EXECUTING DEFERRED TASKS IN DYNAMIC MULTITASKING ENVIRONMENTS

MICHAEL DAVID WILSON

B.A. (Hons)

HUMAN FACTORS AND APPLIED COGNITION LABORATORY

SCHOOL OF PSYCHOLOGICAL SCIENCE

UNIVERSITY OF WESTERN AUSTRALIA

THIS DISSERTATION IS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY OF UNIVERSITY OF WESTERN AUSTRALIA

2019
Dedicated to my daughter,
Torvi Isadora Wren.
“I have gradually come to one negative conclusion about the good life. It seems to me that the good life is not any fixed state. It is not, in my estimation, a state of virtue, or contentment, or nirvana, or happiness. It is not a condition in which the individual is adjusted or fulfilled or actualized. To use psychological terms, it is not a state of drive-reduction, or tension-reduction, or homeostasis… The good life is a process, not a state of being. It is a direction not a destination.”

Carl Rogers - 1961
DECLARATION

This thesis has been substantially accomplished during enrolment in this degree.

This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text and, where relevant, in the Authorship Declaration that follows. It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person.

It has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

This thesis contains published work and work prepared for publication, some of which has been co-authored. My co-authors have approved the inclusion of this work in the thesis (p. XXIII).

Signed: ___________________________________________

Date: ________________________________________________

Signed by Michael David Wilson B.A (Hons)
SUMMARY

In complex and dynamic work environments, such as air traffic control (ATC), operational circumstances sometimes require operators to defer the execution of task actions to a more appropriate future point. The cognitive processes involved in remembering to carry out these deferred tasks are referred to as prospective memory (PM). This thesis aims to better understand prospective memory in complex dynamic work environments by examining how situational factors, such as interruptions, modulate individuals’ ability to complete deferred tasks; and how interruptions and deferred tasks impact the performance of routine operational tasks. In Chapter 1, I provide a general introduction to the theories and methods of the PM and interruptions literature. In Chapter 2, I detail the development of a simulated air traffic control paradigm that enabled the assessment of the impact of interruptions on two ecologically valid deferred tasks. In Chapter 3, I report two empirical studies using this paradigm, comparing a blank screen interruption, an \( n \)-back task interruption, and a secondary ATC task interruption. Results showed the secondary ATC task interruption increased resumption failures, and all interruptions increased resumption time. These findings indicated PM in the simulated ATC task depended on frequent interaction with the display, and that PM was susceptible to interference-based forgetting. In Chapter 4, I focus on a deferred handoff task, and examine how the deferred task retention interval, the presence of interruptions during this interval, and the timing of the interruption relative to deferred task encoding and execution modulated performance. Results showed that increasing retention interval (37 s to 77 s to 117 s) reduces PM accuracy. Consistent with previous PM research in simulated ATC settings, we found that maintaining a PM intention triggered costs to ongoing conflict detection accuracy and routine handoff response time. However, deferred handoff response time and accuracy were, surprisingly, unaffected by interruptions. In Chapter 5, I consolidate these findings, arguing that deferred task performance involves a dynamic interplay between top-down strategic maintenance processes and bottom-up cue driven processes and that the effect of interruptions on PM is dependent upon features of the deferred task. I conclude the thesis with a summary of practical implications and directions for future research.
CONTENTS

DECLARATION.................................................................................................................................. V
SUMMARY ....................................................................................................................................... VII
CONTENTS ................................................................................................................................... IX
LIST OF TABLES ........................................................................................................................ XV
LIST OF FIGURES ................................................................................................................... XVII
LIST OF ABBREVIATIONS AND ACRONYMS ......................................................................... XIX
ACKNOWLEDGEMENTS ........................................................................................................ XXI
INCLUDED PUBLICATIONS .................................................................................................. XXII
CANDIDATE CONTRIBUTION ............................................................................................... XXIII
REFERENCES .......................................................................................................................... 161

1 GENERAL INTRODUCTION .......................................................................................... 1
1.1. FOREWORD ................................................................................................................ 1
1.2. INTRODUCTION ......................................................................................................... 2
1.3. AIR TRAFFIC CONTROL FUNDAMENTALS .......................................................... 3
1.3.1. Future ATC Challenges ...................................................................................... 5
1.3.2. Human Memory in ATC .................................................................................... 6
1.4. SCIENCE OF INTERRUPTIONS ................................................................................ 8
1.4.1. Distinguishing Interruptions, Distractions and Task Switching .................. 9
1.4.2. Interruption Anatomy ...................................................................................... 10
1.5. EMPirical RESEARCH ON THE EFFECTS OF INTERRUPTIONS ......................... 12
1.6. THEORETICAL MODELS OF INTERRUPTED TASK PERFORMANCE ............... 14
1.6.1. Memory for Goals ............................................................................................. 14
1.6.1.1. Empirical Support of Memory for Goals ................................................... 15
1.6.1.2. Conclusions on Memory for Goals ............................................................ 17
1.6.2. Threaded Cognition.......................................................................................... 17
1.7. INTERRUPTIONS IN COMPLEX AND DYNAMIC TASKS ..................................... 20
1.7.1. Long-Term Effects of Interruptions ............................................................... 21
# Preamble: Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7.2.</td>
<td>Dynamics: Changing the Task State</td>
<td>23</td>
</tr>
<tr>
<td>1.8.</td>
<td>SUMMARY OF INTERRUPTIONS</td>
<td>24</td>
</tr>
<tr>
<td>1.9.</td>
<td>PROSPECTIVE MEMORY</td>
<td>25</td>
</tr>
<tr>
<td>1.9.1.</td>
<td>PM in the Laboratory and Underlying Cognitive Mechanisms</td>
<td>26</td>
</tr>
<tr>
<td>1.9.1.1.</td>
<td>Encoding</td>
<td>27</td>
</tr>
<tr>
<td>1.9.1.2.</td>
<td>Retention</td>
<td>27</td>
</tr>
<tr>
<td>1.9.1.3.</td>
<td>Retrieval</td>
<td>30</td>
</tr>
<tr>
<td>1.9.2.</td>
<td>Theoretical Perspectives of Prospective Memory</td>
<td>31</td>
</tr>
<tr>
<td>1.9.2.1.</td>
<td>The Dynamic Multiprocess View</td>
<td>31</td>
</tr>
<tr>
<td>1.9.2.2.</td>
<td>Associative Activation Models</td>
<td>33</td>
</tr>
<tr>
<td>1.10.</td>
<td>PROSPECTIVE MEMORY IN AIR TRAFFIC CONTROL</td>
<td>33</td>
</tr>
<tr>
<td>1.10.1.</td>
<td>Workload and Retention in ATC</td>
<td>35</td>
</tr>
<tr>
<td>1.10.2.</td>
<td>ATC-Lab Research</td>
<td>36</td>
</tr>
<tr>
<td>1.11.</td>
<td>CONNECTING PROSPECTIVE MEMORY AND INTERRUPTIONS</td>
<td>37</td>
</tr>
<tr>
<td>1.12.</td>
<td>OVERVIEW OF THESIS</td>
<td>40</td>
</tr>
</tbody>
</table>

## 2 AIR TRAFFIC CONTROL IN THE LABORATORY

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.</td>
<td>FOREWORD</td>
<td>43</td>
</tr>
<tr>
<td>2.2.</td>
<td>SIMULATIONS OF COMPLEX DYNAMIC TASKS</td>
<td>44</td>
</tr>
<tr>
<td>2.3.</td>
<td>ATC-LAB</td>
<td>45</td>
</tr>
<tr>
<td>2.3.1.</td>
<td>ATC-Lab: Programming Interface</td>
<td>48</td>
</tr>
<tr>
<td>2.4.</td>
<td>CURRENT PARADIGM DEVELOPMENT AND PILOT STUDY</td>
<td>50</td>
</tr>
<tr>
<td>2.4.1.</td>
<td>Rationale for Paradigm Development</td>
<td>51</td>
</tr>
<tr>
<td>2.4.2.</td>
<td>General Scenario Development</td>
<td>52</td>
</tr>
<tr>
<td>2.4.3.</td>
<td>Deferred Task Development</td>
<td>53</td>
</tr>
<tr>
<td>2.4.3.1.</td>
<td>Deferred Handoff Task</td>
<td>53</td>
</tr>
<tr>
<td>2.4.3.2.</td>
<td>Deferred Conflict Task</td>
<td>57</td>
</tr>
<tr>
<td>2.4.4.</td>
<td>Interruption Task Development</td>
<td>59</td>
</tr>
<tr>
<td>2.4.4.1.</td>
<td>n-back Interruptions</td>
<td>59</td>
</tr>
<tr>
<td>2.4.4.2.</td>
<td>External Program Interruption</td>
<td>61</td>
</tr>
<tr>
<td>2.4.5.</td>
<td>Counterbalancing Techniques and XML</td>
<td>62</td>
</tr>
</tbody>
</table>
3 REMEMBERING TO EXECUTE DEFERRED TASKS IN SIMULATED ATC: THE IMPACT OF INTERRUPTIONS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>FOREWORD</td>
<td>70</td>
</tr>
<tr>
<td>3.2</td>
<td>ABSTRACT</td>
<td>70</td>
</tr>
<tr>
<td>3.3</td>
<td>INTRODUCTION</td>
<td>71</td>
</tr>
<tr>
<td>3.4</td>
<td>RESUMING DEFERRED TASKS</td>
<td>73</td>
</tr>
<tr>
<td>3.5</td>
<td>REMEMBERING TO DEVIATE FROM ROUTINE</td>
<td>75</td>
</tr>
<tr>
<td>3.6</td>
<td>EXPERIMENT 1</td>
<td>77</td>
</tr>
<tr>
<td>3.6.1</td>
<td>Method</td>
<td>77</td>
</tr>
<tr>
<td>3.6.1.1</td>
<td>Participants</td>
<td>77</td>
</tr>
<tr>
<td>3.6.1.2</td>
<td>ATC-LabAdvanced Simulator</td>
<td>78</td>
</tr>
<tr>
<td>3.6.1.3</td>
<td>Training Phase</td>
<td>79</td>
</tr>
<tr>
<td>3.6.1.4</td>
<td>Test Phase</td>
<td>80</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Results</td>
<td>81</td>
</tr>
<tr>
<td>3.6.2.1</td>
<td>Deferred Conflict Task</td>
<td>82</td>
</tr>
<tr>
<td>3.6.2.2</td>
<td>Deferred Handoff Task</td>
<td>83</td>
</tr>
<tr>
<td>3.6.2.3</td>
<td>Subjective Workload</td>
<td>85</td>
</tr>
<tr>
<td>3.6.2.4</td>
<td>Ongoing task performance</td>
<td>85</td>
</tr>
<tr>
<td>3.6.3</td>
<td>Discussion</td>
<td>86</td>
</tr>
<tr>
<td>3.7</td>
<td>EXPERIMENT 2</td>
<td>87</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Method</td>
<td>88</td>
</tr>
<tr>
<td>3.7.1.1</td>
<td>Participants</td>
<td>88</td>
</tr>
<tr>
<td>3.7.1.2</td>
<td>Materials and Procedure</td>
<td>88</td>
</tr>
<tr>
<td>3.7.1.3</td>
<td>ATC-Interruption</td>
<td>89</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Results</td>
<td>90</td>
</tr>
<tr>
<td>3.7.2.1</td>
<td>Deferred Conflict Task</td>
<td>90</td>
</tr>
<tr>
<td>3.7.2.2</td>
<td>Deferred Handoff Task</td>
<td>91</td>
</tr>
<tr>
<td>3.7.2.3</td>
<td>Subjective Workload</td>
<td>92</td>
</tr>
<tr>
<td>3.7.2.4</td>
<td>Ongoing Task Performance</td>
<td>92</td>
</tr>
<tr>
<td>3.7.2.5</td>
<td>Interrupting ATC task performance</td>
<td>93</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Discussion</td>
<td>94</td>
</tr>
</tbody>
</table>
3.8. EX-GAUSSIAN DISTRIBUTION MODELLING ......................................................... 95
3.9. GENERAL DISCUSSION ........................................................................................ 99
  3.9.1. Resuming Interrupted Tasks ...................................................................... 100
  3.9.2. Remembering to Deviate from Routine .................................................... 102
  3.9.3. Practical Implications ............................................................................. 104
  3.9.4. Considerations and Conclusions ............................................................. 106
3.10. APPENDIX 3A: BAYESIAN ROBUSTNESS CHECKS ........................................... 108

4 PM IN ATC ROBUST TO INTERRUPTIONS BUT IMPAIRED BY RETENTION INTERVAL ............................................................................................................................... 111
  4.1. FOREWORD ................................................................................................... 112
  4.2. STRUCTURED ABSTRACT .......................................................................... 114
  4.3. INTRODUCTION ............................................................................................ 115
  4.4. THEORETICAL APPROACH ........................................................................ 115
  4.5. CURRENT STUDY .......................................................................................... 118
  4.6. METHOD ......................................................................................................... 119
    4.6.1. Participants ............................................................................................... 119
    4.6.2. Design ...................................................................................................... 119
    4.6.3. Counterbalancing .................................................................................... 120
    4.6.4. ATC-Lab Advanced Simulator .................................................................. 121
    4.6.5. Procedure ................................................................................................. 123
    4.6.6. Questionnaire ........................................................................................... 125
  4.7. RESULTS ........................................................................................................... 125
    4.7.1. Deferred Handoff PM Accuracy ............................................................... 126
    4.7.2. Deferred Handoff PM Response Time ...................................................... 128
    4.7.3. Cost of PM Retention to Ongoing ATC tasks ........................................... 129
  4.8. DISCUSSION .................................................................................................... 133
    4.8.1. Limitations, Future Directions, and Practical Implications ..................... 135
  4.9. APPENDIX 4A: WORKLOAD RESULTS ......................................................... 137
  4.10. APPENDIX 4B: LONG-TERM EFFECTS ...................................................... 137
  4.11. APPENDIX 4C QUESTIONNAIRE ................................................................. 138

5 GENERAL DISCUSSION ............................................................................................ 141
  5.1. SUMMARY OF EMPIRICAL FINDINGS .......................................................... 142
Preamble: Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.</td>
<td>THEORETICAL CONTRIBUTIONS</td>
<td>143</td>
</tr>
<tr>
<td>5.2.1.</td>
<td>Task Resumption Processes</td>
<td>143</td>
</tr>
<tr>
<td>5.2.1.1.</td>
<td>Limitations of Threaded Cognition for SA</td>
<td>145</td>
</tr>
<tr>
<td>5.2.2.</td>
<td>The Impact of Interruptions on Subjective Workload</td>
<td>147</td>
</tr>
<tr>
<td>5.2.3.</td>
<td>Long-Term Effects of Interruptions</td>
<td>148</td>
</tr>
<tr>
<td>5.2.4.</td>
<td>Sector Density, Workload and PM Capacity Sharing</td>
<td>149</td>
</tr>
<tr>
<td>5.2.5.</td>
<td>Further Integrating Prospective Memory and Interruptions</td>
<td>150</td>
</tr>
<tr>
<td>5.2.6.</td>
<td>Modelling Absolute Forgetting</td>
<td>151</td>
</tr>
<tr>
<td>5.3.</td>
<td>PRACTICAL IMPLICATIONS</td>
<td>152</td>
</tr>
<tr>
<td>5.4.</td>
<td>FURTHER RESEARCH DIRECTIONS</td>
<td>154</td>
</tr>
<tr>
<td>5.4.1.</td>
<td>Modelling Absolute Forgetting of Deferred Handoff Task</td>
<td>154</td>
</tr>
<tr>
<td>5.4.2.</td>
<td>The Nature of the Interrupting Task</td>
<td>155</td>
</tr>
<tr>
<td>5.4.3.</td>
<td>Assessing the Impact of Interruptions on Workload</td>
<td>155</td>
</tr>
<tr>
<td>5.5.</td>
<td>CONCLUSION</td>
<td>156</td>
</tr>
<tr>
<td>5.6.</td>
<td>APPENDIX 5A: AN IMPROVED ATC INTERRUPTION PARADIGM</td>
<td>157</td>
</tr>
<tr>
<td>5.6.1.</td>
<td>Appendix Foreword</td>
<td>157</td>
</tr>
<tr>
<td>5.6.2.</td>
<td>Context of Task</td>
<td>157</td>
</tr>
<tr>
<td>5.6.3.</td>
<td>Experimental Details</td>
<td>157</td>
</tr>
<tr>
<td>5.6.4.</td>
<td>Experimental Task</td>
<td>158</td>
</tr>
<tr>
<td>5.6.5.</td>
<td>Application as an Interruption Task</td>
<td>160</td>
</tr>
</tbody>
</table>
LIST OF TABLES

TABLE 2.1. Average proportion (from the pilot study) of deferred handoff aircraft correctly handed off (1 = correct, 0 = error of any type), with standard deviations in parentheses, presented separately for each condition. ..............................56

TABLE 2.2. Reduced Latin-square counterbalancing scheme. In this example, the various scenarios are counterbalanced across the experimental conditions. Participants would be equally distributed across groups A-D to achieve counterbalancing. ....................................................................................................................64

TABLE 3.1. Grand means for all ongoing task performance measures during the post-interruption period, and associated contrast test results (DF = 58). ..........................86

TABLE 3.2 Grand means for all ongoing task performance measures during the post-interruption period, and associated contrast test results (DF = 58). ..........................93

TABLE 3.3. Means and standard deviations for ATC-interruption performance on handoffs, acceptances, and conflict detection ............................94

TABLE 3.4. Summary statistics of Bayesian posteriors of the ex-Gaussian parameters $\mu$, $\Sigma$, and $\Lambda$ and the probability of failure $P_F$. ..........................................................99

TABLE 4.1. The experimental design along with the respective total retention interval.................................................................120

TABLE 4.2 Counterbalancing scheme for the first experimental session ..........................................................121

TABLE 4.3. Means and standard deviations for the three ongoing task performance measures by PM cost condition (i.e., whether the PM intention was to be maintained or not). .................................................................129

TABLE 4.4 Model comparison table for all deferred handoff PM model comparisons (PM errors and RT) .................................................................130

TABLE 4.5. Model comparison table for all ongoing task models .................................................................131
LIST OF FIGURES

FIGURE 1.1. AN EN-ROUTE AIR TRAFFIC CONTROLLER WORKING IN THE COPENHAGEN AREA CONTROL CENTRE (DENMARK)................................................................................................................................. 4

FIGURE 1.2. THE INTERRUPTION ANATOMY ............................................................................................................. 11

FIGURE 1.3. THE MULTITASKING CONTINUUM........................................................................................................ 18

FIGURE 2.1: THE ATC DISPLAY (IDENTICAL TO FIGURE 3.1).................................................................................. 48

FIGURE 2.2. A DISPLAY OF THE ATC INTERFACE, WITH THE AEGISUB SUBTITLES OVERLAID indicating two aircraft pairs have now “entered conflict” (actually refers to a “future conflict”). ......................................................................................................................... 53

FIGURE 2.3. DISTRIBUTION OF DEFERRED HANDOFF RESPONSE TIMES ACROSS THE SIX SUBJECTS IN THE PILOT STUDY ......................................................................................................................... 56

FIGURE 2.4. DISTRIBUTION OF DEFERRED CONFLICT RESUMPTION TIMES (MILLISECONDS) ACROSS THE SIX SUBJECTS IN THE PILOT STUDY ............................................................................. 59

FIGURE 3.1: THE ATC DISPLAY ........................................................................................................................................ 79

FIGURE 3.2: MEAN RESUMPTION TIMES AND FAILURE PROPORTIONS ACROSS THE THREE INTERRUPTION CONDITIONS FOR THE DEFERRED CONFLICT TASK ......................................................................................... 83

FIGURE 3.3: MEAN HABIT-CAPTURE RATE AND RTS ................................................................................................. 84

FIGURE 3.4. AN EXAMPLE OF AN INTERRUPTING ATC SCENARIO ............................................................................. 89

FIGURE 3.5. MEAN RESUMPTION TIMES AND FAILURE PROPORTIONS ACROSS THE THREE INTERRUPTION CONDITIONS FOR THE DEFERRED CONFLICT TASK .......................................................................................... 90

FIGURE 3.6. MEAN HABIT CAPTURE ERROR RATE AND RESPONSE TIMES ACROSS THE THREE INTERRUPTION CONDITIONS FOR THE DEFERRED HANDOFF TASK ........................................................................ 92

FIGURE 3.7. CUMULATIVE RT FUNCTION FOR THE FOUR UNIQUE CONDITIONS IN EXPERIMENT 1 AND 2 ......................................................................................................................................................... 97

FIGURE 3.8. BAYES FACTOR ROBUSTNESS PLOTS FOR DEFERRED HANDOFF ERRORS, DEFERRED CONFLICT FAILURES (MISSES), AND DEFERRED CONFLICT RESUMPTION TIME IN EXPERIMENT 1 ......................................................................................................................................................... 109
Preamble: List of Figures

FIGURE 3.9. BAYES FACTOR ROBUSTNESS PLOTS FOR DEFERRED HANDOFF ERRORS, DEFERRED
CONFLICT FAILURES (MISSES) AND DEFERRED CONFLICT RESUMPTION TIME IN EXPERIMENT 2
.................................................................................................................................................................................. 110

FIGURE 4.1. THE TIME-COURSE OF THE PM HANDOFF TASK (NOT TO SCALE) ...................... 119

FIGURE 4.2. A SCREENSHOT OF THE ATC DISPLAY.......................................................................................... 122

FIGURE 4.3. MEAN DEFERRED HANDOFF ERROR RATE ACROSS THE FOUR TIMING CONDITIONS AND
THE TWO INTERRUPTION CONDITIONS FOR THE DEFERRED HANDOFF TASK......................... 127

FIGURE 4.4. MEAN DEFERRED HANDOFF RT ACROSS THE FOUR TIMING CONDITIONS AND THE TWO
INTERRUPTION CONDITIONS FOR THE DEFERRED HANDOFF TASK................................. 128

FIGURE 4.5. EFFECT DISPLAY PLOTS FOR THE PM OVERLAP PROPORTION................................. 132

FIGURE 5.1 SCREENSHOT OF THE PROTOTYPE ATC TASK INTERFACE. .............................................. 159
**LIST OF ABBREVIATIONS AND ACRONYMS**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATC</td>
<td>Air Traffic Control</td>
</tr>
<tr>
<td>ACT-R</td>
<td>Adaptive Control of Thought Rationale</td>
</tr>
<tr>
<td>C2</td>
<td>Command and Control</td>
</tr>
<tr>
<td>DMPV</td>
<td>Dynamic Multi-process View of Prospective Memory</td>
</tr>
<tr>
<td>MFG</td>
<td>Memory for Goals Theory</td>
</tr>
<tr>
<td>PM</td>
<td>Prospective Memory</td>
</tr>
<tr>
<td>RT</td>
<td>Response Time</td>
</tr>
<tr>
<td>SA</td>
<td>Situation Awareness</td>
</tr>
</tbody>
</table>
Preamble: Acknowledgements

ACKNOWLEDGEMENTS

First and foremost, I owe a great deal of thanks to my wife Bree. I am not sure I will ever be able to repay you for all you’ve done in supporting me throughout this five-year journey. You have been my ship, sail, and anchor, providing relentless support through all the highs and lows. I know with confidence that nobody is happier than you in me achieving this milestone. I am looking forward to the exciting next chapter of our lives to share together with Torvi.

Mum and Dad, thanks for bringing me into this world and giving me the opportunity to learn, grow and explore life. I literally wouldn’t be here without you. Dan and Anna, you’ve both taught me more than I thought I could learn; you both continue to push me onwards and upwards. To my family, you’ve all supported my journey in innumerable ways, thank you.

I am deeply grateful for the tireless efforts of my primary supervisor, Shayne Loft. Your patience, wisdom, and honest pragmatism has guided me through the journey of a PhD, and you got me out in the end in one piece. Thank you for being such a terrific mentor, counselor and friend. I would also to express my gratitude to my co-supervisors, Simon Farrell and Troy Visser. Simon, your knowledge, encouragement, and leadership has forever shaped the way I perceive and practice science — also, thanks for showing me Python. Troy, thank you for providing me with your insight, experience, and expertise on a range of research and life issues over the past five years — every single time with a smile and genuine warmth.

Luke, for reminding me that a job’s a job and to get it done, no matter how torturous.

Andrew, yes.

Fam squad, for knowing not to ask how the thesis is going… until I mentioned it first.

Russ, for not replacing me in those torrid early days, and for the computational modelling book.

Riley, for providing reinforcements when I was out in the trenches.

Raj crew, for all the early Sunday mornings in the temple, and those to come.

There are many individuals who’ve helped to make my PhD so rewarding. To Julian, Huw, Karina, Brad, Nic, all the HuFAC members and ALL friends, Sanderians, and countless people who’ve supported me along the way, I’d like to extend my utmost gratitude!

This research was supported by an Australian Government Research Training Program (RTP) Scholarship and an Australian Postgraduate Award (APA), which I am deeply thankful for.
INCLUDED PUBLICATIONS

This thesis is presented as a combination of journal article manuscripts (i.e., series of papers) and unpublished empirical chapters.

Chapter 3:


Chapter 4:

CANDIDATE CONTRIBUTION

I, the candidate, independently wrote Chapter 1 (general introduction) and Chapter 5 (general discussion) with guidance from my Primary and Coordinating Supervisor, Dr Shayne Loft. I was responsible for the design, development, data collection, and analysis for all experiments and work reported in Chapter 2, 3, and 4; with advice received at various stages from my supervisors Shayne Loft, Troy Visser, and Simon Farrell. I am the first author on the publication associated with Chapter 3, and authored this publication with feedback and advice from Shayne Loft, Troy Visser, and Simon Farrell. I am also the first author on the manuscript associated with Chapter 4, and authored this publication with feedback and advice from Shayne Loft, Troy Visser, and Simon Farrell, and Luke Strickland.

I, the candidate, Michael David Wilson confirm the above statement is true.

Candidate signature: 

Date: 7/11/2019

I, Shayne Loft (coordinating supervisor), certify that the student’s statements regarding their contribution to each of the works listed above are correct. As all co-authors’ signatures could not be obtained, I hereby authorise inclusion of the co-authored work in the thesis.

Coordinating supervisor signature: 

Date: 7/11/2019
Chapter 1: General Introduction

1 GENERAL INTRODUCTION

“What I actually predict will happen is that the lessons of the last 50 years will be repeated in the next 50. Airlines will still prefer to spend $500 on aircraft for every $1 spent on ATC. Will the cost of potential super-systems actually prohibit their introduction, as they prove totally cost-ineffective? If I survive to the age of 93 and I fly somewhere in 2040, I suspect that there will still be a human problem solver on the ground in control of my flight, who will rejoice in the title of ‘the controller.’ And I don’t think that controllers will be there because they are irreplaceable, or because the public wants someone there. I think that, with the right tools to help, the controller will still be there as the most cost effective, flexible system solution to the problem of safely guiding pilots and passengers to their destination. And that is what air traffic control is really all about.” — Levesley (1991, p. 539)

1.1. Foreword

The overall goal of this thesis is to better understand prospective memory (PM) in complex dynamic work environments, such as air traffic control; and to understand what situational factors (e.g., interruptions) may modulate individuals’ ability to prospectively remember. This introductory chapter introduces the reader to the core aims of this thesis and continues with a general overview of air traffic control systems and the PM and interruption science literatures. I introduce the reader to interruption science, focusing on activation-based models of goal suspension and resumption. Then, I outline the various methodologies which have been used to examine the cognitive processes underlying PM in laboratory contexts and introduce the theoretical perspectives of PM founded upon these laboratory-based studies. After this, I provide an overview of the connections between PM research and interruptions research. This chapter concludes with a general overview of the thesis, and an explanation of how the
Executing Deferred Tasks in Dynamic Multitasking Environments

The aforementioned theoretical frameworks and empirical literatures will be applied to the research paradigm implemented in this thesis.

1.2. Introduction

In complex-dynamic work environments, such as air traffic control (ATC), human operators must prioritise and coordinate multiple task objectives in face of a transient and continuously evolving situation to sustain high operational safety and efficiency (Stein, Garland, & Muller, 2009). A common class of task goals which are often reported as being susceptible to forgetting, and a central focus of the current thesis, are deferred tasks. Deferred tasks occur when operators are prohibited from immediately carrying out an intended task, requiring them to remember to perform the action at a more appropriate point in the future (Loft, 2014). For example, an air traffic controller may notice atypical flight parameters, and must remember to inform a fellow controller after attending to other more pertinent demands. The cognitive processes involved with the maintenance, retrieval, and execution of deferred tasks are collectively referred to as prospective memory (PM; Shelton & Scullin, 2017). While forgetting to complete deferred tasks may be a nuisance in everyday life; in many safety-critical workplaces, forgetting to complete such intentions can have serious safety implications. Thus, it is of great importance to understand how situational factors, such as the presence of distractions and interruptions, may impede individuals' ability to prospectively remember. In order to support high performant safety-critical systems, it is important that we are able to predict and understand the situational and cognitive factors that support the execution of goals, as well as the conditions which may inhibit individuals’ ability to complete intended task goals.

The broad aim of this thesis is to examine the how situational factors, such as interruption, can impact individuals’ ability to complete task goals in complex multitasking environments. More specifically, this thesis address three core questions regarding human performance in dynamic multitasking environments.

- To what extent do interruptions interfere with individual’s ability to remember and complete intended task goals?
- To what extent do interruptions impact task performance more generally?
- To what extent do the properties and features of the deferred tasks modulate individuals’ PM performance?
Chapter 1: General Introduction

To address these questions, the current thesis introduces a theoretical framework and an experimental methodology for examining deferred tasks in a laboratory-based ATC simulator. Specifically, I assess how situational contexts such as task interruption, multitasking, and task demand impact the probability of failing to carry out intended behaviours in complex and dynamic task environments. The predictions, experimental designs, and statistical and computational methodologies presented in this thesis are informed directly by the theory and methods from several more fundamental cognitive science literature bodies, including: PM, working memory, interruption science, and human multitasking. Consequently, the empirical outcomes of this thesis not only address practical questions of human factors issues in complex and dynamic environments, but they also speak to the utility and generalisability of basic cognitive theories that are typically examined using simple laboratory tasks. In the following section, I provide a brief overview and task analysis of the ATC task environment, focusing on aspects of the system in which human memory is critical to maintaining overall system safety.

1.3. Air Traffic Control Fundamentals

ATC is a human-machine system that aims to expedite the flow of air traffic as safely and efficiently as possible. The overall safety and effectiveness of ATC systems is dependent on human air traffic controllers who undergo intensive training to learn how to apply appropriate processes and tools to manage air traffic. The principle duties of all air traffic controllers are to enforce separation standards between aircraft and ensure that aircraft reach their destinations in an efficient and timely manner (Nolan, 2009). In Australia, there are three broad categories of air traffic controllers: en-route, terminal and tower controllers (Air Services Australia, 2017). Tower controllers operate from the iconic airport tower and are responsible for all aircraft and vehicles movements on taxiways, runways, and other areas of importance in the airport vicinity. Tower controllers are required to maintain separation between aircraft (and other vehicles) manoeuvring in areas under their control, including airborne aircraft within close proximity to the airport. Terminal controllers (TRACON [terminal radar approach control] in USA) are responsible primarily for aircraft arrivals and departures (e.g., checking altitudes are appropriate at handoff), and the related required transitions to and from cruising altitudes for airports in major cities. Finally, en-route controllers are responsible for all aircraft during the high-speed high-altitude part of flight, known as cruising (above 25,000 ft). In Australia, there are two en-route facilities (located in Brisbane and Melbourne). Within each facility, en-route controllers
are responsible for a smaller subset of airspace, known as a sector. This thesis focuses on simulations of en-route air traffic control, from the perspective of a controller responsible for a single sector.

Figure 1.1. An en-route air traffic controller working in the Copenhagen Area Control Centre (Denmark). The large screen in the centre is the primary air situational display that shows the geographical position of aircraft. The secondary display to the left of the situation display shows flight information in a list format and functions as an electronic flight progress strip. There are three displays underneath the air situational display. The panel of the far left is a communications panel that enables access to all required communication channels (i.e., telephone lines), such as to other ATC sector controllers, control towers, and emergency services. The lower middle display is a dynamic display which can vary depending on situational requirements, and in this example is set to provide channel frequencies for adjacent sectors. The lower right panel is the voice communication system (VCS) interface that is used to communicate with aircraft. The panel also displays VCS diagnostics. Image sourced with permission from Naviair (2007).

The primary objective of en-route controllers is to ensure that all aircraft operating within their sector maintain separation standards (Nolan, 2009). Australian separation standards specify that all aircraft travelling within controlled airspace up to 29,000 ft (8800 m) must maintain at least 1000 ft (305 m) of vertical separation, unless they meet horizontal separation standards. The horizontal separation standard between aircraft flying at the same altitude is 5 nm (9260 m). When flying at over 29,000 ft, the vertical separation increases to 2000
Chapter 1: General Introduction

ft (610 m). Aircraft which violate minimum separation standards (i.e., come to close together), or will do so in the future, are known as aircraft conflicts. One en-route controller has the ultimate responsibility of aircraft separation within a sector, although they may be assisted by other controllers. Secondary to conflict detection, controllers are also required to: maintain and improve airspace efficiency, provide pilots with information and status updates (e.g., weather reports), and provide emergency services as required.

En-route controllers accomplish these objectives with three tools: the primary situation display (also referred to as plan view logical display, or more generally as radar displays), communication devices, and a representation of flight information and progress (Durso & Manning, 2008). In most countries, flight information is represented with electronic flight progress strips that display the aircraft flight plan (e.g., route information; next sector scheduled for handoff). The situation display provides a 2D geographical map of a sector and displays the position of all aircraft travelling within it, with specific flight and aircraft information being presented adjacent to each respective aircraft in a ‘databox’ (e.g., the identity, destination, altitude, speed, and cleared altitude). Typically, controllers use the situation display to assess current and interpolate future aircraft positions; and depending on the specific ATC system, the display can be used to perform other computer-assisted operations (e.g., trajectory estimation and setting reminders). Moreover, the display can signal emergency warnings, such as short-term conflict alerts, area proximity warnings and minimum safe altitude warnings.

1.3.1. Future ATC Challenges

Across all industrialised nations in the past 20 years, airspaces have becoming increasingly crowded due to the increased demand for air traffic services. Air traffic is expected to double by 2031, with a projected growth rate of 4.9% per annum (Airports Council International, 2017). This presents a critical challenge for controllers, who are required to manage this demand while also sustaining operational safety and efficiency. Recent efforts to address this challenge have focused around improving the operating procedures within ATC and developing new automated technologies (Mosier et al., 2013). For example, in the United States, several aviation organizations, including the Federal Aviation Administration, are currently developing the ‘Next Generation Airspace Transportation System’ (NextGen). NextGen is a collection of new technologies characterized by significantly higher levels of automation than current systems,
which may result in controllers serving an increasingly supervisory role and holding responsibility for increased levels of air traffic.

However, at least for the foreseeable future, human operators will continue to remain the backbone of the ATC system. Consequently, technological advancements must keep pace with our understandings of the human cognitive system. Thus, it is imperative that human factors research continues to provide empirical research that can help to ensure controllers are matched with technical systems successfully. In this thesis, I focus on the memory demands posed by the ATC task environment, and what task factors may impede memory performance and pose human systems integration risks in ATC and other similar work environments.

1.3.2. Human Memory in ATC

Retrospective and prospective memory permeate essentially every aspect of the ATC workplace, and are critical determinants of controller performance. Controllers are required to cognitively process an enormous volume of continuously changing transient information (Stein et al., 2009) in an environment characterised by frequent interruptions, distractions, and high workload (Durso & Manning, 2008; Nolan, 2009; Shorrock, 2005). Humans are fallible, and all these factors are known contributors to human error.

Shorrock (2005) conducted an ethnographic field study that documented the different kinds of memory errors that can occur in ATC. The data comprised interviews from 28 UK air traffic controllers, as well as a systematic review of 48 incidents involving loss of separation in the UK, spanning a 3-year period. The interviews with expert controllers revealed that 16 of 42 reported errors (38%) could be attributed to PM errors; with the remainder of errors being attributed to retrospective memory failure. Shorrock further classified these errors by the mechanism underlying the error and found controllers mainly attributed PM errors to memory capacity overload (e.g., when a sector was particularly busy) and distractions or interruptions (e.g., communication requests from a co-worker or pilot). Similarly, the analysis of the incident reports revealed that of the 20 incidents that were caused by memory errors, eight of these (40%) specifically involved PM. Shorrock outlined a variety of specific ATC operations in which PM errors can occur:

- **Forgetting to perform intended task actions.** For instance, marking flight levels and headings on a flight strip; monitoring an identified conflict situation at the appropriate
time; checking that an assistant has carried out the instructed tasks; remembering to transfer an aircraft to the next sector controller; and remembering to inform relief staff of unusual routeings during handovers.

- **Forgetting to monitor errors** involve situations where the controller forgets to monitor a region of the radar display. For instance, Shorrock reported a controller who cleared two aircraft to climb to altitudes within 2000 ft of each other and was relying on aircraft performance to maintain lateral separation. The controller had intended to monitor the situation using predicted vectors, however an interruption (a phone call) had caused the controller to forget to monitor the situation, this resulted in the aircraft entering into conflict.

- **Habit Capture PM errors** occur when a controller intends to deviate from an intended sequence of actions, but fails to do so at the correct time and instead performs the default habitual behaviour (Reason, 1990). For example, air traffic controllers sometimes must remember to deviate from a routine aircraft vectoring procedure and to instead hold an aircraft when it reaches a specific way-point in the future because of heavy traffic in their flight control sector.

One clear finding from Shorrock’s (2005) study is that interruptions were attributed as a key factor underlying PM errors across several operational circumstances. This is not an isolated finding in ethnographic studies. An analysis of 33 voluntary incident reports from air traffic controllers conducted by Jones and Endsley (1996) found that interruptions and distractions occurred in 27.3% of incidents involving memory errors and 53% of incidents where a controller failed to correctly observe data. Further, distractions and interruptions were linked with all forms of errors involving loss of situation awareness.

While ethnographic research such as the above provide practical insights into the general constraints and causes of human error in field settings, they lack the experimental control required to understand the specific psychological mechanisms which underlie the relationship between task factors (such as interruption and distraction) and memory errors. In the following section, I will review the empirical research that has investigated the cognitive effects of interruptions, and the major theories of interrupted task performance.
1.4. Science of Interruptions

Being ubiquitous in life, interruptions naturally have a multitude of definitions. For our purposes, an interruption is: “the suspension of one stream of work prior to completion, [occurring along] with the intent of returning to and completing the original stream of work” (Boehm-Davis & Remington, 2009). Modern work environments are particularly susceptible to interruptions due to their reliance on information and communication technologies, frequent interpersonal communication, and multiple task demands (Trafton & Monk, 2007). Applied psychological research examining interruptions has typically concentrated on work domains characterised by human-computer interaction (McFarlane, 2002; Oulasvirta & Saariluoma, 2004) and safety-criticality such as aviation (Dismukes & Nowinski, 2006; Dismukes, Berman, & Loukopoulos, 2007; Latorre, 1998, 1999), medicine (Grundgeiger et al., 2013; Grundgeiger, Sanderson, MacDougall, & Venkatesh, 2010; Sasangohar, Donmez, Easty, & Trbovich, 2017) and driving (Bowden, Loft, Wilson, Howard, & Visser, 2019; Gregory, Irwin, Faulks, & Chekaluk, 2014). In safety-critical environments failures in managing or recovering from interruptions can be disastrous. Northwest Airlines Flight 255 is a notable example. The crew of the McDonnell Douglas MD-82 were interrupted by air traffic control before completing a pre-flight checklist. Tragically, a failure to return to the checklist following the interruption resulted in an accident killing all on board except one passenger (National Transportation Safety Board, 1988). This incident led to all major airlines mandating that checklists must be restarted if interrupted and installing take-off configuration warning systems on aircraft. The fate of Northwest Airlines Flight 255 serves as a strong reminder of the serious consequences interruptions can have in safety-critical settings.

Interruption science, the formal study of interruptions, can improve safety by providing insights into the cognitive processes underlying interrupted task performance. These insights can then inform work design and training interventions (Boehm-Davis & Remington, 2009; Hodgetts, Tremblay, Vallières, & Vachon, 2015a; Kim, Chun, & Dey, 2015). Before reviewing empirical interruptions literature and the major theoretical frameworks of interrupted task performance, I will introduce the interruption-specific terminology used throughout this thesis, and then introduce the ‘interruption anatomy’ — a general task analysis of typical interruption situations that demonstrates the use of interruption terminology.
1.4.1. Distinguishing Interruptions, Distractions and Task Switching

Historically, the terms *interruption* and *distraction* have been used largely interchangeably, both colloquially and in the academic literature. However, recent work has established that distraction and interruptions are two distinct phenomena, with each having differential effects on performance and underlying cognitive processes (Grundgeiger et al., 2010). Interruptions are considered to be external events that temporarily shift attention entirely away from a primary task to a secondary task. Importantly, a necessary condition for an interruption situation is that the interrupted individual must intend to return to (i.e., resume) the primary task after the interruption. Interruption situations are also referred to as *sequential multitasking*, as multiple tasks are interleaved sequentially (Salvucci & Taatgen, 2008).

Distractions however occur when an event consumes attentional capacity, but the primary task is not entirely suspended (Bowden et al., 2019; Lee, Young, & Regan, 2008), resulting in *concurrent multitasking* (i.e., performing both tasks simultaneously). In distraction situations, there is no intention formed to return the primary task, as it is never suspended. Talking on the phone while driving is a prototypical example of distraction arising from concurrent multitasking. Any sensible driver would not suspend the task of driving entirely in order to talk on the phone (however, driving performance is likely to still be degraded by the distraction). Alternatively, temporarily stopping work on a document to speak to a colleague in another room would be an example of an interruption, as the primary task of writing the document must be suspended (and subsequently resumed) to engage the interrupting task (speaking to the colleague).

Another conceptually related phenomena that should be acknowledged is *task switching*. Task switching experiments involve continuously switching between two short-duration tasks, in a similar manner to switching between a primary and secondary task when interrupted (Alzahabi & Becker, 2013). However, unlike interruption situations (in which there is a clear primary task that is suspended and resumed), task switching paradigms do not require each task to be suspended or resumed, that is, the set of tasks are discrete. Here, the primary variable of interest is switch-costs, which refers to the finding that response times are longer on trials which have just ‘switched’ than on ‘non-switch’ or ‘task-repetition’ trials (Monsell, 2003).

Further discussion of how these distinctions can be conceptualised under a single framework will be provided in the context of Threaded Cognition theory in Section 1.5.4.
Executing Deferred Tasks in Dynamic Multitasking Environments

(Salvucci & Taatgen, 2008); however for present purposes, the key point is that task switching, interruptions and distraction are distinct phenomenon, and that this thesis focuses on how interruptions impact cognitive performance.

1.4.2. Interruption Anatomy

The general stages involved with task interruptions are well established (McFarlane, 2002; Trafton & Monk, 2007). The *interruption anatomy* provides a cognitive task analysis of typical interruption situations and forms the basis of the interruption-specific terminology used throughout this thesis, and the literature more broadly. The defining feature of an interruption is the presence of a primary task, which must be suspended in order to attend to an interrupting task (Monk, Trafton, & Boehm-Davis, 2008). Importantly, following the completion of the interrupting task, the primary task must be resumed, which requires one to retrieve the suspended task goals from memory or reconstruct them from the primary task state. As just mentioned, the presence of primary task that must be returned to distinguishes interruptions research from conceptually related work such as task-switching and concurrent multitasking (in which multiple tasks are performed concurrently).

Figure 1.2 presents the typical task analysis of the time-course of an interruption that is followed by successful task resumption, referred to as the interruption anatomy (Trafton, Altmann, Brock, & Mintz, 2003; Trafton & Monk, 2007). The interruption anatomy provides a temporal and schematic description of the key stages that occur throughout the time-course of an interruption. According to the model, the processes of suspending and resuming a primary task can be divided into two key time points: the *interruption lag*, and the *resumption lag*. The interruption lag is the time gap between being alerted about an upcoming interrupting task and beginning the interrupting task. Increasing the interruption lag length (i.e., delaying the time before switching to the interruption task) has been found to improve the speed at which the primary task can be resumed (for review, see C. D. Wickens, Hollands, Banbury, & Parasuraman, 2015, p. 321). The resumption lag is the time interval between the end of the interruption and resuming the primary task. Resumption lag is typically taken as a dependent variable assumed to indicate the disruptiveness of an interruption – with a longer resumption times indicating greater disruption (Trafton & Monk, 2007). The length of the time that the primary task is suspended for (in many cases, a function of interruption length) is the retention interval. It should be noted that retention interval is not included in the Trafton et al. (2003)
interruption anatomy, however it is a commonly used term throughout the retrospective and PM literatures (Baddeley, 2012; Hicks, Marsh, & Russell, 2000; Tierney, Bucks, Weinborn, Hodgson, & Woods, 2016), and is a core focus of Chapter 4 of this thesis.

Figure 1.2. The Interruption Anatomy

Importantly, each of the unique temporal stages (e.g., interruption lag; resumption lag) in the interruption anatomy require different components of the cognitive system. For instance, encoding (the process of converting stimuli into constructs for memory storage) is important during the interruption lag; maintenance (the process of updating or maintaining stored memory information) is important during the retention interval; and memory retrieval (the process of re-accessing stored memory information) is important when resuming the primary task (Cane, Cauchard, & Weger, 2012; Monk et al., 2008; Trafton et al., 2003). Thus, experimental studies have typically focused on manipulations around these key stages in order to better understand the cognitive processes involved with performance decrements. Further, attempts to ameliorate performance decrements from interruptions through technological interventions have also concentrated on these stages.

To illustrate how the interruption anatomy can be applied in practice, I will describe a situation involving a pilot performing a pre-start (i.e., pre-flight) checklist (G. Wilson, 2018), with the components of the interruption anatomy indicated in parentheses. A pilot is completing a pre-start checklist (the primary task), and then suddenly, the flight deck communication system produces a tone indicating an incoming radio call (warning for an interrupting task). The time gap between the alert for the incoming radio call (warning) and accepting the communication request (starting interruption) would be the interruption lag. The interrupting task would be the proceeding communication – and may take some time to complete. After disengaging communications (ending the interrupting task), the pilot would
Executing Deferred Tasks in Dynamic Multitasking Environments

need to resume the primary task of completing the checklist. The pilot would need to try and remember the step in the checklist they were up to, or may need to reconstruct their progress on the task from contextual information (e.g., the obvious presence of an incomplete checklist)\(^1\). The time taken to resume completion of the checklist would be the resumption lag. The unlikely situation in which the pilot failed to resume the checklist entirely would be classified as a resumption failure, while failing to resume the checklist at the correct re-entry point would be classified as a resumption error.

It is important to note that many real-world situations involving interruptions do not fit into clearly defined boundaries of the interruption anatomy. For instance, some situations may involve several tasks being suspended, and subsequent task resumption may not occur consecutively in the order in which they were suspended (see, Sasangohar et al., 2017). Additionally, in some situations — such when a particularly long interruption occurs — new task demands or situations may arise during the interruption and require immediate attention when returning from the interruption, further deferring the completion of incomplete goals (Hodgetts et al., 2015). Nevertheless, the interruption anatomy has provided a strong set of core terms and assumptions for describing and understanding interruption situations. In the following sections, I will outline the history of empirical interruptions research, and then introduce the two dominant theoretical frameworks for predicting and explaining the impact of interruptions on task performance. I will reference the empirical studies that support the predictions derived from these theoretical frameworks throughout.

1.5. Empirical Research on the Effects of Interruptions

Experimental psychologists have been identifying the outcomes and consequences of interruptions on human cognitive processing for over a century. Throughout the 1920s and 1930s, researchers in the Vygotsky Circle conducted a collection of experiments aimed at identifying the factors which caused individuals to forget to return to unfinished tasks (Zavershneva & van der Veer, 2018). The core finding from the research was that individuals tended to have superior memory recall for the features of incomplete tasks relative to the

\(^1\) As previously mentioned, it is now standard in all modern airlines for pilots to restart the checklist if a significant interruption has occurred during the checklist.
Chapter 1: General Introduction

features of completed tasks, a phenomenon referred to as the Zeigarnik effect (Zeigarnik, 1927). It was theorised that suspending an unfinished task to perform other activities places the suspended task in a state of 'heightened activation', resulting in individuals having superior memory recall for information pertaining to the suspended task. However, subsequent research has largely failed to replicate the Zeigarnik effect (for comprehensive review consult, Bergen, 1968; Patalano & Seifert, 1994). Further, these studies did not require participants to actually resume the suspended task, and thus did not assess how interruptions impact memory performance.

The first research to examine the direct effects of interruptions on measures of human performance was conducted by Kreifeldt and McCarthy (1981) who sought to examine whether different mathematical logic systems (e.g., reverse Polish notation) would show differing levels of resistance to the effects of interruptions. Participants were required to solve four problems using one of two different calculators and were sometimes interrupted for 1 minute with a multiplication task. Interrupted participants were instructed to resume the initial problems after the interruption. The key contribution from this study was the use of six quantitative measures for assessing the impact of interruptions on performance. The two most important of which were: (1) time to press first calculator key after returning from the interruption, and (2) number of incorrect solutions following resumption. It should be noted here that the former measure corresponds to the now standard measure of interruption disruption, resumption time, and the latter measure corresponds to resumption errors. Although Kreifeldt and McCarthy’s study introduced two key dependent variables used throughout the literature to-date, the research was concerned with comparing different designs of mathematical logic systems, not examining the cognitive effects of interruptions themselves. Gillie and Broadbent (1989) made the first attempt to systematically measure the impacts of interruption length, task complexity, and task similarity on cognitive performance in a human-computer interaction task. In their study, participants were required to navigate through a virtual town and collect a specified collection of shopping items. Over four experiments, participants were interrupted by several tasks: mental

---

2 For curious readers, the two compared calculators were traditional algebraic notation and reverse Polish notation. The reverse Polish notation calculated proved to be significantly more robust to the effects of interruptions.
Executing Deferred Tasks in Dynamic Multitasking Environments

arithmetic, memorizing sets of words, and complex arithmetic with alphanumeric representations of numbers. The key finding of the study was that increased complexity of the interrupting task, and greater similarity to the primary task increased disruption to the primary task upon resumption.

Thus, both Kreifeldt and McCarthy (1981) and Gillie and Broadbent (1989) concluded that there were some performance costs associated with performing task after an interruption, relative to performance before the interruption. Both studies also found that people made more errors in the post-interruption time periods. However, while these studies were useful for characterizing the general effects of interruptions on a variety of tasks and provided insights regarding the possible task conditions which may increase the disruptiveness of interruptions; they were not connected by a more general theoretical approach of how individuals cognitively coordinate task goals, information and intentions under interruption situations (Monk et al., 2008). In response, the past 20 years has seen the development of several theoretical models which provide descriptive cognitive accounts of how interruptions impact task performance, which I will review below.

1.6. Theoretical Models of Interrupted Task Performance

1.6.1. Memory for Goals

Memory for Goals (MFG) is an activation-based account of ‘goal-directed’ cognition that is instantiated in the adaptive control of thought – rationale (ACT-R) framework (Anderson & Lebiere, 1998) and was developed by Altmann and Trafton (2002). Broadly speaking, MFG aims to provide a comprehensive account of the cognitive processes involved with the deferral and resumption of task goals. Primarily, it is used for explaining and predicting the effects of interruptions on task performance. Similar to other activation-based models of memory, the central assumption of the model is that task goals are stored in long-term memory within associative networks, and that when the cognitive system queries memory, the goal with the highest level of memory activation will be retrieved and thus direct behaviour (Altmann & Trafton, 2002; Anderson & Lebiere, 1998; Nowinski & Dismukes, 2005). For the correct goal to be successfully retrieved, its activation must be greater than the interference threshold which can be thought of as residual mental clutter from old goals and is formally specified as the activation level of most active non-target goal (i.e., distractor). Task goals are associated with a base-level
Chapter 1: General Introduction

 activation that represent its activation without any associations or cues. Base-level activation decays over time, reducing total goal activation (Altmann & Trafton, 2002; Monk et al., 2008). Specifically, base-level goal activation is determined by the equation below which is adapted from the base-level learning equation in ACT-R (equation 1).

\[
m = \ln \left( \frac{n}{\sqrt{T}} \right)
\]  

(1)

In the equation, \(m\) refers to activation of a goal. \(n\) is how often the goal has been sampled in its history (i.e., retrieved from memory), and \(T\) is the length of the goal’s history, from encoding to present. In effect, this equation defines activation as a function of retrieval frequency and temporal decay.

MFG specifies two constraints which are used to overcome the effects of decay: the strengthening constraint which is driven by a goal’s history (represented by \(n\) in Equation 1); and the priming constraint which is based on a goal’s contextual associations. The strengthening constraint specifies that goal activation can be increased through successive goal retrieval, which is typically just conceptualised as rehearsal. Rehearsal can either be retrospective rehearsal (e.g., “what was I just doing”) or prospective rehearsal (e.g., “what am I about to do”). The more frequently a goal is rehearsed (or retrieved) in memory, the greater its activation (Trafton et al., 2003), thereby decreasing the decay of the goal and increasing the probability of retrieval at the appropriate time. The priming constraint specifies that activation can also increase when attention is directed towards memory items or contexts associated with the task goal (e.g., the obvious presence of a half-completed checklist) via a mechanism referred to as associative-activation (Altmann & Trafton, 2002; Hodgetts & Jones, 2006a).

1.6.1.1. Empirical Support of Memory for Goals

The MFG theory enables several predictions to be made about what variables influence interruption recovery. Firstly, according to MFG theory, in situations where rehearsal is blocked, longer interruption durations should be more disruptive because this will increase the amount of time-based decay (Altmann & Trafton, 2002). To test this assumption, Monk, Trafton, and Boehm-Davis (2008) conducted a series of experiments in which the resumption lag on a primary task was examined after interruptions of several lengths. They found the longer interruption durations were more disruptive to resumption time. Importantly, they found that
Executing Deferred Tasks in Dynamic Multitasking Environments

on average, the relationship between resumption time and interruption length could be described with a log function, consistent with MFGs activation decay function. In descriptive terms, the model specified that the rate of resumption times rose rapidly over short interruption durations (3 – 13 s) before approaching asymptotic stabilisation over the subsequent 45 s. Similar effects of interruption length have been reported by other researchers (Hodgetts & Jones, 2006b).

A second core prediction of MFG is that more difficult or cognitively demanding interruptions are more disruptive because they prevent rehearsal of goals through dual task interference (Altmann & Trafton, 2002). In accordance with MFG, longer and/or more difficult interruptions have been shown to increase the disruptiveness of an interruption (as measured by resumption lag) in a variety of static (i.e., non-dynamic) experimental paradigms (Cades, Davis, Trafton, & Monk, 2007; Hodgetts & Jones, 2006b; Monk, Trafton, & Boehm-Davis, 2008; Trafton & Monk, 2007).

Finally, MFG stipulates that attending to relevant environmental cues can facilitate task resumption through associative activation (i.e., priming the suspended goals). This account would suggest that resumption processes may be aided by simply forming associations between task relevant features (e.g., a particular aircraft) and features of the environmental context (e.g., spatial location), and this may occur in a relative automatic fashion. Several studies have provided support for this assumption. For instance, Hodgetts and Jones (2006a) examined the effects of preparation before an interruption in a Tower of London task. In their paradigm, participants received no warning of impending interruptions, but the interruption lag was manipulated (in one condition) by forcing a 2 s pause prior to starting the interruption. The authors removed any warnings to prevent forcibly eliciting interruption management strategies that participants might not otherwise use in a non-experimental setting. The first study revealed suspended goal retrieval was facilitated when participants were given an opportunity to encode retrieval cues during the interruption lag. A second study was conducted in which the visual display (i.e., task status) was modified during the interruption period, and this was found to impede post-interruption goal retrieval (by the way of task completion time). This finding suggested the benefit to the encoding pause could be attributed to encoding specific contextual cues from the task state. Their final study found that there was no benefit of a brief pause if the retrieval cues changed. Thus, post-interruption goal retrieval is facilitated by encoding specific
cue-action associations prior to the interruption, and impeded when these task-related cues are modified at retrieval. As will be discussed further in section 1.7.2, this finding has important implications for dynamic monitoring tasks (such as ATC) where the environment (and containing cue-action associations) are likely to change during an interruption. In addition, forming contextual associations and increasing the salience of associative cues has been robustly demonstrated to facilitate interruption resumption performance in a range of applied and experimental paradigms (Grundgeiger et al., 2010, 2010; Ratwani & Trafton, 2008; Werner, Foroughi, Baldwin, Youmans, & Boehm-Davis, 2018).

1.6.1.2. Conclusions on Memory for Goals

In summary, MFG is a well-supported and influential theoretical framework that can be applied to work settings in order to predict and understand the effects of interruptions. However, MFG does have a number of limitations. For instance, according to MFG, only one task goal can be active at any given moment (Altmann & Trafton, 2002). While this assumption is convenient for the purposes of modelling within the ACT-R framework, there are many situations where individuals are able to concurrently coordinate multiple task goals to achieve higher order objectives (Neal, Ballard, & Vancouver, 2017; Salvucci & Taatgen, 2008). A further related issue is that MFG does not clearly specify what actually constitutes a ‘single task goal’ or why multiple task goals cannot be concurrently active if the goals do not overlap in cognitive modalities. In the following section, I will introduce a more general theory of multitasking behaviour, Threaded Cognition, which was in part developed to overcome these limitations.

1.6.2. Threaded Cognition

Threaded Cognition is an activation-based model also grounded in the ACT-R framework. Compared to MFG, Threaded Cognition is a general model of human performance that was developed to account for interruption, multitasking, and task switching situations (Salvucci & Taatgen, 2010, 2011a). Salvucci, Taatgen, & Borst (2009) argue that interruptions can be operationalised as a specific form of multitasking called “sequential multitasking”. They argue that all multitasking behaviour exists on a multitasking continuum (see Figure 1.3, below), with
Executing Deferred Tasks in Dynamic Multitasking Environments

Concurrent multitasking (performing two tasks at the same time) on one end, and sequential multitasking (performing two or more tasks interleaved over a longer duration) on the other. The theoretical implication of this is that interrupted task performance can be explained by the same mechanisms as other multitasking behaviours, which is not the case for MFG. Threaded Cognition explicitly accounts for interruptions by extending the MFG framework with two key features: the concept of *threads* (instead of goals); and the *problem state* (a single mental resource that represents the status of a task).

Figure 1.3. The Multitasking Continuum. Concurrent multitasking on the left, through to sequential multitasking on the right. Adapted from Salvucci and Taatgen (2011a).

While MFG and Threaded Cognition share many similar assumptions, there are several important differences between them. Threaded Cognition does not adopt the concept of goal activation, and instead refers to the ways in which multiple tasks can be threaded. According to the Threaded Cognition model, streams of thought (e.g., task goals) can be represented as individual ‘threads’ of processing which are coordinated by a serial *procedural resource* and are executed across available *resources modalities* (e.g., verbal, aural, manual, vocal, or memory) to produce behaviour. Thus, several threads can operate in parallel, but only one thread can use the procedural resource, and threads are constrained by the availability of resources. Simple tasks, such as memorization of a digit span are considered to consume a single thread, whilst complex tasks, such as maritime navigation (Hutchins, 1995), are considered to be a group of threads working in parallel to achieve a higher level goal.

---

3 Concurrent multitasking at the most fine-grained level (i.e., performing two task simultaneously) is generally not a feasible strategy for handling multiple task threads in real-world workplaces, due to the high resource overlapping resource demands (C. D. Wickens, 2002).
Chapter 1: General Introduction

Unlike base ACT-R (Anderson & Lebiere, 1998) and MFG (Altmann & Trafton, 2002), which assume that only one task goal can be active at a time (hence interruptions require the suspension of task goals), the Threaded Cognition architecture allows multiple goals to be concurrently active. Concurrent activation of threads is made possible because rather than specifying that only one task goal can be active, the model assumes that individuals maintain a flexible but easy to maintain set of currently active goals. Of course, it is unlikely that individuals could execute productions (i.e., perform task actions) for all goals within a set concurrently, thus the model makes two assumptions regarding how threads are executed to serve task goals: the resource seriality assumption and the resource usage assumption. The assumption of procedural seriality specifies that the resource processing of threads can operate in parallel (with each other) if the threads do not both require procedural processing at the same time. For example, responding to a simple question and typing a credit card number at the same time would be possible, due to a discrete usage of procedural processes. However, if threads both require procedural processing at the same time, only one thread will proceed (the one with a higher priority). For example, if a task requires an individual to retrieve two separate facts from memory at the same time, each memory retrieval (the threads) must proceed successively, one at a time. In order to ensure that threads release their usage of a resource to allow another thread to continue (i.e., prevent blocking), threads consume resources in a “greedy, polite manner”. Greedy means that threads will acquire and consume all the resources they need; while polite means that threads will free a resource for use by other threads once it does not need it.

The other core component of Threaded Cognition related to interruptions is the problem state — a separate resource which keeps track of all information related to the status of a task that is not available in the external environment (Borst, Taatgen, & van Rijn, 2010; Salvucci & Taatgen, 2011b). For example, an air traffic controller may store the general spatial location of aircraft that requires attention in the problem state while attending to another aircraft in the sector. The problem state construct is closely related to the episodic buffer in Baddeley’s (2000) working memory model that serves the function of a “limited capacity temporary storage system that is capable of integrating information from a variety of sources” (Baddeley, 2000, p. 421). In Threaded Cognition, the problem state is a separate resource that can be activated while engaging with another task (e.g., an interruption). The problem state is a bottleneck because only one problem state may be active at any time (Borst et al., 2010). Thus, any two tasks that have separate task contexts will interfere with each other.
In summary, Threaded Cognition is an alternative account of interrupted task performance that explains: how individuals suspend and resume tasks, what the cognitive bottlenecks are (e.g., problem state) with regards to coordinating concurrently active goals, and the mechanisms by in which multiple task threads can coordinate resource demands. In the following section, I will examine a number of studies that have directly examined the effects of interruptions in complex dynamic task environments. These studies are more applied in nature than those previously reviewed, and I will focus my discussion on some of the key differences in findings, and how dynamic environments may challenge some of the assumptions of MFG and Threaded Cognition with regards to interrupted task performance.

1.7. Interruptions in Complex and Dynamic Tasks

The majority of theoretical models and associated studies that have examined the effects of interruptions have used static sequential primary tasks (e.g., VCR programming, Tower of Hanoi task, trivia questions) that do not change during the interrupting period. However, increasingly, researchers have begun investigating how theories and methodologies from the basic interruption literature can be applied, and generalise, to more complex and dynamic task environments (Sasangohar, Scott, & Cummings, 2014). Broadly speaking, complex dynamic tasks are defined as those in which individuals have to monitor displays that evolve over time (i.e., representing dynamic information; St. John & Smallman, 2008) and where individuals must integrate disparate sources of information to coordinate multiple task objectives (i.e., complexity; Tremblay, Vachon, Lafond, & Kramer, 2012). Prototypical examples of such jobs include unmanned aerial vehicle control, submarine track managers, air traffic controllers, and other military radar operators (these types of tasks are commonly referred to as command and control [C2] tasks). There are a number of obvious practical motivations for studying interruptions in such environments. For one, these task environments are generally safety-critical, placing greater emphasis on identifying the factors underlying performance. Second, studying interruptions in these task environments allows researchers to evaluate the utility of psychological theory, which is often assumed to generalize from basic laboratory tasks to work contexts (Gonzalez, Vanyukov, & Martin, 2005). Finally, such research holds ecological validity: jobs in these work domains require operators to coordinate interruptions frequently due to their heavy dependence on team-based problem solving, information and communication technologies, and multiple task demands (Cooke, Gorman, Duran, & Taylor, 2007).
Chapter 1: General Introduction

following sections, I will review some of the critical findings relating to interruptions in complex and dynamic tasks, and how these findings challenge some of the assumptions from theoretical frameworks based on simple static tasks.

1.7.1. Long-Term Effects of Interruptions

In static task experiments, researchers typically concentrate on the moments immediately following an interruption; for instance, by examining resumption time or post-interruption procedural errors (resuming the interrupted task at the wrong entry point). However, in complex and dynamic task environments, it has been established that interruptions — particularly demanding or long ones — require some time to recover from. This is because recovery from interruption may require several steps and manifest cognitive processes (Loft, Sadler, Braithwaite, & Huf, 2015; Tremblay et al., 2012). There are a number of studies using complex dynamic control tasks have found that interruptions have chronically detrimental effects to post-interruption task performance, well beyond the more immediate few second resumption time cost as measured in more basic tasks.

Tremblay et al. (2012) used a firefighting management task and found that interrupting participants for just 20 s negatively impacted monitoring effectiveness for up to 40 s following the interruption. Furthermore, in contrast with the assumption of MFG that resumption time is an index of interruption recovery quality (Altmann & Trafton, 2002), the results from Tremblay et al. suggest that speed of returning to the primary is not a reliable predictor of quality of recovery. Specifically, despite the differing team size conditions showing similar recovery of contextual information at the one-minute mark, they significantly differed in resumption time. This indicates that in dynamic tasks the recovery process may take time and unfold over a series of multiple steps rather than instantaneously, indicating that resumption time alone is not sufficient to index interruption recovery in dynamic tasks.

Hodgetts et al. (2014) used an air combat control task and found that participants who were interrupted for 24 s were slower to make decisions and exhibited shorter but more frequent eye-fixations up to 40 s post-interruption. This pattern of eye-fixations was argued to reflect

---

4 Monitoring effectiveness was defined as participants’ ability to identify idle vehicles in a timely manner and issue them new instructions.
Executing Deferred Tasks in Dynamic Multitasking Environments

participants rapidly updating information regarding the newly evolved task state (for further discussion of eye-fixations patterns and situation awareness reacquisition consult Gartenberg, McCurry, & Trafton, 2011). Similarly, Loft et al. (2014) found that a 20 s interruption (paper and pencil vessel classification task) condition in a submarine track management simulation resulted in participants (on average) making 16% more vessel heading change detection errors (performance criterion), and were 3.2 s slower detection of such changes, for 93 s after the interruption (on average). Moreover, Loft et al. found that participants took longer to classify a ship based on its history for up to 80 s after the interruption. Collectively, these studies demonstrate that the effects of interruptions in complex dynamic tasks may extend well beyond a more immediate resumption lag delay or a single PM failure.

It is particularly surprising that post-interruption task performance decrements can be so chronic and enduring in complex dynamic environments, as MFG would predict that the task state should be reactivated relatively quickly following the interruption (Trafton & Monk, 2007). One explanation for these findings can be derived from theories of situation awareness (SA) recovery. The most widely accepted definition of situation awareness is Endsley’s (1988) definition, which specifies SA is the “the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future”. According to Endsley’s theoretical framework, situational awareness has three levels: (1) perception of the elements in the environment; (2) a comprehension of the current situation; and (3) a projection of future status. Situation awareness is critical to the safe and efficient operation of many safety critical domains, including ATC, aviation, medicine, industrial control systems, and military operations (Endsley, 1995; Vidulich & Tsang, 2012). Loss of situation awareness has been a key factor underlying numerous disasters. It is plausible that the long decrements to post-interruption task performance indicate that accurately reconstructing or recovering situation awareness in complex tasks may require substantial time and effort. This explanation would be consistent with Threaded Cognition if it is assumed that maintaining situation awareness requires access to the problem state in addition to a range of greedy resource specific threads. I will return to the notion that the problem state could be conceptualised as SA in Chapter 3 and Chapter 5.
Chapter 1: General Introduction

1.7.2. Dynamics: Changing the Task State

Arguably, the most salient difference between dynamic tasks and static tasks is that in dynamic tasks, the objects represented on the display environment evolve and change over the course of the interruption. For static tasks, such as editing a document, the primary task is unlikely to change over an interruption. In static cases, it can be assumed that interruption recovery involves retrieving task goals and intentions from memory or entirely recovering them afresh from contextual information (i.e., a half complete document open on the screen). However, operators in dynamic tasks must additionally regain awareness of a constantly changing situation. Operators therefore may need to assimilate any mismatch between the pre and post interruption display scenes, and this process may require significant cognitive effort itself (Boot, Kramer, Becic, Wiegmann, & Kubose, 2006; St. John & Smallman, 2008; Rensink, O’Regan, & Clark, 2000), introducing a new task demand that needs to be immediately executed upon resumption. Similarly, contextual cues which were encoding during the pre-interruption period may no longer be visible or relevant. Additionally, keep in mind that much like the world it represents, objects on the display are dynamic and changing. Thus, in addition to recovering pre-interruption task goals, operators must also monitor for any changes in the situation that could have occurred during this period, and update goals and plans to respond accordingly (Boot et al., 2006; St. John & Smallman, 2008). All of these factors may also contribute to the longer-term impacts of interruptions in complex dynamic tasks discussed above.

However, simply detecting changes across visual scenes can be a challenging task for humans. This is referred to as change blindness — the difficulty humans have in detecting changes in a visual scene that occur across even brief interruptions (Rensink et al., 2000). When several risk factors align, change blindness poses a real threat to operational effectiveness and safety. For example, Savage-Knepshield and Martin (2005) conducted a usability assessment study on a military global positioning system used for signalling air strike coordinates. In the study, 16 soldiers were instructed to re-enact a scenario in which they were preparing to specify coordinates for an air strike. However, prior to completing the task, the GPS batteries were designed to run out, requiring them to replace the batteries and resume the primary task. Upon resetting the device, the coordinates displayed were reset to the present location (a previously default behaviour of the device), and 38% of participants failed to detect this change and erroneously called in their present position as the airstrike target. Thus, by evaluating a poor GUI design in an experiment motivated by the human susceptibility to not notice unforeseen
visual changes, a superior designed GPS was developed, potentially saving many lives (Savage-Knepshield, 2018).

An intuitively appealing solution to change-blindness is to provide operators with ‘event reviews’ or ‘instant replays’ that summarize the critical changes and events that occurred during the interruption period (Sasangohar et al., 2014). The general idea is that upon returning from an interruption, operators could be trained to quickly interpret information from a decision support tool resulting in improved situation awareness reacquisition compared to unaided performance. However, several studies have failed to find benefits of instant replay or summary features (St. John, Smallman, & Manes, 2005; Scott, Mercier, Cummings, & Wang, 2006); and in fact found paradoxically that interruption recovery aids can have negative effects on interruption recovery (Hodgetts et al., 2015; John, Smallman, & Manes, 2005; Scott et al., 2006). While the studies contained within this thesis do not address the development of recovery aids; the findings from Chapter 3 indicate that the additional visual-spatial demands posed by interruption recovery aids may be one factor that inhibits their utility.

1.8. Summary of Interruptions

One key motivation for this thesis is to further explore the nature of interruptions in complex and dynamic task environments. The studies reviewed above indicate that the time course of recovering post-interruption performance in complex tasks can be substantial. While these studies do not provide a clear account of the exact mechanisms underlying such drawn out recovery windows, they do provide a strong motivation for examining performance on regularly measured dependent variables for extended amounts of post-interruption time. Importantly however, such longer-term measurements are not mutually exclusive with short-term measures, such as resumption time. Thus, in Chapter 2, I will introduce an experimental ATC simulation paradigm capable of measuring both shorter-term (i.e., resumption time) and longer-term interruption effects (i.e., other ongoing tasks that persist throughout a longer period of time).

The interruptions literature reviewed also provides a strong foundation to generate predictions regarding the effects of interruptions (i.e., MFG and Threaded Cognition) under various contexts, and these theories will inform my selection of dependent variables (i.e., resumption time, resumption failures), and methods (e.g., mental demand of the interrupting task) throughout the remaining chapters. However, the primary goal of this thesis is to better
understand the effects of interruptions to deferred tasks in simulated air traffic control. A key issue not addressed by the interruption literature are the cognitive mechanisms by in which individuals remember to perform deferred tasks, and how interruptions might interfere with these processes. In the following section, I will review the PM literature, which concerns the cognitive mechanism that underlie deferred task completion.

1.9. Prospective Memory

PM is the human ability to remember to perform an intended action or task at an appropriate point in the future (McDaniel, Einstein, & Rendell, 2008). For instance, remembering to post a letter on the way home from work requires PM. Throughout this thesis, I use the terms deferred task and PM task to refer to the situations that require individuals to engage PM; and I refer to failures in carrying out intended actions as PM errors. PM plays an integral role in our everyday functioning and wellbeing (e.g., remembering to take medication at a certain time), and atypical PM functioning is present in a range of clinical populations (Burgess, Gonen-Yaacovi, & Volle, 2011; McDaniel et al., 2008). PM tasks also occur in many safety-critical contexts, including: medicine (Grundgeiger, Liu, Sanderson, Jenkins, & Leane, 2008; Grundgeiger et al., 2013, 2010), driving (Bowden, Visser, & Loft, 2017), aviation (Dismukes & Nowinski, 2006; Dismukes, 2012), and ATC (Loft, 2014; Loft, Smith, & Bhaskara, 2011; Shorrock, 2005). In such workplaces, PM errors can have serious safety implications. The Los Angeles runway disaster is a notable example of the possible safety implications of PM errors. On February 1, 1991, a tower controller at Los Angeles International Airport cleared an aircraft to taxi to Runway 24L and hold position (a routine procedure) and intended to instruct the aircraft to take-off once another aircraft cleared the runway. The controller then was distracted by several atypical events, including an aircraft accidently switching off ATC communication frequency and a misfiled flight progress strip. Tragically, the controller forgot that they had not cleared the holding aircraft, and inadvertently instructed another aircraft to land on runway 24L, resulting in a collision killing 23 individuals (National Transportation Safety Board, 1991).

In order to predict and prevent PM errors from occurring, cognitive and human factors psychologists have conducted research into the cognitive mechanisms which underlie PM. Early research typically concentrated on naturalistic studies (e.g., remembering to post a letter to the experimenter after some interval). These studies indicated PM may be predicted by personality
Executing Deferred Tasks in Dynamic Multitasking Environments

or extent of motivational incentives (Meacham, 1982; Meacham & Singer, 1977). However, the uncontrolled nature of these studies limited their insight into the cognitive mechanisms and situational factors underlying PM retrieval and maintenance. More recently, PM has been investigated under controlled laboratory conditions which allow inference of the underlying mechanisms. In this section, I will provide an overview of the categories of PM tasks, followed by an overview of how the forms of PM task relevant to the studies in this thesis are operationalised in experimental settings, and finally discuss the extant theories of PM.

1.9.1. PM in the Laboratory and Underlying Cognitive Mechanisms

In laboratory contexts, the majority of experimental research uses variants of Einstein and McDaniel’s (1990) event-based retrieve-execute paradigm in which participants are busily engaged in a primary ongoing task (e.g., lexical decision making) and are required to remember to execute a special action (e.g., press ‘9’ key) when they encounter a cue matching particular criteria (e.g., a word that contains the syllable ‘tor’). In such event-based PM tasks, the deferred intention is encoded prior to starting the ongoing task, and the PM response must be retrieved and then executed upon the presence of the PM cue (hence why it is referred to as retrieve-execute). Much like event-based PM, in time-based PM, the intention is also encoded prior to beginning the ongoing task. However, instead of the required PM action being associated with a particular event-based cue (e.g., word syllable or situational context), the PM response is required to be performed at particular time intervals (e.g., every 4 mins). In delayed-execute paradigms, a salient event-based cue (e.g., screen turning red) is associated with an intended action (e.g., perform a special keystroke), and critically, the cue is presented in the midst of performing an ongoing task, but the intended action can only be performed after a brief delay (e.g., after completing current ongoing task requirements).

The PM literature typically distinguishes between two forms of PM: event-based, in which an intention is to be performed when a particular situation occurs (e.g., remembering to update a fellow controller on the status of a flight when returning from break); and time-based, in which an intention must be performed at a particular time (e.g., remembering to issue a request to a flight in 5 minutes). However, there are many other situations that require individuals to remember to perform an intended action (i.e., a deferred task). For instance, when interrupted, individuals have to remember to return to, and complete the intended
primary task goal. In fact, deferred tasks in applied settings can even have features of multiple forms of PM task, for instance involving both time-based, event-based PM, as well as interruptions. Despite the diverse situations that require individuals to remember to perform intended actions, they all involve three broad stages of memory processing: encoding, retention, and retrieval.

1.9.1.1. Encoding

The encoding phase of PM can involve different strategies than in retrospective memory encoding. For instance, individuals may encode information pertaining to the construction of plans for when and how the intended deferred task will be completed (McBride & Workman, 2017). Additionally, unlike retrospective memory, encoding for PM is generally intentional (i.e., the individual forms the specific intention to complete the deferred task action). Indeed, both laboratory and applied studies have shown that PM is bolstered by the providing individuals information at encoding regarding the probable future context of PM retrieval (Bowden, Smith, & Loft, 2017; Cook, Meeks, Clark-Foos, Merritt, & Marsh, 2014; Kuhlmann & Rummel, 2014). In light of this information, clinical and social psychologists developed the concept and practice of *implementation intentions*, the process of mentally simulating and visualising the specific context that the deferred task is likely to be executed in, and then rehearsing the association between those cues and the intended action (Gollwitzer, 1999). Instructing individuals to form implementation intentions has been shown to improve PM in a number of laboratory and applied contexts (Chen et al., 2015; A. L. Cohen & Gollwitzer, 2008; McDaniel, Howard, & Butler, 2008; Rummel, Einstein, & Rampey, 2012).

1.9.1.2. Retention

The second phase of PM, retention, concerns how individuals maintain the intended deferred task or cue-task associations in memory throughout the interval between encoding and retrieval. The length of time between encoding new information and the moment in which that information must be retrieved is known as a *retention interval*, which has shown to be a crucial moderator of human retrospective memory (Wixted & Ebbesen, 1991, 1997). Studies in the working memory domain have shown filled retention intervals of a few seconds can lead to substantial rates of forgetting (Lewandowsky, Geiger, & Oberauer, 2008). Two common explanations for the negative effect of extended retention intervals in retrospective memory is
that memory items are subject to time-based decay that can be prevented by continually refreshed the memory trace via rehearsal; or that the events occurring during the retention interval interfere with the stored representations of items in memory (for review consult, Oberauer, Farrell, Jarrold, & Lewandowsky, 2016).

However, research examining the relationship between retention interval and PM is considerably more varied in findings and methodologies. Laboratory studies using retrieve-execute paradigms have found the effects of retention interval vary depending on whether the interval increased as function of the ‘filler task delay’ or ‘ongoing task delay’ (Martin, Brown, & Hicks, 2011; Zhang, Tang, & Liu, 2017). The filler task delay is the length of time between the intention formation and the beginning of the ongoing task; while the ongoing task delay is the length of time between beginning the ongoing task and the first PM stimulus presentation in that ongoing task. Increasing filler task delay has been shown to improve PM (Hicks et al., 2000; Martin et al., 2011) or have no effect (Zhang et al., 2017); while increasing the ongoing task delay has been found to negative impact PM performance (Martin et al., 2011; Zhang et al., 2017). This decrease of performance associated with longer ongoing task delays is assumed to result from extended time-on-take reducing individual’s capability to maintain PM intentions and monitor for PM cues (i.e., reduction in task-focus).

A key challenge in understanding the effects of the retention interval in both retrospective memory and PM is that there is a high correlation between the passage of time, the frequency of events and the duration of cognitive processing (cf., Norris, Hall, Butterfield, & Page, 2018). Indeed, essentially all situations involving PM are likely to involve attention being redirected to other tasks during the retention interval (Dismukes, 2010). Thus, a key challenge of retrieve-execute paradigms is that it can be unclear whether PM failures result from cognitive processing at the retrieval stage (i.e., the PM cue fails to trigger PM retrieval), or due to events occurring during the retention interval. In light of this, Einstein et al. (2003) developed a paradigm to examine factors during a brief delay (or PM retention interval) that could impact PM, known as the delayed-execute paradigm.

In delayed-execute paradigms, individuals are presented with highly salient PM cues that signal an intended action must be performed, but the execution of the intended action is delayed (postponed) until some future point (McDaniel, Einstein, Graham, & Rall, 2004). An everyday example of what this paradigm captures is reflected in the everyday in the everyday office
Chapter 1: General Introduction

situation in which you are told to relay information to an office colleague (encode an intention), but your action is delayed as they talking on the phone, requiring you to wait until they finish the phone call to execute the action. There are several important features of delayed-execute paradigms which distinguish it from the typical retrieve-execute paradigms used in event-based PM research (i.e., retrieve the intention upon cue; then immediately execute the intended action). First, unlike the typical event-based PM paradigm where there is ample time to encode the PM cue and associated action (usually in experimental instructions), in delayed-execute PM tasks the intention is encoded on the fly, in the midst of performing another ongoing task. Additionally, the delay between encoding the intended action the appropriate moment for action execution is typically quite short in delayed-execute paradigms (5 – 40 s in previous research). This enables researchers to make a clear temporal distinction between intention retrieval/formation, which occurs when the cue is encountered, and intention execution, which occurs when the window of opportunity for action arises. Second, the use of highly salient cues ensures that the majority of individuals will notice the target event and retrieve the required action (Cook et al., 2014), whereas in retrieve–execute paradigms PM cues can often be missed. Thus, the features of the delayed-execute paradigm can result in an increased emphasis on intention maintenance processes, rather than cue detection processes (Ball, Knight, Dewitt, & Brewer, 2013; Cook et al., 2014; McDaniel et al., 2004). However, contextual cueing has still be shown to be an important mechanism underlying PM retrieval in delayed-execute paradigms (Schaper & Grundgeiger, 2018).

The most common application of the delayed-execute paradigm has been to study the effects of interruptions on PM. Specifically, during the retention interval, between cue presentation and moment of retrieval, interruptions can be administered and thereby, the effects of interruptions on PM performance can be examined. The majority of research using this paradigm required participants to answer a series of general knowledge questions, and whenever they detect a red screen between questions, they should press a designated key on the keyboard, but importantly not until after they finished the block of questions they were currently working on. Such research has robustly demonstrated that such interruptions disrupt deferred intentions in various contexts (Cook et al., 2014; Einstein, McDaniel, Manzi, Cochran, & Baker, 2000; Einstein et al., 2003; McDaniel et al., 2004). However, reinstating the task context after an interruption, and providing information about the future context for delayed execute responses alleviates the negative effects of the interruption (Cook et al., 2014). A small number of delayed-
execute studies have also explored how different manipulations of the retention interval length effect performance. In these studies, retention interval is referred to as the 'delay length' (i.e., length between cue and moment for action execution). These studies have found that an initial delay of 5 or 10 s substantially impairs PM performance (Einstein et al., 2000), but that longer delays (15 to 40 s) do not further exacerbate errors (Einstein et al., 2003; McDaniel et al., 2004), nor does the length of interruption (McDaniel et al., 2004). Einstein et al. (2003) suggested that the null effect of delay length was likely due to participants using spare capacity to periodically reactivate intentions by attending to relevant contextual cues or engaging strategic maintenance processes.

In summary, the effect of retention interval on deferred task performance is dependent on the extent to which individuals have the capacity to maintain the intended action in focal attention in face of ongoing task demands. Further, although maintenance processes are crucial for successful PM, they are not well understood and many formal models of PM have not incorporated maintenance processes (Strickland, Loft, Remington, & Heathcote, 2018). In Chapter 4, I further explore the role of retention interval length, and the effect of interruptions during that interval to better understand the role of maintenance in deferred task execution.

1.9.1.3. Retrieval
There are two important components of PM retrieval: the prospective component (i.e., remember that you have to do something), and the retrospective component (i.e., the thing that is to be remembered). This distinction is made clear by the occasional situation in which you may find yourself remembering that you had to do something, but you cannot recall what it was you were required to do. It is known that increasing the retrospective demands of the task can decrease probability of PM retrieval, arguably owing to the increased demands on a limited attentional capacity system (Ballhausen, Schnitzspahn, Horn, & Kliegel, 2017). PM retrieval is heavily influenced by the relationship between the specific intention and the target cue (that signals the intended deferred task is to be executed). Cues that are strongly associated with the target action (e.g., via contextual associations) are substantially more likely to trigger PM retrieval (Cook et al., 2014; Loft & Yeo, 2007; McDaniel et al., 2004). Similarly, cues that are distinctive, unique or perceptually salient are also more likely to elicit PM retrieval (Albiński, Kliegel, & Gurynowicz, 2016; Brandimonte & Passolunghi, 1994; McDaniel & Einstein, 2000). As mentioned previously, PM retrieval can also be supported by the prevailing ongoing task
Chapter 1: General Introduction

context, and by associating (at encoding) the probable context of the PM retrieval with the PM

cue (Bowden, Smith, et al., 2017; Cook et al., 2014; Kuhlmann & Rummel, 2014). Finally,
retrieval can also be supported when the processing of PM target features matches the typical
ongoing task processing. When PM targets can be detected with typical ongoing task processing,
the PM task is referred to as focal. For instance, detecting a single PM target word (e.g., press ‘9’
if you see the word ‘tortoise’) is classified as focal to an ongoing lexical decision task, because
identifying a single target word is argued to require similar processing to lexical decision making
(Einstein & McDaniel, 2005; Scullin, McDaniel, Shelton, & Lee, 2010). Conversely, identifying a
substring within a word (e.g., press ‘9’ if the word contains the syllable ‘tor’) is classified as non-
focal to an ongoing lexical decision task, because individually identifying substrings in a larger
letter string is arguably not necessary for lexical decision making. The broader theoretical
explanations for these effects are reviewed in more detail in section 1.9.2, below.

1.9.2. Theoretical Perspectives of Prospective Memory

A great deal of PM theory is based on the presence or absence of the PM cost effect (Smith,
2010; Einstein & McDaniel, 2010) – the finding that correct response times (RT) to the ongoing
task are slowed when individuals need remember to perform a PM task, relative to when no PM
task must be performed. However, there exist broader theoretical frameworks that are not
centred explicitly on PM cost effects, and can provide more general accounts of the variables
involved with PM maintenance and retrieval, and account for variations in human performance.
I will review the two most prominent of such frameworks below: The Dynamic Multiprocess
View of PM and the Associative Activation Theory of PM.

1.9.2.1. The Dynamic Multiprocess View

The Dynamic Multiprocess View (DMPV) is a theoretical framework for understanding PM
that can be applied dynamic multi-tasking contexts (Scullin, McDaniel, & Shelton, 2013). Its
central tenet is that PM is supported by a dynamic interplay between top-down and bottom-up
processes (Shelton & Scullin, 2017). This stands in contrast to isolationist dual-process accounts
of cognition that specify top-down and bottom-up processes operate independently (Wixted,
2007). Top-down processing involves deliberately maintaining an intended PM action in focal
attention or strategic monitoring the environment for PM cues. For instance, should an
individual need to tell a colleague some information, the individual may actively rehearse the
PM intention intermittently (i.e., subvocal repetition of the episodic task action), or checking to see whether each person entering the office is the target colleague (i.e., strategically inspecting environmental features for PM cues). Allocating attention to top-down processing can impede performance on other ongoing tasks (referred to as “PM costs”), which is typically interpreted as evidence that PM maintenance processes consume limited cognitive resources (Einstein & McDaniel, 2005; Smith, 2003). Historically, PM costs were presumed to occur due capacity being devoted to PM cue detection and thus decreased capacity to perform the ongoing task (Einstein & McDaniel, 2005; Nowinski & Dismukes, 2005; Smith, 2003); while recent work applying evidence accumulation models to PM tasks suggests that PM costs arise from individuals strategically delaying ongoing task responses to allow a greater opportunity for the PM response to reach decision threshold (Strickland et al., 2018; Strickland, Heathcote, Remington, & Loft, 2017).

However, PM can also be supported via bottom-up, cue-driven processes (Einstein & McDaniel, 2005) – often referred to as spontaneous retrieval. Specifically, events and environments can become associated with PM intentions, and allocating attention to such contextual cues can prompt retrieval of the intended PM action (Ball, Brewer, Loft, & Bowden, 2015; Bowden, Smith, et al., 2017; Cook et al., 2014). Thus, bottom-up processes are mechanistically driven by the strength of the cue-intention association formed during PM encoding.

The hallmark feature of the DMPV is it specifies that bottom-up and top-down processes function synergistically. If during encoding, a particular PM intention is strongly associated with a particular context (e.g., the PM action will be required in a particular visual-spatial region of an ATC display), then this task context can subsequently trigger the engagement of strategic PM monitoring processes (Scullin et al., 2013; see also, Smith, Hunt, & Murray, 2017). Thus, in an ATC context, controllers may associate a PM action with a particular task context which can remind an operator to begin strategically monitor for PM events. The provision of strong contextual cues at encoding can minimize PM costs associated with top-down processing, and encourage more strategic deployment of monitoring or spontaneous retrieval (Kuhlmann & Rummel, 2014; Lourenço, Hill, & Maylor, 2015; Marsh, Cook, & Hicks, 2006; Shelton & Scullin, 2017; Smith, Hunt, & Murray, 2017).
Chapter 1: General Introduction

1.9.2.2. Associative Activation Models

The Associative Activation Theory of PM (Nowinski & Dismukes, 2005) is an activation-based model of the psychological processes underlying the maintenance and retrieval of deferred task actions and was developed with applied multi-tasking contexts in mind. Like other activation-based theories, such as Memory for Goals Theory (Altmann & Trafton, 2002), and Cowan’s (1995) Attention and Memory Framework; it assumes that humans have two separate information stores: focal attention and long-term memory. Accordingly, after forming a deferred intention individual shift attention to other competing task demands, and the intention moves from focal attention to long-term memory. Activation models assume that probability of retrieving an intention from long-term memory is dependent on its level of activation. The activation of the deferred task goal is determined by the sum of activation from two sources. The first, base-level (or baseline) activation, represent its activation without cues and is subject to decay over time, but this decay can be offset through periodically rehearsing and retrieving the deferred task intention. The second source, source activation, is determined by the attention directed towards relevant cues at the correct time for retrieving the deferred task, and its strength is proportional to the association between the cue and task goal. The task goal that is retrieved from memory at a given moment is the one with the highest total activation (base-level activation + source activation), and it is assumed that this goal directs behaviour. It should be noted that this theory closely resembles the activation-based MFG theory introduced previously, so I will not explore this theory further.

1.10. Prospective Memory in Air Traffic Control

Several studies have applied theories and methodologies from the basic PM literature to simulations of ATC. Before reviewing these studies, I will briefly discuss how the complex and dynamic nature of ATC can change the PM task requirements, relative to more basic task environments.

The first critical difference in complex simulated environments is that the PM intention can be cued by features in the environment due to the rich display context. In basic PM studies, participants generally must retrieve the PM intention with often limited environmental influence (e.g., a single word being displayed on an otherwise sparse white screen). By contrast, ATC displays are rich – and contain many unique features (e.g., spatial locations, sector
geography, points of interest, intersections) which can become associated with, and in turn support, the PM intention. Indeed, there is emerging evidence that individuals are able to "offload" memories into the external environment, and do so to support memory functioning (Gilbert, 2015; Risko & Gilbert, 2016; Todorov, Kubik, Carelli, Missier, & Mäntylä, 2018). These situated theories of cognition assume that individuals overcome memory limitations by associating and storing complex information within environment features, thereby reducing the overall memory load. Spatial offloading is a particularly relevant theorised mechanism of offloading for ATC contexts (Loft, Finnerty, & Remington, 2011; Todorov et al., 2018). In ATC situations, an aircraft feature, such as its spatial location on the display or its position relative to known points of interests may act as a cue to help an operator remember a specific action to be required. Unlike in basic task paradigms (e.g., lexical decision), in ATC, PM-relevant aircraft are likely to remain on the display, accessible for cueing at any moment. Thus, controllers may iterative ‘cycle’ back to that aircraft to reinstate the intended task. This form of behaviour is what would be predicted by the DMPV: individuals strategically coordinate top-down and bottom-up processing in order to achieve PM performance.

A second important difference concerns cognitive load posed by making complex cognitive judgements within a dynamic and uncertain environmental context. As previously mentioned, there is some contention in the literature as to whether the PM cost effect is indicative of strategic delays in responding to targets, or whether it reflects a more general capacity limitation imposed by the task (for further discussion, see Strickland et al., 2017). Recent work applying evidence accumulation models has revealed that in simple tasks, that a strategic delay (i.e., raising response thresholds) is the most likely explanation for the PM cost effect. However, in complex and dynamic tasks, operators may already be encroaching their cognitive capacity limits (referred to as the “red line”). Indeed, several PM studies in simulated ATC have shown that individuals are less accurate at ongoing tasks, such as conflict detection, under PM conditions (Loft, Chapman, & Smith, 2016; Loft, Smith, & Remington, 2013). These findings are hard to reconcile with a PM cost account based purely on elevated response thresholds. Similarly, Strickland et al. (2019) found that when placed at a red-line, the PM cost effects were best explained by capacity limitations, rather than strategic delays. Thus, complex and dynamic environments can lead to strategic shifts in individuals' capacity to engage cognitive control over PM demands, causing costs to ongoing tasks unrelated to the PM task.
1.10.1. Workload and Retention in ATC
Stone, Dismukes, and Remington (2001) used a dynamic visual-spatial task inspired by a subset of duties performed by ATC in order to examine the effects of retention interval, workload, and phonological rehearsal on PM. This study is the only study that has used a delayed-execute paradigm and manipulated the retention interval in an ATC simulation. In this study, participants were engaged in a simulated ATC task which required them to direct aircraft (represented as visual dots) along ‘routine’ flight paths. Participants completed a series of trials, and three times during each trial, a message would appear that instructed participants to divert one of the aircraft to an alternative waypoint instead of the routine waypoint. If participants failed to divert the aircraft at the appropriate time, the aircraft continued along the routine path (and was treated as a PM error). The retention interval between receiving this instruction and the time that the diversion was required was either 1, 3, or 5 minutes (with each time occurring once per trial). The three retention intervals overlapped, meaning that participants had to remember multiple (maximum three) PM instructions concurrently. Participants were assigned to either a low load (three aircraft) or high load (seven or eight aircraft) condition for the ongoing air traffic management task.

Participants were more likely to forget to deviate from routine under higher workload (i.e., number of aircraft on display). Results also showed that ongoing task performance suffered when a PM intention had to be maintained. These findings are consistent with the multiprocess view (Einstein & McDaniel, 2005) and suggest that variations in the demand of the ongoing tasks can directly affect the availability of attentional resources allocated to the PM task. However, the study found no effect of retention interval on PM performance. This null effect was possibly due to the fact that the PM instructions overlapped, meaning that successive PM instruction acted as a reminder for the other incomplete PM tasks. Stone et al. (2001) conducted a second study to examine if inhibiting sub-vocal rehearsal would negative impact PM performance by forcing participants to shadow an auditory message. PM performance was poorer in the high workload condition, but they found no effect of retention interval. Moreover, performance on the ongoing task was improved when no PM task was required, relative to when a PM task had to be maintained. In terms of the verbal rehearsal, their results showed that participants made significantly more PM errors (failing to divert the plane) under the verbal shadowing condition. However, ongoing task performance was unaffected by the verbal
shadowing. These findings are also consistent with activation-based models of PM (and multitasking more generally) that argue rehearsal processes help maintain PM intentions.

In summary, Stone et al.’s (2001) findings suggest that: secondary tasks and distractions may not be an issue for ongoing/routine task performance but may be particularly problematic for PM tasks, that PM is sensitive to shifts in workload, and that PM may be robust to the effects of retention interval over short durations.

1.10.2. ATC-Lab Research

Loft and colleagues have conducted a number of studies investigating PM in simulated air traffic control (for review, see Loft, 2014). The details of these studies differ, but they all share a common core experimental platform, known as ATC-Lab5 (Fothergill, Loft, & Neal, 2009). In ATC-Lab, participants assume the role of an air traffic controller and are required to detect and resolve aircraft conflicts (i.e., violations of minimum separation standards between aircraft), and to accept and handoff aircraft (i.e., coordinate aircraft entering and exiting their airspace). For the PM task, participants were required to remember to press an alternative response key when accepting a target aircraft that fulfilled certain conditions (e.g., altitude > 41,000 ft, specific callsign).

Several robust findings have emerged. First, a consistent finding across all studies is that individuals often failed to perform the deferred task and perform the routine aircraft acceptance action instead. Additionally, several studies have found that the cognitive operations required to maintain and retrieve the deferred task can impede performance on other concurrent air traffic management tasks, such as monitoring for conflicts (Loft et al., 2016; Loft, Finnerty, et al., 2011; Loft, Pearcy, & Remington, 2011; Loft & Remington, 2010; Loft, Smith, et al., 2011, 2013) — that is, PM cost effects (Smith, 2003). The fact the costs were also present on tasks unrelated to the PM instruction (i.e., conflict detection) is an important demonstration of theories of PM which argue that PM consumes cognitive capacity under high operational demands (C. D. Wickens et al., 2015, p. 211). These findings support the notion that capacity sharing between PM and

5 In Chapter 2, I provide a detailed review of ATC-Lab and the modifications made for the purpose of this thesis; but for now, I focus on the practical and theoretical contributions of these previous ATC-Lab studies.
ongoing tasks are likely driven by the highly demanding nature of the task ATC environment (Strickland et al., 2019).

Several ATC-lab studies have also found that the ATC display context can support PM. Loft, Finnerty, and Remington (2011) found that PM performance could be bolstered, and costs to ongoing air traffic management tasks could be reduced, when participants were informed that the PM target aircraft would occur only for aircraft approaching from one region (upper or lower) of the display. Thus, the contextual information enabled participants to spatially localize attentional resources for the PM task to the most relevant display region; or attending that region prompted PM monitoring. This reduction in monitoring costs associated with providing information regarding PM context is consistent with the DMPV. Moreover, this finding is in line with the strategies employed by expert controllers who are trained to manage specific ATC sectors and therefore benefit from the presence of standardised air traffic configurations (Durso & Manning, 2008).

Finally, several studies, have demonstrated that costs to ongoing tasks (non-PM) are reduced with increased ongoing task delays (In ATC, Loft, Smith, et al., 2013), suggesting that longer ongoing task delays lead to reduced allocation of effort or cognitive resources to the PM task (for theoretical explanations consult, Martin et al., 2011; Scullin et al., 2010). This effect could also be accounted for by associated activation theories of PM: in the absence of rehearsal or contextual cues, deferred task goals would be subject to greater temporal decay, thus longer retention intervals should result in a higher error rate (Altmann & Trafton, 2002; Monk et al., 2008; Nowinski & Dismukes, 2005).

1.1.1. Connecting Prospective Memory and Interruptions

A final point to note is that PM and interruptions are highly interrelated constructs, however, most research in the PM and interruptions domains have not acknowledged the conceptual overlaps (Dismukes, 2010; Dodhia & Dismukes, 2009). On a broad level, theories of PM (e.g., DMPV, Associative Activation Theory of PM) share similar assumptions to the activation-based interruption models (e.g., MFG, Threaded Cognition). For instance, the DMPV specifies that PM intention retrieval occurs from either bottom-up contextual cues or in a top-down strategic manner (Scullin et al., 2013; Shelton & Scullin, 2017); whilst similarly, both MFG and Threaded Cognition specify that post-interruption goal retrieval can depend on associative activation.
Executing Deferred Tasks in Dynamic Multitasking Environments

between environment cues and task goals, or through strategic rehearsal of the task goal (Altmann & Trafton, 2002; Salvucci & Taatgen, 2008). However, efforts to consolidate these theoretical frameworks or to identify how both literature bodies can be applied in practice are scarce. Thus, a key motivation for this thesis was to draw from both theories and PM and of interruptions to inform my experimental paradigms, predictions and theoretical accounts. Note that while it would be beyond the scope of the current work to disentangle these two constructs, below I discuss: (1) the research examining overlap between interruptions and PM; (2) how I’ve incorporated this research into the empirical work in this thesis; and (3) a rationale for not explicitly selected one framework over another — choosing to draw from both.

The most significant conceptual overlap is that all interruption situations require some level of PM. Dodhia and Dismukes (2009) argue that all interruption situations implicitly create deferred tasks because (when interrupted) individuals must explicitly form the intention to return to the interrupted task, and must remember to do so at the correct time without prompting — a defining feature of a deferred task (Einstein & McDaniel, 2005). Put simply, interruptions themselves create deferred tasks, requiring PM. To better understand how PM processes operate in the context of task interruption, Dodhia and Dismukes conducted two experiments where the key criterion was a failure to resume an interrupted task. Their paradigm consisted of participants completing blocks of questions, and sometimes being interrupted before they completed a block. If interrupted, participants had to remember to return to the block of questions that was interrupted by pressing a special key. The central finding of the research was that participants frequently failed to return to the interrupted task altogether. Further, participants in a baseline condition (in which there was no time was given before the interruption) performed poorer than conditions that provided a four-second pause before the interruption (i.e., interruption lag). Interestingly, the there was no additional improvement of providing an explicit encoding reminder (i.e., a message displaying “Please remember to return the block that was just interrupted.”) over the blank screen for four-seconds. Drawing from the associative-activation model of PM, reviewed in 1.9.2.2 (Dismukes & Nowinski, 2006; Dismukes, 2010; Nowinski & Dismukes, 2005), the authors interpreted this finding to indicate that pausing prior to an interruption allowed individuals time to recognize, consolidate and encode information that would benefit PM for task resumption.
Chapter 1: General Introduction

There are two implications of this study and its conceptualization of interruptions relevant to the current thesis. Firstly, it implies that resumption failures (i.e., failing to resume an interrupted task) can be considered a form of PM error. In Chapter 2 and 3, I introduce and report the findings from a deferred conflict task that individuals can fail to resume following an interruption (i.e., resumption failure). I draw upon PM theory as well as interruption theory in order to account for the reasons why an increased failure rate may have occurred. Secondly, a more practical implication is in Chapter 4, where I also examine how addition preparation time prior to being interrupted affects individuals’ ability to remember to perform a deferred handoff task in a simulated ATC task.

As stated above, I do not attempt to consolidate PM and interruptions literature in this thesis. Instead, I refer to them quite separately, but do try to apply (or translate) terminologies from both literatures when describing cognitive processes or experimental features. A contentious argument could be made that because the manifest cognitive processes to suspend and resume interrupted tasks are similar to those required to maintain and execute deferred tasks, similar terminology and models should be used. However, there are several reasons why this would not be appropriate. Firstly, PM studies examine situations which elicit memory failures (i.e., forgetting to perform an atypical intended action), while the majority of interruptions studies use primary tasks that are highly unlikely to forgotten and examine resumption time as the primary dependent variable. Arguably, in these situations there are no PM demands as the individual is placed directly back into the task context (though they can still make errors) and resumption time costs are still observed. Thus, while in some contexts resumption time and PM response time may be equivalent, there are many situations where the contributions of PM retrieval to manifest resumption time are negligible. Similarly, the impacts of interruptions in complex dynamic tasks are not limited to simply remembering to resume a single primary task, but (as reviewed above) can be long term and enduring (Hodgetts et al., 2015; Hodgetts, Vachon, & Tremblay, 2013; Loft, 2014; Sasangohar et al., 2014). This indicates there are more cognitive processes involved with interruption recovery (that are describable with interruptions theory) than simply remembering to resume a single interrupted step (as a PM-alone approach would imply). Finally, simply verbally theorising that a cognitive process is involved with interruption situations does not provide any practical benefits to inform prediction. Thus, in the current thesis, I selectively apply the theoretical framework that is most suited to the parameters of the deferred and primary task at hand (in the simulated ATC task).
Executing Deferred Tasks in Dynamic Multitasking Environments

In summary, the experiments and theoretical framework in the current thesis do incorporate the fact that interruption recovery likely required multiple cognitive functions, including PM. However, this thesis is primarily concerned with the direct negative impact of interruptions on deferred tasks, and it is well beyond the current scope to unify PM and interruptions or to map out the situational contexts where the two literature bodies overlap. Instead, this thesis adopts the practical approach of leveraging both theories of interrupted task performance (and multitasking) and PM to inform predictions, experimental designs, and to interpret findings. In Chapter 3 and 4, I demonstrate how these literature bodies can be leveraged synergistically to provide robust theoretical explanations of our observed data.

1.12. Overview of Thesis

The body of this thesis is comprised of one methodological chapter (Chapter 2), and two empirical chapters (Chapters 3 and 4) presented as empirical papers.

In Chapter 2, I outline the design, development and implementation of the ATC simulator paradigm used for the empirical studies reported in this thesis. The chapter begins by reviewing the rationale for the simulation fidelity, and then discusses how the chosen fidelity level coordinates the relative balance between experimental control and ecological validity. The paradigm was designed to be an ecologically valid simulation en-route ATC as reviewed in section 1.3 above, but also possessing sufficiently low correspondence that the results of the current thesis could generalize to other related work domains. In light of the aims of the current thesis, Chapter 2 concentrates on two core components of the simulation paradigm: the capability to administer interruptions during a trial, and the development of two deferred tasks. The chapter also contains technical documentation regarding how counterbalancing procedures typically applied in more basic task environments were applied in the context of a complex ATC simulation framework.

In Chapter 3, I report the first two experiments conducted using this ATC simulation. These experiments include two forms of deferred tasks that correspond to operational circumstances that air traffic controllers face in-situ (Shorrock, 2005): a deferred conflict task that required remembering to resolve a conflict in the future; and a deferred handoff task that required substituting an alternative aircraft handoff action in place of routine handoff action. The predictions from both studies are directly informed by MFG theory, Threaded Cognition,
Chapter 1: General Introduction

and theories of PM. Experiment 1 tests whether inhibiting rehearsal via an \(n\)-back task could increase response time or errors on both deferred tasks. The data from this experiment demonstrated that interruptions do cause a brief resumption time cost on the deferred conflict task, but rehearsal inhibition was not the key factor underlying this performance cost. In Experiment 2, the \(n\)-back task was replaced with an additional ATC sector to monitor and control, and this form of interruption increased the probability of failing to resume the deferred conflict relative to the blank interruption. A key contribution of the study was the development of an Ex-Gaussian model of resumption times which demonstrated how the increased resumption failure rate detected in the study were likely the result of ‘absolute forgetting’ rather than a delayed response time. These findings indicated support for the interference mechanism reported in MFG and Threaded Cognition (Altmann & Trafton, 2002; Borst et al., 2010; Salvucci & Taatgen, 2011); as well as interference-based accounts of working memory (Hurlstone, Hitch, & Baddeley, 2014; May, Hasher, & Kane, 1999; Oberauer, Farrell, Jarrold, & Lewandowsky, 2016). However, neither experiment revealed evidence that the deferred handoff task was impacted by the effects of interruptions. This null effect was taken as evidence that individuals strategically allocated attention to the task at the moment where action execution was required, rather than maintaining it throughout the retention interval.

The primary goal of Chapter 4 was to further explore possible reasons for the null effect of interruptions on the deferred handoff task. To this end, the study was motivated by two core differences between the deferred conflict and handoff tasks: the overall retention interval, and associated with this, the amount of time granted for pre-interruption preparation and post-interruption recovery. Drawing upon the DMPV theory of PM, as well as basic findings from the interruptions literature, I examined three factorial manipulations: (1) the retention interval of the deferred task, (2) the presence of interruptions during this retention interval, and (3) the timing of the interruption relative to deferred handoff encoding and execution. I examined how each of these factors respectively impacted performance on the deferred handoff task. Increasing the deferred handoff retention interval (37 s to 117 s) also decreased the probability of remembering to perform the deferred task. Further, I found costs to ongoing conflict detection accuracy and routine handoff speed were observed when the deferred handoff task goal had to be maintained. These findings indicated that individuals were actively maintaining the intent to deviate from routine. However, interruptions did not affect individuals speed or accuracy on the deferred task. Several possible explanations based on the DMPV are discussed.
2 AIR TRAFFIC CONTROL IN THE LABORATORY

“For most errors, our understanding of the complex interaction between these various causal factors is, and is always likely to be, imperfect and incomplete. Consequently, most error predictions will be probabilistic rather than precise. Thus, they are liable to take the form: "Given this task to perform under these circumstances, this type of person will probably make errors at around this point, and they are likely to be of this variety,” rather than be of the kind: "Person X will make this particular error at such-and-such a time in such-and-such a place.” Nevertheless, predictions of this latter sort can be made in regard to certain types of error when they are deliberately elicited within a controlled laboratory environment.” – James Reason (1990, p. 4).

2.1. Foreword

In this chapter I first provide a brief introduction to the terminology and methods for simulating complex dynamic tasks in laboratory contexts (section 2.2). Next, I describe an existing ATC simulated paradigm which I use in this thesis (section 2.3). Finally, I describe the development of the simulation paradigm implemented in all empirical chapters of this thesis (section 2.4). The full source code for the final version of the simulation, all supporting scripts to generate the sectors, the resulting XML code for all experiments reported in this thesis, the simulator itself, and all counterbalancing materials reported have been made freely available under the GNU General Public License v3.0 and can be obtained from https://github.com/humanfactors/ATC-Lab-Interruptions. Several blocks of code which are central to the reproducibility of the experiments reported in this thesis are reported within this chapter (within enumerated code blocks). This chapter also contains a vignette for Latin square counterbalancing which is useful more broadly to all researchers who implement simulations of
complex tasks in laboratory studies, and also serves as the proof of complete counterbalancing which is reported in reduced form in Chapters 3 and 4.

2.2. Simulations of Complex Dynamic Tasks

While the qualitative and ethnographic studies of ATC in the field, such as those reviewed in Chapter 1, provide valuable insight to the conditions and precursors where human errors may occur; these studies lack the experimental control required to more precisely understand the conditions and cognitive processes that drive performance (Brehmer & Dörner, 1993; Difonzo, Hantula, & Bordia, 1998; Loft, 2014). However, theories of human memory and performance are typically founded on studies that use basic laboratory tasks with static stimuli. For example, the conditions surrounding PM requirements in the Einstein and McDaniel PM paradigm (1990), where individuals make rapid responses to single stimuli that are presented sequentially, are markedly different from the conditions presented in more complex task environments like ATC. Air traffic controllers must prioritise and coordinate multiple task objectives, to multiple stimuli, in the face of a transient, continuously changing situation (Stein et al., 2009). A clear issue is that field research is generally too complex to allow the experimental control required to understand exact cognitive mechanisms, but conversely the realism of laboratory studies is generally too low to generalise conclusions to work settings (Brehmer & Dörner, 1993). Simulated task environments offer one solution to this gap by providing both experimental control and realism. They allow researchers to apply theory and methods from the basic cognitive science literature to simulations of complex dynamic tasks in laboratory contexts.

Realism refers to the degree of overlap between the cognitive processes and experiences encountered in a simulation and those which occur in the field of interest (Difonzo et al., 1998; Ehret, Gray, & Kirschenbaum, 2000; Ehret et al., 2000). An important concept relating to realism is that of correspondence. According to Gray (2002, p. 214): “High correspondence simulated task environments simulate many aspects of one task environment. Low correspondence simulated task environments simulate one aspect of many task environments”.

---

6 Throughout this thesis, I will use the terms static task or basic laboratory task, to indicate tasks in which participants respond to sequential, unchanging (i.e., static) stimuli. For instance, lexical decision making requires individuals to perform a single task goal to a large corpus of stimuli. The task context does not change throughout the experiment.
Chapter 2: Air Traffic Control in the Laboratory

Low correspondence may be a desirable because it allows the research findings to be generalized to other similar systems (Berkowitz & Donnerstein, 1982). For instance, ATC-Lab is a “use-inspired” task (Stokes, 1997) that represents features of many work contexts that requires operators to monitor perceptually dynamic displays, such as ATC, unnamed vehicle control and air battle management. Experimental control refers to the degree to which variables of the simulation (both independent and dependent) can be controlled and manipulated; with higher experimental control, researchers can be more confident that obtained effects are actually the result of experimental manipulations (Brehmer & Dörner, 1993; Fothergill et al., 2009).

Simulation fidelity is a trade-off between realism and experimental control. For instance, high-fidelity simulations (e.g., industrial ATC simulators for training) have very high realism, but poor experimental control, therefore have limited use for identifying the specific mechanisms and conditions underlying errors. Achieving a high degree of simulation fidelity requires participants that possess expertise in the domain of interest (for in-depth discussion consult Sanderson & Grundgeiger, 2015). High fidelity simulations with expert participants is critically important when results of the simulation studies are intended to be applied directly to inform specific operational procedures in the work domain of interest.

Simulations are often used to build theories of the cognitive processes that underlie performance in specific control task contexts (Rantanen & Nunes, 2005). A clear advantage of this is that experimenters can manipulate variables in ways that would be infeasible or dangerous in the ‘real-world’ (e.g., interrupting operators). However, they are also useful for addressing more basic issues in human cognition such as how we perform intended task actions (Loft & Remington, 2010), how uncertainty moderates workload (Corver, Unger, & Grote, 2016), the mechanisms required for the coordination and prioritisation of multiple goals (Ballard, Yeo, Neal, & Farrell, 2016; Ballard, Yeo, Loft, Vancouver, & Neal, 2016), and the effects of temporal pressure on information processing capacity (Hendy, Liao, & Milgram, 1997). In this way, simulated task environments are also useful tools for evaluating the generalisability and practical relevance of theories of human cognition founded in simple laboratory paradigms, thus enabling more comprehensive theories of human behaviour.

2.3. ATC-Lab

In Chapter 1, I reviewed several studies that have applied theories and methods from the PM literature to simulations of ATC. In this thesis, I will focus on variants of the ATC simulator
developed by Fothergill, Loft and Neal (2009) called ATC-Lab\textsuperscript{Advanced} (herein, ATC-Lab), which is based on the Australian Air Traffic Management System. Many of the features included in the simulation were designed based on structured interviews with expert controllers (Fothergill & Neal, 2005) and analyses of the ATC literature (Callantine, 2002; Späth & Eyferth, 2001).

A key feature of ATC-Lab is the high degree of \textit{programming control} it offers, which refers to the extent that the researcher can control what is presented in simulations. Researchers can present standardized air traffic scenarios to control for extraneous and confounding variables; as well as modify the extent of display realism, response system realism, trial presentation, and presentation of rating scales. Indeed, several studies have leveraged this programming control to conduct simulations with high display-realism and high fidelity, using samples of expert controllers (Loft, Bolland, Humphreys, & Neal, 2009). However, this thesis focuses primarily on low to medium fidelity ATC-Lab simulations with student samples (Fothergill et al., 2009). Although the findings that emerge from these lower fidelity simulation studies generally cannot be directly applied to ATC contexts, they do provide important insights into the cognitive processes underlying complex decision-making in dynamic display tasks more generally. Further, the increased statistical power afforded by large student samples enables more definitive conclusions regarding the functional mapping between manipulated simulation conditions (i.e., independent variables) and performance criterion variables.

As previously mentioned, a number of studies have used the ATC-Lab paradigm, resulting in a number of specific variants. Nevertheless, most of the variants share a common core, and in this section (2.3) I will describe these core features and the prototypical PM study design. In the paradigm, participants (generally undergraduate university students) assume the role of an air traffic controller and are required to perform several tasks to maintain operation safety of a flight control sector. The sector is represented as a light grey polygon, with areas outside the participants jurisdiction represented by a dark grey area (see Figure 2.1). Black lines across the sector denote flight paths that aircraft travel. Aircraft are represented by a circle with a leader-line indicating heading, and have an attached information box. The aircraft data-blocks specify call sign, speed, aircraft type, current altitude and cleared altitude. Current altitude and cleared altitude are separated by an arrow that denotes whether the aircraft is climbing ($\wedge$), descending ($\vee$), or cruising ($>$). Aircraft enter the sector from the edges of the display, cross sector boundaries, and then exit the sector. New aircraft can continue to enter throughout trials.
Chapter 2: Air Traffic Control in the Laboratory

A timer can be displayed to show the elapsed time in each trial (shown at the bottom of the display).

When aircraft come within 5 nautical miles of the sector, they flash orange to signal that they need to be accepted into the sector. Participants can accept aircraft by clicking the aircraft icon, and then pushing the “A” key within 15 s. Similarly, as aircraft exit the sector boundary, they flash blue to signal that they need to be handed-off. Participants had to handoff aircraft by clicking the aircraft icon and pressing the “H” key within 15 s. It should be noted that while these tasks realistically simulated the visual features of the Australian Air Traffic Management, expert controllers would have substantially more complex decisional processes involved with coordinating aircraft entering and exiting their airspace. However, the decreased decisional realism serves to increase the simplicity of the tasks such that it can be quickly learnt by undergraduate participants. Previous studies have found that these tasks are sensitive to motivational processes underlying regulation of task-directed effort and PM demands (Loft et al., 2016; Loft, Finnerty, et al., 2011; Loft, Pearcy, et al., 2011). A number of studies also required individuals to complete a deferred task. The deferred tasks are typically based on aircraft acceptance procedure. Specifically, individuals are instructed to substitute an alternative response key in place of the routine key when accepting target aircraft that had certain flight data (e.g., a certain speed, altitude, or call sign).

Participants are also required to detect and resolve conflicts (i.e., violations of minimum separation standards between aircraft). Aircraft conflicts are defined as occurring when an aircraft pair simultaneously violate both lateral (5 nautical miles) and vertical (1,000 feet) separation standards. Participants are required to detect and prevent conflicts from occurring by clicking on the databox of one of the aircraft within the conflict pair, and changing its cleared altitude to a value that resolves the conflict. Within each trial, a number of aircraft pairs are specified to be “near misses”, aircraft pairs that are close to being in future conflict, but actually are safe, allowing calculation of signal detection sensitivity (T. D. Wickens, 2002).
Figure 2.1. The ATC display (identical to Figure 3.1). Inbound aircraft are black (e.g., aircraft C77) as they approach the sector, and flash orange for acceptance (e.g., C31) when they reached within 5 miles of the sector boundary. Aircraft turn green (e.g., C79) when accepted. When outbound aircraft cross the sector boundary, they flash blue (e.g., C15), and then turn white (e.g., C13) when handed off. Aircraft turn yellow (e.g., C19 & C35) if they violate the minimum vertical and lateral separation. The example shows that the individual is required to change the altitude of C82 from 430 to an alternative altitude to avoid a conflict with C68. Previously the individual had failed to change the altitude of either C19 or C35 (illuminated in yellow). The running score (10 points) is presented in the middle right hand side of the display.

To motivate speed and accuracy, participants are awarded points for successfully completing tasks, whilst points are deducted for failure to complete tasks. Typically, 10 points are awarded or deducted for a successful/failed handoff/acceptance. Between 10 and 40 points are awarded for resolving a conflict, depending on how quickly participants resolve the conflict, and 40 points are deducted for failing to resolve a conflict. Forty points are deducted for unnecessary aircraft interventions (i.e., conflict detection false alarms).

2.3.1. ATC-Lab: Programming Interface

I will now briefly describe the generally process of programming an experiment in ATC-Lab. There are two important components for developing simulations: the simulation executable (i.e., the actual simulation software) and experimental scripts. The simulation executable reads and interprets experimental scripts that specify various properties of the simulation. The
simulation executable is programmed in C++, and a copy of the final version and source code of the ATC-Lab (version: 2.4.5.19) resulting from this thesis can be obtained from https://github.com/humanfactors/ATC-Lab-Interruptions. The simulation currently only works on Microsoft Windows (all versions from Windows 2000 to Windows 10).

**Scripts.** The ATC-Lab simulator requires the experimenter to program experimental scripts which specify the events that occur during the simulation. The scripts are written using a custom Extendable Markup Language, Version 1.0 (XML), which is a general-purpose markup language that can be used as a framework for storing text or data in a tree structure. Simulation details in the script include: information regarding the geometry of the sector; the flight paths presented to participant; and details regarding each individual aircraft on the display, including call sign, aircraft type, starting flight level, starting speed, coordinates for simulation start position \((x, y)\), planned route, and information regarding planned climbs and descents (as well as rate of descent). The simulator then displays the information derived from these scripts on the display and will update information dynamically. For example, aircraft data blocks will display the initial information regarding an aircraft status (e.g., altitude, speed, type), but are then dynamically updated upon participant interventions (e.g., changing aircraft altitude or speed will change the information displayed to the participant).

**Scenarios.** Experiments programmed in ATC-Lab generally comprise a series of brief scenarios (rather than one long scenario). The experimenter has control over the duration and number of scenarios presented in each experiment, though typically scenarios last for around 5 mins. Using brief scenarios enables experimenters to assess performance under a variety of experimental conditions. It is important to note the difference in definition between scenarios and trials. In the context of ATC-Lab, a scenario refers to a specific ATC sector that can be re-used under differing experimental conditions (e.g., interrupted or not interrupted). However, the term trial refers to the combination of a specific ATC scenario and an experimental manipulation assigned to that scenario. For example, in the current studies, a trial could either be interrupted or uninterrupted but share the same underlying scenario. In the XML scripts, scenarios are defined in the ‘skys’ data block, while specific trials are defined in the ‘presentation’ data block.

**Data Output.** The data from experiments is output in two forms: a comma-separated values file (CSV; i.e., spreadsheet) and an extended ATC log file. The CSV file contains all
executing deferred tasks in dynamic multitasking environments

experimental data in a human readable format. the ATC log file contains every event that occurred during the simulation, including the dynamic positions and status of every aircraft of this display.

A substantial practical contribution of my work in this thesis was the development of a set of tools that facilitate automatic data extraction from the CSV and ATC log files. In an effort to enhance the reproducibility and replicability of the work presented in this thesis, these tools are provided on the Github repository (link provided above) for the project (Peng, 2011; Plesser, 2018). Firstly, the repository contains a set of Microsoft Excel Macros for analysing Experiments 1 and 2 (Chapter 3). These macros read an entire directory of participant CSVs, and output the data in a wide data format, ready for analysis within a statistical package such as SPSS or R. There are separate macro files for each experiment (1 and 2), and separate macros for the deferred tasks and the ongoing tasks. Secondly, the repository also contains an extensive set of Python scripts which are used to analyse the data from Experiment 3 (Chapter 4). These scripts utilize the complete ATC logs and CSVs in order to gather all experimental data into long form for analysis using mixed-effects regression modelling.

2.4. Current Paradigm Development and Pilot Study

All features reported in this section (2.4) are new to the ATC-Lab ecosystem and were developed for the purpose of the current thesis. The main aim of this section is to (1) introduce the experimental paradigm used for the experiments in this thesis; (2) report key findings from a pilot study conducted using this paradigm; and (3) provide code block examples demonstrating how core experimental design features were implemented. The general structure of this section is as follows. First, I will outline the overall goals and motivations for the current paradigm, referring to the more basic experimental paradigms and the ecological motivations reviewed in Chapter 1. Second, I will describe each of the key components of the experimental task that was developed for this thesis (e.g., primary ATC task, the interrupting tasks, deferred tasks), and the modifications that were made to the software to include these tasks, as well as descriptions of how these tasks were implemented in the simulation software. At the end of each deferred task development section, I will report the deferred task results from a pilot study using the core design from Chapter 3.
It should be noted that this thesis contains three experiments and the experimental paradigm was developed iteratively over the course of these experiments (i.e., each experiment was slightly different). Importantly, while all three experiments utilized the same underlying basic experimental structure, several more substantial modifications were made to the simulation for the third study reported in Chapter 4, resulting in effectively two minor versions of the paradigm. By default, it can be assumed that descriptions of the ATC simulation below are based on the final version used in Chapter 4 and that descriptions of the simulation below are generally applicable to both versions. Any instances of non-trivial differences between the two are explicitly stated (either in footnotes or in-text). The pilot study reported herein was conducted on the first version of the simulator (as used in Chapter 3). The details of the pilot study design are described in detail in Chapter 3. For now, the only information required is that the design featured three within-subjects conditions: an \( n \)-back task interruption, a blank interruption, and a no-interruption condition; and that participants completed a total of 15 trials. The purpose of the pilot study was to ensure simulation scenarios were of achievable difficulty, and to ensure that there was variability in performance on the deferred tasks (i.e., performance not at ceiling or floor).

2.4.1. Rationale for Paradigm Development

To date, no studies have examined the impact of interruptions on deferred task performance in a complex dynamic task environment. Thus, the overarching goal of the current thesis is to examine the extent to which interruptions impact deferred tasks and general task performance in a simulation of ATC. To achieve this, a set of ATC simulation scenarios that were capable of administering interruptions and included deferred tasks were required. Moreover, I required the design of the deferred tasks in all reported studies to have correspondence to the methodologies used the basic interruptions and PM literature (e.g., produce outcome measures such as PM accuracy, resumption time, and resumption failures), whilst also possessing ecological validity to situations faced by controllers in the field. To this end, I developed 16 unique scenarios, each containing a deferred conflict task and a deferred handoff task — all of which are described below. These tasks were motivated by ecological concerns mentioned throughout the ATC literature (Dodhia & Dismukes, 2009; Durso & Manning, 2008; Nolan, 2009; Shorrock, 2005, 2007).
2.4.2. General Scenario Development

As noted above, each ATC-advanced experiment comprises a series of ATC trials, with each trial being associated with a unique ATC scenario. In the current thesis, I developed 16 unique ATC scenarios. There was a common set of required parameters for each scenario, and a systematic methodology for calibrating their difficulty and exact nature. The criteria and associated rationale were as follows:

- Each scenario contained 27-30 aircraft in total, with an average number of 29. This sector density was chosen based on density values from previous ATC-lab studies (Loft et al., 2016; Loft, Smith, et al., 2013). This resulted in an average of 28.63 handoff or acceptance events per scenario, with an average handoff or acceptance event occurring every 10.71 s.

- Each scenario contained three conflicts, one programmed to violate minimum separation prior to the interruption, one after the interruption, and one programmed to enter conflict prior to the interruption and violate minimum separation after the interruption (i.e., overlap). Conflict detection requires identifying lateral separation (the distance between two aircraft in the two-dimensional display), and vertical separation (the difference in the two aircrafts’ altitudes). In Experiments 1 and 2, aircraft were configured to be cruising on flight levels incremented by 1000 ft (i.e., 39,000, 40,000, 41,000). However, this was adjusted in Experiment 3 in order to increase the ecological validity and cognitive demands of the conflict detection task. Thus, aircraft were set to have a 35% chance of remaining at a 1000 ft interval; 25% chance of changing to a 500 ft interval; 40% chance of changing to 100, 200, 300 or 400 ft interval change.

- All scenarios were reviewed by two individuals to detect for excessive and unrealistically demanding scenarios or sector density. For instance, examining whether too many

---

7 The first version used in Experiments 1 and 2 had a mean of 27, with a range of 24 to 34. This was adjusted for the third study such that the average sector density of scenarios was increased, but with a lower variability.

8 The number of conflicts in the first version used in Experiments 1 and 2 varied between 2 and 3. A core issue was that each scenario did not contain an “overlap conflict”, that is, an aircraft pair that first entered future conflict prior to the interruption, but violated minimum separation after. This was adjusted in Experiment 3 to enable more uniform scenarios and comparison between conflict types across all trials.
acceptance and handoff events occurred in a window of time, or whether aircraft were unrealistically close together. To do this, the scenarios were recorded using the open-source recording software, OBS Studio (V 20.0.3; OBS Studio Contributors, 2018). Specifically, scenarios were recorded from start to finish with no human intervention (i.e., nobody controlling the simulation). A subtitle track was placed into the video that provided a message indicating the exact moments conflicts and deferred task events occurred (see Figure 2.2 below). These subtitle tracks were created using Aegisub (V 3.3.2; Aegisub Core Team, 2014). All scenarios were modified based on feedback.

![Figure 2.2](image)

*Figure 2.2. A display of the ATC interface, with the Aegisub subtitles overlaid indicating two aircraft pairs have now “entered conflict” (actually refers to a “future conflict”). Scenario evaluators were instructed that this message meant that unless appropriate intervention is made, the noted aircraft pairs will violate minimum separation in the future and turn yellow. This makes it easy to see when participants are first able to respond to the conflict.*

2.4.3. Deferred Task Development

2.4.3.1. Deferred Handoff Task

The deferred handoff task was motivated by ecological concerns: controllers frequently must remember to deviate from well-practiced behavioural routines. For instance, needing to remember to deviate from a routinely performed aircraft handoff procedure and alternatively hold an aircraft when it reaches a specific way-point in the future because of heavy traffic. This
situation leads to an error known as “habit-capture” (Reason, 1990), where operators fail to perform the intended atypical action, and substitute the routine action instead (Loft & Remington, 2010). To successfully deviate from routine, controllers must inhibit the expectation-driven processing bias cued by automated behavioural routines and notice that the context of the current task indicates an action should be executed (Norman, 1981; Reason, 1990; Vidulich & Tsang, 2012). In these situations, not only must the controller remember a new episodic task, but the intended action must compete for retrieval with task actions strongly associated with primary task goals (Dismukes, 2012; Loft & Remington, 2010).

In addition to these ecological concerns, the specific experimental implementation of the deferred handoff task was motivated by the design of delayed-execute PM tasks that examined the impact of interruptions on PM situations. Specifically, the dependent variables were designed to match those typically used in the PM literature (e.g., PM errors, PM response times). It required participants to encode a non-routine handoff instruction for a target aircraft (i.e., press an arrow key rather than H key), and then remember to handoff the aircraft with the non-routine response key at the appropriate time. Thus, participants encoded (and thus retrieved) a PM intention, but had to maintain this intention over some delay before the opportunity to execute the action arose. In Chapter 3, I examine the impact of interruptions during this delay, and in Chapter 4 I examine whether the length of the delay impacts deferred task performance.

The dependent variables associated with the deferred handoff task are as follows. A PM error occurs when a participant presses the routine handoff key (“H”) instead of the instructed alternative key when handing-off PM aircraft. PM response execution errors occur when a participant remembers to press a different key but presses the incorrect one. Non-response errors occur when participants fail to handoff the aircraft entirely. PM task acknowledgement errors occur when participants fail to acknowledge the PM encoding message and make a PM error. PM response time (PM RT) is the time taken to handoff the aircraft from the time it begins flashing for handoff. Several modifications were made to ATC-Lab to develop the deferred handoff task:

- The simulator was modified such that the experimenter can specify a different handoff key for a particular aircraft. This parameter was configurable for each aircraft via a
configurable parameter in the XML called \texttt{handoffKey}, but only one aircraft specification per trial including this configuration (the deferred handoff aircraft).

- A feature that changed the speed of an aircraft during a climb or descent, previously included to add realism, was disabled. It was critically important to disable this for two reasons. Firstly, the retention interval of each deferred handoff had to be consistent within scenarios (i.e., not varying between participants). Without this modification, if participants changed the altitude of a deferred handoff aircraft, then it would flash for handoff at substantially different times from intended (> 10s) depending on the magnitude of this change and whether it was set to climb or descend. Secondly, scenarios were designed such that no aircraft acceptances or handoffs occurred within a 10 s window either side of the deferred handoff. Disabling speed changes ensured that this criterion was not breached. This is an example of a trade-off between realism and experimental control.

- In early pilot trials, subjects were found to be double clicking on info-boxes (which rotates the position of the data block) to act as an external cue indicating the aircraft requiring the special handoff. Though interesting from a strategic offloading perspective, this feature was removed for the purposes of the current experiment.

- In order to encode the deferred handoff task, a text box and acknowledgement feature were added to ATC-Lab that allowed any number of text boxes to be added to an ATC trial. The text box is yoked to an aircraft (i.e., it moves alongside the aircraft). Font parameters (e.g., font size, face, and weight), distance offset from the aircraft, background color, and border properties can all be specified in the XML. Additionally, the display duration of the textbox can be modified. To check that participants actually encode the text box, an acknowledgement button feature was added that presents a clickable button (default text = “Acknowledge”) below the text box. A parameter enables the button to be unclickable for a given amount of time, such that participants do not inadvertently click it before having a chance to actually encode the text box. The data log files record if the button was clicked, and if so, what time into the trial it was clicked.

Descriptive statistics of performance on this task during the pilot study were examined for two purposes. Firstly, to determine whether the task was at an appropriate level of difficulty (e.g., individuals were not uniformly getting floor or ceiling performance); and secondly, to
determine whether the task was sufficiently demanding to capture individual differences in performance. Table 2.1 shows the mean proportion of PM aircraft correctly handed off in the pilot trial. Figure 2.3 presents the deferred handoff response times for each aircraft in the pilot study. Together, the results indicated that the deferred handoff task satisfied the design requirements: error rates indicated the task difficulty was appropriate and could feasibly be completed by individuals with a limited training history, and there was some variability between individuals on error proportions and RT.

Table 2.1. Average proportion (from the pilot study) of deferred handoff aircraft correctly handed off (1 = correct, 0 = error of any type), with standard deviations in parentheses, presented separately for each condition.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Blank</th>
<th>Nback</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6 (0.55)</td>
<td>0.2 (0.45)</td>
<td>0.2 (0.45)</td>
</tr>
<tr>
<td>2</td>
<td>1 (0)</td>
<td>0.8 (0.45)</td>
<td>1 (0)</td>
</tr>
<tr>
<td>3</td>
<td>0.8 (0.45)</td>
<td>0.6 (0.55)</td>
<td>0.8 (0.45)</td>
</tr>
<tr>
<td>4</td>
<td>0.8 (0.45)</td>
<td>0.2 (0.45)</td>
<td>0.75 (0.5)</td>
</tr>
<tr>
<td>5</td>
<td>0.4 (0.55)</td>
<td>0 (0)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>6</td>
<td>0.2 (0.45)</td>
<td>0.2 (0.45)</td>
<td>0.4 (0.55)</td>
</tr>
<tr>
<td>Total Mean</td>
<td>0.33</td>
<td>0.63</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Note: Blank, Nback and None refer to the three within-subjects interruption conditions used in the pilot study. Information regarding the exact nature of these tasks is reported in Chapter 3.

Figure 2.3. Distribution of deferred handoff response times across the six subjects in the pilot study. Each subject’s mean is presented as a red dot with the exact value adjacent. The box plot presents the inter-quartile range, with a violin plot overlaid (where sufficient data is available) to show the proportional distribution density function.
2.4.3.2. Deferred Conflict Task

The deferred conflict task was motivated by operational circumstances that are reported by Shorrock (2005) as occurring in-situ. Shorrock reported a form of PM error that can frequently occur in ATC, called a “forgetting to monitor” error. In these situations, controllers notice a particular event (normally a conflict) occurring on the radar display, and then forget to execute the intended action of re-monitoring the display and evaluating whether action should be taken. In one of Shorrock’s examples, a controller cleared one aircraft to 17,000 ft and another to 19,000 ft, relying on the performance of each aircraft to maintain separation standards. The controller intended to monitor the situation using automation tools, however an external distraction (communications call) distracted the controller, resulting in a conflict. This is a prototypical example of the situation that the deferred conflict task was designed to capture.

Unlike the deferred handoff task, the deferred conflict task was designed primarily to assess the more immediate effects of interruptions on deferred task performance, and was designed to produce dependent variables that correspond to those typically used in the interruptions literature (e.g., resumption time and resumption errors; Trafton & Monk, 2007). This task requires participants to acknowledge an impending conflict, however, participants are informed that the conflict is not able to be resolved until both aircraft in the pair are cruising. Thus, the participant is forced to defer the execution of the conflict resolution until an appropriate future point. Scenarios were designed such that the two aircraft would both be cruising (and therefore the conflict would be resolvable) at the same moment that the interrupting task ended (under interruption conditions — but equivalent time under no-interruption conditions). Thus, the time taken to detect and resolve the conflict (from the moment they are both cruising) is equivalent to resumption time and failing to resolve the conflict altogether is equivalent to a resumption failure. In Experiments 1 and 2 (Chapter 3), the deferred conflict encoding was set to 20 s prior to the interruption, and the window of opportunity for responding to the conflict (after the interruption) was set to be between 10 and 23 s, depending on the specific scenario.

In order to create the deferred conflict task, a number of steps had to be taken.

---

9 Note that the deferred conflict task was removed from the experiment reported in Experiment 3 (Chapter 4), due to the focus on the deferred handoff task.
Executing Deferred Tasks in Dynamic Multitasking Environments

– In order to prevent participants from resolving the conflict prior to the designated time point, a new functionality was added to ATC-Lab which disallowed conflict resolution features prior to a specified time point.

– The conflict had to be set to unresolvable prior to a particular time point. This new functionality was added to the simulator as an additional XML parameter that could be specified on an aircraft conflict (altitude change time). This restricted participants from changing the altitude of the aircraft.

– The deferred conflict task also required substantial scripting efforts. Specifically, the conflict pair was not permitted to enter conflict with any other aircraft throughout the scenario. Further, no acceptances, handoffs or other conflicts were programmed to occur in the window of time when the conflict was available for response, to ensure participant resumption time could only be attributed to either SA acquisition or the deferred conflict task resumption. These criteria were met in a piecewise fashion, iteratively modifying the conflict parameters, recording the scenario (using the method reported above), and then checking if these criteria were met.

Only one deferred conflict resumption failure was committed during pilot testing. This preliminary evidence indicated that individuals were able to remember to resume the task (as commonly seen in most interruption studies). However, as can be seen in Figure 2.4, there was considerable within-subjects variability in terms of resumption time.
Chapter 2: Air Traffic Control in the Laboratory

Figure 2.4. Distribution of deferred conflict resumption times (milliseconds) across the six subjects in the pilot study. Each subject’s mean is presented as a red dot with the exact value adjacent. The box plot presents the inter-quartile range, with a violin plot overlaid to show the proportional distribution density function.

2.4.4. Interruption Task Development
The most substantial modification to the simulation was the interruption feature. In order to ensure equivalency across interruption and no-interruption conditions, both the deferred tasks were anchored around a pre-programmed interruption epoch. Thus, scenarios were designed such that interruptions occurred at a set time (note that this did vary across scenarios however). The simulation allowed for three forms of interruption: a blank screen interruption, an embedded visual $n$-back task, or an external program. The external program feature allows any executable program to be launched from the simulator.

2.4.4.1. $n$-back Interruptions
The $n$-back task interruption was developed to allow the experimenter to administer a visual-numeric $n$-back task at any point during an ATC trial. In the task, the subject is presented with a sequence of digit stimuli and is required to indicate by pressing spacebar when the current stimulus matches the one from $n$ steps earlier in the sequence (by default, $n = 2$). The participant receives feedback at the bottom of the display in the way of a red or green dot to indicate success or failure, respectively. The load factor $n$ can be adjusted to make the task more or less difficult by changing how many digits back must be remembered. However, there is a restriction such
that the same digit cannot be consecutively presented (i.e., $n = 1$ is not permitted). The duration of stimulus presentation can be modified (default = 2 s), as can the total duration of the task (but it must be divisible by duration of stimulus$^{10}$). The task also allows for an initial stimulus to be displayed in the centre of the display upon loading the task, to allow participants to ready themselves for the task. The probability of a given item (after the initial 2) being an $n$-back target is customisable (default = 40%). Code block 1 shows the various parameters that can be configured in the $n$-back task.

The data output from the ATC-Lab experiment includes a section for $n$-back performance, with a column indicating the stimulus number, the actual stimulus digit presented, and the response outcome. The following response outcomes were used:

1. **Hits** (correctly responded at target stimulus): participant correctly pushes space bar on the bolded number (2 4 2)
2. **Misses** (failure to respond to target stimulus): participant incorrectly does not push space bar on bolded number (2 4 2)
3. **False Alarm** (incorrectly responds to non-target stimulus): participant incorrectly pushes space bar on bolded number (3 5 6)
4. **Correct rejections** (correctly does not respond to non-target stimulus): participant correctly does not push space bar on bolded number (3 5 6)

---

$^{10}$To be clear, the total duration of the task (e.g., 30 s) must be divisible by the stimulus presentation duration (e.g., 2 s), in order to have a valid number of digits presented (e.g., 15).
Chapter 2: Air Traffic Control in the Laboratory

| 113x776 | <atc:nbackTask atc:start='2' atc:end='180' atc:auto_handoff='true' atc:auto_accept='true' atc:background_colour='black' atc:response_key='space' atc:n_factor='2' atc:reptition_probability='40' atc:show_task='true'>
| 750x61 | <atc:initial_display atc:symbol='+' atc:font='Courier' atc:font_size='50' atc:font_colour='white' atc:duration='3' atc:show_timer='true' atc:x='487' atc:y='330'/>
| 721x750 | <atc:stimuli atc:font='SansSerif' atc:font_size='50' atc:font_colour='white' atc:display_duration='2' atc:show_timer='true' atc:x='487' atc:y='330' />
| 698x750 | <atc:correct_feedback atc:symbol='circle' atc:colour='green' atc:width='15' atc:height='15' atc:x='500' atc:y='410'/>
| 640x750 | <atc:incorrect_feedback atc:symbol='square' atc:colour='red' atc:width='15' atc:height='15' atc:x='500' atc:y='410' />
| 652x698 | <atc:nbackTask>

Code Block 1. XML parameters for the n-back task

2.4.4.2. External Program Interruption

The added external program feature requires the experimenter to specify the path to a program stored on the computer. This is an important feature of the simulation, as future researchers intending to use the paradigm would be able explore many different forms of interrupting tasks, and are not limited to any constraints of the ATC-Lab Advanced itself. In Chapters 3 and 4, I use this functionality to launch a secondary ATC simulation trial that lasts 27 s, and I will now outline how this was implemented. It is important to note that the implementation details listed here are generalisable to any other program which an experimenter may wish to launch. It is also possible to launch the secondary application on a secondary device (e.g., a Bluetooth tablet) and have the primary ATC task remain visible.

The most crucial detail of the interruption task is controlling the display of windows within the desktop environment. Specifically, this means ensuring that the original window displaying the ATC simulation does not obstruct any additional programs launched, and the additional programs restore control to the underlying scenario when completed. There are two methods to controlling the window display of the secondary application. The first solution involves a freeware program LaunchOnTop.exe (version 1.1) developed by Savard Software. This
Executing Deferred Tasks in Dynamic Multitasking Environments

program is used via the command line, and when called it waits for the instructed program to load, and then it sets the application window to be always on top of other windows. In order to call the program, a Microsoft Visual Basic script and a batch file are required. The batch file contains the following:

```
start LaunchOnTop "pact.exe" "interruption_scenario.xml"
```

**Code Block 2.1. Microsoft batch file to launch the interruption**

To prevent the command prompt window from appearing, you can call the program directly from a VBS script and suppress any further warnings or popups.

```
Set WshShell = CreateObject("WScript.Shell")
WshShell.Run chr(34) & "interruption_1.bat" & Chr(34), 0
Set WshShell = Nothing
```

**Code Block 2.2. The Microsoft VBScript which calls the batch file to prevent the command line window from appearing upon loading the interruption.**

The secondary interruption window can be controlled by creating an AutoHotkey (AHK) script to coordinate the display of the primary and interrupting applications. Code Block 2.3 below is written in AHK, and demonstrates a simple program that when executed, launches a secondary application and forces this to be at the top of the window stack and under active control. It is beyond the scope of the present guide to list all the possible ways AHK can control window positioning, but it is worth mentioning that AHK is a very flexible tool and this is just a small subset of potential ways in which windows can be controlled. This design feature of the simulation is inherently flexible, allowing future researchers to replace the interrupting task with any other secondary task of their choosing.

```
#SingleInstance, force
SetWorkingDir %A_ScriptDir%
Run, pact.exe "interruption.xml" ;; Launch program
Sleep 250 ;; Allow time for program to run
WinActivate pact ;; Bring window to foreground
ExitApp ;; Stop script
```

**Code Block 2.3. An example of window management for interrupting task written in AutoHotkey.**

2.4.5. Counterbalancing Techniques and XML

Counterbalancing enables experimenters to control for the influence of assumed nuisance variables in experimental designs where participants are exposed repeatedly to conditions,
interventions, or stimuli (e.g., repeated-measures designs). Counterbalancing and randomisation of stimulus presentation are both imperative steps in designing behavioural experiments to minimise risk of stimulus-condition confounds and trial order effects — both serious threats to experimental validity.

However, studies that employ simulations of complex dynamic task environments face significant challenges associated with counterbalancing. Specifically, while the presentation order of scenarios (i.e., stimuli) is generally controlled for, eliminating order effects, the assignment of experimental conditions to scenarios is much more difficult to control for and can potentially cause erroneous unobserved endogenous variables to confound results. Simply put, this issue occurs when experimenters develop simulation scenarios that contain unique events and objects, but the scenarios themselves also comprise the unique experimental manipulations.

This problem is salient for interruption research using complex dynamic task, in which an experimenter seeks to compare interrupted to uninterrupted performance, but each condition is uniquely associated with a particular experimental scenario. Similarly, the issue can arise in experimental designs that feature one long scenario, but the timing of the interruptions is fixed, meaning that performance data is obtained at different parts of the trial. In these situations, statistical analysis is conducted by comparing moments of time after the interruption with “baseline performance” taken in other parts of the scenario (e.g., prior to the interruption). Consequently, there may be extraneous influences unique to specific parts of a scenario which are not controlled for. The risk of this method is that idiosyncratic differences between the scenarios (e.g., higher difficulty, higher workload, increased aircraft density) may confound comparisons across the experimental manipulations.

Thus, a design goal for the current simulation paradigm was to overcome the aforementioned potential control issues and implement a methodology that could ensure ATC scenarios were not uniquely associated with experimental conditions. This methodology enabled the direct comparison between interrupted and non-interrupted task performance (and other important factors), without the confound of scenario. To this end, below I detail a general methodology for achieving complete condition-to-scenario counterbalancing using a reduced Latin square scheme. This can be considered akin to ensuring that stimuli are counterbalanced across experimental conditions in more basic experimental paradigms (e.g., lexical decision). It
is important to note that this methodology is a new approach to developing simulations in ATC-Lab. Moreover, the implementation and design of the method below are generalisable to any study that uses a scriptable interface to program complex simulations. For instance, I have successfully applied this methodology to the Octal Driving Simulator that uses an XML-like interface for programming scenarios (see Bowden et al., 2019).

In the example below, I will focus on the counterbalancing procedure used for Experiment 3, however a similar methodology was used for all experiments in this thesis. The programming language used in the example code blocks is Python 3.7 (Python Core Team, 2018), but, the general methodology could be implemented in many other languages. Experiment 3 featured a 2x2x2 design, resulting in a total of eight within-subjects conditions. Further, there were 16 unique ATC scenarios. In this study, each scenario could be either interrupted or uninterrupted, and there were four different timing conditions for the deferred handoff task. Thus, there was a total of 128 unique possible ATC trials (i.e., 8 conditions x 16 scenarios). The goal was to have each of the unique ATC scenarios have an equal number of observations in each experimental condition across all subjects. Further details of this design are provided in Chapter 4, but the important point here is that this design required the counterbalancing of experimental conditions to scenarios. For the sake of simplicity, in the examples below I replace code specific to ATC-Lab with general text stings to represent the various scenarios and conditions. The overall objective is to generate a reduced Latin square scheme. An example of a reduced Latin square design for a four condition within-subjects design is presented in Table 2.2.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Condition A</th>
<th>Condition B</th>
<th>Condition C</th>
<th>Condition D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group A</td>
<td>Scenario 0</td>
<td>Scenario 1</td>
<td>Scenario 2</td>
<td>Scenario 3</td>
</tr>
<tr>
<td>Group B</td>
<td>Scenario 1</td>
<td>Scenario 2</td>
<td>Scenario 3</td>
<td>Scenario 0</td>
</tr>
<tr>
<td>Group C</td>
<td>Scenario 2</td>
<td>Scenario 3</td>
<td>Scenario 0</td>
<td>Scenario 1</td>
</tr>
<tr>
<td>Group D</td>
<td>Scenario 3</td>
<td>Scenario 0</td>
<td>Scenario 1</td>
<td>Scenario 2</td>
</tr>
</tbody>
</table>

*Table 2.2. Reduced Latin-square counterbalancing scheme. In this example, the various scenarios are counterbalanced across the experimental conditions. Participants would be equally distributed across Groups A-D to achieve counterbalancing.*
In contrast to the example in Table 2.2, the design for Experiment 3 includes an additional complexity common to many simulation situations: there are more scenarios in our experiment (16) than there are experimental conditions (8). Therefore, pairs of scenarios must be grouped together into groups of two, referred to herein as *scenario groups*\(^\text{11}\), meaning that participants see two scenarios per condition.

Code Block 2.4 shows how to create eight columns, where each scenario group occurs once per column and once per row. Specifically, several scenario groups (represented as lists) assigned together in a larger list represented by the variable ‘scenarioGroups’. Below that, a function has been defined called ‘reducedLatinSquare’, which accepts (as an argument) a nested list (i.e., a list of lists), and returns a reduced Latin square. In Code Block 2.4, the ‘scenarioGroups’ is dispatched to the ‘reducedLatinSquare’ function, and the output of this is saved to the variable as ‘cbScheme’. The contents of ‘cbScheme’ can be seen at the bottom of the code block. The eventual goal is to assign conditions to the columns of this scheme and participants to rows.

---

\(^\text{11}\) Recall that Experiment 3 comprises eight within-subjects conditions, and 16 unique ATC scenarios. By creating scenario groups of two, a reduced Latin-square can be created.
#!/usr/bin/env python3
from collections import deque
from random import shuffle
import itertools

scenarioGroups = [[0, 8], [1, 9], [2, 10], [3, 11],
[4, 12], [5, 13], [6, 14], [7, 15]]

def reducedLatinSquare(pairedScenarios):
    '''Generate Latin Square Design Scheme. Takes a nested list of
desired conditions. Returns matrix of reduced latin square.''
    reducedls = []
    conditions = deque(pairedScenarios)
    for i in range(len(conditions)):
        reducedls.append(list(conditions))
        conditions.rotate(-1)  # for left rotation
    return (reducedls)

cbScheme = reducedLatinSquare(scenarioGroups)

Code Block 2.4. Generating the reduced Latin square scheme.

Given the complete counterbalancing scheme contained in 'cbScheme', the next step is
to assign the conditions and participants to this scheme. Note that the examples below will not
implement the XML specific to ATC-Lab, and instead placeholder strings will be used to
represent the various trials.

Code Block 2.5 shows the steps taken to generate the complete set of trial strings, in the
format of 'Interruption-EncodeCondition-ExecuteCondition-ScenarioNumber'. An example of
a trial output from this would be “INTERRUPTION-ShortEncode-ShortExecute-6” —
signalling a trial with an interruption, a short encode and delay, and scenario number six.
Generating the complete combination of conditions from an experiment requires assigning the
Cartesian product for all the conditions and scenario numbers to a list. In Code Block 2.5, the
Cartesian product is assigned to the variable 'trialList', which is then split into chunks of eight,
reflecting the eight different conditions. Finally, each sub-list contains the 16 scenarios all
assigned to one condition, which is denoted by the indexing of 'inst' in the list variable
'SkyConds'.

66
Chapter 2: Air Traffic Control in the Laboratory

```python
trialList = itertools.product(
    ['INTERRUPTION', 'NONE'],      # Interruption
    ['ShortEncode', 'LongEncode'], # Encode Timing
    ['ShortExecute', 'LongExecute'], # Execute Timing
    [str(num) for num in range(1, 17)])  # Number of Scenarios

inst = [''.join(x) for x in list(trialList)]

SkyConds = [
    inst[0:16],               # Int-S-S
    inst[16:32],              # Int-S-L
    inst[32:48],              # Int-L-S
    inst[48:64],              # Int-L-L
    inst[64:80],              # None-S-S
    inst[80:96],              # None-S-L
    inst[96:112],             # None-L-S
    inst[112:128]]            # None-L-L

Code Block 2.5. Generating the list of 128 condition-scenario combinations.

All components required to perform the actual counterbalancing are now instantiated. To review, a counterbalancing scheme that enables the assignment of scenarios to conditions has been created, as has a list containing all specific information for each trial/scenario combination. Thus, the final step is to correctly assign each trial/scenario combination to the corresponding location in the counterbalancing scheme, and assign participants to a specific counterbalancing row. This is performed with a main loop in Code Block 2.6, which works as follows:

— The loop index represents an individual subject number.

— An empty list is created for storing the scenario strings. Next, the modulus operator is used to select a counterbalancing scheme for the indexed subject (i.e., select a row from the reduced Latin square generated in Code Block 2.4), which is stored in ‘subjectScheme’.

— As each scheme (i.e., row) contains a list of scenario groups, a loop is run over each scenario group in that subject scheme, and again to capture each scenario in each group.

— Every piece of information from this information set is utilized to achieve the final counterbalancing. The index of the scenario group (i.e., position in row) is used to lookup the corresponding trial information in skyConds (which contains the ordered
list of all scenario/condition combinations). In other words, the column position of the scenario group is matched to the experimental condition assigned to that column.

Finally, the specific scenario is indexed from the sublist, returning a trial string, which is then saved to the final output list. Before saving our list to the FinalList, we shuffle it to ensure final presentation order is randomised.

```python
iFinalSubjects = {}
for subjid in range(1,9):
    subjectScenarioList = []
    subjectScheme = cbScheme[(subjid-1)%64] # Modulus operator finds which row to use in scheme
    for scenarioGroup in subjectScheme:
        for scenario in scenarioGroup:
            subjectScenarioList.extend([skyConds[subjectScheme.index(scenarioGroup)][scenario]])
    shuffle(subjectScenarioList)
    iFinalSubjects[str(subjid)] = subjectScenarioList
```

*Code Block 2.6. Allocation of the scenario-condition combinations to each participant based on the reduced Latin square scheme.*

The procedure reported here is almost identical for the full XML, except rather than returning single strings, a complete ATC XML Sky tag is returned, which then be written out to a text file and loaded for each participant (see 2.3.1). To summarise, I have demonstrated a general and reproducible methodology for performing counterbalancing of simulation scenarios to experimental conditions, under complex experimental designs. This logic of the programming above can be extended to any simulation paradigm which uses text-based scripts to program simulation scenarios. However, it should be acknowledged that the above description does not provide a complete account of every step required to generate the XML scripts used in Experiment 3, as such a description would be substantially beyond the scope of the current chapter and largely cover basic text manipulation with Python.
This chapter is presented as a journal article manuscript. A version of this manuscript was published in the *Journal of Experimental Psychology: Applied*. The experimental data, all code used to process the data, and all analysis scripts are available for download at [https://github.com/humanfactors/atc-interruptions-impact](https://github.com/humanfactors/atc-interruptions-impact). Note that I use the term ‘we’ throughout this chapter to refer the collective contributions of the manuscript co-authors.

This publication was awarded an American Psychological Association (Division 3) Early Career Award in 2019. The award was presented at the 2019 APA Convention in Chicago.

The publication also received a Higher Degree by Research Achievement Award (Social Science Quantitative) from the University of Western Australia Graduate School of Research in 2019.

3.1. Foreword

In Chapter 1, I introduced and outlined several theories of interrupted task performance and PM that can be leveraged to understand the cognitive mechanisms underlying deferred task execution in complex and dynamic workplace contexts. Further, I outlined how complex and dynamic task environments can result in interruptions having more chronic and long-term effects. In Chapter 2, I introduced a paradigm for measuring the effects of interruptions to two forms of deferred tasks that controllers experience in situ: a deferred handoff task and a deferred conflict task. The paper presented in this chapter presents two experiments which implement this paradigm and interpret my findings in the context of the broader cognitive theories presented in Chapter 1.

3.2. Abstract

Air traffic controllers can sometimes forget to complete deferred tasks, with safety implications. In two experiments, we examined how the presence and type of interruptions influenced the probability and speed at which individuals remembered to perform deferred tasks in simulated air traffic control (ATC). Participants were required to accept/handoff aircraft, detect aircraft conflicts, and perform two deferred tasks: a deferred conflict task that required remembering to resolve a conflict in the future; and a deferred handoff task that required substituting an alternative aircraft handoff action in place of routine handoff action. Relative to no interruption, a blank display interruption slowed deferred conflict resumption, but this effect was not augmented by a cognitively demanding n-back task or a secondary ATC task interruption. However, the ATC task interruption increased the probability of failing to resume the deferred conflict relative to the blank interruption. An ex-Gaussian model of resumption times revealed that these resumption failures likely reflected true forgetting of the deferred task. Deferred handoff task performance was unaffected by interruptions. These findings suggest that remembering to resume a deferred task in simulated ATC depended on frequent interaction with situational cues on the display and that individuals were particularly susceptible to interference-based forgetting.
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

3.3. Introduction

In many workplace and everyday situations, we often need to defer a task until a point in the future when it can be carried out. In the psychological science literature, this is referred to as a prospective memory (PM) task. In air traffic control (ATC) settings, situational contexts such as high workload can prevent controllers from immediately completing task actions, requiring them to defer the action to an appropriate point in the future (Shorrock, 2005). During this retention interval, the controller will almost certainly be interrupted by other ongoing task activities. Data from ATC incident reports, controller interviews, and laboratory simulations has found that controllers sometimes forget to complete deferred task actions (Dismukes et al., 2007; Shorrock, 2005), and such forgetting can have serious safety implications. For instance, in 2013 two A330s violated minimum separation over Adelaide because, in part, the air traffic controller failed to remember to re-evaluate aircraft separation before issuing instructions (Australian Transport Safety Bureau, 2015). To minimize the risk of such occurrences and maintain system safety, we require a better understanding of the cognitive and situational factors that support the execution of deferred actions after a period interposed activity.

With this goal in mind, recent research has applied theories and methods from basic psychological science to simulations of ATC (for review see, Loft, 2014). In these studies, participants assumed the role of a controller responsible for a region of airspace, called a sector, which varies along both vertical (altitude) and lateral dimensions. Participants monitored a two-dimensional display of their sector, with aircraft flight paths being indicated by lines and aircraft indicated by icons with attached flight information (e.g., call sign, altitude, and airspeed). They were trained to detect and resolve conflicts (i.e., identifying whether any two aircraft will violate minimum separation standards in the future and then changing an aircraft’s altitude to resolve that conflict), and to accept and handoff aircraft (i.e., acknowledging responsibility for aircraft entering/exiting their sector by pressing designated response keys). Participants in these studies were also given a deferred (PM) task that required them to remember to deviate from routine operating procedure (by pressing an alternative response key) when accepting ‘target’ aircraft that met certain conditions (e.g., altitude > 40,000 ft.). Multiple studies have shown that participants often failed to perform this deferred task, and performed ongoing tasks, such as aircraft acceptance and aircraft conflict detection, more slowly than if the deferred task was omitted (Loft et al., 2016; Loft, Smith, et al., 2013, 2011; Loft, Finnerty, et al., 2011; Loft, Pearcy,
Executing Deferred Tasks in Dynamic Multitasking Environments

et al., 2011; Loft & Remington, 2010). This suggests that the cognitive control operations required to monitor the ATC environment for cues associated with the deferred task can impair performance (for further theoretical discussions of the psychological mechanisms underlying “PM cost” effects see, Einstein & McDaniel, 2010; Heathcote, Loft, & Remington, 2015; Smith, 2010; Strickland et al., 2017).

In the current research we examine the impact of task interruptions on deferred task performance in simulated ATC in order to further understand the cognitive factors that underlie deferred task performance. Instead of manipulating deferred task demands and interpreting performance on other ongoing tasks as evidence for the cognitive processes underlying deferred task retrieval (Loft, 2014); in the current studies, we manipulate the nature of the ongoing task demands and interpret performance on the deferred task as evidence for the cognitive processes underlying deferred task retrieval. We anticipated that this alternative approach will enable us to further understand the cognitive processes underlying storage of deferred task goals and the role of the external environment in supporting their retrieval.

Specifically, we take the approach of past studies from the basic literature and examine the effect of the nature and presence of interruptions on individual’s ability to resume a primary task (McDaniel et al., 2004; Monk et al., 2008; Trafton & Monk, 2007). In these prior interruption studies, participants are faced with a static primary task (e.g., VCR programming) that is interrupted by a secondary task. Typically, disruption is measured by the resumption time; that is the time taken to remember and to perform the first action on the primary task following an interruption (Trafton & Monk, 2007). However, participants can sometimes forget to resume the interrupted primary task (Dodhia & Dismukes, 2009; McDaniel et al., 2004); and if they do remember to resume, they can have difficulty reconstructing the primary task state, resulting in errors on the primary task (Altmann, Trafton, & Hambrick, 2014; Brumby, Cox, & Back, 2013; Monk et al., 2008).

In the current experiments participants completed a series of ATC trials, some of which were interrupted by a secondary task. We examined how the presence and nature of interruptions — presented between encoding a deferred task action and the correct time to perform that action — influenced the probability and speed at which individuals remembered to perform deferred (primary) tasks. The impact of interruptions was examined on two deferred
tasks which correspond to operational circumstances that air traffic controllers face in situ: remembering to resume a deferred task and remembering to deviate from a routine procedure.

3.4. Resuming Deferred Tasks

Some operational circumstances do not allow controllers to immediately resolve impending conflicts between aircraft (i.e., future violations of minimum aircraft separation; Loft, Bolland, Humphreys, & Neal, 2009). For instance, a controller may identify an aircraft conflict but be unable to immediately issue the conflict resolution because of other air traffic (Loft, Sanderson, Neal, & Mooij, 2007). During this retention interval in which the conflict cannot be resolved, the controller is likely to be interrupted by other ongoing task demands and may have few opportunities for rehearsal of the deferred task (Shorrock, 2005). To what extent does the nature of such interruptions influence a controllers’ ability to remember to resume the deferred task?

To simulate the impact of an interruption on deferred task resumption, we included a deferred conflict task that required participants to encode a temporarily unresolvable but impending conflict, and to form the intention to resolve the conflict at an appropriate time in the future. Importantly, while the deferred conflict task was encoded shortly before the interruption, it was only resolvable immediately after the interruption (or equivalent time under conditions in which participants were not interrupted). In this way, our deferred conflict task resumption measure was comparable to the primary task resumption measures typically used in the interruption literature (Trafton & Monk, 2007).

There are several theoretical frameworks that can be leveraged to predict how interruptions might impact deferred task performance in ATC. According to the Memory for Goals theory (MFG; Altmann & Trafton, 2002), deferred task goals are associated with different levels of memory activation. Task goals with higher activation are recalled more quickly following an interruption than goals with lower activation. MFG specifies that task resumption is accomplished by retrieving the primary task goal and task context from memory, and that this process can be facilitated by attending to contextual cues in the post-interruption environment that are linked to the deferred task goal (Altmann & Trafton, 2002). Salvucci and Taatgen, (2011) extend MFG in their cognitive model of human multitasking called Threaded Cognition, which specifies that each task goal is associated with an individual ‘thread’ which allows for multiple goals to be concurrently activated. Threaded Cognition is a general model of human
Executing Deferred Tasks in Dynamic Multitasking Environments

multi-tasking and treats interruptions as a form of “sequential multitasking”. According to Threaded Cognition, individuals would not only maintain a thread related to the deferred task goal, but also maintain a memory representation of the information associated with and required to perform the goal, called the problem state. The problem state, like other goals, will decay over time unless it is actively rehearsed (Monk et al., 2008). Rehearsal can be either retrospective (e.g., “What was I doing?”), or prospective (e.g., “what was I about to do”), although evidence suggests individuals prefer prospective rehearsal (Monk et al., 2008). Moreover, rehearsal of the problem state is not limited to repetition of the episodic task goal, but can take advantage of the contextual information associated with the deferred task, thereby increasing the effectiveness of situational retrieval cues under proper conditions (Koriat, Ben-Zur, & Nussbaum, 1990).

Based on MFG and Threaded Cognition, we might expect that any interruption which prevents visual inspection of the ATC display for a sufficient time should increase the time taken to remember and resolve the deferred conflict, and possibly increase the probability of forgetting the deferred conflict task goal. This is because participants would no longer have access to the display to prime the deferred task goal. In addition, depending on the extent to which a participant’s problem state has decayed during the interruption, the participant will need to take some time to re-develop situation awareness (SA) for the evolved locations and relationships between aircraft in the post-interruption display. Assimilating the mismatch between the pre-interruption and post-interruption displays can take considerable time (Hodgetts et al., 2013; St. John & Smallman, 2008; St. John et al., 2005).

A second central question is whether the nature of the interruption will modulate its effects on deferred task performance. According to MFG, an interruption which is cognitively demanding would be expected to cause greater costs to deferred task performance because it would block mental rehearsal and thus increase decay of the stored task goal (Cades, Werner, Boehm-Davis, Trafton, & Monk, 2008; Hodgetts & Jones, 2006b; Monk et al., 2008). Similarly, Threaded Cognition would predict that a cognitively demanding interruption should be more disruptive because it would introduce competition with the primary task for access to the problem state representation which causes a cognitive bottleneck because only one problem-state can be active at any given time (Salvucci & Taatgen, 2011).
That said, these predictions rely on the assumption that individuals will encode and retrieve detailed internal representations of the task environment (the problem state). In fact, the ATC task may be too complex or feature rich for the creation of such a representation, thus making memory-based retrieval of the problem state an unfeasible interruption recovery strategy. Salvucci and Taatgen (2011) argue that as task complexity increases, individuals are likely to rely more heavily on reconstructing the problem state rather than on storing and retrieving it from memory. Such reconstructive strategies would likely involve the allocation of attentional resources to the contextual features of the ATC display that are known to be associated with task goals (Hunter & Parush, 2010; Ratwani & Trafton, 2008). For instance, a particular location may become associated with the deferred conflict, and following the interruption, visually scanning that location may reinstate the participant’s intention to resolve that conflict. A reliance on reconstructive strategies might also be fostered by the dynamic nature of the ATC task which increases the likelihood that the problem state formed from the pre-interruption period does not accurately represent the state of the task at task resumption (St. John & Smallman, 2008). Indeed, several theoretical accounts in the SA literature posit that individuals prefer to store partial representations and rely on frequent interactions with their displays to access “highly selective information on an as-needed basis” (Chiappe et al., 2016; Chiappe, Vu, Rorie, & Morgan, 2012; Gray & Fu, 2004).

If participants rely more heavily on interactions with the ATC display to reconstruct their problem state, as opposed to storing and retrieving the problem state from memory, we would still expect interruptions to negatively impact deferred conflict resumption, because of the time needed to recover SA for the locations of aircraft in the post-interruption scene that are related to deferred task goals. However, we would not expect a more demanding interruption to be more disruptive than a less demanding interruption, because participants would not be storing a problem state in memory that needs to be rehearsed and maintained.

3.5. Remembering to Deviate from Routine

Another form of deferred task that controllers perform is to remember to deviate from well-practiced behavioural routines. In these situations, not only must the controller remember a new episodic task, but the intended action must compete for retrieval with task actions strongly associated with primary task goals (Dismukes, 2012; Loft & Remington, 2010). For example, a
controller may need to remember to deviate from a routinely performed aircraft handoff procedure and alternatively hold an aircraft when it reaches a specific way-point in the future because of crossing traffic. This type of situation often leads to “habit-capture” (Reason, 1990), where operators fail to perform the intended atypical action, and substitute the routine action instead (Loft & Remington, 2010). To successfully deviate from routine, controllers must inhibit the expectation-driven processing bias cued by automated behavioural routines and notice that the features of the task indicate that the alternative task action should be executed (Norman, 1981; Reason, 1990; Vidulich & Tsang, 2012). Thus, habit-capture is usually presumed to arise because individuals do not perform the required attentional checks of the task environment at the time that deviation from routine is required.

We simulate this type of situation with a deferred handoff (primary) task in which participants were required to remember to handoff a target aircraft with a non-routine response key. There is reason to suspect interruptions might exacerbate habit capture in simulated ATC. For example, in order to link relevant environmental cues to the corresponding requirement to deviate from routine actions, individuals may create a problem state (Salvucci & Taatgen, 2011) by storing information about their intentions to deviate from routine along with information associated with the task goal on the display. Past studies examining deferred tasks in simulated ATC have found that deferred tasks impair performance on the ongoing tasks, indicating that effort is required to maintain the problem state associated with the deferred task goal (for review see, Loft, 2014). It follows that if participants are interrupted and no longer able to monitor the display, their ability to maintain the problem state associated with the deferred task of deviating from routine could be hindered. It is also possible that interruptions may negatively impact habit capture by increasing workload. In an ATC simulation experiment, Stone, Dismukes, & Remington (2001) found that participants were less likely to remember to deviate from a routine procedure when workload was higher (measured by number of aircraft on the screen), and interruptions have been found to increase subjective workload (Adamczyk & Bailey, 2004; Keus van de Poll & Sörqvist, 2016).

On the other hand, interruptions might not necessarily increase habit-capture. Although basic studies have found that interruptions can decrease PM performance (Dodhia & Dismukes, 2009; McDaniel et al., 2004); this is not the case when the task context can be easily reinstated through the provision of contextual cues (Cook et al., 2014). The ATC display provides rich
information at encoding regarding the probable future context of each deferred handoff event, and this may allow participants to retrieve their intention to deviate from routine “only as needed, in a just-in-time manner” (Braver, 2012). Individuals are more likely to remember to perform deferred task actions if they have been associated with specific ongoing task contexts (Bowden, Smith, et al., 2017; Cook et al., 2014; Nowinski & Dismukes, 2005), including in simulated ATC (Loft, Finnerty, et al., 2011). To the extent this is also the case in the current experiments, we would not expect an interruption during the retention interval (between encoding intention and the appropriate time to deviate from routine) to impact the probability of habit-capture because the interruption would not influence the extent to which relevant information is available on the display at the time for retrieval of the deferred action.

3.6. Experiment 1

We examined the effect of interruptions on two forms of deferred tasks: remembering to resume a deferred conflict task and remembering to deviate from a routine handoff procedure. Participants assumed the role of an air traffic controller responsible for maintaining the safety of aircraft by accepting aircraft entering the sector, detecting and resolving aircraft conflicts, and handing-off aircraft exiting the sector. There were three within-subjects conditions: no-interruption, blank interruption, and \( n \)-back interruption. All interruptions lasted for 27 s, occluded the display, and were presented without warning. The blank interruption was a blank screen. The \( n \)-back interruption comprised a visual numerical 2-back task, with 15 random single-digit numbers (2 s duration each; participant pressed the space bar when the stimuli matched the stimuli that appeared two previously). The no-interruption condition served as the baseline comparison. In this condition, participants were not interrupted during the ATC trial.

3.6.1. Method

3.6.1.1. Participants

Sixty undergraduate students (38 females; median age = 20) from the University of Western Australia participated in the study in exchange for partial course credit or $25 AUD and were tested in groups between one and five.
3.6.1.2. ATC-Lab\textsuperscript{Advanced} Simulator

Figure 3.1 presents a screenshot of the ATC task (Fothergill, Loft, & Neal, 2009). The light grey polygon area is the designated sector, whilst the dark grey area represents flight sectors outside of the participants’ control. The black lines denote flight paths that aircraft travel. Aircraft are represented by a circle with a leader-line indicating heading. The aircraft data-blocks specify the aircraft’s call sign, speed, type, current altitude, and cleared altitude (the altitude an aircraft is cleared to climb to, descend to, or cruise at). Cleared altitude and current altitude are separated by an arrow that denotes whether the aircraft is climbing (\textasciitilde\textasciitilde), descending (\textu2193), or cruising (\textgreater\textgreater). Aircraft enter the airspace from the edges of the display, cross sector boundaries, and then exit the display. New aircraft continue to appear throughout the trial, with aircraft positions updated every second (behavioural measures are recorded with millisecond precision). The timer in the lower centre of the screen showed how much time had elapsed in the trial and was updated each second.

When aircraft approached the sector, they flashed for acceptance and participants had to accept aircraft by clicking the aircraft icon and pressing the A key within 20 s. As aircraft exited the sector boundary, they flashed for handoff. Participants had to handoff aircraft by clicking the aircraft icon and pressing the H key within 15 s. Aircraft conflicts occurred when an aircraft pair simultaneously violated both lateral (5 nautical miles) and vertical (1,000 feet) separation standards. Participants were required to detect and prevent conflicts from occurring by selecting aircraft in future conflict and changing their cleared altitude to prevent that conflict from occurring. Participants were awarded points for successfully completing tasks, whilst points were deducted for failure to complete tasks. The current score was continuously updated and displayed in the middle right of the display. Ten points were awarded or deducted for a successful/failed handoff/acceptance. Between 10 and 40 points were awarded for resolving a conflict, depending on how quickly participants resolved the conflict, and 40 points were deducted for failing to resolve a conflict. Forty points were also deducted for unnecessary aircraft interventions (i.e., conflict detection false alarms).
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

![ATC Display Diagram]

Figure 3.1: The ATC display. Inbound aircraft are black (e.g., aircraft C77) as they approach the sector, and flash orange for acceptance (e.g., C31) when they reached within 5 miles of the sector boundary. Aircraft turn green (e.g., C37) when accepted. When outbound aircraft cross the sector boundary they flash blue (e.g., C15), and then turn white (e.g., C13) when handed off. Aircraft turn yellow (e.g., C19 & C35) if they violated the minimum vertical and lateral separation. The example in the figure shows that the individual is required to change the altitude of C82 from 430 to an alternative altitude to avoid a conflict with C68. Previously the individual had failed to change the altitude of either C19 or C35 (illuminated in yellow). The running score (10 points) is presented in the middle right hand side of the display.

3.6.1.3. Training Phase

The training phase comprised three tasks: a 10 min audio-visual ATC tutorial; two practice n-back trials of 45 s duration; and three 5 min practice ATC trials. Each participant completed one practice trial in each of the three conditions. Results from the training phase (n-back and air traffic control practice trials) were screened prior to the test phase to ensure participants understood the task. To demonstrate competence, participants had to perform both the deferred tasks correctly at least once each and achieve 75% accuracy on the n-back task. Twenty percent of participants repeated the n-back task.
3.6.1.4. **Test Phase**

The test phase comprised 15 five-minute trials, with five trials per within-subject condition (no-interruption, blank interruption, and n-back interruption). A Latin rectangle scheme was used to counterbalance the 15 trials across the three conditions. To create the scheme, trials were divided (with random assignment) into three equal sized groups, and participants into six equally sized groups. The trial groups were counterbalanced across a $3 \times 6$ Latin rectangle, where columns assigned a condition to a group of trials; and rows allocated a counterbalancing scheme to a group of participants. This resulted in each of the three conditions being associated equally with each of the 15 ATC trials. The order in which the trials were presented was randomized for each participant.

Each trial comprised a unique set of aircraft and trials differed with respect to the timing and location of events (e.g., conflicts occurred at different times and locations; interruptions began at different times). However, trials had a similar number of ongoing tasks (three conflicts to resolve; between 13 and 19 acceptances; and between 8 and 14 handoffs) and the basic design of the two deferred tasks was consistent amongst trials. The timing of the two deferred tasks was anchored around the programmed ‘interruption start point’ on each trial (90s to 148s). Every trial began with a period of ongoing ATC tasks (acceptances, handoffs, and conflict detection) with no deferred task requirements.

For the deferred handoff task, approximately 60 s into each trial, a message box would appear adjacent to one aircraft instructing participants to handoff that aircraft with an arrow key that corresponded to the aircraft heading (e.g., ⬆), instead of the routine ‘H’ key. This message was displayed for 10 s, and participants had to acknowledge it by clicking an “Acknowledge” button. The button only became clickable after 3 s from its initial display to prevent accidental acknowledgement. All messages automatically disappeared if they were not acknowledged within 10 s.

For the deferred conflict task, an aircraft would begin a climb or descent to an altitude that would result in a future conflict with another cruising aircraft. The climb/descent began at approximately the same time the deferred handoff acknowledge button timed out if not previously acknowledged. Participants were instructed that if any aircraft in a conflict pair was changing altitude, then the altitude change functionality would be disabled for both aircraft in
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

the conflict pair until both aircraft were cruising. Twenty seconds prior to the 'interruption start point' a message box appeared on the display adjacent to the climbing/descending aircraft instructing the participant to remember to resolve the conflict at the point that both aircraft were cruising in the future. This message also had to be acknowledged within 10 s or it timed out.

At the ‘interruption start point’ of each trial, which varied between trials, participants were either interrupted, or on no-interruption trials, continued ongoing air traffic management. For \( n \)-back interruption trials, the display would be occluded and participants were presented with a fixation point (+) for 3 s, followed by 24 s of the \( n \)-back task in which participants were presented with a series of 15 random single-digit numbers (2 s duration each) and were required to press the space bar when the digit matched the digit that appeared two previously (i.e., 2-back). For the blank interruption trials, the display was occluded by a filled black mask for 27 s. Throughout the interruption interval, aircraft continued to move and the simulator automated aircraft handoff and acceptance. No aircraft were programmed to violate separation during the interruption. After 27s (the end of the interruption), both aircraft in the deferred conflict were cruising. Participants then had between 11 and 23 s to resolve the conflict (this time varied from trial to trial). It is critical to highlight that the deferred-conflict could only be resolved immediately after the interruption or at the equivalent time point in the no-interruption condition. This allowed for a direct comparison of performance between the no-interruption and the two interruption conditions. Approximately 60 s later, the deferred-handoff aircraft would flash for handoff and participants would be required to remember to press the correct arrow key in place of the routine 'H' key.

After each trial, participants answered two workload questions on a 10-point scale ("How mentally demanding was the task during the last test trial?" and "How hard did you have to work to accomplish your level of performance in the last test trial?"). There was a 30 s break between trials, except after the 7th trial in which there was a 5 min rest break.

3.6.2. Results

Data from the one participant whom did not perform any of the deferred handoffs correctly was not analysed. Significance was set at an alpha level of .05. Unless otherwise specified, the data was analysed using two planned contrasts that directly paralleled our hypotheses (Rosenthal &
Executing Deferred Tasks in Dynamic Multitasking Environments

Rosnow, 1985): (1) blank interruption vs no-interruption, and (2) n-back interruption vs blank interruption. Effect sizes are estimated using Cohen’s $d$ (small = 0.3, medium = 0.5, large = 0.8; J. Cohen, 1988).

All analyses are also reported with associated Bayes Factors (BFs), and BFs are used as the primary framework for inference. Bayesian statistics avoid some pitfalls associated with traditional null hypothesis significance testing (Kass & Raftery, 1995; Wagenmakers, 2007). In particular, BFs quantify the relative evidence favouring the null versus the alternative hypothesis. BFs were computed using the Bayes Factor Package (Morey, Rouder, Love, & Marwick, 2015) for the statistical software, R (R Core Team, 2017). Bayesian analyses were conducted using Bayesian paired samples $t$-tests with a medium Jeffreys-Zellner-Siow prior width of $\sqrt{2}/2$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009) and were interpreted using Jeffrey’s (1998) guidelines with adjustments by Andraszewicz et al. (2015). BFs between 1 and 3 indicate “anecdotal” evidence, Bayes factors greater than 3 indicate moderate evidence, Bayes factors greater than 10 are strong evidence, and Bayes factors greater than 30 are very strong evidence for a given hypothesis relative to an alternative. Bayes factors are represented as $BF$, and the subscript indicates whether the model comparison is expressed as favouring the alternative hypothesis ($BF_{01}$) or the null ($BF_{10}$). To test the extent to which BFs for the main hypotheses varied as a function of prior width, Bayes Factor robustness check plots were inspected for the deferred conflict (resumption time and errors) contrasts and deferred handoff (errors only) contrasts. These revealed that interpretations of our Bayes Factors remained stable over a range of priors (see Appendix for further details).

3.6.2.1. Deferred Conflict Task

Participants acknowledged all the deferred conflict task instructions. Resumption time was defined as the time taken to change the altitude of one of the two aircraft in conflict after the point that both aircraft were cruising. A resumption failure occurred if a participant failed to resolve the conflict before the aircraft violated separation. Mean resumption time and resumption failure proportions are presented in Figure 3.2.
3.6.2.2. Deferred Handoff Task

The deferred handoff instruction was not acknowledged on one trial. However, this was not excluded from the analysis as the participant still performed the deferred handoff correctly. We defined a habit-capture as pressing the routine handoff key instead of the instructed alternative arrow key for a target aircraft. Deferred handoff response time (RT) was the time taken to respond to the target aircraft on trials in which the correct PM response was made. No errors of
Executing Deferred Tasks in Dynamic Multitasking Environments

omission were made, that is, all target aircraft were handed off, either correctly (using the correct arrow key: PM response) or incorrectly (using the H key or an incorrect arrow key). False alarms (pressing the arrow key on non-target aircraft) were made to less than 0.5% of aircraft and there was moderate evidence that this did not differ between conditions (smallest $p = .50$, $BF_{01} = 6.09$). PM response execution errors (remembering to press an arrow key but pressing the incorrect key) were made to 3.06% of target aircraft, and there was anecdotal evidence that this did not differ between conditions (smallest $p = .14$, $BF_{01} = 2.71$). PM response execution errors were excluded from the analysis, but the pattern of results reported below did not differ when response execution errors were included. Habit-capture rates and correct response times for each condition are presented in Figure 3.3.

![Figure 3.3: Mean habit-capture rate and RTs across the three interruption conditions for the deferred handoff task. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).](image)

There was moderate evidence that the proportion of habit-capture on the deferred handoff task did not differ between the blank ($M = 29.4\%$, $SD = 26.5\%$) and no-interruption conditions ($M = 26.9\%$, $SD = 25.1\%$), $t(58) = 0.78$, $p = .44$, $d = 0.10$, $BF_{01} = 5.27$, or between the n-back ($M = 30.8\%$, $SD = 25.7\%$) and blank conditions, $t(58) = 0.34$, $p = .73$, $d = 0.04$, $BF_{01} = 6.64$. There was also moderate evidence that deferred handoff RT did not differ between the blank condition ($M = 2645$ ms, $SD = 1047$ ms) and the no-interruption condition ($M = 2711$ ms, $SD = 1115$ ms), $t(55) = -0.41$, $p = .69$, $d = -0.05$, $BF_{01} = 6.34$, or between the n-back ($M = 2517$ ms, $SD = 883$ ms) and blank conditions, $t(55) = -0.87$, $p = .39$, $d = -0.12$, $BF_{01} = 4.77$. 

84
3.6.2.3. **Subjective Workload**

There was strong evidence that the two subjective workload questions were highly correlated, $r(57) = .96, p < .001, BF_{10} = 1.81E+29$ and they were therefore combined (Wetzels & Wagenmakers, 2012). There was moderate evidence that subjective workload did not differ between the no-interruption ($M = 5.11, SD = 1.74$) and blank condition ($M = 5.05, SD = 1.8$), $t(58) = -0.70, p = .486, d = -0.09, BF_{01} = 5.55$; and anecdotal evidence for no difference between the n-back ($M = 5.2, SD = 1.77$) and blank conditions, $t(58) = 1.78, p = .081, d = 0.23, BF_{01} = 1.61$. Interruptions did not increase subjective workload, but this may be due to a central tendency bias in subjective reporting.

3.6.2.4. **Ongoing task performance**

The post-interruption duration varied between two and three minutes across trials. In order to examine whether interruptions had any other impact on ongoing ATC task performance, we assessed six measures of ongoing task performance during the post-interruption period: (1) acceptance RT (mean time taken to accept aircraft), (2) acceptance misses (mean number of failed aircraft acceptances), (3) handoff RT (mean time taken to handoff aircraft), (4) handoff misses (mean number of failed aircraft handoffs), (5) conflict detection time (mean time taken to resolve a conflict), and (6) conflict misses (mean number of aircraft that entered conflict). We did not compare pre-interruption performance to post-interruption performance as systematic scenario design differences between these time periods confound such comparison. For example, the pre-interruption period comprises acknowledgement messages and a period of SA acquisition as the participants have to familiarize themselves with the ATC scenario. Comparing performance in the post-interruption period across conditions was not confounded as our counterbalancing scheme ensured that any differences between scenarios during the post-interruption are controlled for. As can be seen in Table 3.1, there was no evidence of differences between the three conditions on any of the ongoing performance measures. This indicates that interruptions did not have a chronic impact on task performance, and participants were able to recover from the interruptions relatively quickly. The descriptive statistics associated with each condition can be obtained from the online repository.
Table 3.1. Grand means for all ongoing task performance measures during the post-interruption period, and associated contrast test results (df = 58).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Contrast</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
<th>BF01</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance RT (ms)</td>
<td>1</td>
<td>3360</td>
<td>801</td>
<td>1.43</td>
<td>.16</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>1.35</td>
<td>.18</td>
<td>2.96</td>
</tr>
<tr>
<td>Acceptance Misses (%)</td>
<td>1</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>.33</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
<td>.65</td>
<td>6.37</td>
</tr>
<tr>
<td>Handoff RT (ms)</td>
<td>1</td>
<td>2901</td>
<td>810</td>
<td>0.12</td>
<td>.90</td>
<td>6.97</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.64</td>
<td>.53</td>
<td>5.78</td>
</tr>
<tr>
<td>Handoff Misses (%)</td>
<td>1</td>
<td>0.04</td>
<td>0.04</td>
<td>0.51</td>
<td>.61</td>
<td>6.2</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.05</td>
<td>.96</td>
<td>6.2</td>
</tr>
<tr>
<td>Conflict Detection Time (s)</td>
<td>1</td>
<td>56.62</td>
<td>17.30</td>
<td>0.08</td>
<td>.94</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.8</td>
<td>.43</td>
<td>5.18</td>
</tr>
<tr>
<td>Conflict Misses (%)</td>
<td>1</td>
<td>0.13</td>
<td>0.18</td>
<td>0.36</td>
<td>.72</td>
<td>6.59</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.16</td>
<td>.87</td>
<td>6.94</td>
</tr>
</tbody>
</table>

Note: Contrast 1 = between n-back and blank; contrast 2 = between blank and no-interruption.

3.6.3. Discussion

Interruptions slowed deferred conflict resumption time. This suggests that after an interruption, it either took time for participants to retrieve the problem state required to reinstate the deferred task goal, or it took time to develop the sufficient SA after the interruption required to locate and resolve the deferred conflict. However, the moderate Bayesian evidence for no difference between the n-back and blank conditions on conflict resumption time suggests that opportunity for rehearsing the problem state during the interruption retention interval was not the primary factor underlying deferred conflict resumption. This provides some evidence that participants relied on interactions with the ATC display following interruptions as opposed to storing problem states prior to the interruption.

The 30% habit-capture error rate is consistent with previous work demonstrating participant vulnerability to habit-capture in simulated ATC (Loft, 2014). It is likely that habit captures occurred because individuals failed to perform the required attentional checks of the ATC task environment when deviation from routine was required (Norman, 1981; Reason, 1990). However, moderate Bayesian evidence in favour of the null indicated that interruptions
did not further increase habit-capture. This suggests that either individuals did not actively maintain the intent to deviate from routine over the deferred task retention interval, or that the interruption conditions used in Experiment 1 failed to impede the maintenance of the intent.

3.7. Experiment 2

Taken together, the empirical outcomes of Experiment 1 suggest that rather than storing and rehearsing the task problem state and pre-interruption ATC display scene, individuals relied on interactions with the display during the post-interruption period to reconstruct the problem state and information required to perform deferred task goals. However, it may be the case that individuals did store, rehearse, and retrieve problem states but that the n-back task failed to impede that process because it was too dissimilar to the ATC task. Interference accounts of working memory capacity posit that our ability to concurrently hold several memory representations is limited by the mutual interference between these representations (Hurlstone, Hitch, & Baddeley, 2014; May, Hasher, & Kane, 1999; Oberauer et al., 2016). As the similarity between information cues increase, interference between associated memory items occurs within working memory (Bunting, 2006; Norman, 1981). Several studies in the interruptions literature have found that when an interrupting and primary task share high visual similarity or have similar goals or required task actions, resumption time and post-completion errors increase (Borst et al., 2010; Edwards & Gronlund, 1998; Gillie & Broadbent, 1989; Ratwani & Trafton, 2008). Indeed, both MFG and Threaded Cognition specify an interference mechanism, in addition to a decay mechanism, to account for the disruptive effects of interruptions (Altmann & Trafton, 2002; Borst et al., 2010; Salvucci & Taatgen, 2011). On this logic then, an interruption that is more similar to the primary ATC task may be more likely to interfere with stored memory associations.

To test this possibility, in Experiment 2 we replaced the n-back task with a separate ATC sector that participants had to monitor and control during the interruption period. We expected this interrupting ATC task might interfere with participants’ memories for pre-interruption locations and associated episodic task goals (problem state) due to the overlap in the task demands and visuo-spatial memory representations. For example, the interrupting ATC task included conflicts that needed to be resolved, and aircraft that need to be handed-off, in similar locations as the deferred conflict and handoff tasks in the primary ATC task.
Executing Deferred Tasks in Dynamic Multitasking Environments

We expected to replicate the effect of interruption on conflict resumption time when comparing the blank condition to the no-interruption condition. Furthermore, to the extent that problem states are stored and retrieved, we expected slower resumption time in the ATC interruption condition compared to the blank condition. In contrast, if individuals rely primarily on reconstruction strategies, the ATC interruption condition should not differ from the blank condition, replicating the results from the n-back condition in Experiment 1.

For the deferred handoff task, Experiment 1 provided reasonably strong evidence that interruptions do not impact habit-capture. If participants rely on post-interruption contextual cues to prime deferred task goals, then it is unlikely that even an ATC task interruption that causes interference would increase habit-capture. This point notwithstanding, the ATC interruption task provides a stronger test of the alternative hypothesis that some form of internal cognitive control, along with information about the information on the ATC display associated with that task goal, is required to maintain the intention to deviate from routine.

3.7.1. Method

3.7.1.1. Participants
Sixty undergraduate students (33 female; median age = 20) from the University of Western Australia participated in the study in exchange for partial course credit.

3.7.1.2. Materials and Procedure
Experiment 2 differed from Experiment 1 in three ways. Firstly, instead of an n-back interruption condition, participants were interrupted by a separate ATC scenario (ATC-interruption condition). Secondly, to ensure that the onset of the ATC-interruption was distinctive from the primary scenario and to avoid participant confusion about which sector they were currently controlling, a textbox was added to the top-left corner of the display indicating either ‘primary scenario’ or ‘interrupting scenario’. Thirdly, at the end of the test phase, participants completed a brief open-ended short-answer questionnaire that asked: “Describe how you felt when you were interrupted” and “Describe what (if any) strategies you used when you were interrupted”, and were asked “How did this differ between forms of interruptions you were exposed to (i.e., blank vs filled)?”
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

3.7.1.3. ATC-Interruption

The interrupting ATC scenario required participants to monitor an ATC sector which was displayed in place of the primary scenario sector (see Figure 3.4). The task objectives were identical to those of the ongoing ATC task. Each interrupting ATC scenario comprised two or three acceptances, two or three handoffs, and two conflicts requiring resolution. Twenty percent of conflicts violated minimum separation (i.e., turned yellow) during the interrupting period. Conflicts did not always violate minimum separation during the interrupting period because of the limited trial duration of 27 s. Participants were instructed that performance on the interrupting ATC scenario was of equal importance to the primary ATC task but that no deferred task demands would be presented during the interrupting ATC task. The timer normally on the primary ATC display was removed in the interrupting scenario display to ensure it was comparable to the blank and n-back conditions in which participants did not receive feedback regarding the duration of the interruption. There were five unique interrupting scenarios and the order in which they were presented was randomized for each participant.

![Figure 3.4. An example of an interrupting ATC scenario. The differences from the primary scenario to note are: the different sector boundaries (i.e., the shape), different flight paths, the “interrupting scenario” textbox, and the absence of a timer. The two conflicts are between ZF54 and GY58; and DE68 and UK34.](image)

Consistent with the interruption conditions in Experiment 1, the ATC-interruption duration was 27 s. The ATC-interruption comprised three temporal parts: firstly, a crosshair
was presented for approximately\textsuperscript{12} 2500 ms; next participants completed 24 s of ongoing ATC management with no special instructions; finally, participants were presented with a blank screen for approximately 500 ms to provide a brief visual buffer between the interrupting scenario and the primary scenario.

3.7.2. Results

Data from one participant was excluded from all analyses as this participant failed to perform any of the deferred handoff tasks correctly.

3.7.2.1. Deferred Conflict Task

The deferred conflict instruction was not acknowledged on one trial, but this was not excluded from the analysis as the participant correctly resolved the deferred conflict. Mean resumption time and resumption failure proportions for each condition are presented in Figure 3.5.

Figure 3.5. Mean resumption times and failure proportions across the three interruption conditions for the deferred conflict task. Error bars represent 95\% within-subjects confidence intervals (Cousineau, 2005).

\textsuperscript{12}This is an approximation due to minor random variability in the time it took to load the interrupting scenario in the order of up to 100-200 ms. This did not affect interruption end time.
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

There was strong evidence that conflict resumption time was slower in the blank condition \((M = 4501 \text{ ms}, SD = 1789 \text{ ms})\) than in the no-interruption condition \((M = 2369 \text{ ms}, SD = 1214 \text{ ms})\), \(t(58) = 9.93, p < .001, d = 1.28, BF_{10} = 1.84 \times 10^{11}\). There was anecdotal evidence of no difference in conflict resumption time between the ATC \((M = 4951 \text{ ms}, SD = 1937 \text{ ms})\) and blank interruption conditions, \(t(58) = 1.58, p = .12, d = 0.20, BF_{01} = 2.17\). There was moderate evidence that resumption failures were not more likely in the blank condition \((M = 3.4\%, SD = 8.4\%)\) than in the no-interruption condition \((M = 2.4\%, SD = 7.5\%)\), \(t(58) = 0.69, p = .49, d = 0.09, BF_{01} = 5.62\); but there was strong evidence that resumption failures were more likely in the ATC condition \((M = 10.8\%, SD = 14.5\%)\) than in the blank condition \(t(58) = 3.55, p < .001, d = 0.46, BF_{10} = 33.84\).

3.7.2.2. Deferred Handoff Task

Participants failed to acknowledge the deferred handoff instruction on 0.88\% of trials. However, such trials were not excluded from the analysis as in all cases the deferred handoff was still performed correctly. Overall, 0.66\% of responses were errors of omission in which the aircraft was not handed off at all, but there was moderate evidence that this did not differ between conditions (smallest \(p = .79, BF_{01} = 6.81\)). Participants made false alarms to less than 0.5\% of non-target aircraft and there was anecdotal evidence that this did not differ between conditions, (smallest \(p = .09, BF_{01} = 1.86\)). PM response execution errors (remembering to press an arrow key, but pressing the incorrect key) were made to 3.06\% of target aircraft, and there was moderate evidence that this did not differ between conditions (smallest \(p = .82, BF_{01} = 5.02\)). The habit-capture rates and correct RT means are presented in Figure 3.6.

There was moderate evidence that the proportion of habit-capture on the deferred handoff task did not differ between the blank \((M = 34.7\%, SD = 29.6\%)\) and no-interruption conditions \((M = 37.9\%, SD = 30.6\%)\), \(t(58) = -0.86, p = .39, d = -0.11, BF_{01} = 4.92\), or between the ATC \((M = 35.4\%, SD = 32\%)\) and blank conditions, \(t(58) = 0.18, p = .86, d = 0.02, BF_{01} = 6.91\). There was also moderate evidence that deferred handoff RT did not differ between the blank \((M = 2753 \text{ ms}, SD = 951 \text{ ms})\) and the no-interruption conditions \((M = 2918 \text{ ms}, SD = 1374 \text{ ms})\), \(t(49) = -1.03, p = .31, d = -0.14, BF_{01} = 3.94\), or between the ATC \((M = 2899 \text{ ms}, SD = 1268 \text{ ms})\) and blank conditions, \(t(49) = 0.80, p = .43, d = 0.11, BF_{01} = 4.8\).
### 3.7.2.3. Subjective Workload

There was strong evidence that the two subjective workload questions were highly correlated, \( r(57) = .96, p < .001, BF_{10} = 7.89E+17 \), and thus they were combined. There was moderate evidence that subjective workload did not differ between the no-interruption (\( M = 5.31, SD = 1.68 \)) and blank condition (\( M = 5.37, SD = 1.67 \)), \( t(58) = 0.75, p = .454, d = 0.10, BF_{01} = 5.36 \), or between the ATC (\( M = 5.45, SD = 1.71 \)) and blank conditions, \( t(58) = 0.83, p = .410, d = 0.11, BF_{01} = 5.06 \).

### 3.7.2.4. Ongoing Task Performance

Of all the ongoing task measures for the post-interruption period, only post-interruption aircraft acceptance response time was significantly faster following a blank interruption compared to no-interruption (mean difference = 124.5 ms), however there was only anecdotal Bayesian evidence favouring this effect. The results of all other ongoing task performance measures are reported in Table 3.2. The descriptive statistics associated with each condition can be obtained from the online repository.
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

Table 3.2 Grand means for all ongoing task performance measures during the post-interruption period, and associated contrast test results (df = 58).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Contrast</th>
<th>Mean</th>
<th>SD</th>
<th>t</th>
<th>p</th>
<th>BF_{01}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance RT (ms)</td>
<td>1</td>
<td>3262.17</td>
<td>772.89</td>
<td>0.98</td>
<td>.33</td>
<td>4.45</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-2.45</td>
<td>.018</td>
<td>0.46</td>
</tr>
<tr>
<td>Acceptance Misses (%)</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.04</td>
<td>.30</td>
<td>4.18</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.08</td>
<td>.94</td>
<td>7.46</td>
</tr>
<tr>
<td>Handoff RT (ms)</td>
<td>1</td>
<td>3313.54</td>
<td>842.79</td>
<td>-1.65</td>
<td>.10</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.33</td>
<td>.74</td>
<td>6.67</td>
</tr>
<tr>
<td>Handoff Misses (%)</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>0.24</td>
<td>.81</td>
<td>6.84</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-0.72</td>
<td>.47</td>
<td>5.48</td>
</tr>
<tr>
<td>CDT (s)</td>
<td>1</td>
<td>54.12</td>
<td>16.39</td>
<td>-0.49</td>
<td>.63</td>
<td>6.26</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>0.28</td>
<td>.78</td>
<td>6.77</td>
</tr>
<tr>
<td>Conflict Misses (%)</td>
<td>1</td>
<td>9</td>
<td>15</td>
<td>0.49</td>
<td>.63</td>
<td>6.26</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-0.8</td>
<td>.43</td>
<td>5.18</td>
</tr>
</tbody>
</table>

Note: Contrast 1 = between ATC and blank; contrast 2 = between blank and no-interruption.

3.7.2.5. Interrupting ATC task performance

Performance across the five ATC-interruption trials was aggregated for each participant. Three aircraft which flashed for acceptance in the last 2 s of two trials were excluded from analysis. Table 3.3 presents the means and standard deviations of ongoing task measures. While the proportion of missed conflicts, accepts and handoffs was considerably higher during the ATC-interruption trials than on the primary ATC scenarios, it is important to note that due to the abrupt onset and short duration of the interrupting scenario, this difference is not likely indicative of reduced task effort. In primary scenarios participants had considerably more time to detect conflicts before they violated minimum separation ($M = 82.28$ s) and had more time to develop their SA of the sector. Additionally, the conflict performance measures reported in Table 3 include conflicts that did not violate minimum separation during the interruption (i.e., did not turn yellow), therefore participants may have detected such conflicts if given more time.
Table 3.3. Means and standard deviations for ATC-Interruption performance on handoffs, acceptances, and conflict detection

<table>
<thead>
<tr>
<th>Variable</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Handoff Miss Proportion</td>
<td>6.39%</td>
<td>7.72%</td>
</tr>
<tr>
<td>Handoff Response Time</td>
<td>2691 ms</td>
<td>736 ms</td>
</tr>
<tr>
<td>Accept Miss Proportion</td>
<td>8.57%</td>
<td>6.55%</td>
</tr>
<tr>
<td>Accept Response Time</td>
<td>4124 ms</td>
<td>720 ms</td>
</tr>
<tr>
<td>Conflict Miss Proportion</td>
<td>45.67%</td>
<td>19.43%</td>
</tr>
<tr>
<td>Conflict Detection Time</td>
<td>14.57 s</td>
<td>2639 ms</td>
</tr>
<tr>
<td>Conflict False Alarms</td>
<td>0.17</td>
<td>0.59</td>
</tr>
</tbody>
</table>

3.7.3. Discussion

Experiment 2 replicated the finding of Experiment 1 that interruptions slowed deferred conflict resumption time. In addition, we found strong Bayesian evidence that individuals failed to detect more deferred conflicts following an ATC task interruption compared to a blank interruption. We further explore the nature of this effect below using Ex-Gaussian distribution modelling, but at a minimum, the increased resumption errors indicate that storing and retrieving the problem state plays a role in conflict detection task resumption.

In Experiment 1, there was a significant difference in conflict detection resumption failures between the blank and the no-interruption conditions, but the Bayesian evidence revealed anecdotal evidence for the null hypothesis. In Experiment 2, there was no significant difference between these two conditions for conflict detection resumption failures, and there was moderate Bayesian evidence in favour of the null hypothesis. Given that the blank and no-interruption conditions were identical across the two experiments (other than the presence of the "primary" scenario textbox), the resumption failure data for the blank and no-interruption conditions were combined for meta-analysis. There was no significant difference in resumption failures between the combined blank ($M = 4.58\%, SD = 9.93\%$) and combined no-interruption ($M = 2.54\%, SD = 7.64\%$) conditions, $t(117) = 1.92$, $p = .057$, $d = 0.18$, $B_{F_{01}} = 1.66$, and the Bayesian evidence favoured the null hypothesis with anecdotal support. Thus, there is no clear
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

evidence one way or the other as to whether the blank interruption increased resumption failures over and above the no-interruption condition.

In Experiment 2, we replicated the moderate Bayesian evidence from Experiment 1 that interruptions do not increase habit-capture. Either participants were not actively maintaining the intent to deviate from routine over the deferred task retention interval, or both the $n$-back and ATC interrupting tasks failed to impede the maintenance of that intent. We return to these possibilities in the General Discussion.

3.8. Ex-Gaussian Distribution Modelling

Resumption failures and resumption time are not independent measures of performance. A resumption failure occurs when a participant fails to ‘resume’ the deferred conflict task (i.e., fails to redetect and/or resolve the conflict) before violation of minimum separation standards (herein, referred to as the cutoff time, which varied between 11–23 s post interruption). It is possible that some participants would have eventually remembered to resolve the deferred conflict, given more time. If so, the increased resumption failures in the ATC condition may reflect a delayed retrieval of the problem state, rather than complete forgetting of the deferred conflict task goal. We can adjudicate between these possibilities by inspecting and modelling the entire deferred conflict RT distribution (i.e., resumption time distribution) and assessing whether the probability of correctly resolving the deferred conflict has plateaued prior to the cutoff time (consistent with forgetting), or whether the cutoff has effectively censored some slower responses that would otherwise have been made if there was further time available to resolve the conflict. In addition, consideration of the entire distribution of RTs allows us to more specifically identify the nature of any differences between all the conditions (Ratcliff, 1979; Wixted & Rohrer, 1994).

Figure 3.7 plots the empirical distribution function for deferred conflict RTs in the four unique conditions in Experiment 1 and 2. For each time, the plot shows the probability that the deferred conflict has been resolved. Given the data were right-censored (RTs are not observed for responses that did not occur prior to the cutoff), the functions were estimated using the Kaplan-Meier estimator (Kaplan & Meier, 1958), using the ‘survival’ package in R (Therneau, 2015). The crosses denote individual trials that were censored; that is, they indicate the cutoff times for cases where the conflict was not resolved prior to the cutoff. The crosses indicate that
Executing Deferred Tasks in Dynamic Multitasking Environments

in some situations the cutoff may have prevented a slower response from being executed as the
distribution functions continue to gradually rise to the right of the earlier (left-most) cutoffs.
However, that rise is gradual, and it is not clear whether it can fully explain the differences in
failure probabilities between the conditions.

We modelled the data with an exponentially modified Gaussian distribution. The ex-
Gaussian distribution is frequently used to model RT distributions (Balota & Yap, 2011; Luce,
1986; Ratcliff, 1979) and recall times (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994) in
psychology. Here, we use the ex-Gaussian model in a manner similar to its application to free
recall (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). We assume that participants constantly
monitor the display, and in doing so have a constant probability of resolving the deferred
conflict. If it is assumed that locations in the display are sampled with replacement, then an
exponential distribution of resumption times will be observed. In addition, if the time for non-
decisional processes outside of this sampling process (e.g., responding to the deferred conflict
when detected) are normally distributed, the resulting distribution is ex-Gaussian in nature. We
model complete resumption failure as the asymptote of the cumulative ex-Gaussian distribution
function, which is estimated as a separate parameter in the model.
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

Figure 3.7. Cumulative RT function for the four unique conditions in Experiment 1 and 2. The lines plot out the Kaplan-Meier estimator of the cumulative density for right-censored data and the crosses indicate the censored data (deferred conflict resumption failures).

The ex-Gaussian model was applied to latencies from all four unique conditions in both experiments, and parameters were estimated using ‘Stan’ (Carpenter et al., 2017), which uses the no-U-turn sampler to obtain Bayesian posterior distributions on parameters. Critically, the truncated responses were fit using a censored distribution: the known cutoff for each trial was used to estimate the probability, under the model, that a resumption time would exceed that cutoff for those trials where the conflict was not resolved, and the resulting probability entered in the likelihood calculation. Specifically, the log-likelihood was given by:

\[
\ln L = \sum_{i=1}^{N_c} \ln(1-p_f) + \log f(y_i|\mu, \sigma, \lambda) + \sum_{j=1}^{N_e} \ln(p_f + (1-p_f)(1-F(cutoff_j, \mu, \sigma, \lambda)),
\]

where \(N_c\) is the number of correct responses on the deferred conflict task, \(N_e\) is the number of resumption failures, \(f\) is the ex-Gaussian density function, and \(F\) is the cumulative distribution function for the ex-Gaussian. The parameter \(\mu\) is the shift of the ex-Gaussian function, and captures how long it takes for the cumulative functions in Figure 3.7 to kick up from 0. The \(\sigma\) parameter is the standard deviation of the Gaussian component of the ex-Gaussian, and is
treated here as a nuisance parameter not of theoretical interest. The $\lambda$ parameter is the rate parameter of the exponential component of the ex-Gaussian, and captures the rate of increase towards the asymptote in Figure 3.7. Finally, $p_f$ is a parameter capturing the probability of complete resumption failure, represented by the asymptote of the empirical cumulative distribution functions shown in Figure 3.7. As explained earlier, for each of the observed resumption failures, it is unclear whether the intention to resume the deferred conflict task was forgotten, or whether a correct conflict resolution that would have occurred was prevented from occurring by the imposition of a cutoff for that trial. The right term of Equation 1 models and weights each of these possibilities: either a resumption failure was due to complete forgetting (with probability $p_f$), or was due to early termination, obtained by calculating the area under the tail of the ex-Gaussian lying above the cutoff for that trial.

As there was a limited number of observations per participant, all observations in a condition were fit using a common set of parameters (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). The data were fit using the rstan package (Stan Development Team, 2016) in R (R Core Team, 2017). Weakly informative priors were specified for $\mu$ ($N(0, 10)$), $\lambda$ ($U(0, 10)$), and $p_f$ ($U(0, 1)$). A stronger prior was placed on $\sigma$ ($N(0, 0.4)$) to prevent inflated estimates $\sigma$. The parameters $\sigma$ and $\lambda$ were bounded at 0, and $p_f$ was restricted to lie between 0 and 1.

Table 3.4 gives the posteriors for each of the parameters by condition. The top part of the table shows that the estimates of $p_f$ correspond to the empirical frequencies of resumption failures plotted in Figure 3.7. In other words, the model fits suggest that the empirical distributions plotted in Figure 3.7 have reached asymptote, and that resumption failures in our experiments were almost solely due to participants forgetting to return to the deferred conflict task (rather than not having time to resolve the deferred conflict). The $\mu$ estimates suggest that participants were slower to initiate responding to the deferred conflict in the three interruption conditions compared to the no-interruption condition, and that this slowing was greater for the ATC and $n$-back interruption conditions compared to the blank interruption condition. Finally, the $\lambda$ estimates at the bottom of Table 3.4 show a higher sampling rate in the no-interruption condition, with little evidence of difference between the three interruption conditions. In other words, at each time step, if the deferred conflict task had not yet been resolved on that trial, the conflict was more likely to be resolved in the no-interruption condition, producing lower (faster) and less variable resumption times.
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

Table 3.4. Summary statistics of Bayesian posteriors of the ex-Gaussian parameters $\mu$, $\sigma$, and $\lambda$ and the probability of failure $p_f$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Condition</th>
<th>Mean</th>
<th>.025 quantile</th>
<th>.975 quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_f$</td>
<td>None</td>
<td>0.97</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>Blank</td>
<td>0.95</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>$n$-back</td>
<td>0.96</td>
<td>0.94</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>0.90</td>
<td>0.86</td>
<td>0.93</td>
</tr>
<tr>
<td>$\mu$</td>
<td>None</td>
<td>1.04</td>
<td>1.01</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>Blank</td>
<td>1.87</td>
<td>1.80</td>
<td>1.96</td>
</tr>
<tr>
<td></td>
<td>$n$-back</td>
<td>2.38</td>
<td>2.24</td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>2.09</td>
<td>1.93</td>
<td>2.25</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>None</td>
<td>0.08</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Blank</td>
<td>0.23</td>
<td>0.17</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>$n$-back</td>
<td>0.28</td>
<td>0.17</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>0.31</td>
<td>0.20</td>
<td>0.46</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>None</td>
<td>0.66</td>
<td>0.60</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Blank</td>
<td>0.38</td>
<td>0.35</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>$n$-back</td>
<td>0.40</td>
<td>0.03</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>0.33</td>
<td>0.02</td>
<td>0.38</td>
</tr>
</tbody>
</table>

3.9. General Discussion

We conducted two experiments investigating how interruptions impact deferred primary task performance in simulated ATC. We examined two forms of deferred tasks which represented tasks that controllers perform in-situ: remembering to resume a task at a later point in time and remembering to deviate from routine. The empirical findings of the present study can be summarized as follows. The blank interruption slowed the time taken to resume and resolve the deferred conflict. In Experiment 1 we did not find the more demanding $n$-back interruption increased resumption time over and above the blank interruption. In Experiment 2, we also did not find evidence favouring a difference in resumption time between the ATC-interruption and
the blank interruption; however, the ATC-interruption was associated with an increased likelihood of resumption failures — that is, forgetting to resolve the deferred conflict altogether — compared to the blank condition. On the deferred handoff task, we replicated findings from previous work that individuals are vulnerable to habit-capture in simulated ATC (Loft, 2014). However, habit-capture was not affected by interruptions. Interruptions also did not affect post-interruption ongoing task performance or ratings of subjective workload.

3.9.1. Resuming Interrupted Tasks
Controllers must sometimes defer task actions and remember to resume them at the appropriate point in the future (Shorrock, 2005). We simulated this situation with a deferred conflict task that required participants to acknowledge a temporarily unresolvable conflict, and then to remember to resolve it at the appropriate time. We expected that an interruption which occluded the display would increase the time taken to remember and resolve the deferred conflict and possibly increase the probability of resumption failure — that is forgetting the task goal altogether. We reasoned that this would occur because individuals would no longer have access to the ATC display to prime the deferred task goal and because individuals would require some time to re-develop SA for the post-interruption display. In line with this, the data from both experiments revealed strong evidence that the time to resolve the deferred conflict was slower under blank interruption conditions than when not interrupted. This suggests that participants used the information in the display to perform the deferred task goal, and individuals required some time to re-develop SA for the post-interruption display after an interruption. This finding supports Salvucci’s (2010) argument that in complex tasks, an important factor underlying interruption recovery is the time required for deliberate and strategic reconstruction of the task environment (i.e., problem state).

However, the meta-analysis revealed no clear evidence one way or the other as to whether the blank interruption increased conflict resumption failures. One possible reason for this may be because there was substantial variability in how individuals utilized the blank interruption time interval. We had anticipated \textit{a priori} that the blank interruption would provide individuals the opportunity to rehearse their deferred conflict task goal; however, the results from our post-experiment open-ended survey revealed there was considerable variability in participants’ strategic response to the blank interruption. For instance, 37.5% of participants reported finding the blank interruption to be restful, enjoyable, or quite manageable (compared
to only 4.36% for the ATC-interruption); whilst 58% of participants reported feelings of anxiety, stress, or frustration (compared to 62.50% for the ATC-interruption).

We also expected that an interruption that placed significant cognitive demands on participants would further slow conflict task resumption, because it would impede rehearsal of the problem state and the deferred task goal (Altmann & Trafton, 2002; Borst et al., 2010). However, in Experiment 1, resumption time did not further increase following a cognitively demanding \( n \)-back interruption, relative to the blank interruption where individuals were presumably free to rehearse deferred task goals. We reasoned this null effect might have occurred because individuals were not storing or rehearsing the task problem state and pre-interruption ATC display scene. However, the fact that the ATC-interruption increased resumption failures in Experiment 2 suggested instead that individuals did store and actively maintain representations of the ATC task problem state, but that the \( n \)-back task failed to impede this process. This contrasts with previous research that found an \( n \)-back task interruption increased resumption time and errors relative to unfilled or undemanding interruptions on a VCR programming task (Cades et al., 2008; Monk et al., 2008).

One possible explanation for this discrepancy is that more basic static tasks may be more sensitive to the effects of rehearsal inhibition because the problem-state can be more effectively stored in memory and task resumption can be achieved with fewer contaminant cognitive operations (Salvucci & Taatgen, 2011). In comparison, resumption in the more complex and dynamic ATC task involves prioritizing attention across several competing goals whilst also reacquiring SA of the updated display scene. This increase in the number of cognitive operations and problem-state complexity in ATC likely limits the effectiveness of memory-based rehearsal strategies, and this increased range of possible resumption strategies could mean that isolating the influence of rehearsal on a given primary task resumption process is more difficult.

In Experiment 2, while there was no clear evidence as to whether the ATC-interruption increased resumption time compared to the blank condition or not, there was strong evidence that it increased resumption failures (by 7.46%). Modelling the entire distribution of RTs provided evidence that resumption failures were not simply due to a slowed resumption process, but were due to complete forgetting of the deferred intention. This finding extends basic
laboratory studies that have found that interruptions can sometimes cause forgetting to return to a primary task or execute deferred tasks (Dodhia & Dismukes, 2009; McDaniel et al., 2004).

Resumption failures constitute a substantial error, as they indicate that the contextual cues present in the environment (e.g., the presence of the conflict on the display) failed to prompt retrieval of the deferred task goal through associative memory cueing (Cook et al., 2014). Consistent with our motivation for Experiment 2, a likely explanation for increased resumption failures in the ATC-interruption (but not in the n-back condition) is that the visual-spatial memory representations and task goals were similar to those of the deferred conflict task (Borst et al., 2010; Bunting, 2006; Norman, 1981). This argument is based on interference accounts of working memory which posit that our capacity to concurrently hold several memory representations is limited by the mutual interference between representations (May et al., 1999; Oberauer et al., 2016). Alternatively, resource-based theories posit that the efficiency of a cognitive function is monotonically related to the amount of cognitive resources allocated to it (Ma, Husain, & Bays, 2014; Tombu & Jolicœur, 2003). As a result, tasks will interfere with each other to the extent that they require the same processing resource at the same time (Logie, 2009; Logie, Zucco, & Baddeley, 1990; Navon & Gopher, 1979; Ratwani & Trafton, 2008). Based on these accounts, it is possible that the increased resumption failures occurred not due to high similarity, but because the ATC-interruption required access to the same resources that are required to maintain the primary task goals or problem state (i.e., overlapping visual-spatial memory demands). Future research might attempt to tease apart the independent contributions of similarity and resource demands by examining the effect of a dissimilar interrupting task which placed significant demands on visual-spatial working memory, for instance, a visual-spatial n-back task.

3.9.2. Remembering to Deviate from Routine

Operators sometimes must remember to deviate from firmly reinforced behavioural routines. We simulated this situation with a deferred handoff task in which participants had to remember to handoff a target aircraft with a non-routine response key. Consistent with previous research examining deferred handoff tasks in simulated ATC (Loft, 2014), participants failed to remember to deviate from routine on a significant proportion of occasions (31% and 37.9% habit capture rate in Experiments 1 & 2, respectively). Habit-capture likely occurred because participants failed to adequately attend to the relevant features of the task at the time that
deviation from routine was required (Norman, 1981; Reason, 1990). Findings of PM costs to ongoing ATC tasks in prior research suggests that individuals may maintain some form of cognitive control over the deferred task retention interval in order to remember to attend to the ATC task features associated with the intention to deviate from routine (Loft, 2014). Thus, we reasoned interruptions could interfere with individual’s ability to maintain the problem state associated with the deferred task goal, particularly when the interruption was more cognitively demanding.

In contrast to this prediction, none of our interruptions, whether cognitively demanding or not, affected habit-capture. One possible explanation is that participants engaged in a cue driven, “just-in-time” reactive strategy, allocating their attention to the deferred handoff tasks only at the time that deviation was required (Braver, 2012). This strategy would have been particularly effective for the deferred handoff task because not only was the task associated with a specific future context, but the target aircraft flashed blue when the handoff was required which provided a clear cue regarding the exact context of the deferred task handoff event. This explanation is consistent with basic and applied PM research that has found that PM tasks are more likely to be completed if they are reinstated through the provision of contextual cues (Bowden, Smith, et al., 2017; Cook et al., 2014; Loft, Finnerty, et al., 2011). If the ability to deviate from routine is indeed impacted more by the attentional demands at retrieval compared to those over the PM retention interval, it will be important for future research to examine whether distractions or increased workload at the time the deviation from routine is required increases habit-capture (e.g., Stone et al., 2001).

An alternative possibility is that participants did proactively maintain the problem state associated with their intention to deviate from routine, but interruptions simply had no effect on their ability to do so. This could have well been the case because the deferred handoff task was performed a reasonably long time after the interruption had ended, thereby allowing participant’s sufficient time to re-engage with the deferred task goal. An important avenue for future research will be to examine the effects of manipulating the time that the act of deviating from routine is required, relative to the end of the interruption. Another possibility is that the intention to deviate from routine was forgotten during the interruption, but was also forgotten during the no-interruption condition at a similar rate. In any case, the implications of our findings are clear: overcoming a habitual response poses a significant cognitive challenge, and
the cognitive processes required to do this are not influenced by interruptions in this version of
the ATC task.

One factor that might help to explain the higher habit-capture rate in this study compared to previous simulated ATC studies was the fact that the deferred conflict task was nested within the retention interval of the deferred handoff task, whilst in previous studies participants only completed one deferred task per trial (Loft, 2014; cf. Stone et al., 2001). The extent to which this would have triggered performance costs to one or both of the deferred tasks is unclear. However, in an electronic order-entry task, Sasangohar, Donmez, Easty, and Trbovich (2017) examined the effects of nested interruptions where individuals had to resume multiple tasks and reported that this nesting further increased the impact of interruptions. In field operations, it is likely individuals would be faced with remembering multiple future intentions. An important next step will be to extend current cognitive theory to account for the costs associated with holding multiple deferred intentions, and for human factors practitioners to determine whether these costs translate to increased risk of human error and under what situational contexts.

3.9.3. Practical Implications

ATC is a complex, safety-critical task characterized by time pressure, and the concurrent monitoring and execution of multiple tasks. Interruptions are frequent, and examples include communication requests from aircraft or other controllers, unexpected events requiring immediate attention, or the performance of other routine tasks (Kontogiannis & Malakis, 2009). Observational and ethnographic studies examining interruptions in ATC have suggested that interruptions may increase the likelihood of controllers forgetting deferred tasks and failing to detect or monitor data (Jones & Endsley, 1996; Shorrock, 2005). For instance, a common form of radar monitor error is when controllers form an “intention to monitor” a situation, but then fail to remember to do so due to other concurrent task demands or interruptions (Shorrock, 2005). The findings from Experiment 1 suggest the negative effects of interruption on memory may be a result of attention being shifted away from the situational cues in the ATC display that could prompt retrieval of intended actions. The fact that the n-back interruption was not associated with higher levels of disruption than the blank interruption suggests that increased opportunity for problem state rehearsal may not be sufficient to decrease the negative impact of interruption.
Chapter 3: Remembering to Execute Deferred Tasks in Simulated ATC: The Impact of Interruptions

The increased resumption failures in the ATC-interruption condition suggests that under some circumstances, monitoring multiple displays that share information processing requirements could have safety implications. On July 1st, 2002, a Tupolev-Tu-154M and a Boeing 757 collided in mid-air over Überlingen in Germany. Subsequent investigation revealed that a major contributor to the accident was the fact that the controller was monitoring a secondary ATC display at the time and had failed to subsequently divert attention back to the impending conflict on the primary display in sufficient time (Shorrock, 2007). The present work suggests that one element in the Überlingen accident may have been the effect that switching between multiple similar ATC displays has on memory for deferred tasks. Additionally, the evidence from the ex-Gaussian model suggested that resumption failures were likely due to complete forgetting, rather than a slowed resumption process. This suggests that if a controller fails to resume an interrupted task at the correct re-entry point due to memory interference, they may not eventually recover the intention from the perceptual elements of the display, no matter how much time they have. Indeed, several of Shorrock’s (2005) case reports specifically note that in most “intention to monitor” memory errors, controllers only recovered the information with explicit prompting from either automated ATC systems or other controllers on duty.

The findings from Experiment 2 also may be particularly relevant to research on interruption recovery aids which explore whether ‘event reviews’ and ‘change logs’ on secondary displays can improve recovery from interruption. These tools are designed to assist operators restore SA after an interruption by visually representing important changes that occurred in the environment during the interruption (St. John & Smallman, 2008; Sasangohar et al., 2014; Scott et al., 2006). While our findings support the idea that helping operators locate and attend to important items after an interruption could be useful; it is possible that these tools could cause interference-based forgetting if operators are switching between a primary task and recovery tool which share visually similar interfaces. Indeed, this may explain why several studies have found paradoxically negative effects of interruption recovery aids on interruption recovery (Hodgetts et al., 2015; St. John et al., 2005; Scott et al., 2006).
3.9.4. Considerations and Conclusions

Several studies using dynamic control tasks have reported that interruptions can negatively impact ongoing task performance (Hodgetts et al., 2015), and that it can take considerable time to return to baseline performance (Loft, Sadler, Braithwaite, & Huf, 2015). However, we failed to detect any effect of interruptions to ongoing task performance in the post-interruption period. One possible reason for this finding is that our analysis of ongoing task performance was limited in resolution to the entire post-interruption period, but the effects of the interruption may have only occurred for a shorter duration, more immediately after the interruption. Therefore, it is possible that there was an effect of interruptions to post-interruption ongoing task performance, but due to the low frequency of events in the ATC task, we were unable to detect it as we did not have enough ongoing task events to analyse post-interruption performance as a function of time after interruption.

The use of a student sample with limited training does limit our ability to generalize the results to expert controllers with lengthy training histories. Experienced controllers receive intensive training with well-defined performance standards and as such are far less likely to forget to perform deferred tasks as often as we observed in our experiments. Nonetheless, given the high frequency of aircraft and concomitant controller actions required each day, even small changes in error probabilities associated with deferred intentions could translate into large differences in incidents (Dismukes, 2012; Loft et al., 2013). There is a critical need to determine whether current display technology/aids are equipped to prevent the source of human error identified in the current study (Loft et al., 2016).

Additionally, the interruption manipulations we examined in the present study do not represent the range of interruptions ‘in the wild’. For instance, controllers may face interruptions in the order of several minutes. We selected the interruption duration of 27 s to ensure that items on the display were not entirely different from the pre-interruption state. If this was the case, participants would have to fully reconstruct (rather than recover) SA of the ATC display scene. Additionally, controllers in operational settings would often receive warnings about impending interruptions, allowing some strategic control over when to engage the pending interruption (McFarlane, 2002). In our studies, interruptions were presented with no explicit warning which may have inhibited participants from using interruption preparation strategies (Trafton et al., 2003). Nevertheless, several participants reported in follow-up
interviews that they strategically memorized the location of important aircraft when the deferred conflict encoding message appeared (note however, an interruption only followed this message 2/3 of the time, and participants did not know in advance whether they would be interrupted). Future research should examine whether preparation strategies during the ‘interruption lag’ can minimize deferred task performance decrements and whether strategic preparation is associated with performance costs to other ongoing primary tasks.

It is pertinent that both researchers and practitioners understand the factors underlying deferred task performance in safety-critical complex work environments. The present study has used a “use-inspired” basic research paradigm (Stokes, 1997) that allowed the systematic investigation of the effects of interruptions on multiple forms of deferred task. The results of this paper suggest that while interruptions can negatively impact deferred tasks under some circumstances, the presence and extent of the negative effects will depend heavily on the characteristics of both the interrupting task and deferred task at hand.
3.10. Appendix 3A: Bayesian robustness checks

The Bayesian $t$-tests reported in Experiment 1 and 2 (Chapter 3) were conducted using a default prior distribution on the effect size for the alternative hypothesis. Specifically, we used a Cauchy distribution with a medium Jeffreys-Zellner-Siow prior that specifies a scale parameter of, $r = \sqrt{2}/2$ (Rouder, Speckman, Sun, Morey, & Iverson, 2009). Increasing the scale parameter $r$ widens the prior distribution, reflecting a prior expectation of larger effect sizes; whilst $r$ values closer to zero make the distribution narrower, reflecting prior expectations of smaller effect sizes.

An important question is whether the findings related to the deferred task performance measures were robust to the choice of prior selected. To answer this question, we conducted a Bayes Factor robustness check by recalculating and plotting BFs under a range of different priors. Specifically, each panel in Figure 3.8 and Figure 3.9 contain a robustness check for one Bayesian $t$-test, and shows a range of scale parameters on the $x$ axis and the corresponding recalculated Bayes Factors on the $y$ axis. Across all panels, it can be seen that our interpretations of the BFs based on the user selected (default) prior remained consistent across a wide range of possible priors, indicating that our findings are robust to the choice of prior we selected.
### Deferred Handoff Errors (n-back ≠ Blank)

- User prior: $BF_{10} = 6.039$
- Wide prior: $BF_{10} = 9.234$
- Ultrawide prior: $BF_{10} = 12.943$

Evidence: Anecdotal → Strong

### Deferred Conflict Failures (n-back ≠ Blank)

- User prior: $BF_{10} = 2.528$
- Wide prior: $BF_{10} = 3.421$
- Ultrawide prior: $BF_{10} = 4.722$

Evidence: Anecdotal → Strong

### Deferred Conflict Resumption Time (n-back ≠ Blank)

- User prior: $BF_{10} = 4.470$
- Wide prior: $BF_{10} = 6.148$
- Ultrawide prior: $BF_{10} = 8.564$

Evidence: Anecdotal → Strong

### Deferred Handoff Errors (None ≠ Blank)

- User prior: $BF_{10} = 5.269$
- Wide prior: $BF_{10} = 7.280$
- Ultrawide prior: $BF_{10} = 10.168$

Evidence: Anecdotal → Strong

### Deferred Conflict Failures (None ≠ Blank)

- User prior: $BF_{10} = 1.072$
- Wide prior: $BF_{10} = 1.418$
- Ultrawide prior: $BF_{10} = 1.931$

Evidence: Anecdotal → Strong

### Deferred Conflict Resumption Time (None ≠ Blank)

- User prior: $BF_{10} = 28.708.939$
- Wide prior: $BF_{10} = 28.080.054$
- Ultrawide prior: $BF_{10} = 246754.830$

Evidence: Anecdotal → Strong

---

Figure 3.8. Bayes Factor robustness plots for deferred handoff errors, deferred conflict failures (misses), and deferred conflict resumption time in Experiment 1.
Figure 3.9. Bayes Factor robustness plots for deferred handoff errors, deferred conflict failures (misses) and deferred conflict resumption time in Experiment 2.
This chapter is presented as a journal article manuscript. A version of this manuscript is published in *Human Factors: The Journal of the Human Factors and Ergonomics Society*. The version of the manuscript presented in this thesis has several changes from the shorter version in *Human Factors*. Namely, the methods section is expanded and includes specific details on the counterbalancing procedure, the questionnaire protocol, and the workload assessment procedures. Several analyses concerning workload and long-term effects of interruptions are reported in the appendix but are not further discussed until my general discussion (Chapter 5). Throughout this chapter, I cite Chapter 3 in its published form: Wilson, Farrell, Visser, and Loft (2018). Also note that I use the term ‘we’ throughout this chapter to refer the collective contributions of the co-authors in the manuscript preparation (consistent with Chapter 3). Finally, I’ve changed terminology and wording throughout the chapter to make it more consistent with the other work in this thesis.

4.1. Foreword

The surprising finding from Chapter 3 was the null effect of interruptions on the deferred handoff task, and this served as the primary motivation for the study reported in Chapter 4. In Chapter 3, I discussed several possible explanations for this finding based on the differences between the two deferred tasks. The first possibility was that the deferred handoff task was simply robust to the effects of interruptions, possibly due to the fact that it was associated with a perceptually salient situational cue, allowing participants to engage a reactive control strategy at the time the deviation from routine was required (Braver, 2012; Cook et al., 2014; Friedrich, Biermann, Gontar, Biella, & Bengler, 2018; Loft, Finnerty, et al., 2011). This account would suggest that interruptions, no matter how demanding, would be unlikely to impact deferred handoff task performance. However, this explanation also would suggest that individuals do not actively maintain the intended task over the deferred task retention interval, indicating that retention interval is unlikely to impact deferred handoff performance and that costs should not be expected throughout the retention interval on ongoing tasks. The experiment reported in this chapter directly addresses these questions, by identifying whether interruptions can impact deferred task performance, and whether performance varies under a range of different retention intervals.

A second and interlinked explanation is that the negative effect of interruptions on deferred tasks might be mitigated by the provision of time prior to being interrupted (i.e., the PM encoding delay), or the provision of time after the interruption (i.e., PM execution delay) to orientate to the updated visual scene and use contextual cues to trigger deferred task intention retrieval. Indeed, there were clear differences between the deferred conflict and deferred handoff task in terms of amount of preparation time prior to the interruption, and time permitted for recovery after the interruption, either of which could account for differences in performance. Similarly, the deferred conflict task had to be resolved immediately after the interruption in all trials — the time period where tasks are most susceptible to negative interruption effects (Bowden et al., 2019; Brumby et al., 2013; Cane et al., 2012). Importantly however, manipulating the relative encoding and retrieval times of a delayed-execute PM task also results in changing the total retention interval of the task. Thus, to address these questions in this chapter, I implement an experimental design and analytical methodology (with mixed-effects models) capable of determining if the data suggest (1) any evidence for unique roles of encoding or
execution delays in affecting PM performance, and (2) whether these delays can account for
interrupted task performance over and above the effects of retention interval alone.

A final point to be addressed is that it was possible that the deferred handoff task
actually suffered some form of PM cost effects (i.e., reduced performance) from the deferred
conflict, owing to the fact it was nested within its retention interval (Sasangohar et al., 2017). Put
simply, after encoding the deferred handoff, individuals also had to encode and retrieve the
defered conflict task, in addition to completing other routine task demands. The study reported
in this chapter removes the deferred conflict task entirely in order to assess the deferred handoff
task under less confounded conditions. Thus, I examined how the (1) retention interval of the
defered task, (2) the presence of interruptions during this retention interval, and (3) the timing
of the interruption relative to deferred handoff encoding and execution, all respectively
impacted performance on the deferred handoff task.
4.2. Structured Abstract

Objective: To examine the effects of interruptions and retention interval on prospective memory for deferred tasks in simulated air traffic control.

Background: In many safety-critical environments, operators need to remember to perform a deferred task, which requires prospective memory. Laboratory experiments suggest that extended prospective memory retention intervals, and interruptions in those retention intervals, could impair prospective memory performance.

Method: Participants managed a simulated air traffic control sector. Participants were sometimes to perform a deferred handoff task, requiring them to deviate from a routine procedure. We manipulated whether an interruption occurred during the prospective memory retention interval or not, the length of the retention interval (37 s to 117 s), and the temporal proximity of the interruption to deferred task encoding and execution. We also measured performance on ongoing tasks.

Results: Increasing retention intervals (37 s to 117 s) decreased the probability of remembering to perform the deferred task. Costs to ongoing conflict detection accuracy and routine handoff speed were observed when a prospective memory intention had to be maintained. Interruptions did not affect individuals speed or accuracy on the deferred task.

Conclusion: Longer retention intervals increase risk of prospective memory error and of ongoing task performance being contaminated by cognitive load; however, prospective memory can be robust to effects of interruptions when the task environment provides cuing and offloading.

Application: To support operators in performing complex and dynamic tasks, prospective memory demands should be reduced, and the retention interval of deferred tasks should be kept as short as possible.
4.3. Introduction

The cognitive processes involved in the maintenance, retrieval, and execution of deferred tasks are referred to as *prospective memory* (PM). Individuals often need to remember to perform deferred tasks in safety-critical work contexts, such as air traffic control (ATC), healthcare, piloting, and unmanned aerial vehicle control. For example, air traffic controllers sometimes must remember to deviate from a routine aircraft vectoring procedure and to instead hold an aircraft when it reaches a specific way-point in the future because of heavy traffic. This requires the controller to defer the execution of the action and to remember to execute it at an appropriate time. Unfortunately, controllers can forget to complete deferred tasks, an outcome referred to as PM error (Shorrock, 2005). Such PM errors can, in turn, have serious safety implications (Dismukes, 2012; Loft, 2014).

In order to reduce PM errors, it is important to understand the psychological processes underlying PM, and how task characteristics affect those processes. Previous laboratory research has identified two important factors likely to affect the performance of deferred tasks: (a) retention interval (see, Martin, Brown, & Hicks, 2011), which refers to the amount of time between the encoding of the PM intention and the opportunity to execute it; and (b) interruptions arising from competing task demands, which can occur frequently during PM retention intervals (e.g., Cook, Meeks, Clark-Foos, Merritt, & Marsh, 2014; Schaper & Grundgeiger, 2018). In the current study, we examine how the length of the PM retention interval, and the presence of interruptions during the retention interval, impact the probability and speed at which individuals remember to deviate from a routine aircraft handoff procedure in a simulated ATC task. Additionally, we measure performance on concurrent ongoing ATC tasks to examine performance costs associated with a PM load. The study aims to illuminate how individuals maintain deferred task goals and use situational cues to support PM in safety-critical work contexts such as ATC.

4.4. Theoretical Approach

The Dynamic Multiprocess View (DMPV) is a useful theoretical framework for understanding PM in applied dynamic multi-tasking contexts (Scullin, McDaniel, & Shelton, 2013). Its central tenet is that PM is supported by the dynamic interplay between top-down and bottom-up cognitive processes (Shelton & Scullin, 2017). Top-down processing involves deliberately
maintaining the intention to perform the PM action in focal attention, or strategically monitoring the environment for PM cues (e.g., inspecting aircraft call signs). Evidence for such top-down processing in ATC has been demonstrated in several studies showing that maintaining PM is detrimental to performance on other ATC tasks (referred to as “PM costs”; for a review see Loft, 2014). However, PM can also be supported via bottom-up, cue-driven processes (Einstein & McDaniel, 2005). For instance, if events or environments become associated with a PM intention, attending to them can prompt retrieval of the PM action. A key feature of the DMPV is that bottom-up and top-down processes can interact. For instance, task context can trigger bottom-up processes that subsequently result in the engagement of strategic PM monitoring processes (Scullin et al., 2013; see also, Smith, Hunt, & Murray, 2017). Thus, external cues can trigger an operator’s intention to strategically monitor for PM events.

Controllers often report that PM tasks with long retention intervals are the most susceptible to PM error (Loft, Smith, et al., 2013). According to the DMPV, this would occur because limited-capacity top-down monitoring and maintenance processes are difficult to sustain for extended durations. In line with this, studies of PM in basic laboratory tasks (e.g., lexical decision making) have shown that increasing retention interval decreases PM performance (Martin et al., 2011; Scullin et al., 2010; Tierney et al., 2016; Zhang et al., 2017). Furthermore, PM costs to other ongoing tasks have been found to decrease over longer retention intervals suggesting that individuals decrease top-down monitoring over time (Loft, Kearney, & Remington, 2008; McBride, Beckner, & Abney, 2011). As such, there is reason to suspect that in ATC, PM tasks with longer retention interval could be at higher risk of not being completed. However, it is also possible that the continued presence of the PM relevant aircraft on the ATC display could act as a persistent contextual cue to prompt PM monitoring over the retention interval (Todorov et al., 2018). In line with this, Stone, Dismukes, and Remington (2001) reported no effect on PM error when the PM retention interval increased from 1-min to 5-min in simulated ATC. To our knowledge, this is the only study that has manipulated retention interval in simulated ATC to date.

According to the DMPV, people are less likely to engage in PM maintenance and monitoring if they exit the environmental context that is associated with the PM intention. This is consistent with many laboratory studies in which shifts in ongoing task context result in decreased evidence of PM monitoring (Bowden, Smith, et al., 2017; Kuhlmann & Rummel,
In ATC, such situations are particularly likely to arise due to task interruptions, which can be defined as situations in which an individual must suspend a primary task (e.g., display monitoring) in order to perform a secondary interrupting task (e.g., answering a pilot communication), with the explicit intention to return to the primary task after the interruption (Trafton & Monk, 2007). Indeed, several laboratory studies have found that interruptions during PM retention intervals can impair PM performance (Cook et al., 2014; McDaniel et al., 2004; Schaper & Grundgeiger, 2018). However, in the first experimental investigation of interruptions in ATC, Wilson, Farrell, Visser, and Loft (2018) found that interrupting participants for 27 s during the PM retention interval had no effect on the speed or accuracy of performing a PM task that required deviation from a routine aircraft handoff procedure.

The DMPV offers two explanations for Wilson et al. (2018)'s results. One is that PM retrieval in ATC may largely depend on bottom-up, cue-driven processes (Einstein & McDaniel, 2005), in which case any effect of interruptions on top-down monitoring processes would be irrelevant. However, previous ATC studies show that PM load (i.e., having an active PM intention) impairs concurrent air traffic management tasks, such as conflict detection (Loft et al., 2016; Loft, Finnerty, et al., 2011; Loft & Remington, 2010; Loft, Smith, et al., 2011, 2013), indicating reliance on PM monitoring processes. Another possibility is that top-down PM processes might have been reinstated in the interval between the interruption ending and the PM action being required. In Wilson et al. the PM task had to be performed approximately 1 min after the interruption had ended, thus permitting time to process contextual cues associated with the PM action, which in turn may have re-engaged top-down monitoring.

This second option suggests that it is important to further consider how the temporal relationship between interruptions and the correct time for PM retrieval influences PM errors. In the interruptions literature, studies generally examine tasks where the resumption lag (interval between end of interruption and primary task resumption) is effectively zero - individuals must "resume" the intended primary task immediately after returning from the interruption. For PM tasks, however, the resumption lag (i.e., PM execution-delay) is likely to be heterogeneous, and this may impact how well individuals can orientate to the updated visual scene and use contextual cues to trigger PM retrieval. Research on visual working memory has shown that memory representations are most volatile in the immediate moments following an
interruption, resulting from insufficient time for attention to recover (Wang, Theeuwes, & Olivers, 2018). Similarly, PM can be improved by the provision of time prior to being interrupted (i.e., the PM encoding-delay). An increased encoding-delay allows more time to rehearse and consolidate intentions, and an opportunity to strengthen associations between intentions and contextual cues (Boehm-Davis & Remington, 2009). This can improve PM (Dismukes & Nowinski, 2006) and resumption time (Hodgetts & Jones, 2006a) in simple tasks, and improve decision-making in complex dynamic tasks (Labonté, Tremblay, & Vachon, 2019).

To be clear, this interruption-delay hypothesis specifies that increasing encoding delay or execution delay may improve PM when interrupted. However, increasing these delay intervals also increases the total retention interval, which may be expected to impair PM. In the current study, we compare these hypotheses by testing whether encoding-delays and execution-delays function in a different manner to what would be expected from an effect of retention interval when interrupted.

4.5. Current Study

We examined how PM in a simulated ATC is affected by retention interval (as well as the separable contributions of differences in encoding and execution delay to this interval), and by interruptions. Participants assumed the role of an air traffic controller responsible for accepting aircraft entering their air sector, detecting and resolving aircraft conflicts, and handing-off aircraft exiting their sector. Participants completed a number of trials, and in each, a PM task occurred that required them to acknowledge an instruction to perform an alternative action (e.g., press right arrow key) instead of a routine action (H key) when handing off a target aircraft (i.e., deferred handoff). During some PM retention intervals participants were interrupted by an additional ATC task (for 27 s) that comprised the same task objectives as the primary scenario, but with different aircraft and flight paths. The interruption and primary ATC tasks overlapped in visual appearance and processing modality, which has previously been linked to interference-based PM errors in ATC (Wilson et al., 2018).

Figure 4.1 shows how the timing of the deferred task relative to the interruption yielded variations in encoding and execution delays. The square dots represent key stages in each trial (trial start, encode PM task, execute PM task, and trial end), while the two triangles indicate the point that the PM task was encoded, and need to be executed, relative to the interruption. The
Chapter 4: PM in ATC Robust to Interruptions but Impaired by Retention Interval

PM retention interval consisted of the encoding delay, duration of the interruption, and the execution delay. The combination of two encoding-delays (10 or 50 s), two execution-delays (0 or 40 s) and a 27 s interruption yielded three retention intervals (37 s, 77 s, 117 s).

We examined: (1) whether the interruption decreased PM performance (i.e., decreased accuracy and increased PM RT); (2) whether longer encoding or execution delays improved PM performance; (3) whether longer PM retention intervals were associated with decreased PM performance; (4) whether there were costs to aircraft acceptances or handoffs (non-response errors and RT) or conflict detection accuracy during the PM retention interval.

![Figure 4.1. The time-course of the PM handoff task (not to scale).](image)

4.6. Method

4.6.1. Participants
78 undergraduate students (female = 31; median age = 20) from the University of Western Australia participated in the study in exchange for course credit or $50 AUD. This research complied with the American Psychological Association Code of Ethics and was approved by the Human Research Ethics Office at the University of Western Australia. Informed consent was obtained from each participant.

4.6.2. Design
The experiment used a 2 (interruption) x 2 (encoding-delay) x 2 (execution-delay) within-subjects design. The interruption manipulation was either ‘uninterrupted’ in which participants were not interrupted, or ‘interrupted’ in which they had to manage an additional ATC sector for 27 s. The timing of the encoding and execution-delay manipulations was anchored around the interruption. The encoding-delay was either short in which the deferred handoff PM task was
encoded 10 s prior to the interruption; or long in which the PM task was encoded 50 s before interruption. Similarly, the execution-delay was either short, in which the PM aircraft flashed for handoff immediately after the interruption ended; or long in which the PM aircraft flashed for handoff 40 s after the interruption ended. These timings were identical on uninterrupted trials, except that there was no interruption. As shown in Table 4.1, this resulted in eight within-subjects conditions, associated with three retention interval durations (37 s, 77 s, 117 s).

Table 4.1. The experimental design along with the respective total retention interval.

<table>
<thead>
<tr>
<th>Interruption</th>
<th>Encoding-Delay</th>
<th>Execution-Delay</th>
<th>Total Retention Interval</th>
<th>N trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interrupted</td>
<td>S (10 s)</td>
<td>S (0 s)</td>
<td>37 s</td>
<td>256</td>
</tr>
<tr>
<td>Interrupted</td>
<td>S (10 s)</td>
<td>L (40 s)</td>
<td>77 s</td>
<td>258</td>
</tr>
<tr>
<td>Interrupted</td>
<td>L (50 s)</td>
<td>S (0 s)</td>
<td>77 s</td>
<td>262</td>
</tr>
<tr>
<td>Interrupted</td>
<td>L (50 s)</td>
<td>L (40 s)</td>
<td>117 s</td>
<td>254</td>
</tr>
<tr>
<td>None</td>
<td>S (10 s)</td>
<td>S (0 s)</td>
<td>37 s</td>
<td>259</td>
</tr>
<tr>
<td>None</td>
<td>L (10 s)</td>
<td>L (40 s)</td>
<td>77 s</td>
<td>253</td>
</tr>
<tr>
<td>None</td>
<td>L (50 s)</td>
<td>S (0 s)</td>
<td>77 s</td>
<td>256</td>
</tr>
<tr>
<td>None</td>
<td>L (50 s)</td>
<td>L (40 s)</td>
<td>117 s</td>
<td>254</td>
</tr>
</tbody>
</table>

Note. The encoding-delay and execution-delay conditions are relative to the ‘interruption start point’ and ‘interruption end point’, respectively. The strike-through for the delay manipulations indicates that while timing was equivocal across interruption conditions, uninterrupted trials sum to the retention interval.

The total retention interval includes the 27s of either continued ongoing air traffic management (uninterrupted trials) or the interrupting ATC task.

The total number of observed PM trials per condition after the specified exclusion criteria (see results section).

4.6.3. Counterbalancing

In order to counterbalance all 32 ATC trials across the eight conditions, two 8 x 8 Latin square schemes were used (one for each session). Specifically, for the first session, we divided the 16 scenarios into eight scenario groups (containing two scenarios each). Next, the eight scenario groups were counterbalanced in an 8 (columns) x 8 (rows) Latin square. The columns of the Latin square were used to allocate the eight within-subjects conditions across the eight scenario groups; whilst rows were used to allocate equal sized groups of participants to a specific
counterbalancing scheme. Note that the final order in which trials were presented in was randomized for each participant. This process was repeated for the second session, however, in order to minimize the influence of the prior exposure to the scenarios, the order of the columns in the Latin-square was reversed, resulting in participants seeing the opposite interruption condition and delay condition relative to the first day. This counterbalancing scheme resulted in the interruption and delay conditions being associated equally with each scenario across the 32 simulation trials (and all subjects). Table 4.2 below shows the full counter balancing scheme for both days.

Table 4.2 Counterbalancing scheme for the first experimental session

<table>
<thead>
<tr>
<th></th>
<th>ATC-SS</th>
<th>ATC-SL</th>
<th>ATC-LS</th>
<th>ATC-LL</th>
<th>None-SS</th>
<th>None-SL</th>
<th>None-LS</th>
<th>None-LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>(1,9)</td>
<td>(2,10)</td>
<td>(3,11)</td>
<td>(4,12)</td>
<td>(5,13)</td>
<td>(6,14)</td>
<td>(7,15)</td>
<td>(8,16)</td>
</tr>
<tr>
<td>Group 2</td>
<td>(2,10)</td>
<td>(3,11)</td>
<td>(4,12)</td>
<td>(5,13)</td>
<td>(6,14)</td>
<td>(7,15)</td>
<td>(8,16)</td>
<td>(1,9)</td>
</tr>
<tr>
<td>Group 3</td>
<td>(3,11)</td>
<td>(4,12)</td>
<td>(5,13)</td>
<td>(6,14)</td>
<td>(7,15)</td>
<td>(8,16)</td>
<td>(1,9)</td>
<td>(2,10)</td>
</tr>
<tr>
<td>Group 4</td>
<td>(4,12)</td>
<td>(5,13)</td>
<td>(6,14)</td>
<td>(7,15)</td>
<td>(8,16)</td>
<td>(1,9)</td>
<td>(2,10)</td>
<td>(3,11)</td>
</tr>
<tr>
<td>Group 5</td>
<td>(5,13)</td>
<td>(6,14)</td>
<td>(7,15)</td>
<td>(8,16)</td>
<td>(1,9)</td>
<td>(2,10)</td>
<td>(3,11)</td>
<td>(4,12)</td>
</tr>
<tr>
<td>Group 6</td>
<td>(6,14)</td>
<td>(7,15)</td>
<td>(8,16)</td>
<td>(1,9)</td>
<td>(2,10)</td>
<td>(3,11)</td>
<td>(4,12)</td>
<td>(5,13)</td>
</tr>
<tr>
<td>Group 7</td>
<td>(7,15)</td>
<td>(8,16)</td>
<td>(1,9)</td>
<td>(2,10)</td>
<td>(3,11)</td>
<td>(4,12)</td>
<td>(5,13)</td>
<td>(6,14)</td>
</tr>
<tr>
<td>Group 8</td>
<td>(8,16)</td>
<td>(1,9)</td>
<td>(2,10)</td>
<td>(3,11)</td>
<td>(4,12)</td>
<td>(5,13)</td>
<td>(6,14)</td>
<td>(7,15)</td>
</tr>
</tbody>
</table>

Note. Participants were distributed equally across the counterbalancing groups (i.e., rows). ATC scenarios were grouped in pairs of two (shown in parentheses) and distributed across the counterbalancing scheme. Experimental conditions were assigned column-wise. For the second session, the column order was reversed (e.g., Group 1 would complete scenarios 8 and 16 under ATC-SS, and scenarios 2 and 10 under None-LS).

4.6.4. **ATC-Lab Advanced Simulator**

Figure 4.2 presents a screenshot of the ATC task (Fothergill, Loft, & Neal, 2009). The light grey polygon area is the flight control sector, whilst the dark grey area represents sectors outside the

---

14 For example, if in the first session a particular subject completed ‘Scenario 5’ in the “uninterrupted + long encode-delay+ long execute-delay”; then in the second session they would complete ‘Scenario 5’ in the “interrupted + short encode-delay + short execute-delay” condition.
participants’ control. The black lines denote flight paths that aircraft travel. Aircraft are represented by a circle with a leader-line indicating heading. The aircraft data-blocks specify call sign, speed, aircraft type, current/cleared altitude. Current altitude and cleared altitude are separated by an arrow that denotes whether the aircraft is climbing (\(^\uparrow\)), descending (\(^\downarrow\)), or cruising (\(\rangle\)). Aircraft enter the sector from the edges of the display, cross sector boundaries, and then exit the sector. New aircraft enter throughout the trial, with aircraft positions being updated every second (although behavioural measures are recorded with millisecond precision). Time elapsed in each trial was displayed on the bottom of the display.

![Figure 4.2](image)

**Figure 4.2.** A screenshot of the ATC display. Inbound aircraft are black (e.g., aircraft GA85) as they approach the sector, and flash orange for acceptance (e.g., EK69) when they reached within 5 miles of the sector boundary. Aircraft turn green (e.g., MH44) when accepted. When outbound aircraft cross the sector boundary they flash blue (e.g., EK29), and then turn white (e.g., JQ79) when handed off. The example in the figure shows that the individual is required to change the altitude of EK63 (cruising at 340) to an alternative altitude to resolve the conflict with JQ68 (cruising at 395). Aircraft turn yellow (e.g., QR04 & BA01) if they violate the minimum vertical and lateral separation. The running score (-40 points) is presented in the middle right hand side of the display. Note the ‘primary scenario’ text box in the upper left to distinguish the primary from interrupting scenarios.

When aircraft approached the sector boundary, they flashed for acceptance and participants had to accept aircraft by clicking the aircraft and pressing the A key within 15 s. Similarly, as aircraft exited the sector boundary, they flashed for handoff and participants had to click the aircraft and press the H key within 15 s. Conflicts occurred when an aircraft pair
violated both lateral (5 nautical miles) and vertical (1,000 feet) separation standards, and were indicated by the pair of aircraft turning yellow. Participants were required to prevent conflicts from occurring by clicking on one of the aircraft they believed to be in future conflict and changing the cleared altitude. Each trial comprised 3 conflicts, and 25 to 30 acceptances and handoffs. Participants were instructed to respond as quickly and accurately as possible.

Participants received points for successfully completing tasks and the current score was continuously updated on the right of the display. Ten points were awarded or deducted for a successful/failed handoff/acceptance. Between 10 and 40 points were awarded for resolving a conflict, depending on the speed of resolution, and 40 points were deducted for failing to resolve a conflict or for unnecessary interventions (i.e., altering the altitude of aircraft not in conflict).

4.6.5. Procedure
The experiment consisted of two, 2.5 hour sessions. Session one comprised a training and a test phase. Session two comprised a test phase and a brief questionnaire.

Training Phase. Training started with completing an instructional website (~30 min). The website provided explanations of basic ATC concepts (e.g., what a conflict is; how to read aircraft data-blocks), instructions for completing the ongoing tasks (handoffs, acceptances, and detecting/resolving conflicts), instructions regarding the deferred-handoff task, and information regarding the interruptions and point scoring. Participants then completed two 5 min practice trials, which followed the same structure as the test trials.

Test Phase. Each test phase comprised 16 5 min trials, resulting in four PM task observations per within-subject condition per participant. The sequence of events was identical in every trial, but trials differed with respect to event timings and locations (e.g., conflicts had occurred at differing times and locations; interruption onset times differed). The same 16 trials were used for both sessions, however on the second session aircraft call-signs were randomized and experiment conditions were opposed (e.g., interrupted to uninterrupted; short to long delay). Experimental conditions were counterbalanced across trials and subjects with two 8 x 8 Latin square schemes (one for each session). Final trial presentation order was randomized.

The interruption start time was fixed for each of the 16 trials (between 1m 30s and 2m 30s). On each trial either 50 s (long encode-delay) or 10 s (short encode-delay) before the
interruption, a message box would appear adjacent to one aircraft instructing participants to handoff that aircraft with an arrow key that corresponded to the aircraft heading (e.g., ↑), instead of the routine ‘H’ key. This message was displayed for 10 s, and participants had to acknowledge it by clicking a button marked "Acknowledge" that became active after 3 s to prevent accidental acknowledgement. Messages disappeared if they were not acknowledged within 10 s. Encoding delay was manipulated by changing the onset of this message.

After the interruption, the deferred handoff target aircraft would immediately flash for handoff (i.e., cue for performing the PM action), or participants resumed the primary ATC tasks for 40 s, and then the PM aircraft flashed for handoff. This was our execution-delay manipulation. No aircraft flashed for acceptance or handoff within 10 s of the PM aircraft flashing for handoff. After the PM aircraft was handed-off (or recorded as missed), participants continued ongoing ATC tasks until the trial ended. Event timings were identical for uninterrupted trials, but participants only performed the primary ATC task.

The interrupting ATC task required monitoring a different sector that was displayed in place of the primary sector, with task objectives being identical to those of the primary task. Each interrupting trial comprised two or three aircraft acceptances, two or three aircraft handoffs, and two conflicts requiring resolution. The interruption began with a 1.7 cm crosshair presented in the centre of the display for 2500 ms, a 24 s ATC scenario, and a black visual buffer for approximately 500 ms. Participants were instructed that the interrupting ATC task was equally important as the primary ATC, but no “special handoff aircraft task” (i.e., deferred handoff) occurred during the interruption. The timer was removed from the interrupting sector display. There were eight unique interruption trials which varied with regards to event timing and locations. The presentation order was randomized for each participant per session.

There was a 15 s break after trials, except for the 8th trial which was followed by a 180 s break. Participants could then begin the following trial by pressing spacebar and could take a longer rest break if required by pausing before pressing spacebar. For the second session only, immediately after each trial, participants were presented with two subjective performance questions to be answered on an 8-point scale using a text entry box:

1. “How content, relaxed, and complacent versus irritated, stressed, and annoyed did you feel during the task? (with 0 being I felt content, relaxed, and complacent, and 7 being I felt irritated, stressed, and annoyed).”
2. “How hard did you have to work to accomplish your level of performance in the last test trial? 0 being not at all, and 7 being extremely hard.”

4.6.6. Questionnaire

A 19-question exit questionnaire (see Appendix 4C) was used to measure three constructs: (1) participants overall perceptions and affect regarding the interruption trials; (2) self-reported interruption management strategies; (3) the task perceived as most difficult task in the experiment, and (4) the task perceived as most impacted by the interruption. Additionally, to determine the extent to which participants recognized the similarity in trials between the first and second session, at the conclusion of the experiment, the experimenter asked each participant to report any similarities or differences they noticed between the trials on the first and second session. Only one participant mentioned similarity in the locations of conflicts. The majority of participants reported that they found the scenarios in the second session more difficult, despite the fact the scenarios were exactly the same (remember however that participants experienced the scenarios in the inverse conditions to the previous session).

4.7. Results

Four participants who did not complete the second experimental session were excluded, as were four participants who only correctly perform the deferred handoff task on less than 10% of trials (final n = 70). Hypothesis testing was conducted using mixed-effects modelling, implemented in the lme4 R package (Bates, Mächler, Bolker, & Walker, 2015) for the R programming language (R Core Team, 2017). Continuous dependent variables (e.g. mean RTs) were analysed with linear mixed models, and binary dependent variables (e.g., PM errors) with generalized linear mixed models using a logistic link function. Mixed effects modelling enables control of variance associated with random factors (e.g. participant) without data-aggregation (Baayen, Davidson, & Bates, 2008). Models were compared with likelihood-ratio tests. Specifically, for each dependent variable, a null model was specified that included only the dependent variable of interest and a random intercept across participants. The impact of each experimental factor was evaluated by comparing a model that included the fixed effect of interest to the null model. Interaction effects were tested by comparing a full model specifying the interactions with a model containing identical predictors but no interaction. Reported p values were obtained with
the Satterthwaite approximation by conducting chi-square tests ($\chi^2$) on the log-likelihoods of the respective models (Kuznetsova, Brockhoff, & Christensen, 2017). Coefficients ($\beta$) and standard errors (SE) for each effect in question are presented in-text. The results associated with each model tested can be seen in Tables 4.4 and 4.5.

4.7.1. Deferred Handoff PM Accuracy

To assess whether the number of excluded trials based on error type differed across conditions, Pearson’s Chi-squared tests for count data were conducted. PM response execution errors (remembering to press an arrow key but pressing the incorrect arrow key) were made on 3.17% ($n$ trials = 74) of trials and did not significantly differ between conditions, $\chi^2(7) = 3.84, p = 0.8$. Non-response errors occurred on 0.56% of trials ($n$ trials = 13) and did not significantly differ between conditions, $\chi^2(5) = 1.31, p = 0.93$. PM false alarms (pressing the arrow key on non-PM aircraft) were made on 3.17% of trials ($n$ trials = 74) and did not significantly differ between the conditions, $\chi^2(7) = 6.65, p = 0.47$. PM task acknowledgement errors (failing to acknowledge the PM encoding message and making a PM error) occurred on 1.67% of trials ($n$ trials = 39) and did not significantly differ between conditions, $\chi^2(7) = 5.51, p = 0.6$. All these trial types were excluded from final analysis (final $n$ observations = 2052).

PM errors were defined as pressing the routine handoff key (H) instead of the instructed key when handing off PM aircraft. PM error rates by condition are presented in Figure 4.3. First, we examined whether PM errors increased as a function of retention interval in the uninterrupted condition. There was a significant main effect of retention interval, $\beta = 0.64, SE = 0.17, \chi^2(1) = 14.75, p = < 0.001$, and this was associated with a significant polynomial linear contrast, $z = 3.85, p = < 0.001$, indicating PM errors increased over longer retention intervals. This analysis was repeated with both interruption conditions which revealed a similar pattern of results.
Chapter 4: PM in ATC Robust to Interruptions but Impaired by Retention Interval

Figure 4.3. Mean deferred handoff error rate across the four timing conditions and the two interruption conditions for the deferred handoff task. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).

PM error rates did not significantly vary by interruption condition, $\beta = 0.01$, $SE = 0.11$, $\chi^2(1) = 0.01$, $p = 0.94$. To test whether delays prior to, or after, an interruption affected PM errors, we examined the interaction between interruption condition and the encoding and execution delay conditions, respectively. A non-significant interaction would indicate that the effect of encoding or execution delay is equivalent between uninterrupted and interrupted conditions, indicating that only retention interval impacts PM task performance. A significant interaction would indicate the effect of delay differs for each interruption condition, suggesting a unique role of delay for interrupted trials. There was no significant interaction between encoding-delay and interruption condition, $\beta = -0.16$, $SE = 0.22$, $\chi^2(1) = 0.51$, $p = 0.47$, or between execution-delay and interruption condition, $\beta = -0.32$, $SE = 0.22$, $\chi^2(1) = 2.07$, $p = 0.15$. 

127
Figure 4.4. Mean deferred handoff RT across the four timing conditions and the two interruption conditions for the deferred handoff task. The non-significant resumption time effect can be seen by the increase in RT for the interrupted SS and LS conditions. Error bars represent 95% within-subjects confidence intervals (Cousineau, 2005).

4.7.2. Deferred Handoff PM Response Time

PM task response time (RT) was defined as the time taken to correctly handoff the PM aircraft after it flashed for handoff. Trials with RTs more than 3 SDs from a participant’s grand mean were excluded from analysis (1.43% of RTs). Mean RTs are presented in Figure 4.4, separated by retention interval. There was no significant effect of retention interval for the uninterrupted condition, $\beta = 134.64$, $SE = 89.52$, $\chi^2(1) = 2.24$, $p = 0.33$. There was also no significant main effect of interruption, $\beta = -97.25$, $SE = 62.11$, $\chi^2(1) = 2.45$, $p = 0.12$, and no significant interaction between encoding-delay and interruption condition, $\beta = -25.74$, $SE = 123.48$, $\chi^2(1) = 0.04$, $p = 0.83$, or between execution-delay and interruption condition, $\beta = -181.86$, $SE = 123.62$, $\chi^2(1) = 2.16$, $p = 0.14$. Finally, we conducted a planned contrast to determine whether RT was slower for PM aircraft which had to be responded to immediately following an interruption (i.e., the ATC condition with short execution delay, relative to all other conditions), but this was not significant, $\beta = 123.09$, $SE = 71.01$, $\chi^2(1) = 3$, $p = 0.084$. Thus, we did not find that retention interval or interruption impacted PM RT, and we did not find a resumption time effect.
4.7.3. Cost of PM Retention to Ongoing ATC tasks

To determine the impact of PM on ongoing task performance, we examined three tasks: aircraft handoffs, acceptances, and conflict detection. Descriptive statistics for these tasks are presented in Table 4.3 below. Handoff RT was slower for handoffs that occurred during the PM retention interval relative to outside it, $\beta = 62.05$, $SE = 24.41, \chi^2(1) = 6.36, p = 0.012$, but there was no significant difference in non-response errors, $\beta = 0.27$, $SE = 0.16, \chi^2(1) = 2.86, p = 0.09$. Aircraft acceptance RT did not significantly differ for acceptances occurring during versus outside the PM retention interval, $\beta = 46.21$, $SE = 29.9, \chi^2(1) = 2.38, p = 0.12$, nor was there a significant difference in non-response errors, $\beta = -0.27$, $SE = 0.15, \chi^2(1) = 3.27, p = 0.071$.

Table 4.3. Means and standard deviations for the three ongoing task performance measures by PM cost condition (i.e., whether the PM intention was to be maintained or not).

<table>
<thead>
<tr>
<th>Measure</th>
<th>Type</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept Response Time</td>
<td>No PM</td>
<td>2818 ms</td>
<td>552 ms</td>
<td>140</td>
</tr>
<tr>
<td>Accept Response Time</td>
<td>PM</td>
<td>2864 ms</td>
<td>567 ms</td>
<td>140</td>
</tr>
<tr>
<td>Correct Accept Proportion</td>
<td>No PM</td>
<td>96.9%</td>
<td>17%</td>
<td>24178</td>
</tr>
<tr>
<td>Correct Accept Proportion</td>
<td>PM</td>
<td>97.4%</td>
<td>16%</td>
<td>5887</td>
</tr>
<tr>
<td>Handoff Response Time</td>
<td>No PM</td>
<td>3044 ms</td>
<td>658 ms</td>
<td>140</td>
</tr>
<tr>
<td>Handoff Response Time</td>
<td>PM</td>
<td>3106 ms</td>
<td>746 ms</td>
<td>140</td>
</tr>
<tr>
<td>Correct Handoff Proportion</td>
<td>No PM</td>
<td>98.9%</td>
<td>11%</td>
<td>18678</td>
</tr>
<tr>
<td>Correct Handoff Proportion</td>
<td>PM</td>
<td>99.13%</td>
<td>9.3%</td>
<td>5608</td>
</tr>
<tr>
<td>Conflict Detection Proportion</td>
<td>No PM</td>
<td>88%</td>
<td>32%</td>
<td>2720</td>
</tr>
<tr>
<td>Conflict Detection Proportion</td>
<td>PM</td>
<td>82%</td>
<td>39%</td>
<td>3634</td>
</tr>
</tbody>
</table>

Note: N = number of observations. Conflict detection proportion type includes any aircraft with any degree of overlap with the PM aircraft, and is reported here for descriptive purposes only.
### Table 4.4: Model comparison table for all deferred handoff PM model comparisons (PM errors and RT).

<table>
<thead>
<tr>
<th>Dependent Variable (y)</th>
<th>Model Specification</th>
<th>k</th>
<th>AIC</th>
<th>BIC</th>
<th>Deviance</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PM Errors (Full Dataset)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y ~ $\beta_0$</td>
<td></td>
<td>1</td>
<td>2073.71</td>
<td>2084.96</td>
<td>2069.71</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Retention)</td>
<td></td>
<td>2</td>
<td>2075.70</td>
<td>2092.58</td>
<td>2069.70</td>
<td>.94</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(RetentionPolicy)</td>
<td></td>
<td>3</td>
<td><strong>2060.53</strong></td>
<td><strong>2083.04</strong></td>
<td><strong>2052.53</strong></td>
<td>&lt; .001</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Encoding-Delay) + $\beta_2$(Retention)</td>
<td></td>
<td>3</td>
<td>2074.21</td>
<td>2096.71</td>
<td>2066.21</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Execution-Delay) + $\beta_2$(Retention)</td>
<td></td>
<td>3</td>
<td>2064.22</td>
<td>2086.72</td>
<td>2056.22</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1 + \beta_2 + \beta_3$(Encoding-Delay × RetentionPolicy)</td>
<td></td>
<td>4</td>
<td>2074.14</td>
<td>2102.27</td>
<td>2064.14</td>
<td>.47</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1 + \beta_2 + \beta_3$(Encoding-Delay × RetentionPolicy)</td>
<td></td>
<td>4</td>
<td>2064.22</td>
<td>2086.72</td>
<td>2052.53</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1 + \beta_2 + \beta_3$(Encoding-Delay × RetentionPolicy)</td>
<td></td>
<td>4</td>
<td>2064.22</td>
<td>2086.72</td>
<td>2052.53</td>
<td>—</td>
</tr>
<tr>
<td><strong>PM Errors (Uninterrupted)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y ~ $\beta_0$</td>
<td></td>
<td>1</td>
<td>1032.68</td>
<td>1042.54</td>
<td>1028.68</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Retention)</td>
<td></td>
<td>3</td>
<td><strong>1021.93</strong></td>
<td><strong>1041.65</strong></td>
<td><strong>1013.93</strong></td>
<td>&lt; .001</td>
</tr>
<tr>
<td><strong>PM RT (Full Dataset)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y ~ $\beta_0$</td>
<td></td>
<td>1</td>
<td>8599.44</td>
<td>8612.26</td>
<td>8593.44</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Retention)</td>
<td></td>
<td>2</td>
<td>8598.99</td>
<td>8616.08</td>
<td>8590.99</td>
<td>.12</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Resumption)</td>
<td></td>
<td>2</td>
<td>8598.44</td>
<td>8615.53</td>
<td>8590.44</td>
<td>.08</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Encoding-Delay) + $\beta_2$(Retention)</td>
<td></td>
<td>3</td>
<td>8598.76</td>
<td>8620.13</td>
<td>8588.76</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Encoding-Delay) + $\beta_2$(Retention)</td>
<td></td>
<td>3</td>
<td>8600.99</td>
<td>8622.35</td>
<td>8590.99</td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1 + \beta_2 + \beta_3$(Encoding-Delay × RetentionPolicy)</td>
<td></td>
<td>4</td>
<td>8600.72</td>
<td>8626.36</td>
<td>8588.72</td>
<td>.83</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1 + \beta_2 + \beta_3$(Encoding-Delay × RetentionPolicy)</td>
<td></td>
<td>4</td>
<td>8600.83</td>
<td>8626.47</td>
<td>8588.83</td>
<td>.14</td>
</tr>
<tr>
<td><strong>PM RT (Uninterrupted)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y ~ $\beta_0$</td>
<td></td>
<td>1</td>
<td><strong>4338.89</strong></td>
<td><strong>4349.63</strong></td>
<td><strong>4332.89</strong></td>
<td>—</td>
</tr>
<tr>
<td>y ~ $\beta_0 + \beta_1$(Retention)</td>
<td></td>
<td>3</td>
<td>4340.65</td>
<td>4358.55</td>
<td>4330.65</td>
<td>.33</td>
</tr>
</tbody>
</table>

*Note. Bolded model names indicate selected models. k = number of fixed effect parameters. $\beta_0$ = intercept. All models included a participant-level random intercept term. Interaction models included main effects for both factors in the interaction, indicated by the unlabelled terms ($\beta_1, \beta_2$). Reported values were obtained from Chi-square tests on the log-likelihoods via Satterthwaite approximation.*
Table 4.5. Model comparison table for all ongoing task models.

<table>
<thead>
<tr>
<th>Dependent Variable (y)</th>
<th>Model Specification</th>
<th>k</th>
<th>AIC</th>
<th>BIC</th>
<th>Deviance</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Handoff RT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y \sim \beta_0 )</td>
<td></td>
<td>1</td>
<td>4050.85</td>
<td>4061.76</td>
<td>4044.85</td>
<td>—</td>
</tr>
<tr>
<td>( y \sim \beta_0 + \beta_1(During PM Retention Interval) )</td>
<td>2</td>
<td>4046.49</td>
<td>4061.03</td>
<td>4038.49</td>
<td>.012</td>
<td></td>
</tr>
<tr>
<td><strong>Handoff Errors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y \sim \beta_0 )</td>
<td></td>
<td>1</td>
<td>2630.39</td>
<td>2646.58</td>
<td>2626.39</td>
<td>—</td>
</tr>
<tr>
<td>( y \sim \beta_0 + \beta_1(During PM Retention Interval) )</td>
<td>2</td>
<td>2629.53</td>
<td>2653.82</td>
<td>2623.53</td>
<td>.09</td>
<td></td>
</tr>
<tr>
<td><strong>Acceptance RT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y \sim \beta_0 )</td>
<td></td>
<td>1</td>
<td>4093.17</td>
<td>4104.07</td>
<td>4087.17</td>
<td>—</td>
</tr>
<tr>
<td>( y \sim \beta_0 + \beta_1(During PM Retention Interval) )</td>
<td>2</td>
<td>4092.79</td>
<td>4107.33</td>
<td>4084.79</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td><strong>Acceptance Errors</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y \sim \beta_0 )</td>
<td></td>
<td>1</td>
<td>3421.29</td>
<td>3437.87</td>
<td>3417.29</td>
<td>—</td>
</tr>
<tr>
<td>( y \sim \beta_0 + \beta_1(During PM Retention Interval) )</td>
<td>2</td>
<td>3420.02</td>
<td>3444.9</td>
<td>3414.02</td>
<td>.07</td>
<td></td>
</tr>
<tr>
<td><strong>Conflict Detection</strong></td>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y \sim \beta_0 )</td>
<td></td>
<td>1</td>
<td>5184.92</td>
<td>5198.44</td>
<td>5180.92</td>
<td>—</td>
</tr>
<tr>
<td>( y \sim \beta_0 + \beta_1(PM Overlap Proportion) )</td>
<td>2</td>
<td>5135.15</td>
<td>5155.42</td>
<td>5129.15</td>
<td>&lt;.001</td>
<td></td>
</tr>
</tbody>
</table>

Note. Bolded model names indicate selected models. \( k \) = number of fixed effect parameters. \( \beta_0 \) = intercept. All models included a participant-level random intercept term. Interaction models included main effects for both factors in the interaction, indicated by the unlabelled terms \( (\beta_1, \beta_2) \). Reported values were obtained from Chi-square tests on the log-likelihoods via Satterthwaite approximation.
A conflict detection failure occurred when two aircraft violated minimum separation. Because conflicts evolve over time, their degree of overlap with the PM retention interval differed. To examine the cost of the PM load to conflict detection accuracy, we calculated an ‘overlap proportion’ measure that indexed the proportion of time that the PM retention interval overlapped with the time aircraft pairs involved in a conflict were in the flight control sector. An overlap proportion of 0% indicates that the PM retention interval did not overlap with the evolving conflict. An overlap proportion of 100% indicates that the entire time the conflict pair was evolving occurred during the PM retention interval. Figure 4.5 shows predicted detection probability by overlap proportion. There was a significant effect of overlap proportion, $\beta = -0.94$, $SE = 0.13$, $\chi^2(1) = 51.78$, $p = < 0.001$, with higher overlap being associated with poorer conflict detection accuracy. Conflict response time was not examined as it varied systematically as a function of conflict duration (i.e., how long it takes the aircraft pair to violate separation from when they were first both on the display), and thus did not allow unconfounded comparison.

![Figure 4.5. Effect display plots for the PM overlap proportion. The line represents the predicted detection probability means for each overlap proportion value (x-axis). Error bars reflect 95% confidence intervals for the fixed effect.](image)
Chapter 4: PM in ATC Robust to Interruptions but Impaired by Retention Interval

4.8. Discussion

PM errors can have serious safety implications in complex work domains such as ATC (Dismukes, 2012; Loft, 2014). The current study examined how PM in simulated ATC was affected by the PM retention interval and sought to identify any contributions of encoding and execution delays under conditions where participants were interrupted. Our choice of experimental manipulations was motivated by previous research and ecological concerns, and our predictions motivated by the DMPV of PM (Scullin et al., 2013) which suggests that PM is supported by top-down strategic monitoring and maintenance, bottom-up cue-driven processes, and their interaction. We found that PM errors increased with longer retention intervals, but were not affected by the presence of an interruption during the retention interval. Further, there was no evidence that encoding or execution delays influenced PM error or RT. Conflict detection accuracy and routine aircraft handoff RT were both impaired during the PM retention interval, suggesting that participants relied on top-down PM maintenance and monitoring processes.

The finding that PM error rates and PM RTs were unaffected by interruptions sits in contrast to findings in basic tasks, where interruptions have been found to negatively impact PM (Dodhia & Dismukes, 2009; McDaniel et al., 2004). Further, there was no increase in resumption time for handoffs immediately after the interruption. Interruptions may have failed to affect PM because shifting back to the pre-interrupted context cued participants to reinstate PM monitoring, as suggested by DMPV (Martin et al., 2011; McDaniel et al., 2004). Participants may have done so by utilizing a meta-cognitive ‘offloading strategy’ (Risko & Gilbert, 2016), associating the PM intention with spatial and contextual features of the ATC display (Todorov et al., 2018). Offloading strategies can eliminate the costs of interruption on PM tasks in basic paradigms (Gilbert, 2015; Risko & Gilbert, 2016). Interestingly however, 70% of subjects reported in a post-experiment questionnaire that the deferred handoff task was the task made most difficult by the interruption. This discrepancy between subjective reports and our findings highlights the importance of conducting empirical studies: subjective intuition obtained through qualitative methods may not align with objective performance data.

Although the lack of effect of interruptions on PM handoff errors is surprising, this is consistent with recent finding from Wilson et al. (2018). Although Wilson et al. found interruptions increased RT and resumption errors on a deferred conflict detection task that
required a response immediately following an interruption, they did not find that interruptions impacted PM handoff errors. They reasoned that the effects of interruptions on PM might vary depending on the temporal proximity of PM encoding and execution to the interruption, whereby PM is improved by consolidation and recovery, respectively. However, we found no evidence that the effects of encoding or execution delay differed between our interruption conditions, indicating that the magnitude of the retention interval alone underpinned the observed PM error rate. Perhaps the differences between the deferred conflict task and handoff task can be attributed to the complexity of conflict detection, or the reliability of cuing of the deferred handoff task (i.e., aircraft flashes blue for handoff at a predictable future time). It would be valuable for future research to examine what properties of deferred tasks promote robustness to interruptions in applied contexts. The results here indicate that the timing of the deferred task relative to an interruption was not an important factor determining PM performance in simulated ATC. This contrasts with research suggesting that time for consolidation before an interruption, and time for recovery afterwards, may benefit PM (Dismukes & Nowinski, 2006; Hodgetts & Jones, 2006a; Wang et al., 2018; Labonté et al., 2019). One possibility is that encoding and execution delays may simply not impact PM when contextual cues can be quickly reinstated. However, because PM performance was unaffected by the interruptions in the current study, it is possible that individuals did not have to mitigate any disruptive effects through preparation or recovery strategies. Thus, future research needs to examine the effect of encoding and execution delays on PM under conditions where interruptions negatively impact PM.

In line with subjective reports from experienced controllers (Loft et al., 2013), and with the DMPV, we found PM errors increased over longer retention intervals, reflecting the challenge associated with sustaining PM monitoring over long durations. This result was not guaranteed, given that the continued presence of the PM aircraft on the display could potentially have overcome the negative effects of PM retention interval. Assuming individuals were engaging in strategic offloading as suggested above, the negative effect of retention interval might indicate that the associations between cues and the intended action diminished over time. This would also explain why Stone et al. (2001) found no effect of retention interval on PM. Participants in Stone’s study performed multiple PM tasks with overlapping retention intervals,
thus each consecutive PM instruction may have facilitated both recall of the remaining PM tasks, and strengthened the PM cue-intention associations.

Holding a PM intention also produced robust costs to ongoing ATC tasks. During the PM retention interval, participants were slower to handoff routine aircraft and had poorer conflict detection accuracy. Costs to routine handoff RT likely may indicate that individuals were strategically inspecting aircraft when handing them off for PM features (e.g., callsign of aircraft; relative spatial location). By contrast, there were no costs to aircraft acceptances as they were unrelated to PM and thus there was no PM features to inspect. However, conflict detection was also PM unrelated, but was susceptible to the effects of PM load. Conflict detection is likely sensitive to shifts in allocated resources due to the high degree of attentional and cognitive demand posed by the conflict detection task (see Loft & Remington, 2010). The acceptance task is unlikely to be sensitive to such an attentional burden, because acceptance events were perceptually salient and did not require complex decision making. Thus, conflict detection PM costs may have occurred because participants were engaging in some form of active maintenance (e.g., PM rehearsal) that consumed limited cognitive resources (70% of participants reported using a rehearsal strategy in the post-experiment questionnaire), or because they were spending increased time attending to the PM aircraft (i.e., triggering bottom up cues).

4.8.1. Limitations, Future Directions, and Practical Implications

The use of a student sample with limited training does constrain our ability to generalize the results to expert controllers. In addition to the differences in cognitive skill and motivation between experts and students, controllers learn to recognize specific events that occur routinely at certain sector locations (Bowden & Loft, 2016; Stein, Garland, & Muller, 2009), greatly reducing demands on their executive processing. The results of the current study may hold greatest relevance to situations where there aren’t predictable patterns that controllers could rely upon for automatic processing (e.g., the sector is not highly familiar to controllers). Another limitation was that we were unable to examine how PM load impacted performance on the interrupting ATC task (as a PM intention was active in all interruption scenarios). Future research could also manipulate the retrospective memory demands, that is what must be
remembered (i.e., the action to-be perform), and *when* it must be performed (i.e., PM cue features).

There are several practical implications of the current research. First, the presence of PM tasks and their respective retention intervals should be considered when conducting cognitive work design in complex jobs. We demonstrated that longer retention intervals not only lead to higher rates of PM error, but increasing the retention interval of the PM task also increases the risk that performance on other ongoing tasks will be contaminated by the PM load in simulated ATC. Thus, in applied contexts where operators are required to monitor dynamic displays and make complex cognitive judgements, it is crucial to attempt to minimize demands on PM and the retention interval of contaminant PM tasks. This recommendation is supported by our finding of a dose-dependent between PM overlap and proportion of conflicts missed.

In conclusion, the current study showed that longer retention intervals caused PM deficits and increased the risk of costs to ongoing conflict detection accuracy. As automation solutions emerge in ATC and other complex dynamic work tasks, it will be critical to continue to evaluate the nature of the memory load placed on human operators to prevent excessive memory demands competing with the overall safety of the human-machine system. Practitioners must examine whether automation solutions themselves inadvertently increase PM demands, for instance, by increasing the interleaved monitoring of greater numbers of concurrent tasks (Loft et al., 2019).
4.9. Appendix 4A: Workload Results

As noted in the method, two workload items were captured on the second day. The two workload items, effort and frustration, were highly correlated $r(68) = 0.74, p < 0.001$, and thus collapsed into a single composite score. Composite workload was significantly associated with sector density (i.e., number of aircraft within a trial), $[\beta = 0.1, SE = 0.03, \chi^2(1) = 13.01, p = < 0.001]$. In turn, composite workload significantly predicted deferred handoff errors (over and above effects of retention interval), with increased workload ratings being associated with poorer PM performance $[\beta = 0.13, SE = 0.09, \chi^2(1) = 14.61, p = < 0.001]$. However, additionally including sector density (number of aircraft on display) did not account further improve model fit over the effects of subjective workload $[\beta = 0.13, SE = 0.09, \chi^2(1) = 2.11, p = 0.15]$. In summary, we found that increasing the objective task demands of each trial (via sector density) did increase subjective workload, and in turn, this subjective workload increase decreased PM performance, indicative of limited-capacity induced error.

4.10. Appendix 4B: Long-Term Effects

In order to examine the possible long-term effects of interruptions on routine task performance, we compared the no-interruption and interruption conditions on aircraft or routine tasks that occurred in the post-interruption period. This included, non-response errors and response times for both acceptances and handoffs, as well as proportion of conflicts that first entered conflict after the interruption correctly detected. There was no significant difference between the interruption conditions on post-interruption acceptance errors, $[\beta = -0.04, SE = 0.09, \chi^2(1) = 0.22, p = 0.64]$, or on acceptance RT, $[\beta = -23.19, SE = 40.78, \chi^2(1) = 0.32, p = 0.57]$. There was no significant difference for handoff errors $[\beta = 0.13, SE = 0.18, \chi^2(1) = 0.56, p = 0.45]$, but there was a significant difference for handoff RT $[\beta = 76.33, SE = 38.57, \chi^2(1) = 3.88, p = 0.049]$, however it was in the opposite direction to what would be expected (i.e., slightly higher for the no-interruption condition). Finally, there was no significant difference for proportion of post-interruption conflicts detected, $[\beta = -0.04, SE = 0.12, \chi^2(1) = 0.11, p = 0.74]$. In summary, there were no long-term effects of the interruption on any of the routine task measures.
4.11. Appendix 4C Questionnaire

The qualitative questionnaire and survey conducted at the end of the second session. The questionnaire data was not analysed other than the descriptive statistics presented in the discussion above, and the general discussion of this thesis.

<table>
<thead>
<tr>
<th></th>
<th>strongly disagree</th>
<th>disagree</th>
<th>neither agree or disagree</th>
<th>agree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>I felt stressed and/or anxious during an interruption.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2.</td>
<td>I felt stressed and/or anxious in the time soon after an interruption.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3.</td>
<td>I was irritated, frustrated and/or annoyed by the interruptions.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4.</td>
<td>Being interrupted had no effect on my stress or anxiety.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5.</td>
<td>Being interrupted made it harder to complete tasks on the primary scenario.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>6.</td>
<td>I got better at dealing with interruptions over time (i.e. with more experience).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>7.</td>
<td>I had to work harder to maintain my level of performance if I was interrupted.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>8.</td>
<td>Interruptions made me feel more fatigued, tired and/or exhausted.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>9.</td>
<td>Interruptions made remembering task goals more difficult.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>10.</td>
<td>I felt the interruptions were boring, irrelevant and/or unimportant.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>11.</td>
<td>Being interrupted made me feel confused and/or disoriented.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>
Chapter 4: PM in ATC Robust to Interruptions but Impaired by Retention Interval

The next 8 questions relate to any strategies or techniques which you may have used to remember task goals (such as the arrow key handoff task). When answering these questions, think back to your experiences in both the first and second day. Answer the open ended questions in the space provided. The scales below range from strongly disagree, to strongly agree. Please indicate how much you agree or disagree with each of the following statements by ticking the appropriate box:

<table>
<thead>
<tr>
<th></th>
<th>strongly disagree</th>
<th>disagree</th>
<th>neither agree or disagree</th>
<th>agree</th>
<th>strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I used rehearsal techniques to accomplish task goals (e.g. repeating objectives or aircraft names in my head).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2</td>
<td>I did not use any special techniques or strategies.</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3</td>
<td>I used spatial or visual memory to help me remember important task goals (e.g. remembering the approximate location of an aircraft).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4</td>
<td>I used physical space to help me remember task goals (e.g. putting a finger on the special key or moving hands/feet to serve as reminders).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5</td>
<td>I was more cautious or conservative in my behaviours (e.g. taking more time to respond to an aircraft handoff).</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

6. If you used another form of strategy, please specify:

..................................................................................................................................................

7. What was the most difficult task for you to do (overall)?

..................................................................................................................................................

8. What task did you feel was made most difficult by the interruption?

..................................................................................................................................................
Chapter 5: General Discussion

5 General Discussion

“The notions of intention and error are inseparable. Any attempt at defining human error or classifying its forms must begin with a consideration of the varieties of intentional behaviour.” – James Reason (1990, p. 5).

In complex and dynamic task environments, operators are often required to remember to complete an intended deferred task, requiring PM (Dismukes, 2012). Failures to carry out PM intentions in safety critical work environments can have serious safety implications, and thus it is of high priority to determine the factors which impact PM in these work environments. The aim of this thesis was to examine what situational factors (e.g., interruptions) may modulate the ability of individuals to remember to perform deferred tasks (i.e., to prospectively remember), and to examine how interruptions influence task performance in simulated ATC more generally. In Chapter 2, I described several extensive modifications to Fothergill et al.’s (2009) ATC-Advanced simulator which enabled the administration of interruptions and I detailed the motivation for, and implementation of, two ecologically valid deferred tasks that corresponded to operational circumstances faced by air traffic controllers: a deferred handoff task and a deferred conflict task. In Chapter 3, I examined how different forms of interruption impacted both these deferred tasks and ongoing ATC task performance. In Chapter 4, I focused on whether interruptions impacted the deferred handoff task, and the extent to which the retention interval of the deferred task influenced performance. In this chapter, I will begin by briefly reviewing the motivation for this thesis and summarising the core empirical findings. Then, I will discuss how the current work contributes to the broader psychological theory introduced in
142

Chapter 1, and I will address the practical implications of this work, as well as the limitations and opportunities for future research.

5.1. Summary of Empirical Findings

In the first empirical chapter (Chapter 3), I report two experiments that examined the impact of interruption type (no-interruption, n-back task interruption, secondary ATC task interruption) on individuals’ ability to complete a deferred handoff and conflict task, as well as on routine ongoing primary task performance. On the basis of MFG theory and Threaded Cognition, it was predicted that any interruption which occluded the display would increase the time taken to remember and resolve the deferred conflict (i.e., resumption time), and possibly increase the probability of resumption failure – that is, forgetting the task goal. Further, for the deferred handoff task, I reasoned that interruptions could interfere with individual’s ability to maintain the problem state associated with the deferred task goal, thereby reducing deferred task accuracy, particularly if the interruption was cognitively demanding or caused memory interference. Indeed, findings of PM costs to ongoing ATC tasks from prior research indicated that individuals likely engage cognitive control throughout the deferred task retention interval in order to maintain the intended action and associated problem state (Loft, 2014). Relative to no interruption, a blank display interruption slowed deferred conflict resumption, but this effect was not augmented by a cognitively demanding n-back task or a secondary ATC task interruption. However, the ATC task interruption increased the probability of failing to resume the deferred conflict relative to the blank interruption. An ex-Gaussian model of resumption times revealed that these resumption failures reflected true forgetting of the deferred task. This indicated that the resumption failure effect for the ATC interruption was driven by interference with the primary task due to the overlapping demands on both visual-spatial memory and ATC task goals (Altmann & Trafton, 2002; Borst et al., 2010; Logie et al., 1990; Salvucci & Taatgen, 2011). However, deferred handoff task performance was unaffected by interruptions. I reasoned that participants may have engaged in a cue-driven, “just-in-time” reactive strategy, and allocated attention to the deferred handoff tasks only at the time that the routine deviation was required (Braver, 2012). These findings provided evidence that PM for deferred tasks in our simulated ATC task depended on frequent interaction with situational cues on the display, and that prospective memories were particularly susceptible to interference-based forgetting.
The surprising null effect of interruptions on the deferred handoff task in Chapter 3 served as the primary motivation for Chapter 4. Specifically, it was hypothesised that the negative effect of interruptions on deferred tasks may be mitigated by the provision of time prior to being interrupted (i.e., the PM encoding delay) owing to increased preparation time, or by the provision of time after the interruption (i.e., PM execution delay) which could be used to orientate to the updated visual scene and use contextual cues to trigger PM retrieval. However, increasing these delay intervals also increased the total retention interval, which prior theory indicated should impair PM. Therefore, I examined how performance on the deferred handoff task was impacted by (1) retention interval of the deferred task, (2) the presence of interruptions during this retention interval, and (3) the timing of the interruption relative to deferred handoff encoding and execution. Results showed that increasing retention interval (37 s to 117 s) reduced PM accuracy. This finding was consistent with DMPV of PM, suggesting that sustaining PM monitoring processes over long durations is challenging. Moreover, in line with previous ATC PM studies, costs were also detected to ongoing conflict detection accuracy and routine handoff speed when a deferred task intention had to be maintained. However, in replication of the findings from Chapter 3, interruptions did not affect speed or accuracy on the deferred handoff task. Further, there was no evidence that the increased recovery and preparation offered by increased encoding and execution delays modulated PM accuracy or RT. This null finding indicates that deferred task performance involves a dynamic interplay between top-down strategic maintenance and bottom-up cue-driven processes, and the rich contextual displays in complex task environments can be leveraged by individuals to mitigate the effects of interruptions. I discuss further explanations for this null effect, and how these findings fit more broadly with the literature, throughout this general discussion chapter.

5.2. Theoretical Contributions

5.2.1. Task Resumption Processes
The findings from this thesis offer several contributions concerning the cognitive processes involved in resuming an interrupted task in complex dynamic task environments. A key contribution, consistent with Labonté et al. (2019), is support for the notion that interruption recovery in dynamic environments likely involves two cognitive strategies: memory-based retrieval strategies and reconstruction processes. In Chapter 3, the slowed resumption time for
the deferred conflict under the blank interruption conditions suggested that individuals utilized display information to perform the deferred task goal, and individuals required time to re-develop SA for the post-interruption aircraft locations and status after the interruption (as individuals presumably could rehearse during this period). Initially, I took this finding as evidence that individuals were relying primarily on situational cues to trigger recovery-based reconstruction strategies. However, the fact that the ATC interruption caused an increase in resumption failures (likely owing to interference processes) indicated that individuals did store and actively maintain representations of the ATC task problem state, and therefore they likely utilized both cue-based reconstructive and active-maintenance memory strategies. Further, in Chapter 4, the finding of both a retention interval effect and the costs to ongoing task performance indicated either that individuals were actively maintaining the intent to deviate from routine and interruptions did not influence this process, or that they were able to successfully utilize bottom-up cues to recover the deferred task goal.

These findings have implications regarding the application and utility of MFG and Threaded Cognition for predicting the effects of interruptions in complex and dynamic task environments. According to MFG, maintaining task goals over the course of interruption is supported by rehearsal and thus negatively impacted by rehearsal inhibition, because situational cues cannot be attended to during an interruption (Altmann & Trafton, 2002; Hodgetts & Jones, 2006b). Thus, MFG would assume that individuals engage rehearsal of the episodic deferred task goal throughout the interrupting period (if an opportunity exists) in order to boost its activation; and interfering with these rehearsal processes with demanding interrupting tasks should lead to poorer resumption performance (Boehm-Davis & Remington, 2009; Monk et al., 2008). However, the findings from this thesis are difficult to reconcile with these primarily memory-based predictions from MFG. I did not find any additive decrements from the cognitively demanding n-back task and found that only the ATC task interruption resulted in a higher proportion of resumption failures. Threaded Cognition, however, specifies two mechanisms that provide a better account of the observed data: the problem state and reconstructive processes. Threaded Cognition specifies that individuals not only maintain individual threads associated with a specific task goal, but also maintain a memory representation of the information associated with and required to perform the goal, called the problem state. Consistent with other memory representations, the problem state can be
maintained in memory through prospective or retrospective rehearsal, as well as through the utilization of relevant contextual retrieval cues (i.e., environmental). Further, Threaded Cognition specifies that for long-duration interruptions or situations involving a highly complex task state, task resumption relies more heavily on reconstructing the problem state than on storing and retrieving it from memory (Borst et al., 2010; Salvucci, 2010; Salvucci & Taatgen, 2011a). From this perspective, the interruption failure induced by the ATC interruption indicates that individuals were likely maintaining a problem state, and that the mechanisms underlying interference of the problem state were consistent with broader psychological theory (e.g., working memory interference). It would be interesting to examine how the trade-off between memory-based strategies and reconstruction operates more specifically as a function of time. Two factors likely to contribute to an increased reliance on reconstruction strategies are the total interruption time and the speed that display objects evolve at (i.e., the extent of change in a display over a period of time). As interruption length and object evolution speed increase, naturally individuals would be more likely to utilize reconstruction strategies. A critical challenge however is to adjust these experiment parameters at increments such that memory-based strategies are still useful, because if the extent of change is too great, individuals would have no choice but to reconstruct the task context entirely.

5.2.1.1. Limitations of Threaded Cognition for SA

Despite Threaded Cognition’s appeal as a verbal account of interruptions in complex dynamic tasks, it should be acknowledged that a limitation of applying it in such contexts is the underspecified nature of the problem state resource. In Chapter 3, I argued that the problem state could be operationalised in a similar manner to SA. This explanation was motivated by Salvucci and Taatgen’s (2011) argument that complex tasks require deliberate and strategic reconstruction of the task and problem state (which overlaps with definitions of SA recovery), as well as previous work conceptualising the problem state as a placeholder for SA (Bai, Jones, Moss, & Doane, 2014; Hodgetts et al., 2015; Labonté et al., 2019; Schoelles & Gray, 2012). However, formally within the Threaded Cognition section of the ACT-R framework, the problem state is implemented as an “imaginal module” — a short-term store that allows the memorization of chunks, and dispatching of chunks to declarative modules (i.e., holds intermediate representations required in solving a problem or performing some task). Critically, however, the information contained in each module is assumed to be static. This
contrasts with theoretical models of SA that argue SA is dynamic, is updated by dynamic interactions with the environment, and can contain long-term interpolated visual representations (Chiappe et al., 2016; Endsley, 1995; Vidulich & Tsang, 2012). For example, as controllers’ SA increases, they can project future aircraft positions, and then update their task goals in light of these projections. Indeed, this interpolation is how deferred conflict task resumption was assumed to operate: individuals would use the problem state to guide visual orientation at resumption (perhaps interpolating flight trajectories), and in turn, this would serve to orientate individuals to bottom-up cues to retrieve the task goal and update their problem state. Thus, there is a discrepancy between Threaded Cognitions static problem state implementation and its common application as a placeholder for SA.

A further concern regarding the formal implementation of Threaded Cognition, as well as the verbal theory which surrounds its application, is that a problem state can only be associated with a single task thread at a time (Borst et al., 2010; Salvucci & Taatgen, 2011a, 2011b). This has serious limitations when assuming the problem state can be conceptualised as SA, because it is conceivable that general SA information would be shared across tasks and require multiple threads. For instance, the location of the deferred handoff aircraft could be used as both a location-based PM retrieval cue, as well as for ensuring safe separation between aircraft (i.e., conflict detection). Schoelles and Gray (2012) attempted to overcome this single-thread allocation limitation. They applied Threaded Cognition to an aviation simulation and implemented a “situation buffer” — a store similar to the problem state, but additionally allows information to be shared among threads (Schoelles & Gray, 2012). This enabled them to accommodate for the fact that monitoring flight speed (conducted routinely by pilots) requires knowledge about speed limits, and navigating requires knowledge about position, and both of these units of information must be shared by a range of other subtasks, such as navigating to the appropriate runway. Despite this more advanced instantiation of Threaded Cognition, the reported model was still limited by the fact that problem state information was static (and only updated explicitly) and did not account for the role of the operator offloading to the external environment (Chiappe et al., 2016, 2012; Gilbert, 2015). I argued in Chapter 4 that participants may have overcome the effects of interruption by utilizing a meta-cognitive ‘offloading strategy’ (Risko & Gilbert, 2016), associating the PM intention with spatial and contextual features of the ATC display. Threaded Cognition’s problem state does not sufficiently account for how these
Chapter 5: General Discussion

spatial and contextual features are maintained in a dynamic evolving environment. Thus, a clear direction for future theoretical development is in refining the concept of the problem state and identifying how it can better reflect the cognitive processes that occur in complex dynamic tasks.

5.2.2. The Impact of Interruptions on Subjective Workload

Interruptions (of all forms) had no effect on subjective workload measures in any of the reported experiments. This indicates that for the present ATC task environment, judgements of mental demand and effort are not influenced by interruptions. Interestingly, however, the questionnaire results from Experiment 3 indicated that 47.83% of participants agreed or strongly agreed that interruptions made it more difficult to complete primary scenario task goals (item 4), and 55.07% of participants agreed or strongly agreed that they had to work harder to maintain their level of task performance when interrupted.

Initially, it was hypothesised that the interruptions would increase ratings of mental demand and exerted effort when interrupted, due to the interruption recovery process placing substantial demands on the limited-capacity information processing system (Speier, Valacich, & Vessey, 1999). These findings add to the mix of results in the literature regarding the impact of interruptions on perceived mental workload. For instance, Kirmeyer (1988) found that interruptions can increase workload, and Loft et al. (2015) demonstrated that interruptions can increase subjective workload in complex dynamic tasks (as measured by the NASA-TLX). In contrast, Hodgetts, Tremblay, Vallières, and Vachon (2015) did not find that interruptions increased subjective mental workload in a complex dynamic task.

One explanation for this finding is that the condensed workload measures used in the present study and by Hodgetts et al. (2015) were not sensitive enough to detect the brief increases in workload posed by the interruptions. The post-scenario two-item workload questionnaire (in our case, taken up to 3 minutes after the end of the interruption) may have been insensitive to the momentary increase in task load posed by the interruption recovery process. This may be in part due to retrospective questionnaires being fallible to self-report biases (Muckler & Seven, 1992), participants’ being poor at reporting their own cognitive limitations (Levin, Momen, Drivdahl, & Simons, 2000), or some facets of increased workload being inaccessible to consciousness (Matthews, Reinerman-Jones, Barber, & Abich, 2015). This explanation points towards a more generalized problem of examining workload using subjective
measures in interruption paradigms: the expected workload increase is likely to occur for only a brief portion of the test trial, yet retrospective subjective measures require introspection over a long period of time. One possible method is the detection response task, in which participants are required to press a button in response to an easily detected stimulus (e.g., tactile vibration) that occurs randomly every 3–6 s. Castro, Strayer, Matzke and Heathcote (2019) found (using an accumulation model approach) that the detection response task was sensitive to dynamic fluctuations in cognitive workload (i.e., shifts in capacity-limited attentional processing resources).

5.2.3. Long-Term Effects of Interruptions

A noteworthy finding is that interruptions had no long-term effects on performance in any of the studies reported in this thesis. Previous research has found that interruptions in complex and dynamic task environments require some time to recover from, as recovery from interruption may require several steps and manifest cognitive processes (Loft et al., 2015; Tremblay et al., 2012).

In Chapter 3, I argued that a possible reason for this finding is that the analysis of ongoing task performance was limited in resolution to the entire post-interruption period, but the effects of the interruption may have decayed, and been temporally limited. However, exploratory analyses conducted for Chapter 4 did not show any clear pattern of changes in acceptance or handoff RT as a function of time. Instead, a more likely reason for the lack of long-term effects is related to the complexity and design of our ongoing tasks. Specifically, both acceptance and handoff tasks are behaviourally very simple, requiring the detection of the aircraft followed by a mouse click and single key stroke. Even though conflict detection is more complex, there was only a single conflict per trial in the post-interruption period. By contrast, long-term effects reported elsewhere in the literature are typically seen on criterion variables related to complex decisions, such as: making classification/engagement decisions and identifying changes in vessel direction (Loft et al., 2015), and decision-making times (Hodgetts et al., 2015, 2013; Tremblay et al., 2012). Thus, the current results suggest that long-term effects likely arise due to the time required to recover SA of the post-interruption display, and costs will only be observed on tasks which heavily rely on high level SA. This could be examined in the
current ATC paradigm by modifying the acceptance and handoff tasks to involve more complex decision criteria (e.g., involve the aircraft’s vectoring history, or based upon flight parameters).

5.2.4. Sector Density, Workload and PM Capacity Sharing

An important finding from Chapter 3 is that increased workload, by way of increased air traffic, reduced deferred handoff performance. Specifically, sector density was identified as the primary factor that drove high workload ratings for a given scenario, and it was in these high workload scenarios that individuals were most likely to forget the deferred handoff task. This finding is largely consistent with Stone et al. (2001), who found that PM performance in their ATC task was significantly worse when workload (number of aircraft in sector) was higher. However, Stone et al. (2001) did not assess workload ratings, thus could not attribute a specific mechanism between sector density and PM performance. There are two likely cognitive explanations as to why increased sector density and subsequent high workload could reduce PM performance.

The first explanation can be derived from the notion that workload ratings are a monotonical function of the perceived ratio between task demands and capacity available to perform the task (Estes, 2015; Hockey, 1997; Vidulich & Tsang, 2012). Individuals reporting higher levels of stress and task-directed effort to achieve overall performance are likely encroaching the ‘red-zone’, the region of workload where additional task load can compromise performance on single tasks without adjusting task processing strategies (Loft et al., 2007; Strickland et al., 2019). In these situations, the addition of a PM load may breach cognitive capacity limitations, leaving individuals no option but to draw capacity from the routine task operation to support PM. This is in line with Strickland et al. (2019), who quantitatively modelled PM in a complex dynamic unmanned maritime surveillance task using an evidence accumulation framework. Their results provided clear evidence for capacity sharing between PM monitoring and ongoing task performance resulting from their task placing demands near to the ‘red-zone’. Further evidence for capacity sharing in our experiments is given by the PM cost effects identified in Experiment 3. Specifically, the costs to ongoing aircraft handoffs and conflict detection accuracy indicate that in order to support PM maintenance processes, individuals had to trade off performance on the ongoing tasks.

The second explanation, which is by no means mutually exclusive with the first, is that increased workload and sector density also trigger an increase in subtasks to be performed. With
each aircraft added to the sector, participants not only have to accept and handoff more aircraft, but within the sector there are more potential conflicts, and conflict resolution solutions increase in complexity. For instance, individuals must ensure that any altitude alterations issued do not create further conflicts, and increased aircraft require increased checks to ensure this. Each of these additional manifest cognitive processes required may function as distractors to deferred task maintenance, iteratively increasing the probability of dual-task interference (Craik, 2014; Graydon & Eysenck, 1989; Loft & Remington, 2010; May et al., 1999). For expert controllers in the field, it would be important to establish the extent to which PM is negatively impacted by increased complexity of tasks during the retention interval, relative to simply the increased frequency.

It should be acknowledged that while sector traffic density (number of aircraft) is commonly used in organisations to estimate workload, it does not fully capture the complexity of air traffic situations (Corver et al., 2016; Durand, Gotteland, & Matton, 2018; Malakis, Psaros, Kontogiannis, & Malaki, 2019). For example, a large set of aircraft with similar performance travelling orderly along a standard route may not impose high workload on a controller compared with a smaller number of aircraft with variable performances flying in complex geometric patterns (Boag, Neal, Loft, & Halford, 2006). Nevertheless, the implications of the current result are clear: deferred tasks are at a higher risk of being forgotten when workloads are high, and the evidence in this thesis indicates sector density is a key mediator of this relationship.

5.2.5. Further Integrating Prospective Memory and Interruptions

A key contribution of this thesis was the application and integration of both activation-based models from the interruption literature as well as theories of PM. In both Chapter 3 and Chapter 4, I drew upon these frameworks and synergised their respective predictions to each deferred task. As discussed in Chapter 1, the interruption and PM literatures share many common underlying assumptions. In real world contexts, it is challenging to imagine situations that interruptions and PM would not co-occur. Specifically, interruption situations almost always involve some level of PM, as an individual must remember to return to an intended task; and PM tasks in most real-world contexts are likely to involve interruptions during the retention interval (Dodhia & Dismukes, 2009). Furthermore, the major theories of both literatures share a
similar set of common assumptions. For instance, interruption activation-based models and PM’s DMPV all assume that in order for individuals to remember to resume a deferred task, they can engage some form of top-down strategic process, such as maintenance or monitoring or rehearsal; and that memory can be supported by some form of bottom-up cue-driven process, such as attending to contextual cues or noticing environmental features (Altmann & Trafton, 2002; Scullin et al., 2013). However, literature documenting how best to consolidate PM and interruptions theory is extremely limited. The work in this thesis shows how both theories can be applied concurrently to better understand deferred task performance and interruptions. In Chapter 3, I demonstrated how applying both PM and interruptions theories enhanced our overall understanding of the possible mechanisms likely to be driving deferred task performance. Further, in Chapter 4, I demonstrated how concepts from the interruption’s literature can be conceptualised in PM-specific terminology to better understand the factors driving performance (e.g., resumption lag to execution-delay).

5.2.6. Modelling Absolute Forgetting
In Chapter 3, I introduced a computational model of resumption failures that used an exponential Gaussian distribution to model the probability that the observed resumption failures were a result of forgetting and not truncated (but intended) response times. The ex-Gaussian model may be useful to researchers interested in any interruption situation in which individuals must respond to an event immediately after an interruption. In particular, the $\mu$ parameter of the model represents the shift of the ex-Gaussian function and captures how long it takes for the sampling process to initiate. The $\mu$ parameter estimates showed that participants were slower to initiate responding to the deferred conflict in the three interruption conditions compared to the no-interruption condition, and that this slowing was greater for the ATC and $n$-back interruption conditions compared to the blank interruption condition. Given the ATC task context, it is possible that this slowing indicates time taken to perform initial scanning of the task display.

However, it is important to acknowledge several limitations of the model. The main limitation associated with the ex-Gaussian model of resumption failures was that it was applied to group level data (i.e., assumed the data had all been obtained from a single subject). While numerically convenient, this approach assumes that cognitive processes operate in the exact
same way for all participants. This can lead to the properties of the average model not being representative of any given individual to which it is applied (Ly et al., 2017). Put simply, a ‘one size fits all’ approach is unable to account for performance at the individual level. This has been shown to be problematic in applied contexts where there can be substantial variation across individuals. The key challenge in extending our model to the individual level is that a great amount of data is required to obtain model convergence. Computational models are widely applied in cognitive science, but these numerical and sampling constraints underlie why they are less frequently applied in human factors practice. It should be acknowledged that the ex-Gaussian model has been criticized because its parameters do not show unique patterns of association with the well-established psychological processes from the diffusion model (Matzke & Wagenmakers, 2009). However, such criticism is limited only to its application in two-choice response tasks (e.g., lexical decision), whereas in Chapter 3 it was applied in a single-choice task similar to free-recall memory research (Rohrer & Wixted, 1994; Wixted & Rohrer, 1994). Although further model exploration would be beneficial, these criticisms of the ex-Gaussian model do not apply to interruptions researchers modelling only resumption time distributions.

5.3. Practical Implications

Qualitative and ethnographic studies of ATC, such as Shorrock’s (2005) field study, have highlighted that PM is required in a range of ATC tasks, including in conflict detection, and controllers do sometimes forget to complete PM tasks (Dismukes, 2012; Shorrock, 2005). While these studies provide valuable insights regarding the general conditions and precursors to PM errors, they lack the experimental control required to more precisely understand the conditions and cognitive processes that drive performance. The current thesis demonstrates how a “use-inspired” basic research paradigm (Stokes, 1997) can be used to begin bridging the gap between the theory and methods from cognitive science to applied contexts. Though concentrated on ATC, this thesis provides several practical contributions to all workplace environments in which operators monitor complex dynamic displays. Moreover, this thesis has shown that for cognitive science to be relevant to applied contexts, researchers cannot simply assume that findings from more basic task environments will generalise to more complex real-world situations. Applied researchers must also test theories of cognitive processing in tasks that are sufficiently ecologically valid to determine how cognitive processes might operate under the constraints of a complex work system (Morrow, 2018; Stokes, 1997). Ultimately, such studies should be...
proceeded by higher-fidelity experimental protocols involving expert participants from the work domain in question (Sanderson & Grundgeiger, 2015).

The observed dose-dependent relationship between PM retention interval and cost to conflict detection in Chapter 4 has important implications for any applied context involving PM. It indicates that longer retention intervals not only result in higher rates of PM error, but increasing the retention interval of the PM task also increases the risk that performance on other ongoing tasks will be contaminated by the PM load. Previous research has demonstrated that PM costs can be almost eliminated through strategically positioned and timed display aids that signal when the PM action is required. However, designing such tools in real world systems is not as trivial as it is for laboratory conditions. Ideally, the emergence of new automation technologies across a range of safety-critical industries will be designed with the cognitive limitations of humans in mind. It is also important to consider that environmental context faced by expert controllers is likely to offer a plethora of memory offloading strategies not readily captured through experimental simulation research.

Finally, the finding that increased workload was associated with increased PM failures points towards the importance of evaluating the likely workloads of deferred task situations when conducting cognitive task analyses. This applies to the design assurance of future C2 work systems irrespective of the changing landscape of automation technologies. While new automation technologies promise increased performance capabilities and reductions in operator workload, these changes may place C2 operators in an increasingly supervisory role, with greater fluctuations in workload (Metzger & Parasuraman, 2005). The Law of Stretched Systems states that for each advancement in technological capability, operational personnel typically are placed under pressure to do more tasks, more efficiently, and in more complex ways (Woods & Dekker, 2000). In other words, operator task demands are generally stretched and ‘fill’ any capacity freed by technological interventions. Thus, it is important to isolate what deferred tasks will remain the responsibility of operators in new complex C2 systems, and not assume automation will result simply increase capacity for maintaining PM performance.
5.4. Further Research Directions

5.4.1. Modelling Absolute Forgetting of Deferred Handoff Task

A fruitful direction for future research would be to model the probability of absolute forgetting on the deferred handoff task. Previous work examining PM in simulated ATC (Loft, 2014) has assumed that failing to remember to perform an intended atypical action in place of a routine action occurs due to habit capture (Reason, 1990). Habit capture errors are thought to occur when individuals fail to inhibit an expectation-driven processing bias cued by automated behavioural routines and thus fail to notice that the features of the task indicate that the alternative PM task action should be executed in place of the routine action (Norman, 1981; Reason, 1990). This account suggests that deferred handoff errors occurred in our ATC simulation because participants failed to perform attentional checks of the task environment at the time that deviation from routine is required, but critically, the deferred task intention was still active. That is, the intention itself had not been forgotten, but failed to activate at the appropriate time. However, it is possible that at least some proportion of the deferred handoff errors in our task occurred because the PM decision process failed to trigger entirely – that is, the intention was forgotten and individuals failed to even monitor the aircraft for cues indicative of the deferred handoff aircraft (e.g., its callsign or spatial location), and thus failed to perform the routine action. Indeed, given that deferred handoff errors increased over longer retention intervals (Chapter 4), it does seem likely that some proportion of these errors could have been attributable to absolute forgetting, rather than simply performing the incorrect response.

Failures of the PM decision process to activate are referred to as “trigger failures” in stop-signal task paradigms (Matzke, Hughes, Badcock, Michie, & Heathcote, 2017; Matzke, Love, & Heathcote, 2017), and several computational modelling techniques are available to estimate the probability that errors occurred due to trigger failures. “Prospective Memory Decision Control” (PMDC; Strickland et al., 2018) is one such framework, and has recently been successfully applied to studies of PM in lexical decision tasks, and to lower-fidelity simulations of ATC (Boag, Strickland, Neal, & Loft, 2019). Thus, a clear next step is to implement the trigger-failure model, using a relatively simple evidence accumulation model (e.g., exponentially shifted Wald distribution) to better understand the different ways that PM errors can manifest.
5.4.2. The Nature of the Interrupting Task

In the experiments presented within this thesis, I examined three forms of interruption. However, real-world situations likely involve a wide range of different interruptions. A promising direction for future research would be to enable properties of the ATC interrupting task to be further defined and manipulated. For example, enabling variations in the perceptual, motor, and cognitive difficulty of the scenario, or manipulating interrupting task duration while controlling the number of activities performed, as is often possible when manipulating interruptions in more basic task paradigms (e.g., Gillie & Broadbent, 1989; Hodgetts & Jones, 2005; Monk et al., 2008). In Appendix 5A of this chapter, I present a brief overview of an ATC-Advanced paradigm that I developed in a distinct research stream. The task was designed to be run as a primary task. Participants engaged in a conflict detection task that required them to identify which aircraft pair from a set of aircraft pairs was in conflict. Across conditions, participants monitored a sector containing 2, 3 or 4 conflicts, and the difficulty of the conflict detection was manipulated to be easy (clear violation of lateral separation) or hard (difficult perceptual judgment to ascertain lateral separation violation). Results of the preliminary study suggested clear and highly reliable accuracy and speed costs associated with more difficult trials and with a greater number of pairs on the display. Including this as an interrupting task would enable a better understanding of how factors, such as cognitive load or task difficulty, affect deferred task maintenance, retrieval and encoding in complex and dynamic task environments. It would also be of value to develop manipulations using this paradigm in terms of interruptions likely to be experienced from expert controllers.

5.4.3. Assessing the Impact of Interruptions on Workload

Future research may benefit from concentrating workload measurement around the brief interruption recovery window by using online workload measures (i.e., asking during the task, rather than the end of a trial). Several methods exist to support online measurement of workload specifically during the interruption recovery window with high levels of sensitivity, including DRT (e.g., Castro et al., 2019) and eye tracking (e.g., Hodgetts et al., 2015a).

However, recent work has stressed the importance of not oversimplifying and conflating measures of cognitive workload with the construct of workload itself (De Winter, 2014). Workload may not be a unitary latent construct: when information processing demands
increase as a function of increased task load, there may be multiple neurocognitive responses depending on the task environment and individual (Matthews et al., 2015). In turn, only certain facets of workload may be captured by particular measures (e.g., those available to conscious awareness for subjective measures; or attentional processes for DRT) and only certain task environments may be sensitive to these fluctuations, such as those which elicit long recovery processes (e.g., Loft, Sadler, Braithwaite, & Huf, 2015). Therefore, it is prudent that future research examining workload and interruptions include multiple physiological measures of workload, such as eye tracking, EEG, and heart rate. This could help explain what facets of workload are increased by interruption in specific task environments and enable a more diagnostic approach to predicting the impact of interruptions. Moreover, such measures would be particularly useful in interruptions research, as they can provide continuous data, thereby revealing how workload changes in the different parts of the interruption process (interruption lag, during the interruption, resumption lag).

5.5. Conclusion

This thesis examined interruptions and deferred tasks in simulated air traffic control — a complex and dynamic task environment. Chapter 2 introduced a simulation paradigm capable of assessing the effects of interruptions on two forms of deferred task with robust counterbalancing procedures. Chapter 3 examined three forms of interruptions and found evidence for interference to deferred task resumption resulting from an additional ATC sector interruption, while there was limited evidence favouring rehearsal mechanisms. Chapter 4 explored the extent to which the timing of deferred tasks (relative to an interruption), interruptions, and retention intervals during the interruption impacted a deferred handoff task. Results indicated that retention interval was the primary factor driving deferred handoff performance, indicative of top-down control mechanisms. In conclusion, the empirical contributions of this thesis demonstrate that (1) the effects of interruptions in complex dynamic tasks are sensitive to the nature of the interrupting task and the deferred task itself, (2) the importance of producing simulations capable of robust quantitative models, and (3) that PM and interruptions theory mutually benefit from application to complex task contexts. It is anticipated that findings and methodologies from this thesis, all made openly available through various platforms, will prove useful as a research base for future work and support the overall safety and efficiency of complex and dynamic task environments.
Chapter 5: General Discussion

5.6. **Appendix 5A: An Improved ATC Interruption Paradigm**

5.6.1. **Appendix Foreword**

In the general discussion of this thesis, I outlined the design of an ATC interrupting task which would enable the experimenter to make fine adjustments to the cognitive and visual processing demands. This appendix reports a brief methodology that I developed in collaboration with Andrew Neal, Andrew Heathcote, Luke Strickland and Shayne Loft. I programmed and implemented this task independently for purposes external to the scope of this thesis. However, the task itself provides a promising example of how an ATC interrupting task could be refined to better understand the exact mechanisms underlying observed performance costs. Results from the trial are omitted here and not relevant to the intended application in the current thesis.

5.6.2. **Context of Task**

Broadly speaking, the purpose of the study was to model how evidence accumulation processes related to conflict detection would change under varying levels of difficulty and number of aircraft pairs. Participants were presented with a series of ATC trials, and in each trial, they were required to detect which aircraft pair (out of a set of \( n \) pairs) was in conflict. The number of pairs was manipulated as 2, 3 or 4. Trials were either easy, in which the difference between conflict and non-conflict pairs was more obvious, or difficult, in which the difference was more difficult to detect. Trial length was set to maximum of 40 s. Aircraft were hidden by a mask and were able to be revealed one at a time. History dots (trajectory tails) were present on the aircraft and evolved over a 10 s period, and after 10 s they became static (i.e., did not grow further than the 10 s of movement history). The task was piloted with 24 undergraduate psychology students from the University of Western Australia in exchange for partial course credit.

5.6.3. **Experimental Details**

The experiment uses a 3 (number of stimuli) x 2 (difficulty) within-subjects design. Difficulty was manipulated as the relative difference in distance of minimum separation (DOMS) between the conflict and non-conflict pairs. DOMS refers to the minimum lateral separation between two aircraft throughout their trajectories. A DOMS of zero indicates the two aircraft will directly collide. For the present experiment, aircraft conflicts occurred if the two aircraft came within 5 NM (or less) of each other. For the easy trials, the conflict had a DOMS of 2.5 NM and the non-
conflicts had a DOMS of 7.5 NM. For the hard trials, the conflict had a DOMS of 3.75 NM and the non-conflicts had a DOMS of 6.25 NM. Difficulty was manipulated as distance of min separation: 3.75/6.25 for hard conflicts and 2.5/7.5NM for easy conflicts.

5.6.4. Experimental Task
The conflict detection task was implemented in the ATC-Lab\textsuperscript{Advanced} simulator, version 0.4.28.Bell1.52 (Fothergill, Loft, & Neal, 2009). Figure 5.1 presents a screenshot of the task display. The black lines denote flight paths that aircraft travel upon (note, however, that in this experiment aircraft only travel on a subset of the flight paths). Aircraft are represented by a white-filled circle and an attached flight data block. A flight data block specifies an aircraft’s call sign, type, current altitude, and cleared altitude (the altitude an aircraft is cleared to climb to, descend to, or cruise on). As no aircraft were programmed to climb or descend in this experiment, cleared altitude and current altitude were the same and always separated by a $>$ symbol indicating the aircraft was cruising.

The sector comprised 12 aircraft crossing points, each placed at an edge of a dodecagon (i.e., a clockface structure) and an equal distance from the centre of the display. The lines forming the base of the dodecagon can be seen emanating from the centre of the display in Figure 5.1. The radius of the dodecagon (i.e., distance from center of screen to crossing point) was set to be equal to 2/3 of the distance between the center of the sector and the edge of the sector. The lines forming the crossing points of the dodecagon are those which do not emanate from the centre of the display. This resulted in a series of 12 crossing points, each an equal distance from the centre of the display, and the flight paths that intersected a crossing point were positioned at 60 degrees to each other. This meant that aircraft pairs could be positioned symmetrically around the screen, and all aircraft travelled in a clockwise direction.
Figure 5.1 Screenshot of the prototype ATC task interface. Participants click the large box to unveil the aircraft underneath it and hide the currently visible aircraft. They must click the smaller “CONFLICT” button on the target pair they believe to be in conflict.

The experiment comprised 192 trials of conflict identification. In each trial, participants were presented with either two, three, or four aircraft pairs, and each trial was either of hard or easy difficulty. Participants were required to identify which one of the aircraft pairs was in conflict (i.e., would violate the minimum lateral separation standard of 5NM). Importantly, only one aircraft pair could be inspected at a time: the other conflict(s) on the display were masked by a large button which read ‘REVEAL AIRCRAFT PAIR’. Clicking this button would unveil the aircraft pair underneath the mask, and at that time, a new mask would appear over the previously visible aircraft pair. Consequently, participants were required to iteratively unmask and inspect each pair one at a time. When participants had decided which of the aircraft pairs was in conflict, they were required to click the button underneath this pair, which was labelled ‘CONFLICT’. Upon clicking the conflict button, the response was recorded, the current trial ended, and the participant was presented with the next trial. The conflict button was programmed to only become active after 5 s to prevent accidently skipping trials. If a decision was not made within 40 s, the trial terminated, and this was recorded as a non-response.

Conflict detection was aided by the presence of history dots (white tails shown in Figure 5.1). These represented the movement of the aircraft over the previous 10 s. Importantly, the
Executing Deferred Tasks in Dynamic Multitasking Environments

History dots evolved gradually from the start of each trial and took 10 s to fully form. This was to encourage participants to inspect each conflict gradually. Participants were instructed to first inspect each aircraft pair in a clockwise fashion, and after all pairs had been inspected at least once each, then they were free to inspect aircraft in any order they liked until they had made a decision. Participants were instructed to respond as quickly and as accurately as possible.

5.6.5. Application as an Interruption Task
As outlined in the general discussion body, the general design of this task would be well suited as an interrupted ATC task. Specifically, it enables the manipulation of two parameters with direct correspondence to mental demand (difficulty and number of pairs). It would also allow examination of hypotheses specified in the general discussion of Chapter 3 regarding visual-spatial and task interference. Specifically, increasing the number of aircraft pairs on the display would increase the to-be-processed visual-spatial representations, assessing the role of interference (Oberauer et al., 2016; Ratwani & Trafton, 2008), while increasing the difficulty of the conflict detection (or even difficulty of the resolution decision criteria) would assess how the overall cognitive demand of the task undergirds interruption disruption (Hodgetts & Jones, 2005; Monk et al., 2008).
REFERENCES


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press on Demand.


Executing Deferred Tasks in Dynamic Multitasking Environments


172
References


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Executing Deferred Tasks in Dynamic Multitasking Environments


References


Executing Deferred Tasks in Dynamic Multitasking Environments


