There were two differences in the design of experiment 7 and that was that there were longer retention intervals and participants made RK judgments on list 1 and list 4. It is possible that participants in the interrupted condition, without having practice on the RK task during a list discrimination task, approached the task in the same manner they did on list 1. That is if participants relied on recollection in the absence of familiarity on list 1 then this may have been useful considering that any content recalled would be likely to be
criterial. However on list 4 most recollected content would have been noncriterial so maintaining the list 1 approach would not have been helpful. In addition to this, as has been stated previously, the longer retention interval may have revealed the effect more clearly as this would have made the task more difficult.

Another alternative to consider is that making RK judgments may encourage participants to evaluate their recognition decisions, and this extra cognitive effort may be negatively affecting performance. Presumably, greater effort expended on evaluating the subjective experience of a recognition decision would be required on first encountering the RK task. On subsequent lists in the continuous condition the task may become less arduous. Mather et al., 1997 found that when memory was more evaluated then it tended to improve recognition. In Mather et al., (1997) memories were more evaluated in the MCQ condition than the RK condition, as the MCQ forced participants to consider a number of specific attributes of their memory for the item, while the RK task asked for more global information. However we found that when there was greater evaluation in the justification condition, recognition performance was less accurate. It may be that when recollected content is not useful then putting greater effort into that recollection may be harmful. This process may be similar to what Reder et al., (2000) assumed was responsible for their discovery that list discrimination performance was reduced when participants were asked to make a list discrimination judgment after saying old compared to a group who were only asked to make a list discrimination judgment after responding R. Reder et al., (2000) proposed that when the requirement to make a list discrimination when there was no memory trace to support it added interference to the task. In the case of my results, this might be considered similar to the notion of recollection without familiarity.

The effects of the RK paradigm in the experiments reported here are likely to be inflated due to the difficulty of the task and the effects may not be as large in single list studies. However it is not the size of the effect that is of interest, but rather that an effect can be detected. By utilising a demanding task it was possible to expose small effects that may be present but not always obvious in less demanding tasks. In support of this in some of the experiments it was sometimes observed that performance on the first list was diminished by the inclusion of the RK task. It is reasonable that the smaller difference observed for the first list is due to the fact that recollection is more likely to be criterial than on the final list and there the effect was much larger.
The results at the very least indicate boundary conditions for the RK paradigm. That it was an issue in the list-specific task I employed in this thesis adds to the previous issues listed with the RK task in the introduction. While finding a new limitation for the task is not unexpected, it is also not insignificant. Changes brought about by the inclusion of the RK task indicate certain things about the nature of content recollection. That is, that it is not always relied upon in order to complete an episodic task, that it may not be helpful in the absence of familiarity and further to this, participants can possibly be trained to rely upon it. Note that the assumption is that R judgments are mostly a reflection that content has come to mind, the failure of the task that I refer to is that the process of making RK judgments appears to not be a simple reflection of two processes as has been previously assumed, rather it appears to be much more active than that. This holds some interesting implications for previous findings, that is, that it may simply be task attributes that direct in what manner participants rely on content recollection or not, and that content recollection is not always required for a particular task over another.

*Exploring a role for acontextual familiarity and other methods for completing the task.*

I assumed that my task manipulations would greatly reduce the utility of acontextual familiarity, however it may be argued that acontextual familiarity was greater for more recently viewed items. Although possible, the current assumptions and data regarding the role of acontextual familiarity in recognition memory does not support the idea that acontextual familiarity decays rapidly enough to be of use in the paradigm I used. An assumption commonly made in the literature is that one can manipulate the usefulness of acontextual familiarity by manipulating frequency, (i.e., Heathcote, 2003; Weeks et al., 2007, Yonelinas, 1999). For example, Weeks et al. (2007) assumed that more frequently occurring words may reduce the subjects reliance on familiarity to perform a recognition task. They repeated items across three lists randomly interchanging old and new items. All words were studied in pairs, while at test some items were presented for a single item recognition test, while other items were tested in pairs (intact or rearranged). For the single test items, they found similar results to ours, in that d-prime values tended to decrease while the false alarm rate significantly increased from list 1 to list 2. However, in pair recognition false alarm rates declined across lists. This result supports the idea that for list 1, participants were using the familiarity of the items in the pairs as a basis for
responding intact even though item familiarity was the same in intact and rearranged pairs. However, this use of familiarity decreased in subsequent lists as all of the items became more familiar. Thus participants appear to be aware that manipulations of frequency impair the ability to use familiarity in a list discrimination task.

It further does not appear that others would expect familiarity to decay quickly in my task. In his third experiment Yonelinas (1999) had participants study two lists with items repeated in the first list so as to equate familiarity across both lists. This was done in order to make it difficult for participants to use familiarity to differentiate between the lists and thus it was expected that participants would rely on recollection in order to complete the task (Yonelinas, 1999). Yonelinas (1999) also examined a situation where he thought that participants would use acontextual familiarity to recognize the words from the most recent list. However, to do this he used an interlist interval that was far longer than ours. In experiment 4, participants studied list 1 on day 1 and then list two on day 5. The source-monitoring test was presented immediately after learning list 2 on day 5. He found a curvilinear ROC, which he interpreted as evidence that acontextual familiarity was relied upon.

The assumption made that acontextual familiarity could not have decayed fast enough to be useful in the list specific task I employed is further supported by Tulving’s (1985) assumptions about his retention interval manipulation. In his second experiment, Tulving (1985) found that R judgments were relatively less likely than K judgments after a 7-day retention interval. Tulving’s interpretation was that correct recognition judgments at this delay were based on familiarity not recollection. Although this assumption may be necessary if one wants to maintain that familiarity has a large semantic component this does not necessarily rule out a short lasting component. To the best of my knowledge there is no current consensus regarding how quickly acontextual familiarity might decay.

While the predominant opinion is that acontextual familiarity does not decay rapidly enough to play a role in our experiments very little has been published about how quickly acontextual familiarity might decay. Experimental data has supported both that acontextual familiarity decays slowly and rapidly. Yonelinas and Levy (2002) using a source monitoring framework with process dissociation instructions found that familiarity declined while recollection remained stable over short retention intervals. However, Yonelinas (1994) found that familiarity remained stable while recollection declined. These inconsistencies have not been addressed and until they
are it is difficult to make predictions about how quickly acontextual familiarity decayed in our task.

One further consideration, is that there is an assumption evident in discussions on the usefulness of familiarity as recency, and that is that one process must decay faster than the other (ie familiarity or recollection). In this regards, an alternative to considering a remaining role for familiarity is to consider that in general there was forgetting for both targets and distractors. In response to the different pattern of forgetting in Yonelinas (1994) and Yonelinas and Levy (2002), Weeks et al., (2007) tested the Yonelinas and Levy (2002) assumption that a rapid loss of item familiarity was responsible for the decline in pair false alarm rate. As previously outlined, the results indicated that the use of familiarity decreased in subsequent lists as all of the items became more familiar. In addition to this, Weeks et al., (2007) noted no evidence of reduced forgetting in source monitoring because both hits and false alarms declined at the same rate, indicating that both targets and distractors were being forgotten. Thus conclusions about rates of forgetting when both hits and false alarms are declining are problematic. While a role for acontextual familiarity can not be completely ruled out there are other alternatives for how participants may have performed the task.

A possible role for contextual familiarity

The difference between contextual familiarity and acontextual familiarity is that for the latter the decay process is closely tied to the retention interval, whereas contextual familiarity may not be closely tied to the retention interval if there is an opportunity to reinstate the study context (i.e., Humphreys, Bain and Pike, 1989). Contextual information may be provided by the instructions (‘do you remember the list of words I showed you last week?’). Participants may also spontaneously reinstate aspects of the study context such as the physical environment (Smith and Vela, 2001) or a face paired with the study word (Starns & Hicks, 2013). In Dennis and Humphreys (2001) it was proposed that the current context could be useful under conditions where reinstating a list context might be difficult. Participants may have been able to reinstate a previous context or to use the current context in order to complete the list specific task. Thus, items in the last list may be more familiar because the context stored with those items matches the context the participant is using, not simply due to the length of the retention interval.
An alternative to a decay process is considering a role for contextual change. Using a semi-continuous distractor paradigm, Averall and Heathcote (2011), noted that under explicit and implicit retrieval instructions the observed rate of forgetting was rapid but the curve was almost identical for both instruction conditions. That is they were forgotten at the same rate. While forgetting was rapid in Averall and Heathcote (2011), Dennis and Humphreys (2001) noted almost no forgetting in their list length experiment. In the long condition participants received three lists of study items with 3 minutes of puzzle occurring before list 1, more of the same puzzle activity between lists, and still more following the last list. This was followed by a recognition test for items in list 1 and list 3. There was very little difference in performance between items from the two lists, despite the fact that one list was a lot more recent than the other. These discrepant results suggest a role for contextual change. That is Dennis and Humphreys tried to make contextual reinstatement easy by inserting the same distractor task before and after each of the three lists. In contrast in the semi-continuous distractor paradigm employed by Averall and Heathcote (2011) participants have no idea about how far back in time the test item occurred so they cannot reinstate an earlier context.

I do not have a lot of direct evidence to support the idea that contextual familiarity was responsible for the remaining ability for participants to perform the recognition task. However there is some indirect evidence from other studies that participants could use relational information (between the item and context) to discriminate between conditions when recall could not possible be useful. If this is correct then the relationship between an item and the context it occurred in may also be available in the absence of recall/recollection. There are two papers which show that results that had previously been interpreted in terms of recollection (Diana & Reder, 2006; Reder et al., 2000) occurred when recollection was unlikely. Using the maintenance rehearsal paradigm, Humphreys et al (2010) showed that participants could distinguish intact from rearranged pairs and had higher hit rates for low frequency words than for high frequency words in single item recognition. This occurred with a level of learning that was so low that recall/recollection seemed highly unlikely. These results were further supported and expanded upon by McFarlane and Humphreys (2012). The most interesting finding for our purpose was that there was a significant decrease in the FAR with no decrease in the HR. The standard explanation for a decrease in the FAR is the use of recall to reject (Xu and
Malmberg, 2007). However, that was not possible in McFarlane and Humphreys (2012) because a subsequent experiment showed that recall was almost non-existent. Instead McFarlane and Humphreys (2012) proposed that under some conditions the FAR would decrease when the strength of the parent pairs increased.

A role for context is further supported by the results of Humphreys and Bowyer (1981). Humphreys and Bowyer (1981) found that recognition after studying weakly associated pairs of words was more similar to recall than it was to recognition after studying single words. The main implication was that participants came to rely on relational information (being able to recall the word which the target had been paired with during the study phase) and thus ignored item information or familiarity. So the act of recalling is a potential basis for responding old but it is not the content of what is recalled that is driving the response. Humphreys and Bowyer (1981) proposed that the reason participants rely on recall us because recall is contextually dependent so that there is a strong tendency to either recall the paired word or nothing is recalled. They also noted that with weak associates recall is very good so that little is lost by ignoring item information or familiarity.

The results from experiment 7 provide evidence that supports the Humphreys and Bowyer (1981) interpretation and also the idea that context is important. That is, my conclusion from the results of that experiment was that recollection was only accurate when it was accompanied with a contextual familiarity. The idea that content recollection happens yet is not pivotal to a list discrimination task is at odds with the dominant idea behind dual process theory, that is, that the content of what emerges into consciousness enables one to make a reasonably valid inference that the word occurred in the list or in a particular source. The SMF also makes this inference. Further the inference from the PDP is that recollection is driving the recognition decision, however it appears from experiment 7 that contextual familiarity can also drive the recognition decision.

\textit{Autonoesis as a defining characteristic of episodic memory.}

There are a number of considerations to be made regarding Tulving’s (1983) theory of declarative memory. Our results in conjunction with the Humphreys et al. (2010) results and the McFarlane and Humphreys (2012) results indicate that there may be a feeling of familiarity (knowing) that incorporates relational information without the subjective experience of remembering (self-knowing). It has been argued that autonoesis must be present, as simply ‘knowing’ the contents of an episode,
would not be counted as episodic (Tulving, 2005; Suddendorf and Corballis, 2007). Since there was no indication that autonoesis was helpful in our task, the type of contextual familiarity I have indicated may be operating in our task may not be considered truly episodic. However, I have already argued that our task was episodic though it does not follow that all episodic tasks are solved in the same way or by the same memory system. If this is the case then it may be, that like Clayton, Yu and Dickenson’s (2001) scrub-jays, participants in our study utilised ‘episodic-like’ memory (knowing ‘www’ without autonoesis), however it is unclear at this point how helpful that distinction is. That is, I might wonder whether there is a principled reason for calling a memory, accompanied by autonoesis, episodic while not calling a memory that distinguishes between episodes but which is not accompanied by auotnoesis, episodic.

Can the Recollection of Semantic Content be assigned to a Specific Episode?

The possibility that participants may have confused their semantic reaction to a test word with a memory for having that same reaction at study does not fit well with Tulving’s (1983; 2005) theory. An examination of the written responses made by participants in the justification conditions of experiment 6 illustrates that these responses can include semantic content. Semantic content can come to mind any time the word that produces that response is encountered. For example the following is just a sample of some of the associative words that participants reported thinking about at the time of study: patrol – soldiers, leather – cows, elbow – forearm, chaos - apocalypse, surgeon – Grey’s Anatomy. The involvement of semantic memory occurring as recollection raises the question as to whether participants can discriminate their semantic reaction when that word is being tested from the memory for having a semantic reaction when that word was studied. There is indirect evidence from our data that indicates this type of confusion may occur. R responses were often made to false alarms in our study and this is not so surprising considering that participants are not equipped with many cues to enable them to judge whether they had reactions to items in the immediately prior list or items appearing in any of the prior lists. That is, there are few cues available that might help them judge the time of their reaction and participants don’t seem to be well equipped to always know when their reaction to the word occurred and sometimes falter. This is further supported by the results of Higham and Vokey (2004), who found that R responses increased when participants didn’t first make a recognition decision. This indicates that participants...
may be recollecting from other lists or previous encounters with that item, or confusing a reaction at test with a reaction at study.

It is possible these intrusions were reactions that participants had at test rather than memories they had for the reaction at study. There is no direct evidence of this from the data, however Conway (1995) reported on one patient (AKP) who appeared to confuse thoughts he had for the first time when he encountered a test item with a memory for those thoughts from the time of study. On a recognition task AKP (along with control participants) was asked to justify his recognition memory responses and describe what in his experience led to the recognition judgment. AKP, like other participants could make appropriate use of the response categories (remember, familiar and guess), however AKP experienced false alarms as old items that had been studied. In particular it was concluded that on viewing a word at test AKP was confusing his reaction at test (the association AKP made to that word) as one he had had at study. At this time it is not possible to say whether this patient suffered from a recollection problem, as assumed by Conway, or from a lack of differential familiarity. That is a failure of familiarity may result in an overreliance on recollection and an inability to determine whether a reaction to the test item is or is not a memory from a prior experience with that item. As I indicated earlier, without familiarity, recollection may be inaccurate.

*Future Research*

There are a number of questions and possibilities raised in the discussion that require further research. It has previously been pointed out that it is possible that acontextual familiarity in the form of recency may still be useful in the list specific task and further research may be necessary in order to clarify this. It has also been indicated that participants may confuse a reaction had at study with a prior reaction. It may be beneficial to run a study in which new items are added on the final test list as this would enable us to measure to what extent a reaction had at test may be confused with a reaction that may have occurred at study. If it was found that participants were incorrectly responding R to new items, then this would be difficult for Tulving’s (1983; 2005) theory to accommodate as this would indicate that when cues to a particular context are deficient, then autonoesis may not suffice.
Another avenue for future research might be to explore different manipulations of the RK procedure. This may help to clarify previous effects. For example, it may also be helpful to include a guess measure as the results of Gardiner et al., (1997) indicate that this may take the pressure off participants in choosing between a R and K judgment. It would be interesting to see if participants reduce the number of R judgments to a substantial degree (as they did in the justification condition of experiment 6 or if participants still report R, because they might still not realize the change between criterial and noncriterial information. If we found that R judgments decreased in number without affecting recognition performance then this may be a more reasonable method of obtaining RK judgments.

Further to this, some more understanding about whether or not RK judgments are encouraging participants to recall content may be gained by asking for RK responses initially in a similar manner to Higham and Vokey (2004). Such a manipulation may more dramatically affect our results. In particular this may further clarify the proposal that participants are relying on recollection without familiarity in the final condition and learning to rely on recollection only in the presence of familiarity in the continuous condition. Introducing the RK paradigm initially may encourage participants to rely on recollection more, thus possibly negatively affecting performance in the continuous condition, and more greatly in the final condition. Such an experiment would involve four conditions, with all RK judgments made immediately. Half the participants would perform the task continuously or on the final list as they have previously, however the other half would make RK judgments initially without making a prior recognition judgment.

In addition, it is possible that asking participants to justify their responses was unhelpful since we did not inform participants about criterial recollection. An improvement in recognition performance with the addition of the RK task might be observed if instructions ask participants to make R judgments only for criterial information. This might be quite similar to the RK instructions used by Yonelinas (2002). While more effortful, this is also focusing participants’ attention on more helpful information, so this may further clarify the previous proposal that effortful recollection may hinder performance only when noncriterial information is available. In Mather et al., (1997) they asked for very specific evidence of encoding attributes in the MCQ, in a sense asking for only criterial information would be synonymous to this. Of course, it may be that criterial information is simply not available to
participants – in this case, R judgments may decrease, similar to what was observed in the justification condition of experiment 6. However it is also possible that participants in the continuous condition may change the way they study the items in order to be able to attribute recollections to the occurrence of the test word in the last list. This relies on the assumption that the instructions are suggestive, however when performance is changed by such measures, then the role for suggestion would appear entirely likely.

Other applications

Exploring the ‘butcher on the bus’ experience

Mandler’s (1980) ‘butcher on the bus’ example illustrated the experience of recognizing someone on the bus, while you may feel like you know the person you may not recollect anything about who that person is or where you know them from. Mandler (1980) indicated that a feeling of knowing the person might direct a search to possible contexts that you might be likely to know that person from, in his example this resulted in recognizing the man on the bus as a butcher. In this case, at first this might be considered a feeling of familiarity without recollection, which then might lead to recollection of a particular context. A majority of studies examining RK subjectivity have incorporated words as stimuli, however for the few that have deviated Gruppuso, Lindsay and Masson (2007) note that results tend to reflect that memory for faces are more likely to elicit K rather than R judgments.

Gruppuso et al., (2007) explored whether or not switching contexts paired with faces would create a dissociation such that R responses were affected rather than K responses. Participants studied face-context pairs, with all faces and contexts unique, they were also required to rate the pairs in an orienting task in order to ensure deep encoding of the pairs. Context was created by pairing a picture of a face with a unique landscape such as building interiors or travel scenery. Participants were presented with eight unique study/test lists – including pairs of the following combinations: studied face/studied context, studied face/switched context; studied face/new context, new face/old context and new face/new context. Each pair type was presented at test and participants were told to make a judgment about the face only. R responses were more likely to be made when faces were tested with the context pictures they had been studied with, rather than with new or different contexts. Thus R responses were correlated with increased contextual information.
In Gruppuso et al., (2007), individual and unique information was paired with faces which, of course, were also unique. The results of Gruppuso et al., (2007) indicated that R judgments could be induced if faces were viewed in similar contexts. I was curious as to what may happen to RK judgments if faces were presented instead of words in the list specific task. This task would test participants ability to identify ‘what, where and when’ for faces that all become increasingly familiar over time. It might seem improbable that a R judgment would occur, however this may occur if participants try to rely on recollection with the introduction of the RK paradigm on the final trial. I undertook a small study with only 8 participants included in each condition and therefore it is difficult to conclude anything from such a study, however initial results are interesting. I selected 40 faces (controlling for gender (all were female), colour (gray-scale), size and attractiveness (all were previously rated as being of average attractiveness) to be included in the study, the design was the same as experiment 5 from the thesis (a practice trial was conducted initially). Again targets and distractors were randomized across four lists and participants either made RK judgments on every list (continuous) or on the final list. Performance across the lists on the recognition task was extremely consistent and quite good, despite participant’s insistence that they had not performed well on the task. While there was some increase in hits and false alarms across lists, this was marginal, but may indicate a reliance on familiarity.

Potentially of more interest was the result of the RK manipulation. Interestingly, unlike previous studies, participants in the continuous condition were not necessarily more accurate, but rather, were more likely to say old and were more likely to accompany hits with a K response, not a R response – although this leveled out as they continued the task. However, in the final condition when participants made RK judgments on the final list they were more likely to respond R than K. It is further interesting to note that a decrement to performance was noted in the final condition on the last list compared to the continuous condition and that this correlated with an increase in R responses. Thus possibly participants were again trying to recollect content and therefore possibly relying on recollection in the absence of familiarity, resulting in poorer performance on the recognition task.

Conclusion
It can be concluded that the addition of the RK task does not produce reasonable results when it is applied to multiple recognition lists where on each list a small set of items is randomly reassigned to be old and new items. However, the issues observed may pose more general problems for the procedure. That is, participants did not notice the change from criterial to noncriterial information. Further to this, there is a possible learning process to the task, indicating that RK judgments may not always be a simple reflection of underlying processes. In addition the first time RK decisions are introduced, recognition performance can falter. At the very least, it would seem wise to be cautious in interpreting RK data, especially when there is an absence of converging evidence from another procedure.

Our results further raise some issues regarding the role of content recollection and familiarity in an episodic task. Content recollection was not helpful in our task, however it may be that there was a role for contextual familiarity which may have enabled participants to perform the task. I can not know this for certain until it can be understood how quickly acontextual familiarity might decay. Our results also indicate that recollection may not always be accurate if it is unaccompanied by familiarity and that with the addition of the RK task, participants may learn to report recollection only in the presence of familiarity. Furthermore I raised the possibility that participants may not be able to distinguish between their semantic reaction to a test item and a memory of having that reaction at study. All of these results taken together have implications for how episodic memory is viewed. That is our task is episodic, however this is not the same thing as saying that participants were tapping a unitary episodic memory system. These possibilities raise doubts about whether a principled definition of episodic memory is possible.
REFERENCES


Klatzky, R. L. (1984). Armchair theorists have more fun. *Behavioral and Brain Sciences, 7*(02), 244-244.


Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal


