VALIDATING INTEGRATED HUMAN PERFORMANCE MODELS INVOLVING TIME-CRITICAL COMPLEX SYSTEMS

by

Brian Francis Gore

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

© Copyright by Brian F. Gore (2010)
Abstract

Validating Integrated Human Performance Models Involving
Time-Critical Complex Systems
Doctor of Philosophy, 2010
Brian Francis Gore
Graduate Department of Mechanical and Industrial Engineering
University of Toronto

Abstract

The current research sets out to demonstrate a comprehensive approach to validate complex human performance models as applied to time-sensitive tasks. This document is divided into 4 sections. Section 1 (Chapters 1 – 3) outlines previous efforts in the literature that have attempted to validate complex human performance models in the field with an emphasis on manual control models, task network models, cognitive models and integrated architectures. Section 2 (Chapters 4 – 7) elaborates on a validation approach and applies it to a baseline model of a complex task in the air traffic control domain. Section 3 (Chapters 7-12) outlines the importance of adopting an iterative model development-model validation process and reports on the three model iterations in an attempt to improve the validity of the baseline model. Each model augmentation was validated using the same validation approach and measures that were defined in Section 2. Section 4 (Chapters 13-14) provides a discussion and interpretation of the model results and highlights contributions to the field of both model validation and the field of human performance modelling of complex systems.
Acknowledgments

ACKNOWLEDGEMENTS

I would like to express my appreciation to my thesis committee, Professor Paul Milgram, Professor Daniel Frances and Professor Mark Chignell for their input and guidance throughout my research process. I learned a great deal from our spirited discussions about the philosophy of human performance modelling and model validation. I also extend my appreciation to Dr. Ronald Laughery and Professor Baris Balcioglu for their time and effort in serving on my thesis committee.

Mr. Ken Leiden (currently with MOSAIC, formerly of Alion Science and Technology, Micro Analysis and Design Operations) was a critical resource for the model development effort. Without Ken’s invaluable help, the model’s performance would not have been as refined as it is.

I would like also like to thank a number of researchers at NASA Ames Research Center for their support throughout the thesis effort. Firstly, Ms. Sandra Hart provided guidance on time estimation and workload; Dr. Walter Johnson provided additional information on time estimation; Dr. Parimal Kopardekar allowed the use of Micro Saint Sharp and permitted the FEWS HPM simulation to be modified for this thesis research.

Ms. Jill Kamienski (formerly of Alion Science and Technology, Micro Analysis and Design Operations; retired Denver Center Controller) and Mr. Vernol Battiste (retired ATC/NASA CS; SJSURF) were each instrumental as ATC SMEs for the ATCo model logic and behaviours.

Finally, I extend my greatest thanks to my wife Becky and my 2-year-old son Jason; Becky, for her constant support and advice without which, this effort would not have been possible, and Jason, for keeping everything grounded in reality.
Table of Contents

CHAPTER 1: INTRODUCTION AND OBJECTIVES ......................................................................................... 2
1.1. INTRODUCTION .......................................................................................................................... 2
1.2. OBJECTIVES ............................................................................................................................ 3
   1.2.1. Validating Complex Human Performance Models ............................................................... 3
   1.2.2. Modelling Time Management Behaviour ........................................................................... 3
1.3. APPROACH ............................................................................................................................... 4
1.4. EXPECTED RELEVANCE .......................................................................................................... 5

CHAPTER 2: THE EVOLUTION OF HPMS .................................................................................................. 7
2.1. OVERVIEW OF HUMAN PERFORMANCE MODELLING APPROACHES ........................................... 7
   2.1.1. Simple Engineering Models ................................................................................................. 7
   2.1.2. Task Network Models ......................................................................................................... 8
   2.1.3. Cognitive Models ................................................................................................................ 9
   2.1.4. Integrated Models .............................................................................................................. 9
2.2. SUMMARY – CHALLENGES TO MODELLING HUMAN PERFORMANCE IN COMPLEX SYSTEMS ....... 11

CHAPTER 3: VERIFICATION AND VALIDATION ..................................................................................... 13
3.1. DEFINING MODEL VERIFICATION AND VALIDATION ................................................................. 13
3.2. APPROACHES TO MODEL VALIDATION ...................................................................................... 13
   3.2.1. Qualitative Approach to Model Validation ......................................................................... 13
   3.2.2. Quantitative Approach to Model Validation ....................................................................... 14
3.3. HISTORICAL REVIEW OF MODEL VALIDATION ....................................................................... 15
3.4. CHALLENGES IN VALIDATING COMPLEX HPMS ..................................................................... 16
3.5. EXAMPLES OF LARGE-SCALE VALIDATION EFFORTS OF HPMS OF COMPLEX SYSTEMS .......... 17
   3.5.1. AFRL Cross-Model Comparison and Validation Effort ......................................................... 18
   3.5.2. NASA Cross-Model Comparison and Validation Effort ....................................................... 19
3.6. VALIDATION LESSONS LEARNED .............................................................................................. 21
   3.6.1. Models Must be Developed and Validated for a Specific Purpose ..................................... 21
   3.6.2. Developing Valid Models is an Iterative Process .................................................................. 22
   3.6.3. Validation Requires Multiple Measures ............................................................................. 23
   3.6.4. Data Granularity .................................................................................................................. 23
3.7. SUMMARY OF THE CURRENT VALIDATION APPROACHES AND THE WAY FORWARD .............. 24

CHAPTER 4: AIR TRAFFIC CONTROL: A REAL-WORLD TIME-CRITICAL COMPLEX ENVIRONMENT ................................................................................................................................. 26
4.1. ATC HITL STUDY ....................................................................................................................... 27
4.2. DESCRIPTION OF FEWS STUDY .............................................................................................. 27
   4.2.1. Handoff Receipt .................................................................................................................. 29
   4.2.2. Conflict Detection and Resolution ....................................................................................... 32
4.3. DESCRIPTION OF DATA AVAILABLE FROM FEWS STUDY ...................................................... 33
   4.3.1. Receive Handoff Duration (RHD) ....................................................................................... 33
   4.3.2. Workload ........................................................................................................................... 34

CHAPTER 5: DEVELOPING THE BASELINE MODEL ............................................................................... 37
5.1. ENVIRONMENT MODEL ........................................................................................................... 37
5.2. GENERAL HUMAN OPERATOR MODEL ................................................................................... 38
5.3. ATC ACTIVITY MODEL ............................................................................................................ 40
5.4. TASK PRIORITIZATION ............................................................................................................ 40
5.5. EXERCISING THE BASELINE SIMULATION MODEL ................................................................. 42

CHAPTER 6: INITIAL VALIDATION OF THE BASELINE MODEL ............................................................. 44
6.1. ANALYSIS APPROACH .............................................................................................................. 44
Table of Contents

6.1.1. Validation Measures for the Model’s Specific Purpose ........................................... 44
6.1.2. Iterative Approach ............................................................................................................ 45
6.1.3. Multiple Measures .......................................................................................................... 45
6.1.4. Aggregate Versus Fine-Grained Level of Analyses ......................................................... 46

6.2. THE VERIFICATION AND VALIDATION VARIABLES .......................................................... 46
6.2.1. Measure 1: Workload ...................................................................................................... 46
6.2.2. Measure 2: Queue Length .............................................................................................. 47
6.2.3. Measure 3: Receive Handoff Duration (RHD) ................................................................. 48
6.2.4. Measure 4: Window Open Time .................................................................................... 49
6.2.5. Measure 5: Handoff Complete Time ............................................................................. 49

6.3. VERIFYING AND VALIDATING THE BASELINE MODEL ..................................................... 49
6.3.1. Measure 1: Baseline Model Workload ......................................................................... 49
6.3.1.1. Mean Workload for Baseline Model ......................................................................... 49
6.3.1.2. Workload Trend Output for Baseline Model .............................................................. 50
6.3.2. Measure 2: Queue Length for Baseline Model ............................................................... 51
6.3.3. Measure 3: Receive Handoff Duration (RHD) for Baseline Model ............................... 52
6.3.3.1. RHD t-test and Results for FEWS versus Baseline Model .......................................... 52
6.3.3.2. RHD Chi-Square Test and Results for FEWS versus Baseline Model ....................... 53
6.3.3.3. RHD Histogram for FEWS and Baseline Model ......................................................... 54
6.3.4. Measure 4: Analysis of Handoff Window Open Times for Baseline Model ................ 56
6.3.5. Measure 5: Adjusted Handoff Complete Time for Baseline Model ............................. 59

6.4. CONCLUSION OF BASELINE MODEL VALIDATION EFFORT ........................................... 61

CHAPTER 7: IMPROVING TIME MANAGEMENT IN COMPLEX HUMAN PERFORMANCE MODELS .......................................................... 64

7.1. THE STATE OF TIME MANAGEMENT IN COMPLEX HUMAN PERFORMANCE MODELS .......... 64
7.2. BACKGROUND LITERATURE ON TIME MANAGEMENT .................................................... 65
7.2.1. Task Management Literature .......................................................................................... 65
7.2.1.1. Task Management and Information Processing ......................................................... 65
7.2.1.2. Task Management and Control Modes ........................................................................ 66
7.2.2. Time Estimation Literature ........................................................................................... 68
7.2.3. Time Estimation Research Methods .............................................................................. 68
7.2.3.1. Human Estimation of Time Available (Tₐ) ................................................................. 70
7.2.3.2. Human Estimation of Time Required (Tᵣ) ................................................................. 72
7.2.3.3. Human Estimation of Time to Task Onset (Tₒ) ......................................................... 72
7.2.4. Workload and Time Management Literature ................................................................. 73
7.2.5. Summary of Literature on Time Management ................................................................. 74

7.3. A FRAMEWORK FOR MODELLING TIME MANAGEMENT IN COMPLEX SYSTEMS .............. 74
7.4. DEVELOPMENT OF A LINKING FRAMEWORK OF TIME MANAGEMENT ............................. 77
7.4.1. Task Managing .............................................................................................................. 78
7.4.2. Time Estimating ............................................................................................................ 79
7.4.2.1. Relationship among Tₐ, Tᵣ, and Tₒ in the Time Estimating Model .............................. 82
7.4.2.2. Impact of Tₐ, Tᵣ, and Tₒ on Task Performance ......................................................... 83
7.4.3. Illustration of Model Interactions within the Time Management Framework ................ 84

CHAPTER 8: VERIFYING AND VALIDATING THE TASK MANAGEMENT (TM) MODEL ................. 86

8.1. AUGMENTATION OF BASELINE TO “OPPORTUNISTIC-ONLY” MODE ............................... 86
8.2. VALIDATION OF THE TASK MANAGEMENT (TM) MODEL .................................................. 87
8.2.1. Measure 1: TM Model Workload ................................................................................. 87
8.2.1.1. Mean Workload for TM Model ................................................................................. 87
8.2.1.2. TM Model Workload Trend Output .......................................................................... 88
8.2.2. Measure 2: Queue Length for TM Model ...................................................................... 88
8.2.3. Measure 3: RHD for TM Model ..................................................................................... 89
8.2.3.1. RHD t-tests and Results for FEWS versus TM Model ................................................. 90
8.2.3.2. RHD Chi-Square Test Results for FEWS versus TM Model ...................................... 90
8.2.3.3. RHD Histograms of FEWS and TM Model ................................................................. 91
8.2.4. Measure 4: Handoff Window Open Time for TM Model .............................................. 92
8.2.5. Measure 5: Adjusted Handoff Complete Time Analysis for TM Model ........................ 95

8.3. CONCLUSION OF THE TM MODEL VALIDATION EFFORT .................................................. 98
## Table of Contents

**CHAPTER 9: DEVELOPING AND VERIFYING THE TIME ESTIMATION (TE) MODEL AUGMENTATION** ................................................................. 99

9.1. Modelling Estimates of Time Available ($T_A$) ................................................................. 99
9.2. Developing an Algorithm for Estimated $T_A$ ................................................................. 100
9.3. Workload Demands ......................................................................................................... 102
9.4. Verifying the Time Estimation Model Algorithm ............................................................ 104
9.5. Verifying the Time Estimation Algorithm Model Using a Simplified ATCo Task ............ 106
  9.5.1. Generic Model Development .................................................................................... 106
  9.5.2. Exercising the Generic Model ............................................................................... 110
9.6. Summary of Time Estimation Model Development ......................................................... 115

**CHAPTER 10: VERIFYING AND VALIDATING THE TIME ESTIMATION (TE) MODEL** ........ 117

10.1. Augmentation of Baseline Model with Time Estimation ................................................ 117
  10.1.1. Measure 1: TE Model Workload ........................................................................... 118
  10.1.1.1. Mean Workload For TE Model ...................................................................... 118
  10.1.1.2. Workload Trend Output For TE Model .......................................................... 119
  10.1.2. Measure 2: TE Model Queue Length ................................................................... 119
  10.1.3. Measure 3: TE Model RHD ............................................................................... 120
    10.1.3.1. RHD t-test and Results for FEWS versus TE Model ...................................... 120
    10.1.3.2. RHD Chi-Square Test and Results for FEWS versus TE Model .................... 121
    10.1.3.3. RHD Histogram of FEWS, TE Model .......................................................... 122
  10.1.4. Measure 4: TE Model Handoff Window Open Time .............................................. 123
  10.1.5. Measure 5: TE Model Adjusted Handoff Complete Time .................................... 126
10.2. Results SUMMARY FOR TE MODEL AUGMENTATION ............................................... 128

**CHAPTER 11: VERIFYING AND VALIDATING THE TASK MANAGEMENT + TIME ESTIMATION MODEL** ................................................................ 129

11.1. Final Iteration of Baseline Model Augmentation: TM+TE .......................................... 129
  11.1.1. Measure 1: TM+TE Workload ........................................................................... 130
    11.1.1.1. Mean Workload For TM+TE Model ............................................................. 130
    11.1.1.2. Workload Trend Output For TM+TE Model ................................................. 131
  11.1.2. Measure 2: TM+TE Queue Length .................................................................... 132
  11.1.3. Measure 3: TM+TE Receive Handoff Duration (RHD) ..................................... 133
    11.1.3.1. RHD t-test and Results for FEWS versus TM+TE Model ............................ 133
    11.1.3.2. RHD Chi-Square Test and Results for FEWS versus TM+TE Model ........... 134
    11.1.3.3. RHD Histogram of FEWS, TM+TE Model ................................................... 135
  11.1.4. Measure 4: TM+TE Handoff Window Open Time ............................................. 136
  11.1.5. Measure 5: TM+TE Adjusted Handoff Complete Time .................................... 139
11.2. Conclusion of TM+TE Augmentation Model Validation Effort ..................................... 141

**CHAPTER 12: A DISCUSSION OF MODEL PERFORMANCE** ............................................ 142

12.1. Model Performance Review ....................................................................................... 142
12.2. The Task Management (TM) Augmentation ............................................................... 142
12.3. The Time Estimation (TE) Augmentation ................................................................ 143
12.4. The TM+TE Augmentation ....................................................................................... 144

**CHAPTER 13: DISCUSSION** ......................................................................................... 147

13.1. Multiple Human Performance Measures .................................................................... 148
13.2. Impact of Resolution of Analysis on Validation ......................................................... 150
  13.2.1. Use of Aggregate Data ....................................................................................... 150
  13.2.2. Use of Fine-Grained Data for Validation Efforts ............................................... 151
13.3. Model Generalisability .............................................................................................. 152
13.4. Model Transparency .................................................................................................. 153
13.5. Data Availability for real-world complex Tasks ......................................................... 154

**CHAPTER 14: CONCLUSION** ..................................................................................... 157

14.1. Significant Contributions ......................................................................................... 157
Table of Contents

14.1.1. The Validation Approach........................................................................................................... 157
14.1.2. The TC Measure.......................................................................................................................... 157
14.1.3. The Task Management Model................................................................................................... 158
14.1.4. The Time Estimation Model....................................................................................................... 158
14.1.5. The Normative Model Results................................................................................................... 158

14.2. Limitations and Recommendations for Future Research................................................................. 158
14.2.1. Workload Projecting Component of the Time Management Model........................................ 159
14.2.2. Augment Time Management Model............................................................................................ 160
14.2.3. Augment Empirical Data of Time Estimation Model................................................................. 160
14.2.4. Enhance Time Estimation Bivariate Model Relationship.......................................................... 161
14.2.5. Expand Time Estimation Model to Account for T_R and T_O.................................................. 161
14.2.6. Assess Cross-domain Applicability............................................................................................... 161

CHAPTER 15: REFERENCES.......................................................................................................................... 162

APPENDIX A – COGNITIVE MODELS...................................................................................................... 177
APPENDIX B – INTEGRATED ARCHITECTURES..................................................................................... 178
APPENDIX C - HUMAN IN THE LOOP DATA FOR VALIDATION ............................................................ 179
APPENDIX D - EXAMPLE/SNAPSHOT OF ENVIRONMENT MODEL CODE .......................................... 181
APPENDIX E - BASELINE MODEL ENVIRONMENT SNAPSHOT ............................................................... 182
APPENDIX F - BASELINE MODEL VERIFICATION PROCESS............................................................... 183
APPENDIX G - TASK ANALYSIS AND WORKLOAD TABLE................................................................. 186
APPENDIX H - TASK NETWORK MODEL ............................................................................................. 187
APPENDIX I – ADDITIONAL IP INFORMATION AND TASK SCHEDULES........................................ 189
APPENDIX J – OVERVIEW OF TIME ESTIMATION LITERATURE......................................................... 190
APPENDIX K - HUMAN’S ESTIMATION OF TIME REQUIRED............................................................ 194
APPENDIX L - WORKLOAD AND TIME MANAGEMENT LITERATURE............................................... 196
APPENDIX M - LIST OF ACRONYMS..................................................................................................... 197
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Descriptive Statistics for FEWS RHD for 100% DL, 133% DL, and 133% No DL</td>
<td>34</td>
</tr>
<tr>
<td>2</td>
<td>Mean and SD of ATWIT WAK Workload Ratings by Traffic Condition</td>
<td>35</td>
</tr>
<tr>
<td>3</td>
<td>FEWS Baseline Model Taskload, Number of Handoffs and Data Points Removed</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Significance Tests on Adjusted Workload for Traffic Level Increases for FEWS and Baseline Model</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>RHD Chi-square and Correlation Analyses for Traffic Levels</td>
<td>53</td>
</tr>
<tr>
<td>6</td>
<td>Number, Relative Timing and Order for FEWS and Model Window Open Times</td>
<td>57</td>
</tr>
<tr>
<td>7</td>
<td>Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times</td>
<td>59</td>
</tr>
<tr>
<td>8</td>
<td>Relationship among T_A, T_R, and T_e</td>
<td>84</td>
</tr>
<tr>
<td>9</td>
<td>Predicted Interactions Between Levels of Workload and Resulting Conception of Time, Effect on Task, and Effect on Model Behaviour</td>
<td>85</td>
</tr>
<tr>
<td>10</td>
<td>Significance Tests for FEWS Adjusted, BL and TM Model Workload by Traffic Level</td>
<td>88</td>
</tr>
<tr>
<td>11</td>
<td>Chi-square and Correlation Analyses for FEWS, BL, and TM Model by Traffic Level</td>
<td>90</td>
</tr>
<tr>
<td>12</td>
<td>Number, Relative Timing and Order for FEWS and TM Model Window Open Times</td>
<td>93</td>
</tr>
<tr>
<td>13</td>
<td>Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times</td>
<td>96</td>
</tr>
<tr>
<td>14</td>
<td>Brown and Boltz (2002) Constant Error Duration Time Estimates</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>Generic Model Inner Loop Assumptions</td>
<td>108</td>
</tr>
<tr>
<td>16</td>
<td>The Number of Aircraft Generated by Simulation Model and Entry Time Spacing, as a Function of Workload Level</td>
<td>110</td>
</tr>
<tr>
<td>17</td>
<td>Descriptive Statistics of the Time Error Distributions for Baseline and TE Models</td>
<td>111</td>
</tr>
<tr>
<td>18</td>
<td>Significance Tests for Traffic Level Increases on Workload for FEWS and the TE Model</td>
<td>119</td>
</tr>
<tr>
<td>19</td>
<td>Comparison of FEWS, Baseline, TE Model t-tests</td>
<td>121</td>
</tr>
<tr>
<td>20</td>
<td>Chi-square and Correlation Analyses of FEWS, Baseline and TE Models</td>
<td>124</td>
</tr>
<tr>
<td>21</td>
<td>Number, Relative Timing and Order for FEWS and Model Window Open Times</td>
<td>124</td>
</tr>
<tr>
<td>22</td>
<td>Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times</td>
<td>126</td>
</tr>
<tr>
<td>23</td>
<td>Significance Tests for FEWS Adjusted, BL and TM+TE Model Workload by Traffic Level</td>
<td>131</td>
</tr>
<tr>
<td>24</td>
<td>Comparison of FEWS, Baseline, TM+TE Model t-tests</td>
<td>133</td>
</tr>
<tr>
<td>25</td>
<td>Chi-square and Correlation Analyses of FEWS, Baseline, and TM+TE Model RHDs</td>
<td>134</td>
</tr>
<tr>
<td>26</td>
<td>Number, Relative Timing and Order for FEWS and Model Window Open Times</td>
<td>137</td>
</tr>
<tr>
<td>27</td>
<td>Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times</td>
<td>139</td>
</tr>
<tr>
<td>28</td>
<td>Summary of Cognitive Model Architectures, Common Uses and References</td>
<td>177</td>
</tr>
<tr>
<td>29</td>
<td>Summary of Integrated Architectures, Common Uses and Key References</td>
<td>178</td>
</tr>
<tr>
<td>30</td>
<td>Low Taskload</td>
<td>181</td>
</tr>
<tr>
<td>31</td>
<td>High Taskload</td>
<td>181</td>
</tr>
<tr>
<td>32</td>
<td>Task Analysis and Workload (Modified) Scale Values</td>
<td>186</td>
</tr>
<tr>
<td>33</td>
<td>Overview of Time Estimation Literature including the Effects of Over/Under estimation on Timing Behaviour</td>
<td>190</td>
</tr>
</tbody>
</table>
List of Equations

EQUATION 1. RECEIVE HANDOFF DURATION (RHD) CALCULATION .......................................................... 34
EQUATION 2. INSTANTANEOUS WORKLOAD CALCULATION COMPLETED IN THE BASELINE MODEL .................. 39
EQUATION 3. RECEIVE HANDOFF DURATION (RHD) CALCULATION FOR THE HPM DATA ........................................ 49
EQUATION 4. POWER FUNCTION EQUATION FOR THE TIME ESTIMATES IN THE BASELINE WORKLOAD CONDITION . 101
EQUATION 5. POWER FUNCTION EQUATION FOR TIME ESTIMATES IN THE HIGH WORKLOAD CONDITION ........... 101
EQUATION 6. LINEAR INTERPOLATION FOR DETERMINING THE GAIN, G, FOR THE TIME ESTIMATION MODEL ....... 103
EQUATION 7. LINEAR INTERPOLATION FOR DETERMINING THE EXPONENT, P, FOR THE TIME ESTIMATION MODEL . 103
EQUATION 8. THE TIME ESTIMATION MODEL LINEAR INTERPOLATION ......................................................... 103
EQUATION 9. THE TIME ESTIMATION MODEL’S FUNCTION ................................................................. 104
EQUATION 10. RATIO OF INNER-TO-OUTER LOOP SCAN UPDATES FOR THE GENERIC MODEL ................................. 108
List of Figures

FIGURE 1. TIMELINE OF VALIDATION EFFORTS CONDUCTING IN HUMAN PERFORMANCE MODELLING ............ 16
FIGURE 2. HANDOFF TIMINGS (NOTIONAL) FOR THE SENDING/RECEIVING CONTROLLERS ....................... 30
FIGURE 3. WINDOW OF OPPORTUNITY AND HANDOFF RECEIPT TIME PERFORMED BY THE RECEIVING CONTROLLER. 32
FIGURE 4. MEAN FEWS WAK WORKLOAD AND SE BY TRAFFIC CONDITION ...................................... 35
FIGURE 5. MEAN FEWS WAK WORKLOAD AND SE BY TRAFFIC CONDITION ...................................... 36
FIGURE 6. BASELINE MODEL OF TASK MANAGEMENT BEHAVIOUR .................................................... 42
FIGURE 7. MEAN (A) FEWS ADJUSTED AND (B) BASELINE MODEL WORKLOAD BY TRAFFIC CONDITION ........ 50
FIGURE 8. FEWS (TOP) COMPARED TO BASELINE MODEL (BOTTOM) WORKLOAD OUTPUT OVER SIMULATION, FOR (A) 100 DL CONDITION, (B) 133 DL CONDITION, (C) 133 No DL CONDITION ........................................ 51
FIGURE 9. MAXIMUM NUMBER OF TASKS TO BE SERVICED IN THE BASELINE MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ............................................................. 52
FIGURE 10. AVERAGE NUMBER OF TASKS TO BE SERVICED IN THE BASELINE MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ............................................................. 52
FIGURE 11. AVERAGE RHD TIMES FROM THE FEWS DATA AND BASELINE MODEL .................................. 53
FIGURE 12. RHD HISTOGRAM FOR (A) FEWS AND (B) BASELINE MODEL 100% DL ............................... 55
FIGURE 13. RHD HISTOGRAM FOR (A) FEWS AND (B) BASELINE MODEL 133% DL ................ ................ 55
FIGURE 14. RHD HISTOGRAM FOR (A) FEWS AND (B) BASELINE MODEL 133% No DL .......................... 56
FIGURE 15. TC GRAPH FOR FEWS AND BASELINE MODEL WINDOW OPEN TIMES 100% DL ................. 57
FIGURE 16. TC GRAPH FOR FEWS AND BASELINE MODEL WINDOW OPEN TIME 133% DL ................. 58
FIGURE 17. TC GRAPH FOR FEWS AND BASELINE MODEL WINDOW OPEN TIMES 133% No DL ............... 58
FIGURE 18. TC GRAPH FOR FEWS AND BASELINE MODEL ADJUSTED HANDOFF COMPLETE TIME 100% DL 60
FIGURE 19. TC GRAPH FOR FEWS AND BASELINE MODEL ADJUSTED HANDOFF COMPLETE TIME 133% DL .... 60
FIGURE 20. TC GRAPH FOR FEWS AND BASELINE MODEL ADJUSTED HANDOFF COMPLETE TIME 133% No DL .... 61
FIGURE 21. REPRESENTATION OF Hollnagel’s (1993) CONTROL MODES ................................................. 67
FIGURE 22. THE COCOM CYCLICAL MODEL OF HUMAN ACTION WITHIN A SYSTEM CONTEXT BASED ON THE 3C’S: CONSTRUCTS, CONTROL AND COMPETENCE .................................................. 76
FIGURE 23. RELATIONSHIP BETWEEN SYSTEM LEVEL MODELS AND INFORMATION PROCESSING MODELS .... 77
FIGURE 24. GENERIC RELATIONSHIP AMONG WORKLOAD PROJECTING, TIME ESTIMATING, AND TASK MANAGING IN THE TIME MANAGEMENT FRAMEWORK .................................................... 78
FIGURE 25. RELATIONSHIP OF TIME PRESSURE TO ELAPSED TIME, AS A FUNCTION OF UNDER- OR OVER-ESTIMATION OF TIME AVAILABLE PRIOR TO BEGINNING AN ACTIVITY ........................................ 81
FIGURE 26. CONCEPTUAL REPRESENTATION OF KEY TIME ESTIMATION FACTORS .................................. 83
FIGURE 27. THE TASK MANAGEMENT MODEL ............................................................... 86
FIGURE 28. MEAN (A) FEWS ADJUSTED, (B) BASELINE, AND (C) TM MODEL WORKLOAD PREDICTIONS BY TRAFFIC. ............................................................... 87
FIGURE 29. FEWS COMPARED TO TM MODEL WORKLOAD OUTPUT FOR (A) 100 DL CONDITION, (B) 133 DL CONDITION, (C) 133 No DL CONDITION ............................................................. 88
FIGURE 30. MAXIMUM NUMBER OF TASKS TO BE SERVICED IN THE TM MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ............................................................. 89
FIGURE 31. AVERAGE NUMBER OF TASKS TO BE SERVICED IN THE TM MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ............................................................. 89
FIGURE 32. AVERAGE RHD TIME AS A FUNCTION OF TRAFFIC LEVEL FOR THE FEWS, THE BASELINE MODEL, AND THE TM MODEL .......................................................... 90
FIGURE 33. RHD HISTOGRAM FOR (A) FEWS AND FOR (B) TM MODEL 100% DL ............................... 92
FIGURE 34. RHD HISTOGRAM FOR (A) FEWS AND FOR (B) TM MODEL 133% DL ............................... 92
FIGURE 35. RHD HISTOGRAM FOR (A) FEWS AND FOR (B) TM MODEL 133% No DL ........................... 92
FIGURE 36. TC GRAPH FOR FEWS AND TM MODEL WINDOW OPEN TIME 100% DL .......................... 94
FIGURE 37. TC GRAPH FOR FEWS AND TM MODEL WINDOW OPEN TIME 133% DL .......................... 94
FIGURE 38. TC GRAPH FOR FEWS AND TM MODEL WINDOW OPEN TIME 133% No DL ..................... 95
FIGURE 39. TC GRAPH FOR FEWS AND TM MODEL ADJUSTED HANDOFF COMPLETE TIMES 100% DL ...... 96
FIGURE 40. TC GRAPH FOR FEWS AND TM MODEL ADJUSTED HANDOFF COMPLETE TIME 133% DL .......................... 97
FIGURE 41. TC GRAPH FOR FEWS AND TM MODEL ADJUSTED HANDOFF COMPLETE TIME 133% No DL .......................... 97
FIGURE 42. PLOT AND POWER FUNCTION OF BASELINE WORKLOAD BROWN & BOLTZ DURATION ESTIMATES .......................... 101
FIGURE 43. PLOT AND POWER FUNCTION OF HIGH WORKLOAD BROWN AND BOLTZ DURATION ESTIMATES .......................... 101
FIGURE 44. ILLUSTRATION OF TIME ESTIMATION MODEL FUNCTION ................................................ 104
List of Figures

FIGURE 45. ESTIMATES OF TIME ELAPSED AS A FUNCTION OF CHANGING WORKLOAD VALUES AND ACTUAL TIME ELAPSED ................................................................. 105
FIGURE 46. TASK NETWORK MODEL OF A SIMPLIFIED ATC ENVIRONMENT ................................................................. 109
FIGURE 47. MEAN TIME ERROR OF 10000 GENERIC MODEL RUNS TO VERIFY TE MODEL PERFORMANCE .................. 112
FIGURE 48. HISTOGRAM FOR (A) BASELINE AND (B) TE MODEL TIME ERRORS FOR LOW WORKLOAD. .................. 112
FIGURE 49. HISTOGRAM FOR (A) BASELINE AND (B) TE MODEL TIME ERRORS FOR LOW-MODERATE WORKLOAD. 113
FIGURE 50. HISTOGRAM FOR (A) BASELINE AND (B) TE MODEL TIME ERRORS FOR MODERATE WORKLOAD ......... 114
FIGURE 51. HISTOGRAM FOR (A) BASELINE AND (B) TE MODEL TIME ERRORS FOR MODERATE-HIGH WORKLOAD. 114
FIGURE 52. HISTOGRAM FOR (A) BASELINE AND (B) TE MODEL TIME ERRORS FOR HIGH WORKLOAD .................. 115
FIGURE 53. BASELINE MODEL WITH TE AUGMENTATION .................................................................................. 117
FIGURE 54. MEAN (A) FEWS ADJUSTED, (B) BASELINE, AND (C) TE MODEL WORKLOAD BY TRAFFIC CONDITIONS. 118
FIGURE 55. FEWS (TOP) COMPARED TO THE TE AUGMENTATION (BOTTOM) WORKLOAD OUTPUT OVER SIMULATION, FOR (A) 100 DL CONDITION, (B) 133 DL CONDITION, (C) 133 NO DL CONDITION .......................................................... 119
FIGURE 56. MAXIMUM NUMBER OF TASKS TO BE SERVICED IN THE TE MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ................................................. 120
FIGURE 57. AVERAGE NUMBER OF TASKS TO BE SERVICED IN THE TE MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ................................................................. 120
FIGURE 58. COMPARISON OF THE AVERAGE RHD IN THE FEWS, BASELINE, AND TE MODELS ............................................ 121
FIGURE 59. RHD HISTOGRAM FOR (A) FEWS AND (B) TE MODEL 100% DL ............................................................. 123
FIGURE 60. RHD HISTOGRAM FOR (A) FEWS AND (B) TE MODEL 133% DL ............................................................. 123
FIGURE 61. RHD HISTOGRAM FOR (A) FEWS AND (B) TE MODEL 133% NO DL ...................................................... 123
FIGURE 62. TC GRAPH FOR FEWS AND TE MODEL WINDOW OPEN TIMES 100% DL ................................................ 124
FIGURE 63. TC GRAPH FOR FEWS AND TE MODEL WINDOW OPEN TIMES 133% DL ................................................ 125
FIGURE 64. TC GRAPH FOR FEWS AND TE MODEL WINDOW OPEN TIME 133% NO DL ......................................... 125
FIGURE 65. TC GRAPH FOR FEWS AND TE MODEL WINDOW OPEN TIME 133% COMPLETE HANDOFF COMPLETE 100% DL ............................................................... 127
FIGURE 66. TC GRAPH FOR FEWS AND TE MODEL WINDOW OPEN TIME 133% COMPLETE HANDOFF COMPLETE 100% NO DL .............................................................. 127
FIGURE 67. TC GRAPHS FOR FEWS AND TE MODEL WINDOW OPEN TIME 133% COMPLETE HANDOFF COMPLETE 133% NO DL .............................................................. 128
FIGURE 68. BASELINE MODEL WITH TM+TE AUGMENTATION ................................................................................ 130
FIGURE 69. MEAN (A) FEWS ADJUSTED, (B) BASELINE, AND (C) TM+TE MODEL WORKLOAD BY TRAFFIC CONDITIONS .................................................................................. 130
FIGURE 70. FEWS (TOP) COMPARED TO TM+TE MODEL (BOTTOM) WORKLOAD OUTPUT, FOR (A) 100 DL CONDITION, (B) 133 DL CONDITION, (C) 133 NO DL CONDITION .......................................................... 131
FIGURE 71. MAXIMUM NUMBER OF TASKS TO BE SERVICED IN THE TM+TE MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ................................................. 131
FIGURE 72. AVERAGE NUMBER OF TASKS TO BE SERVICED IN THE TM+TE MODEL QUEUE AS A FUNCTION OF THE EXPERIMENTAL CONDITION ................................................................. 132
FIGURE 73. AVERAGE RHD TIMES FROM FEWS AND THE FOUR MODEL ITERATIONS .................................................... 133
FIGURE 74. RHD HISTOGRAM FOR (A) FEWS AND (B) TM+TE MODEL 100% DL ............................................................. 136
FIGURE 75. RHD HISTOGRAM FOR (A) FEWS AND (B) TM+TE MODEL 133% DL ............................................................. 136
FIGURE 76. RHD HISTOGRAM FOR (A) FEWS AND (B) TM+TE MODEL 133% NO DL ...................................................... 136
FIGURE 77. TC GRAPH FOR FEWS AND TM+TE MODEL WINDOW OPEN TIME 100% DL ................................................ 137
FIGURE 78. TC GRAPH FOR FEWS AND TM+TE MODEL WINDOW OPEN TIME 133% DL ................................................ 138
FIGURE 79. TC GRAPH FOR FEWS AND TM+TE MODEL WINDOW OPEN TIME 133% NO DL ................................................ 138
FIGURE 80. TC GRAPH FOR FEWS AND TM+TE MODEL WINDOW OPEN TIME 133% COMPLETE HANDOFF COMPLETE 100% DL ............................................................... 140
FIGURE 81. TC GRAPH FOR FEWS AND TM+TE MODEL WINDOW OPEN TIME 133% COMPLETE HANDOFF COMPLETE 133% NO DL .............................................................. 140
FIGURE 82. TC GRAPH FOR FEWS AND TM+TE MODEL WINDOW OPEN TIME 133% COMPLETE HANDOFF COMPLETE 133% NO DL .............................................................. 141
FIGURE 83. REPRESENTATION OF THE MODELED ENVIRONMENT .............................................................................. 179
FIGURE 84. DETAILED VIEW OF ZGN AIRSPACE SECTOR WITH FIX, ASCENT, AND DESCENT WAYPOINTS ............ 180
FIGURE 85. ENVIRONMENT MODEL SNAPSHOT .................................................................................................. 182
FIGURE 86. TASK NETWORK REPRESENTATION OF THE BASELINE MODEL ................................................................. 187
SECTION 1

INTRODUCTION TO HUMAN PERFORMANCE MODELLING AND VALIDATION

The goal of Section 1 is to outline the efforts that have been undertaken in validating complex human performance models (HPMs). As such:

Chapter 1 introduces the motivation of the research as being the need for an ever-increasing level of fidelity to accurately model human behaviour in the context of complex human-system operations.

Chapter 2 describes the evolution of HPMs from simple engineering models, through to task network models, and onwards to more complex integrated HPMs.

Chapter 3 outlines Verification and Validation methods, and provide examples of validation techniques that have been applied to integrated HPMs of complex tasks, ending with general lessons learned from the literature on verifying and validating complex models that will be applied in Section 2 of this dissertation.
CHAPTER 1: INTRODUCTION AND OBJECTIVES

1.1. INTRODUCTION

Human performance modelling is the process whereby human characteristics are embedded within a computer software structure that represents a simulated human operator interacting with a simulated operating environment. Integrated human performance models (HPMs) simulate and predict emergent behaviour based on multiple, interacting sub-models of human behaviour, such as perception, attention, working memory, long-term memory and decision-making. This is accomplished typically by incorporating sub-models of human performance that feed both forward and back to other constituent models within the human information processing system.

The use of appropriate and validated integrated HPMs can thus support the basic human factors principle of predicting the impact of alternative design options early in the system design process. These HPMs also provide many advantages over human-in-the-loop (HITL) studies, especially for complex systems (Gore & Corker, 1999).

Complex systems are those that include human operators interacting with actual technology and automation, to carry out multiple interacting, and often conflicting, tasks. These systems often involve time-critical tasks. Time critical tasks typically have a specific onset time – that is, a time after which the task may be commenced – and a specific time by which the task needs to be completed. Together these define a window of opportunity for the action to take place. For such systems, the dynamic interactions among system elements often form critical couplings for control of the system by the human (Corker, Gore, Fleming, & Lane, 2000; Gore & Corker, 2000). Examples of complex environments involving such time critical tasks include the Air Traffic Control (ATC) domain, Military Command and Control (C2) environments, and Space Shuttle operations. In all cases, responsibility for system control resides in a centralized location, while responsibility for operations exists in multiple interacting agents that share information across (large) distances and engage in complex tactical decision making behaviour in the face a rapidly changing operational environment. In such complex systems, one of the advantages of a HPM is the ability to model critical events that cannot be studied fully with HITL subjects, due to safety concerns, cost considerations, or practical difficulties associated with the simulation of very rare events.
One of the most significant hurdles facing modellers is the challenge of validating these integrated HPMs, a goal without which the credibility of any model predictions will clearly be greatly reduced. In the recent past, there have been a small number of attempts to validate integrated HPMs (as will be discussed in Chapter 3), however, most validation efforts to date have been in the area of simpler engineering models and cognitive architectures. Furthermore, of the validation efforts that have been conducted for integrated models, there is little agreement as to what constitutes appropriate validation techniques and measures. The development of these integrated HPMs is in its infancy, and so too are the validation techniques. There is a real need for the advancement of techniques and approaches for validating complex models.

1.2. OBJECTIVES

The present research seeks to address two major challenges in the field of modelling human performance in time critical tasks: 1) Validating complex HPMs, and 2) Modelling time management behaviour.

1.2.1. Validating Complex Human Performance Models

Modelling human behaviour in complex systems such as air traffic control is complicated, particularly when the human’s tasks are highly cognitive in nature and they interact in a closed-loop fashion with other operator and environmental characteristics. Since cognitive tasks are not directly observable, it is very difficult to objectively validate these complex models. As a field, our ability to model these complex tasks and demonstrate that such models of human behaviour validly represent actual human behaviour is in its infancy. Many HPM validation efforts often rely only on subjective or qualitative measures as opposed to objective, quantitative measures. Thus, one major objective of this work is to focus on quantitative validation techniques that can be used to demonstrate that the model adequately represents human cognitive processes.

1.2.2. Modelling Time Management Behaviour

Another common weakness in complex HPMs is the failure to properly account for how human operators manage time. It is proposed that time management behaviour is comprised of two elements: a task manager and a time estimation element, both of which are impacted by the projection of workload. As workload increases, the manner in which tasks are managed or scheduled changes from ‘strategic’ to ‘opportunistic’. That is, operators shift along a continuum
from planning for future events using choice behaviour in the ‘strategic’ mode of behaviour to
operating with limited planning and making decisions based on the environment in, as Hollnagel
terms, the ‘opportunistic’ mode (see Hollnagel, 1993, 1998; Hollnagel & Woods, 2005).
However, most current HPMs assume some kind of rigid, nominally optimal, strategic task
scheduling and fail to adequately capture the more realistic ‘opportunistic’ scheduling often
adopted by operators of high-workload tasks.

The second element of time management behaviour is time estimation. Humans are
frequently biased in their estimation of time – tending to underestimate the passage of time
(thinking that less time has passed than actually has passed) when they become very busy
(Block, Zakay, & Tsal, 1997; Boltz, 2005; Brown & Boltz, 2002; Buehler, Dawes, 1988; Griffin,
Dunning & Ross, 1990; Griffin, & Ross, 1994; Johnson & Sherman, 1990; Kahneman &
Tversky, 1979; Kahneman & Tversky, 1982; Michon & Jackson, 1984; Zakay, 1990; Zakay,
1993; Zakay & Block, 1995, 1996). This underestimate of time passage leads naturally to an
overestimate of the remaining time available (operators thinking that they have more time than
they actually do) in any window of opportunity of task completion. However, most models
assume perfect estimation of time, and execute tasks accordingly. Furthermore, this time
estimation bias is impacted by workload, with larger time estimation errors occurring as
workload increases. No HPMs to date account for this bias in human estimation of time as a
function of workload in a closed-loop manner. Thus, a second objective of this work is to
develop a closed-loop representation of time management behaviour that accounts for changes in
task scheduling and time-estimation as a function of workload.

1.3. APPROACH

The above objectives will be accomplished by extending an HPM of a complex, time-critical
environment, namely the air traffic control environment, which will be used as a test-bed with
which to develop and exercise validation techniques that concentrate on validating the time-
relevant aspects of the model. Using an iterative develop-validate process, the test-bed model
will be augmented with submodels that are internal models embedded within the test-bed model
that represent the processes required to execute a series of procedures (Baron, Kruser, & Huey,
1990), in this case, time management procedures. The submodels can be modified individually,
and thus any differences in model output can be attributed to the submodel under investigation.
These submodels are based on an analytical synthesis of existing literature of the manner in
which humans’ time management (task management and time estimation) changes in the face of the dynamics of the operational environment, as a function of time pressure, and thus of perceived workload.

To accomplish this requires the following steps:

A. Identify and develop a baseline model by identifying a relevant time-critical task and HITL data source with which to demonstrate and exercise validation techniques (Section 2)

B. Apply validation techniques to validate the time-relevant aspects of the model (Section 2)

C. Develop submodels of time management that encompasses the available literature on time management, time estimation, and workload, and integrate these submodels in an iterative develop-validate process to attempt to improve the validity of the model (Section 3)

1.4. EXPECTED RELEVANCE

Three communities will benefit from the results of this analytical model development and validation effort: (1) the model developer community; (2) the research community; and (3) the aviation system design community. Human Performance Model developers will be able to use the validation approach applied in Section 2 of the present research to complete more successful validation efforts, especially for complex, time-critical environments. Researchers will be able to use the submodels developed in Section 3 to understand the relationships that time has on influencing behaviours and performance, by specifically exploring the effect that workload possesses on impacting time estimates and task management. Aviation System Designers will be able to use the complex ATC model to quantify, predict and evaluate simulated human operator workload and performance, to design and compare human-automation interface designs and alternative design options. The results of this research effort, which include applying validation techniques to time-related submodels of performance, are also expected to benefit system designers of complex systems in several domains involving complex environments requiring time critical responses and for which concepts, technologies, and automation cannot be tested in empirical simulations because the system concepts are too new, too difficult, or overly dangerous
for the human operator (e.g. space operations, technologies for advanced aircraft system concepts, advanced medical technologies, and advanced surface transportation technologies, among others).
CHAPTER 2: THE EVOLUTION OF HPMS

This chapter presents a brief historical overview of the evolution of HPMs and is intended to illustrate the differences in complexity from the early simple engineering models (such as manual control models), which have relatively observable output through to the more complex, integrated architectures, which possess many embedded, interacting submodels. Four classes of HPMS will be reviewed: 1) Simple Engineering Models, 2) Task Network Models, 3) Cognitive Models, and 4) Integrated Models.

2.1. OVERVIEW OF HUMAN PERFORMANCE MODELLING APPROACHES

2.1.1. Simple Engineering Models

The first known HPM, developed close to 60 years ago, were the quasi-linear and optimal manual control models (Craik, 1947; Tustin, 1947), which still today represent the state-of-the-art in manual control modelling (Pew, 2007). These early models were limited to human tracking behaviour in closed-loop person-machine systems (Craik, 1947). These ‘engineering’ models were derived from an assumption that the operator’s control behaviour in perceiving an error and translating this error into a response can be modelled as a linear transfer function. The term ‘quasi-linear’ derives from the recognition that linear behaviour is only an approximation to actual human behaviour (Pew, 2007).

Data from early tracking studies led Craik to conclude that the human operator behaves basically as an intermittent correction servo, or intermittent correction machine. Concurrently, Tustin (1947), taking lessons from gun operators in World War II, applied the servomechanism concept to human control of a massive gun turret. This was motivated by the difficulties that were apparent in aiming and turning large guns towards distant aircraft. Although the approaches utilized during those times were suitable primarily for only linear and time invariant dynamic systems, the resulting models, although simple, have proven to be quite generalisable and robust.

Manual Control models have been shown to be consistently capable of describing and predicting human performance in a wide variety of manual control tasks, by examining the conditions necessary for system stability. (These models therefore operate best in conditions where there exists the need for inner loop control and, as such, are less appropriate for systems
that allocate inner loop control activities to automation.) In manual control it is assumed that the human operator is tracking an error of a signal varying continuously in time, to produce a response to that error which is also a continuous signal varying in time. The models have been applied to a vast number of experimental paradigms, including aircraft and other vehicle control display problems, determining stressor effects on performance, evaluating simulator requirements, and assisting in experimental and simulation planning where manual control was included (Baron, Kruser, & Huey, 1990).

2.1.2. Task Network Models

Early task network models emerged in 1969 from the reliability work that was being completed in the nuclear power community by Siegel and Wolfe (1969). The task network models use a decomposition of tasks into low-level units of action, or "nodes", connected together by a series of branches to represent a network or a sub-network. This network can be degraded by performance shaping factors (PSFs) applied at the action unit level or at the global level, which laid the foundation for representing degradation from optimal performance.

The Systems Analysis of Integrated Network of Tasks (SAINT) and its more recent PC counterpart Micro Saint are discrete event simulation engines that were spawned from the original Siegel and Wolfe work. While they are currently general-purpose simulation packages, they were initially developed to represent human behaviour as a set of interrelated tasks and were the first to capture the PSFs developed by Siegel and his colleagues (Wortman, Pritsker, Seum, Seifert, & Chubb, 1974). These task network models predict human behaviour in complex systems using stochastic probabilistic models of behaviour with outputs of task timelines, task accuracy data, and estimates of operator workload and task loads. Task network models are set up to operate according to a series of task triggers identified through a task decomposition of some behaviour of interest. These tasks can be defined at varying levels of abstraction. External models can be incorporated to impact the tasks defined in the task network model along various dimensions, including workload and time. The task network model does not include models of operator perception, attention, decision-making, reaction / response time, or outside environment models. All of these external models need to be imported through function calls in the task network model that is being created for the specific context under evaluation.
2.1.3. Cognitive Models

Cognitive models are representations of the knowledge that a human possesses about the world and their response characteristics to that world information. Generally, computational models of cognition are derived from a conceptual framework based on empirical data. For example, Wickens’ computational model of Attention / Situation Awareness (A/SA), was derived, in part, from his conceptual framework and decades of empirical research that established how visual attention was driven by four parameters: Salience, Expectancy, Effort and Value (SEEV; see Wickens, Goh, et. al, 2003; Wickens & McCarley, 2008).

Cognitive models and their architectures are generally considered human process models (Elkind, Card, Hockberg, & Huey, 1989), and are commonly represented as an information processing system, with an input, central processing and output component. Card, Moran and Newell (1983) are credited for developing the first cognitive architecture of this kind, known as the Model Human Processor (MHP). They describe this information processing system to be a set of memories linked together and a set of principles that drive the set of memories. The MHP is divided into three segments: the perceptual system, the motor system and the cognitive system. Each of these systems has their own set of memories and processes that work to buffer, encode, decay, and act on the information retrieved from the memory and process structure. MHP has been applied to the low-level tasks of keyboard input, basic reading, and responding to auditory inputs (Card, Moran, & Newell, 1983).

Since the development of the MHP many other cognitive architectures have emerged. Appendix A provides a brief description of each, with examples of common uses, and key references. As can be seen in Appendix A, these tend to focus on fairly simple, limited-in-scope aspects of human behaviour, such as memory and pattern matching. However, a few have been applied to more complex tasks (e.g. SOAR has been applied to tasks such as car-driving and job-shop scheduling).

2.1.4. Integrated Models

Integrated models, also known as integrated architectures, combine a number of individual process models (also referred to as submodels) of human perceptual, cognitive and motor systems into a coordinated representation of interacting elements within the context of an environment, in an attempt to predict behaviours of higher complexity. Some of these link
models of human anthropometry, biomechanics, and human cognition together with an environment to determine whether the human can perform to a criterion level with new technological concepts. These integrated architectures link task networks, cognitive architectures and control models in order to predict integrated human performance. Examples of such integrated architectures include MicroSaint Sharp, COGNET/iGEN, Soar/iGEN, Soar/EPIC, D-OMAR, IMPRINT/WinCrew, IPME, and MIDAS. These are summarized in Appendix B.

Most of these models utilize a task analysis or cognitive task analysis (CTA) process to create their procedural models. The outputs from such HPMs have traditionally taken the form of workload predictions, situation awareness predictions, and procedural and task timelines, each of which is often referred to as an emergent characteristic.

As models increase in complexity, both in terms of the embedded submodels contained within their architecture and the environments to which their results are being applied, it is increasingly difficult to: (1) interpret the empirical basis behind the algorithms driving the modelling architecture; (2) determine which submodel is operating at a given time in the simulation; (3) observe many of the behaviours generated from the integrated submodel, and particularly the embedded cognitive submodels, directly; and (4) trace which submodels will compete or conflict with other operating models due to conflicting model assumptions in the interacting submodels.

As discussed by Leiden and Best (2008), in some HPMs, individual submodels can be engaged or disengaged for many reasons. For example, if behaviours become too constrained by some rule, the model developer can turn off this constraint, thereby allowing the model to proceed to completion. Although determining and tracking the complexity of model inputs and outputs is one challenge that can be met with an appropriate level of transparency in the model’s architecture (Gore et al, 2008), interactions among submodels makes it more difficult to validate these model.

One class of integrated model, those created using Micro Saint Sharp, merits further discussion, as it is used in Sections 2 and 3 of this dissertation. It is a discrete-event simulation-modelling environment that can simulate complex processes, to solve problems spanning a number of complex application domains. Micro Saint Sharp uses the same task network structure as Micro Saint and Saint, but can also include process models and environment models, depending on the modeller’s needs. Micro Saint Sharp can be considered an integrated model.
simulation engine, as opposed to a task network model, because Micro Saint Sharp integrates these lower level submodels of operator performance and provides a significantly increased capability of integrating outside models into the Micro Saint Sharp operational environment and testing them. Examples of these external models include such perceptual models as MIDAS (Gore et al., 2008) and the A/SA SEEV Attention model (Gore et al., 2009). They also include external environment models (Hooey et al., 2008), event detection, response times, a task “scheduler”, a runtime output capability that generates a physical rendition of the operator and the environment in addition to task timeline information and operator workload, and a host of new functional models to simulate Air Traffic Control operations in the next generation of air travel termed NextGen (Leiden & Kamienski, 2006).

Inputs to models like Micro Saint Sharp include the operating procedures in a task format and a sequence of nested activity trees embedded under higher-level tasks. The output measures of interest for integrated HPMs have traditionally included task demands, workload, task load, information load, attention demands, stress, procedural timing measures, and the human’s contributions to system errors. These measures have been used to identify when, where, and how often errors occurred within a specific job design and, combined with the load measures, how procedures can be re-organized to reduce time and load demands (Corker et al., 2003; Foyle & Hooey, 2008; Gore, 2002; Gore & Corker, 1999; Gore et al., 2001; Gore & Smith, 2006; Shively, et al., 2000; Siegel & Wolfe, 1969; Smith & Corker, 1993).

It is important to note that Micro Saint Sharp is a software coding language, much like C++ or LISP. In other words, it is not the Micro Saint Sharp coding software that must be validated, but the models built with the language, as will be the focus of Section 2.

2.2. SUMMARY – CHALLENGES TO MODELLING HUMAN PERFORMANCE IN COMPLEX SYSTEMS

Modelling complex systems brings with it unique challenges not present in such early models as the manual control models, or even in most of the modelling efforts conducted with more recent cognitive architectures, which typically represent aspects of human performance within very limited domains. Generally speaking, those limited models are the products of laboratory research on very specific human tasks developed to model human information processing, rather than human-machine interaction. Those models tend to ignore some of the complexity of human-
system performance because they often ignore aspects of the environment or tasks that might modify the model's predictions. For example, Baron et al. (1990) point out that models of human reaction time typically predict response time primarily as a function of the number of possible signals or their relative probability, and give secondary consideration to such physical factors as how far apart the response keys are, whether eye movements are needed to monitor signal occurrence, or anatomical dimensions of the operator that might affect performance. Several of the human information processing models have been adapted from engineering models to represent human behaviour, but these have traditionally been based upon a single theory or technique. Such models invoke, among others, information theory, the theory of signal detection, sequential decision theory, theories of reaction speed and accuracy, sampling theory, psychophysical scaling theory, and fuzzy set theory (Sheridan & Ferrell, 1974).

As models have evolved from simple engineering models to integrated architectures, their complexity has increased many-fold due to the increased number of submodels and the interactions among them. It is often impractical to validate complex HPMs in the same manner as for simple models of limited scope (Baron et al., 1990). In fact, it is often the case that "pure, formal validation" is simply not possible because of the complexity of human behaviours and their interactions with the environmental elements (in a closed-loop fashion) that exist in the complex systems that are being modelled today.

The relatively simple manual control models have clear, observable output that can be used for validation, whereas the more complex integrated architectures possess many more embedded mechanisms, many of which are cognitive in nature and thus not observable. This increased complexity adds significantly to the challenges of validating complex HPMs, as will be discussed next in Chapter 3.
CHAPTER 3: VERIFICATION AND VALIDATION

This chapter begins by defining each of the terms Verification and Validation, but then focuses on validation methods, and provides examples of validation techniques that have been applied to integrated HPMs of complex tasks. This chapter ends with lessons learned from the literature on validating complex models that will be applied in Section 2 of this dissertation.

3.1. DEFINING MODEL VERIFICATION AND VALIDATION

Model verification and validation (V&V) are essential elements of any modelling effort, particularly when attempting to generate predictive behaviours of human performance in complex environments. Model verification is the process of determining whether a simulation model and its associated data behave as intended by the model developer / analyst (Sargent, 1980). Model validation is the process of determining the degree to which a model or simulation and its associated predictions are an accurate representation of the real world, from the perspective of the intended users of the model or simulation (Balci, 1998; Law & Kelton, 2000; Sargent, 1980). Both model verification and model validation must be considered when attempts are made to validate a model, particularly as models increase in complexity (DMSO, 2001).

3.2. APPROACHES TO MODEL VALIDATION

Model validation can take many forms, ranging from common qualitative approaches to quantitative approaches (Campbell & Bolton, 2005). Each is described next.

3.2.1. Qualitative Approach to Model Validation

Qualitative approaches are those validation processes that ask subject matter experts (SMEs) to make judgements about the content and/or behaviour of a human behaviour model. The qualitative approach thus essentially tests a model’s face validity, ensuring that simulation results are consistent with the expected system behaviour (Law & Kelton, 2000). Campbell & Bolton (2005) indicate that qualitative approaches are one of the most commonly used to validate a model. Given, however, that judgements are subjective and SME knowledge is not perfect, they are prone to many biases and no concrete conclusions can be drawn about the accuracy of relatively complex models.
The subjective judgements are best when they follow formal standardized and structured stages (e.g. questionnaires, interviews, etc.), and when they are made independently by SMEs who were not involved in creating the model. One commonly used subjective approach is known as the Turing test. The Turing test requires that subject matter experts (SME’s – defined as people knowledgable about the system under development) compare one or more sets of system data to one or more sets of data from the model in a double blind fashion (Law & Kelton, 2000). The SME’s review the data and judge it on the grounds of reasonableness. If the simulation results are consistent with the SME’s expectations, then the model is judged to have face validity. One disadvantage of qualitative techniques like the Turing test, however, is that they can be subject to SME biases, which may result in inappropriate model design or inappropriate validation criteria.

As Balci, Nance, Arthur & Ormsby (2002) outline, qualitative methods often possess insufficient accuracy for relatively large (complex) HPMs and, as a result, qualitative approaches should be used only in concert with quantitative approaches, and should be conducted iteratively throughout the model development effort.

3.2.2. Quantitative Approach to Model Validation

Obtaining a quantitative measure of the similarity between a model’s behaviour and empirically determined human behaviour is a complement to the qualitative approach. More explicitly, a quantitative test for a model’s validity is the degree to which the model’s output resembles the output that would be expected from the real world. Quantitative approaches are traditionally statistical in nature and attempt to measure the degree to which a model’s data are similar to an empirically collected set of data. The recommended statistical tests used to measure the similarity between the data sets are goodness-of-fit tests, involving the calculation of $r^2$ to assess trend consistency, ANOVAs to compare human and model data sets, root mean squared scaled deviations to assess the exact match, and chi-square analyses to assess whether the underlying distributions of the two data sets (model, real world) can be regarded as being the same (Campbell & Bolton, 2005; Law & Kelton, 2000). Balci (1990) also outlines that graphical comparisons are an effective model validation approach, particularly as a first validation phase in testing the performance of a model. The graphs of values of model variables over time are compared with the graphs of values of system variables, to investigate similarities in periodicities, skewness, number and location of inflection points, logarithmic rise and linearity,
Chapter 3 – Model Verification and Validation

phase shift, trend lines, or exponential growth constants. The histogram is an estimate of the density function and is an effective graphical technique for showing the symmetry, the skewness and kurtosis of a data set. Chapter 6 describes how a qualitative approach will be combined with a goodness-of-fit approach in the current effort.

3.3. HISTORICAL REVIEW OF MODEL VALIDATION

A review of the literature has been summarized here graphically as a historical timeline in Figure 1. The literature generally suggests that, as the field of human performance modelling has shifted its emphasis over time from modelling manual control towards more cognitive modelling, it has continued to make a concerted effort to validate their models. However, at the same time, as can be seen, the validation efforts undertaken thus far, especially for integrated architectures, have been limited. Campbell and Bolton (2005, p 365) note that “human behaviour representation validation is a difficult and costly process [and] most in the community would probably agree that validation is rarely, if ever done”. They point out that there is no general agreement on exactly what constitutes an appropriate validation of a cognitive architecture. Since cognitive architectures are developed for a wide variety of reasons, there is a correspondingly wide set of validation (and evaluation) objectives and metrics and associated methods. A lack of established benchmarks and criteria exacerbates this problem. This reinforces the postulate that validating integrated models is an important and relevant topic and that the current state of the art in modelling human performance possesses a number of challenges that are directly associated with determining their validity.
Figure 1. Timeline of Validation Efforts Conducting in Human Performance Modelling. 
Note: The models have been organized here into one of four categories following the categories summarized in Chapter 2. Some model architectures transcend more than one category however. Material based off of Pew & Mavor (1998), and Pew (2005).

### 3.4. CHALLENGES IN VALIDATING COMPLEX HPMS

Zacharias, MacMillan and van Hernel (NRC, 2008) state that validation remains one of the most challenging aspects of cognitive architecture research and development. The complexity of today’s integrated models makes traditional ‘proof of correctness’ techniques a difficult undertaking; in fact, the proof of correctness concept is often not even an attainable goal (DMSO, 1991; 2001). This is especially true when examining the performance of cognitive models and the more complex behavioural models because of the non-linearity of human behaviours, which prevents simple causal relationships to be drawn between situational elements and resulting actions (DMSO, 2001).
As we move from the simple engineering models towards the integrated architectures, validation has become increasingly difficult, due to the cognitive nature of the models and their interactions as discussed in Chapter 2. The relatively simple manual control models have clearly observable outputs that can be used for validation; however, the more complex integrated architectures possess many more embedded submodels, many of which are cognitive in nature. Validation of complex human cognitive processes and their subsequent impact on human task completion is a large challenge because many of the cognitive processes can not be readily observed; only the behavioural output can be seen and the cognitive aspects must be inferred from the observed output. However, it is further challenged, because the existence of overt behaviours does not necessarily mean that a specific cognitive action has taken place. For example, eye movements are often used to infer visual attention. It can be the case that the eye moves to an area of interest, but the human does not attend to information in that location. Or in contrast, it may be the case that a human can attend to information in his/her periphery without moving the eyes to foveate on that region.

Furthermore, the interactive nature of these embedded submodels and the corresponding assumptions built within many HPMs (that have the potential of occasionally contradicting one another) makes validating these integrated models a difficult challenge (AGARD, 1998). This having been stated, it is frequently impractical to validate complex HPMs in the same manner as done for simple models of limited scope (Baron et al., 1990).

### 3.5. **EXAMPLES OF LARGE-SCALE VALIDATION EFFORTS OF HPMS OF COMPLEX SYSTEMS**

Two large-scale integrated model development and validation efforts undertaken in the very recent past, one conducted by the Air Force Research Laboratory (AFRL) (Gluck & Pew, 2005; Pew, Gluck, & Deutsch, 2005) and one by the National Aeronautics and Space Administration (NASA) (Foyle & Hooey, 2008), are summarised below. They are reviewed here both because they involve integrated HPMs of complex tasks (mostly aviation-related), and because they illustrate the challenges associated with validating complex models that use integrated cognitive mechanisms. Both efforts included multiple model architectures modelling the same real-world aviation problem. Comparisons across the model architectures provided insight into the different validation techniques that could be used to validate a model’s output.
3.5.1. AFRL Cross-Model Comparison and Validation Effort

An integrated model development and validation effort was undertaken by the Air Force Research Laboratory’s (AFRL) Agent-Based Modeling and Behavior Representation (AMBR) project in the 2000-2004 timeframe (Gluck & Pew, 2005). The AFRL funded four modelling groups to model a simplified ATC task, and generate predictions of performance using eight variables along three dimensions: accuracy, response time, and workload. The modelling groups were:

i) Atomic Component of Thought – Rationale (ACT-R)
ii) Cognition as a NETwork of tasks (COGNET)/iGEN,
iii) Distributed Cognition (DCOG),
iv) EASE (an Integration of ACT-R, Soar, and EPIC)

The teams were all required to model the same problem and were required to validate using the same objective measures. The models’ predictions in the AMBR project were validated using sum of squared error (SSE) for continuous variables and \( G^2 \) statistics for categorical or counted data. \( G^2 \) is a log-likelihood ratio statistic designed to measure the fit between expected and observed data, to measure how well the model predicted the human data (Gluck & Pew, 2005). \( G^2 \) analysis yields the same result as the \( X^2 \) test for most comparisons.

Specifically, the models’ validation results using eight primary validation variables related to learning and accuracy/response time are summarized as follows:

- ACT-R accounted for 74.1% of the variance averaged across the 8 variables;
- COGNET/iGEN accounted for 77.8% of the variance averaged across the 8 variables, and;
  possessed the best \( G^2 / SSE \) on four of eight performance measures relative to the other models tested;
- DCOG accounted for 51.9% of the variance averaged across the 8 variables, with most of its incorrect performance predictions due to unobserved interactions between primary (learning/response time) and secondary effects (workload);
- EASE accounted for 63% of the variance averaged across the 8 variables.

One of the key findings in that project was that, while a number of the integrated architecture models were able to be validated on the occurrence of very low level behaviours using eight primary validation variables related to learning, some degree of “learning” occurred
as measured by accuracy/response time, but only in specific, known environments (Zachary et al., 2005). A number of the integrated models failed along the dimension of response time as a function of workload (with the majority of models responding too slowly relative to the observed human data) as illustrated by the lack of a perfect match between the model and the human observed data. The AFRL study is highlighted as it provides insights into one style of an approach for modelling and validating complex models in which models are subjected to a single, quantitative, validation effort only upon completion of a model development phase.

### 3.5.2. NASA Cross-Model Comparison and Validation Effort

A large-scale HPM development and validation effort was undertaken by the National Aeronautics and Space Administration (NASA) in the 2002-2005 time frame (Foyle & Hooey, 2008). Compared to the AFRL approach, this was a more iterative approach that employed a range of both qualitative and quantitative techniques throughout the model development process. NASA funded five modelling groups to develop HPMs of flight deck performance on the airport surface (i.e. taxi operations) and on aircraft final approach using augmented displays (termed the Synthetic Vision System, or SVS). The five models were:

i) ACT-R / IMPRINT
ii) ACT-R /X-Plane Flight Simulator
iii) Air MIDAS
iv) Distributed-OMAR (D-OMAR)
v) Attention/Situation Awareness (A/SA)

HITL data were provided in a manner intended to be representative of data that might be available to modellers addressing a real-world aviation design and evaluation problem. These included existing HITL data in the published literature and a small study (three pilots) intended to guide model development. Each modelling team adopted different validation techniques and measures.

The ACT-R / IMPRINT team created a model of the performance of the crew in a cockpit and populated this cockpit with a number of display locations as a series of interconnected tasks, with dependencies between the task network representation of the displays (Lebière, Archer, Best, & Shunk, 2008). IMPRINT focuses on the task-level of operator performance and breaks high level functions down into smaller-scale tasks. ACT-R on the other hand targets the atomic level of thought - the individual cognitive, perceptual and motor actions that take place at the
sub-second level (Lebière, Archer, Best, & Shunk, 2008). ACT-R / IMPRINT validated their model on the high level outcome of whether the modelled aircraft and pilot behaviours occurred in the same way as the HITL simulation broken down by phase of flight. The next step in their validation effort was to examine the detailed model traces of the pilot’s activities during the specific phase of flight. The third step was to compare the amount of time that attention was allocated to particular display (defined as dwell time) on specific areas of interest as a percentage of total time.

Another team integrated the ACT-R model with the X-PLANE flight simulator, which was used to model the environment (Byrne, Kirlik, & Fleetwood, 2008; Byrne & Kirlik, 2005). This model integration effort was designed to increase the fidelity of the simulation being generated so that the results would closer approximate reality. Their model was first validated on the aggregate measures of dwell percentage and dwell time and attained an \( r^2 \) of .978. They then validated their model of attention allocation on the basis of percentages of fixation times on each display in the cockpit but this time looking at a fine-grained level of the visual scan transition matrix. This fine-grained analysis yielded an \( r^2 \) value of .772, still high but somewhat lower than the .978 attained with the aggregate analysis. The fine-grained analyses provided more insight into the underlying process of eye movements to various display locations than the aggregate-level analysis. This validation process revealed that the model captured the total time on each display well, but it did not do as well at capturing how the pilots moved their visual attention within the cockpit.

The Air MIDAS team used a phased validation in which they built a model of pilots flying a nominal approach-and-land scenario with a Synthetic Vision System (SVS) using two data sets – one to develop and tune the model, and a second independent data set to validate the model. They then extended their validated model of a nominal approach-and-land scenario to new contexts (different visibility conditions, cockpit displays, and approach procedures) to determine whether the model produced valid output in the new scenarios (Corker, Muraoka, Verma, Jadhav, & Gore, 2008). Their validation approach was to compare their model output to HITL data on the parameters of visual attention, specifically visual fixation rates at an aggregate level of analysis (percent time on each cockpit display). For the nominal approach and land scenario, the regression analysis revealed extremely high correlations, accounting for 99% of the variance. The regression values dropped significantly (accounting for only 31% to 58% of the
Chapter 3 – Model Verification and Validation

variance) when the model was subjected to additional validation analyses in the new scenarios on which the model had not been tuned.

The D-OMAR team carried out a qualitative Turing-type test validation on dwell time percentage on the internal and external environments (Deutsch & Pew, 2008). They reported that the model predictions were judged to be reasonable by SME evaluation when compared to HITL dwell percentage of the three pilots used in the empirical simulation. This qualitative model validation was deemed a reasonable approach given the complexity of the behaviours being predicted by the models (Foyle & Hooey 2008).

The A/SA team’s validation effort compared their model predictions against the NASA-provided HITL data on the percent dwell time on areas of interest both internal and external to the cockpit (Wickens et al., 2008). The A/SA team undertook model fits to the data with and without components from their theoretically built A/SA model. During the validation approach, the A/SA team added one component at a time to determine whether the component had an impact on the model’s performance. They determined that certain variables (e.g., effort required to initiate a visual scan) did not add anything to the model’s prediction so they were removed from the model. In addition to supporting the principle of simplicity (Occam razor – Jefferys, Williams, & James, 1991) this illustrates that the iterative approach is a very effective approach to determine the impact that variables have on the model’s performance.

3.6. VALIDATION LESSONS LEARNED

A review of the above model validation efforts lead to four important insights: 1) models must be developed and validated for a specific purpose; 2) developing valid models is an iterative process, 3) as models increase in complexity, validation requires multiple measures, and 4) validation needs to consider different levels of data granularities. These insights, discussed in further detail below, were used to guide the validation approach in Section 2.

3.6.1. Models Must be Developed and Validated for a Specific Purpose

Simulation models are created with specific purposes or criteria in mind, and it is against these criteria that a model must be judged to be accurate (Balci, 1998; DMSO 2001; Law & Kelton, 2000; Sargent, 1980). It is not reasonable to assert that an integrated HPM model is “valid” on a general level. For example, it is not appropriate to make the claim that a particular model (such
as Air MIDAS or ACT-R) is valid, but it might be quite appropriate to claim that a specific simulation model of pilot performance on approach and landing made with that particular architecture is valid for that specific application. Each new application domain, and model purpose, may require a new “validation” exercise for the specific model context. Since HPMs in general are developed for a wide variety of reasons, there is a correspondingly wide set of validation objectives and metrics and associated methods that have been used to determine the suitability of a candidate model. The NASA modelling project illustrated the fact that different models that were developed with different purposes all had different validation approaches. For some, qualitative Turing tests were appropriate, while others required more quantitative results validation (Foyle & Hooey, 2008).

Modelling of time estimates, workload, and schedules of performance (and the interactions among them) is in its infancy within the field of human performance modelling. As such, no comprehensive validation approaches have been developed or applied specifically for this purpose. This research aims to advance the development of valid models of time-sensitive tasks, by demonstrating an approach to validate complex HPMs that is specific to models of time-sensitive behaviours.

3.6.2. Developing Valid Models is an Iterative Process

To address the challenge in Section 3.4 related to the interactive nature of the embedded submodels and the corresponding assumptions built within many HPMs that have the potential of occasionally contradicting one another, validation needs to be considered as part of the process of developing the integrated model. By iteratively creating the model by adding one model parameter at a time, one can assess the impact of each model manipulation to determine whether the model developed operates verifiably and validly. Validation is not simply a phase or a step to be conducted at the end of a model development exercise; rather, it is a process undertaken throughout the modelling and simulation lifecycle (Balci, 1998). The interactive relationship between iterative model development and model validation is a key component when developing models, particularly models that will be applied in complex environments.

Wickens et al. (2008) conducted such an iterative model development and validation effort. They developed a conceptual model and then added one variable at a time to identify where the impact was within the model. Those variables that did not improve predictions of the model were removed. The model iterations proved to be a successful approach to determining
the model’s validity prior to incorporating the model into the full operational scenario. An iterative model development effort can illustrate the impact of various elements on operator performance and can provide greater insights into the underlying processes that cause the model’s performance.

Model validation appears to be improved by ensuring that model iterations are used, particularly because a model could be influenced by a number of contributing factors, not necessarily the one expected by the model developer. The only way to know the influence of any individual component of a model is through an iterative model development-validate effort to identify the contributing factors of the various model parameters.

### 3.6.3. Validation Requires Multiple Measures

The review of validation efforts also revealed that many of the integrated HPM validation efforts considered only a single output measure. However, to validate cognitive behaviours in complex HPMs (a challenge associated with validating complex HPMs as discussed in Section 3.4), validity must be judged using multiple measures (Law & Kelton, 2000). Often a model can produce valid specific output along one dimension; however, in so doing, this may create a model that may not accurately represent human behaviour along another dimension, thus producing invalid conclusions. The model analyst becomes less vulnerable to generating incorrect conclusions by comprehensively analysing multiple measures to determine operation of the model. As part of the AFRL effort, Diller et al. (2005) outline that model interpretability is a significant issue for complex human performance models because it is difficult to know precisely what the model is doing at any given time. Multiple measures assist in uncovering the causes of the model’s performance. For example, the AFRL effort used both accuracy and response time because looking at one in isolation would ignore the potential of an important speed-accuracy tradeoff. This calls for an approach to model validation that looks at the model’s validation from multiple different levels, and hence that uses multiple measures.

### 3.6.4. Data Granularity

As raised in Section 3.4, validating complex human cognitive processes is a challenge because while the human’s overt behaviour is directly observable, many of the underlying cognitive processes (such as time estimation and workload) cannot be readily observed. As such, different levels of data granularity need to be considered to determine not only that the modelled
behaviour represents the human’s behaviour, but also that the cognitive processes used to generate the behaviour are valid. As outlined by Pew, Gluck, and Deutsch (2005), a comprehensive validation approach must consider the appropriate granularity, or resolution, of the analyses – that is, whether the data are analyzed at an aggregate level or a fine-grained level. Aggregate measures are computed by averaging across an entire simulation or segment of a simulation. These are in contrast to the fine-grained measures, which attempt to consider the randomly distributed error by using individual data points (Campbell & Bolton, 2005). Aggregating data removes some of the performance variability and is an effective approach for comparing outputs across multiple models, but is less useful for evaluating specific activity sequences. Relying solely on the aggregate measure may provide the analyst with an incomplete picture of the operations of the models as it does not capture the underlying processes that are causing the model to perform. As discussed by Campbell & Bolton, (2005) and Estes (2002), aggregated data may not be the most appropriate for assessing validity of an integrated model, because different cognitive processes could be operating to produce a given output. It is necessary therefore to possess a suitable breadth of aggregate analyses and fine-grained analyses.

In the review of the model validations effort above, it was shown that the A/SA and Air-MIDAS team above evaluated only aggregate measures, whereas the ACT-R team (Byrne, Kirlik, & Fleetwood) looked at both aggregate and fine-grained level performance and found much stronger relationships for aggregate performance than for fine grained performance; however the latter was more informative regarding the underlying processes operating in the model.

3.7. SUMMARY OF THE CURRENT VALIDATION APPROACHES AND THE WAY FORWARD

The analysis of the current state-of-the-art in integrated HPM validation has served to identify a set of objectives that should be addressed to successfully validate an integrated model of a complex environment. In Section 2 (Chapter 6), a validation approach for a complex environment will be outlined that a) considers the specific purpose of the model, b) uses an iterative develop-validate approach, c) incorporates multiple measures, d) considers different levels of data granularities. The validation principles will be applied to determine the validity of a time-sensitive HPM.
AN APPROACH FOR VALIDATING AN INTEGRATED HPM OF A COMPLEX ENVIRONMENT

The goal of Section 2 is to demonstrate a validation approach applied to a model of a complex time-sensitive task. Following from the validation lessons learned identified in Section 1, it is clear that one cannot define a validation approach without consideration of the purpose of the model. As such:

**Chapter 4** will describe the complex domain of ATC operations that will be used as a test case for which to assess a validation approach.

In **Chapter 5**, a baseline model of the ATC environment will be developed using MicroSaint Sharp, a commercially produced, discrete event simulation, integrated human performance modelling tool with several interacting models, including a model of the environment and the operator.

In **Chapter 6**, a validation approach will be defined and applied to the ATC baseline model.
CHAPTER 4: AIR TRAFFIC CONTROL: A REAL-WORLD TIME-CRITICAL COMPLEX ENVIRONMENT

Understanding the manner in which human operators engage in time sensitive behaviours is critical in highly complex real world operational environments such as air traffic control, piloting, Space Shuttle mission control, and military operations. These command and control (C2) environments are often difficult to study empirically due to cost, safety, and logistical constraints. As such, HPMs of these environments are increasingly relied upon for system design and evaluation issues. Appropriate validation approaches and techniques are needed to ensure that the conclusions drawn from these HPMs are valid. This research focuses on elaborating a validation approach, specifically for complex systems in which human responses are of a time-sensitive nature. The ATC environment was selected as a suitable test bed for applying the validation approach presented in this thesis, because ATCos manage both the strategic operations of an airspace system and the tactical activities completed by individual aircraft.

This chapter describes an empirical HITL simulation of the ATC environment to be modelled. The purpose of the HITL simulation was to evaluate new technologies and procedures expected to be introduced into the current system to improve system capacity while maintaining the safety of the airspace system (JPDO, 2006). The improved efficiency is expected to result in closer coordination among aircraft, and between aircraft and ATCos, which will result in a much more highly synchronized system.

The ATC environment is characterized by multiple, time critical, competing ongoing tasks requiring the ATCo’s attention. The need for accurate models of the timeliness of ATCos responses is critical for effectively predicting future operations. The chapter will begin with a description of the HITL study conducted by the FAA Tech Center. The description will outline the ATC environment, the ATCos task, and the dependent variables that were available for model validation.
4.1. ATC HITL STUDY

A HITL study, recently conducted as a joint effort by NASA and the FAA, called the Future En-route Workstation Study, or FEWS (Willems, 2005) has been identified as an appropriate dataset source with which to inform model development, and against which to validate the model output.

The FEWS study (Willems, 2005) was conducted to examine ATCo performance in an en-route sector as a function of three traffic loads and two alternative workstation designs. The two workstation designs investigated in the FEWS study were (1) the current day operations that utilize paper strips and verbal communication amongst ATCos and between ATCos and the flight crew, and (2) a future workstation design that introduces datalink communication\(^1\) into the en-route operations.

En-route operations are those operations that involve aircraft flying through an airspace sector while en-route from their point of departure to their point of arrival. The ATCo generally services the multiple aircraft in his/her sector, primarily by maintaining safe separation between them, while concurrently communicating with adjacent sectors. En-route operations face many high workload events, in particular the handing off of aircraft from one airspace sector to another. This event requires an accurate knowledge of the passage of time so that the ATCo can correctly sequence their tasks as a result of competing demands.

4.2. DESCRIPTION OF FEWS STUDY

As a HITL experiment to test advanced displays and operational concepts on air traffic controller performance, the FEWS study tested 16 Full Performance Level controllers who had actively controlled traffic at level 11 and 12 Air Route Traffic Control Centres for at least 16 hours in the month preceding the experiment. These controllers were asked to perform 32 scenarios in en-route ATC simulations distributed over four task load conditions: (1) 100% of current traffic levels with datalink (DL); (2) 133% of current traffic levels with datalink; (3) 133% of current traffic levels with no datalink (No DL); and (4) 166% of current traffic levels with datalink\(^2\).

ATCos were presented with a version of the national aerospace system (NAS) that has a mixed fleet (75% newly equipped aircraft, 25% traditionally equipped aircraft). Procedures in

\(^1\) Datalink (DL) is an electronic communication protocol, like email, between the flight crew in the air and the ground based ATCos. No Datalink (No DL) uses current day, voice communication protocol.

\(^2\) While the FEWS study included a 166% traffic condition, the data output from this condition was incomplete so it is not included in the current analysis.
this NAS have changed to provide maximum automated flight capabilities to the fully equipped aircraft while the other aircraft continue to use 2004 procedures.

ATCos worked in sector teams of two, with either conventional (current-day) or future en-route workstation capabilities. The controllers controlled aircraft piloted by six simulation pilots flying in the Genera Control Centre environment, a generic airspace developed by the FAA. During the simulation, the weather conditions required instrument flight rules (IFR) were in effect.

The airspace and the scenarios used the high fidelity ATC simulator at the FAA WJH Research and Development Human Factors Laboratory. The simulation used an integrated system that included the Target Generation Facility to generate targets and air space and a Display Suite Replacement (DSR) emulator. The ATCo was familiarized with the airspace and the Letters of Agreement (LOAs) and Standard Operator Procedures (SOPs) were available on two adjacent controller stations equipped with a (high-resolution) radarscope, a DSR keyboard, and either a trackball or an alternative input device.

The simulation pilots were tasked to manoeuvre the aircraft according to ATCo clearances and issue “ghost” ATCo commands (a ghost sector is a sector that was included to load the ATCo with ambient traffic). A landline allowed inter- and intra-facility communications between Air Traffic Management (ATM) for planning flow routes and control. A Keyboard Selection Device and a Computer Readout Display were available for the ATCo to enter and observe relevant traffic information. ATCo communication data (recording times and frequencies) were collected from a system that possessed communication links between the ATCo, simulation pilots, and experimenters, and push-to-talk recording. Real-time video recordings were also made of the ATCos at their positions.

The FEWS study included eight teams of controllers, each of which included a Radar Controller (also known as a R-side controller) and a Flight Data Controller (also known as a D-side Controller). The R-side Controller is responsible for receiving aircraft into the sector that are handed-off by adjacent sectors, for sending aircraft into adjacent sectors, to communicate with the D-side Controller, and to monitor and manage traffic to ensure conflict-free trajectories as described in the handoff section below. The D-side ATCo organizes the flight data strips for the R-side ATCo. The present HPM effort focuses on the R-side receiving ATCo, for whom the role of time estimation is expected to have the largest impact.
ATCo performance was evaluated by both objective and subjective measures. Participants were trained to criterion through eleven scenarios that integrated use of the airspace and the DSR emulation (no paper strips were available). Each training and experimental scenario lasted between 40 and 45 minutes3.

The FEWS study investigated the effect of a change in workstation design and traffic levels on controller performance and behaviour and the effect of adding a conflict probe to the conventional R-side position. Both ATCo teams also sequence aircraft that join the flight paths and descend aircraft from the flight paths to surrounding airports in parallel to the handoff and conflict monitoring tasks.

4.2.1. Handoff Receipt

As an aircraft progresses through the system, it is to be handed off from one sector to the next sector or facility. The sending R-side ATCo initiates a handoff when an aircraft approaches a new airspace boundary - that is, as an aircraft is about to exit the sector. Once the handoff is initiated, the icon for the aircraft to be transferred begins to flash on the receiving R-side Controller’s display. The receiving R-side ATCo must first acknowledge that the handoff needs to occur, then carry out a series of other activities and return to the handoff aircraft to accept the handoff either by voice (in current day operations) or through the entry of a computer message (consistent with datalink operations).

It is detrimental to overall system performance if the receiving ATCos accept hand-offs early, because the aircraft remains the responsibility of the sending sector ATCo until the point in time that the aircraft crosses the sector boundary. However, in cases where the aircraft is accepted by the receiving ATCo too early, both the sending and receiving ATCos become vulnerable to increased workload as they may need to regain control in terms of the sending ATCo or initiate a “return-send” in terms of the receiving controller of the aircraft if the aircraft requires any flight plan modification prior to crossing into the new sector.4

3 The fact that the FEWS simulation did not have a predefined experimental scenario time, resulted in inconsistent datapoints that were available for analysis and comparison with the HPM output.
4 Note that the sending ATCos also tend to balance their desire to hand off an aircraft early, and thus, be free of responsibility for it, with the consequences of handing it off too early. Conversely, they must also weigh the consequences of potentially handing the aircraft off to the new sector too late, that is, before the receiving ATCo has time to confirm acceptance. Sending ATCos thus try to avoid allowing an aircraft to enter a new sector without being under control of the receiving sector because it is a violation of FAA rules.
Figure 2. Handoff Timings (Notional) for the Sending/Receiving Controllers

Notes: Handoffs should be completed at 60s and is considered in violation if not completed by 120s.

Figure 2 represents a simplified example of two airspace sectors, each with a controller team maintaining the safe operation of the airspace – ZGN Sector 08 and ZGN Sector R22. In this simplified example flight AC0199 has been received and is under control of the ZGN Sector 08 ATCo, who is also the sending ATCo for flight NWA002. After handoff, the receiving controller in ZGN Sector R22 accepts flight NWA002.

NWA002 is hatched in the figure to indicate that this data block has been put into the “flash state” to prepare the data block for handoff to the receiving sector. The data block will remain flashing until the receiving controller, in this case receiving controller for R22, accepts the aircraft. No data block should remain flashing when it enters into a new sector according to FAA Order #7110.65R (FAA, 2006). If the sending ATCo notices that the aircraft remains flashing, s/he is to regain control of the aircraft and contact the other sector ATCo or their supervisor. Note that the sending ATCo can put NWA002 into the “flash state” at any point when s/he has control of it, but FAA Order #7110.65R states that the ATCo shall do this 150s before the aircraft crosses into the new sector.

In addition to the above interactions, Figure 3 depicts a normative model of when the
ATCo should receive the handoff, specifically the temporal relationships between the time that the aircraft goes into the “flash state”, the window of opportunity, as well as the times that the receiving ATCo take actions to control the aircraft. As shown in Figure 3, the handoff condition possesses a clearly defined *handoff period, or” windows of opportunity”* for handoff. The window opens when the aircraft enters the "flash" state - meaning that part of the flight data block is flashing – and this happens 150s prior to the sector boundary. This means that the sending ATCo is about to relinquish control of the aircraft to the receiving ATCo, who must subsequently make the decision to accept the handoff. The optimal time for the receiving ATCo to return to the flashing datablock is 60s, however the handoff must be completed by 120s from flash onset. If the receiving ATCo fails to accept the aircraft by 120s it is labelled a Late Handoff and if it exceeds the sector boundary (150s), the aircraft is in violation of FAA rules.

At the point of flash onset, the receiving controller undertakes other tasks while estimating the passage of 60s of time. The receiving ATCo will periodically re-evaluate his progress relative to the optimal 60s window. For example, at 40s he will re-estimate whether he has sufficient time available to undertake other actions and return the flashing datablock within the window of opportunity. According to this example, this leaves the receiving ATCo with an optimal time available of 20s and a maximum time available of 80s to receive the aircraft into his sector. The receiving ATCo is not idly sitting counting time increments as he/she controls their airspace; rather they go on to complete other tasks and periodically estimate how much time has passed since the flash onset. The window of opportunity closes 30s prior to the sector boundary because if an aircraft has not been handed off by this point, ATCo’s handoff is in violation of FAA rules, and supervisors are required to intervene.

It is anticipated that if the receiving ATCo overestimates the available time to complete other tasks and return to the handoff task, they will return to the handoff task too late, in violation of FAA handoff rules. The consequence of this violation is that more aircraft will cross into adjacent sectors without being properly received by the receiving ATCo. In addition, as explained below, incorrectly estimating $T_A$ values may result in aircraft not being descended in time to cross the metering fix, and, in extreme cases, may even result in loss of separation between aircraft.

---

5 Note that in the jargon of Air Traffic Control, the word “descend” is typically used as a *transitive* verb, essentially meaning “to issue an order to descend.”
4.2.2. Conflict Detection and Resolution

In addition to performing the handoff functions, controllers are also responsible for resolving potential conflicts between aircraft. Without advanced technologies, conflict resolution is a process that requires the controller to analyze, understand and characterize any potential conflict, determine the outcomes physically possible, and select an optimal resolution for that conflict situation between aircraft (Eurocontrol, 2002). Conflict detection in the present and future generation air transportation system relies on automated conflict detection systems that aim to automate strategic resolutions to airspace conflicts (Wickens, Mavor, & McGee, 1997). Whether resolution is completed via automated means or by active processing on the part of the ATCo, the controller must communicate with the aircraft to indicate a suitable resolution to the potential conflict. This would likely involve requesting that aircraft make modifications to their flight plans along the lateral, vertical or speed dimensions (in that order of preference) to avoid an airspace conflict (Eurocontrol, 2002). The automated approaches in use today possess a conflict resolution advisory function that works in conjunction with an underlying conflict probe function to provide controllers with control action advisories that will resolve existing conflicts without causing additional conflicts.
4.3. DESCRIPTION OF DATA AVAILABLE FROM FEWS STUDY

The FEWS is a rich data source that includes environmental scenario information\(^6\), objective HITL data (e.g., communication, button presses, control input, eye tracking, workload, situation awareness, task timing, aircraft performance, etc.), and subjective data (e.g., post-trial subjective ratings and questionnaires). An evaluation of the teams’ performance revealed that Team 3 possessed more complete data for all the traffic load scenarios than the other teams and, as such, was chosen as the data source upon which to compare the model output. Compared to Team 3, the data for the other ATCo Teams (4, 5, 6, 7, and 8) had a larger number of extreme response times (either immediate or very extended), instances of missing data, and idiosyncracies associated with subject input such as double button presses. The fact that the data integrity was questionable for Teams 4, 5, 6, 7 and 8 led to the conclusion that the data for those teams should not be used, in favour of a team that had a complete data set. Overall, the data for Team 3 was deemed to be more reliable than for the other teams.

4.3.1. Receive Handoff Duration (RHD)

The FEWS HITL simulation revealed that the 100% traffic load condition required 79 handoffs to be received and the 133% conditions (both DL and No DL conditions) both required 102 handoffs to be received by the receiving ATCo within approximately the same period of time. One case out of the 79 handoffs received in the 100% condition had handoff data that were beyond acceptable limits (the controller button press did not activate correctly), while 2 cases out of the 102 handoffs received in the 133% condition had handoff data that were bad.

The FEWS provides empirical data on the time that a generic window of opportunity opens and closes. The nominal window open state is defined as being 150s before the sector boundary (when the sending ATCo attempts to put the data block into the “flash state”). The nominal window closed state is defined as being at the sector boundary (generally when the receiving controller receives the aircraft).

For each handoff, Receive Handoff Duration (RHD) was calculated as per Equation 1 by subtracting IHO, the time that the aircraft data blocks started flashing (initiate handoff time) from RHO, the time that the aircraft was received by the receiving controller (receive handoff time) for those aircraft entering the sector. An average RHD was calculated for each of the DL

\(^6\) Described in further detail in Appendix C. Consult Willems (2005) for a complete description of the experiment.
and traffic load conditions; 100%DL, 133%DL, and 133% No DL. Referring to Figure 3, the optimal RHDs should have values close to 60s.

Equation 1. Receive Handoff Duration (RHD) Calculation.

\[
RHD = RHO - IHO;
\]

where

- \( RHD \) = receive handoff duration time;
- \( RHO \) = receive handoff window time;
- \( IHO \) = initiate handoff window time

An average RHD was calculated for each of the DL and traffic load conditions; 100%DL, 133%DL, and 133% No DL. Table 1 presents the descriptive statistics (means, standard deviations, minima, and maxima) associated with the FEWS RHD data for the handoff task. It is important to note that, although the FAA mandates that the ATCo should strive to meet a 60s time window, the mean RHD times were much lower than this value, ranging from 27 to 43s. Additionally, the minimum response times appear to be almost instantaneous (2-3 s), while there were some very long response times, up to 200s.

Table 1. Descriptive Statistics for FEWS RHD for 100% DL, 133% DL, and 133% No DL. Times are in Seconds.

<table>
<thead>
<tr>
<th>Traffic Load/Condition</th>
<th>Mean RHD</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>27.18</td>
<td>28.87</td>
<td>1.85</td>
<td>120.87</td>
</tr>
<tr>
<td>133% DL</td>
<td>30.24</td>
<td>28.82</td>
<td>3.47</td>
<td>146.75</td>
</tr>
<tr>
<td>133% No DL</td>
<td>43.80</td>
<td>52.36</td>
<td>2.01</td>
<td>200.87</td>
</tr>
</tbody>
</table>

4.3.2. Workload

The workload results collected comprised the subjective evaluations of the experienced workload using the Air Traffic Workload Input Technique (ATWIT), a technique developed to evaluate mental workload in real time, using a reliable and unobtrusive input device known as the Workload Assessment Keypad (WAK) (Stein, 1985). In response to either an auditory (tone) or visual (illumination) cue, ATWIT prompts a controller/participant to press one of ten buttons within a specified time to indicate their subjectively experienced workload. In the FEWS study, participants provided ratings on the WAK device every 2 minutes using a scale ranging from 1 (low workload) to 10 (high workload).

Table 2 summarises the WAK values collected in the FEWS simulation. The data processing scripts averaged the 22 WAK data values for each simulation run. Although the R-
side controller data comprise the data of interest, the D-side controller data are also included in Table 2 for completeness. The mean R-side controller results are shown in Figure 4. The data in Table 2 and Figure 4 reveal, not surprisingly, that the controllers experienced increasing workload as the number of aircraft under active/positive control increased.

<table>
<thead>
<tr>
<th>Traffic Condition</th>
<th>R-Side Mean</th>
<th>D-Side Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>1.7</td>
<td>4.68</td>
</tr>
<tr>
<td>133% DL</td>
<td>3.56</td>
<td>5.84</td>
</tr>
<tr>
<td>133% No DL</td>
<td>2.91</td>
<td>5.05</td>
</tr>
</tbody>
</table>

What is surprising, however, is that the HITL workload for the 133 DL condition is greater than the 133 No DL condition. This is counterintuitive because addition of datalink is supposed to reduce the workload experienced by the ATCos. However, the result becomes less surprising when one considers the number of missed responses to the workload query, in other words the number of cases for which ATCo’s did not respond to the experimental request to rate their instantaneous workload on the ATWIT. There were 2 missed workload responses in the 100% DL condition, 6 on the 133% DL condition, and 11 on the 133% No DL condition. SME consultation (personal communication, J. Kamienski, ATCo Denver Center, November 9, 2006) suggested that workload was most likely too high in those cases for the ATCo to effectively deal with the tasks required. An additional ATCo SME was used to substantiate this workload characterization (personal communication, V. Battiste, ret. ATCo, August 23rd, 2007).

Figure 4. Mean and SE for FEWS WAK Workload Ratings for R-Side Controller for Different Traffic Condition.
The ATCos’ main task is to maintain efficient operation of their airspace, and when their task load becomes too great, any tasks outside of controlling traffic, such as inputting information into the ATWIT, are not performed, in favour of reserving resources for the airspace and its possible events. Assuming that the actual workload number is reduced in the No DL condition, when the number of missed tasks is factored into the equation and a 90th percentile workload score (Chen & Sun, 2008) obtained from the distribution of observed workload scores is entered in the dataset for those cases (a reasonable assumption given the communication from the two ATCo’s noted above), the workload does increase as a function of traffic load increases. These adjusted data are plotted in Figure 5.

![Figure 5. Mean FEWS WAK Workload and SE by Traffic Condition with the Missing Data Replaced.](image-url)
CHAPTER 5: DEVELOPING THE BASELINE MODEL

A baseline model of the FEWS ATC environment was developed by Leiden & Kamienski (2006) using Micro Saint Sharp, to evaluate the timeliness of ATCo handoffs between airspace sectors with a variety of technologies and procedure sets. Although this baseline model was deemed to be a valid representation of the FEWS performance in efforts conducted separate to this thesis (Leiden & Kamienski, 2006), that validation effort focused solely on environmental variables such as traffic flow and the number of aircraft within sector boundaries, and not on human behaviour variables, or on the task sequences. Model performance in Leiden & Kamienski’s modelling effort was validated using a Turing test approach, in which a SME completed an unstructured evaluation of the output of the model’s traffic sequence and the tasks included in the model. Leiden & Kamienski’s validation effort focussed on the environment model, but ignored the challenge of validating the performance of the human operator.

The focus of the present research effort was to elaborate on a validation approach to determine how well the baseline model of the FEWS ATC environment represented human behaviour, particularly as it pertains to the timeliness of ATCo’s handoffs between airspace sectors. Prior to describing the validation approach in Chapter 6, an explanation of the baseline model and its three integrated components follows.

The baseline model has three integrated components
1. Environment model
2. Human operator model
3. Activity model

5.1. ENVIRONMENT MODEL

The “environment model” refers to a numerical simulation model that represents the physical dimensions of the airspace, including its geometric representation, the various aircraft flying within it, the aircraft flight plans, and the number of flights passing through the airspace. The environment model triggers the onset of the activities (see Section 5.3) within the overall model. All of the code driving the environment model is based on latitude/longitudinal positions of the aircraft and the aerodynamics of the aircraft. Complete details are presented in Appendix D.
A separate environment model was developed for each of the traffic levels of the FEWS study. As per the FEWS simulation, the 100% traffic condition included a mean of 21 aircraft flying in the airspace sector at any given time, while the 133% traffic condition included a mean of 25 aircraft. The modelled flight plans were the same as those observed by the HITL simulation teams.

The environment model that was used to drive the baseline behaviour model was created to match the FEWS HITL simulation. A verification of the environmental model was undertaken as part of this effort, where the number of aircraft flying in the airspace along specific flight paths, the number of handoffs, the specific aircraft datablock flash state, and the handoff receipt success was verified. Leiden and Kamienski used the verification effort in developing their final environment model. Whenever the model diverged from the expected aircraft behaviour, the model was manipulated. The steps taken for this verification are presented in Appendix F.

Table 3 presents the number of aircraft, the number of handoffs that existed in the data streams and the number of cases removed for both the FEWS and the model output. Note that not all of the cases were used for comparison with the HPMs because of the mismatch between the data sets. The cases listed below as having been removed were removed because these data were considered outside the acceptable possible response values. The data set included for certain analyses reported later were only those data points that could be aligned between data sets.

Table 3. FEWS Baseline Model Taskload, Number of Handoffs and Data Points Removed.

<table>
<thead>
<tr>
<th>Taskload</th>
<th>FEWS</th>
<th>Baseline Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Taskload (# of A/C)</td>
<td>Number of Handoffs</td>
</tr>
<tr>
<td>100 DL</td>
<td>21</td>
<td>79</td>
</tr>
<tr>
<td>133 DL</td>
<td>25</td>
<td>102</td>
</tr>
<tr>
<td>133 NO DL</td>
<td>25</td>
<td>102</td>
</tr>
</tbody>
</table>

5.2. GENERAL HUMAN OPERATOR MODEL

Micro Saint Sharp has a general operator model embedded within it, which was held constant across model iterations. It makes use of standard ‘primitive behaviours’, such as visual monitoring, visual fixations, visual recognition, auditory recognition, reach, recalling

---

7 A screen snapshot of the baseline model simulation environment is presented in Appendix E.
information, sequencing aircraft through spatial calculations, and communication activities. For each of these ‘primitive behaviours’, standard four-channel VACP (visual, auditory, cognitive and psychomotor; VACP) workload values have been determined. Each workload value has been previously validated based on the Task Analysis and Workload Index (TAWL); (McCracken & Aldrich, 1984) and Mitchell’s (2000) Revised Workload scales. Detailed workload values are presented in Appendix G. Recall that the present modelling effort is of the receiving ATCo operating within the environment described above.

The model outputs workload along each of the VACP channels and the total workload is defined and calculated as “instantaneous workload,” in a manner consistent with Equation 2. This equation states that the workload is the average of each of the workload values collected per task \((t=1-15)\) per resource channel (i.e. divided by the total number of resource channels used). Fifteen tasks were selected for the moving workload window on the basis of SME guidance as it is a reasonable representation of the number of tasks in the future that the ATCo is planning at any given time (personal communication, J. Kamienski, ATCo Denver Center, November 9, 2006). This method of collecting workload differs from the FEWS simulation (ATWIT technique described above), which used real humans doing real tasks. It is unreasonable to expect human subjects to rate workload across 4 channels while engaged in the experimental task, and soliciting subjective estimates after every event would be too disruptive to the main purposes of the FEWS experiment.

Equation 2. Instantaneous Workload Calculation Completed in the Baseline Model.

\[
W = \frac{1}{15} \left( \sum_{t=1}^{15} \left( V + A + C + P \right) / 4 \right) / 15
\]

\(W\) – Workload

\(V\) – Visual demand; \(A\) – Auditory demand; \(C\) – Cognitive demand; \(P\) – Psychomotor demand.

Each channel is measured on a 0 – 7 scale.

Primitive activities are outlined by Boff and Lincoln (1988) in their Engineering Compendium, which summarizes existing research on basic human performance capabilities. For example, in order to model tasks such as the reading of textual messages, the data of interest must be in focus, attended to, and within the foveal field of view for 200 ms. This human process agent computes the environment or cockpit objects that are imaged on the operator's retina, tagging them as in/out of the peripheral and foveal fields of view (90º and 5º respectively), in/out
of the attention field of view (variable depending on the task), and in/out of focus relative to the fixation plane (distance from the eye).

These basic human performance capabilities are converted into computational algorithms, representing human process models, using relevant research identified in either the Boff and Lincoln (1988) Engineering Compendium or by Boff, Kaufman and Thomas (1986), or by other generally accepted human performance data. These standard models, which are commonly considered to be validated models, are then embedded within the modelling tool, Micro Saint Sharp. These basic human behaviours become triggered by the modelled sequence of tasks that, in turn, get triggered by the environment model. Since the behaviours being modelled are fundamental, they can be applied to ATC or to any other domain. In the ATC case, all aspects of the simulation, which include the onset of human activities, are driven by sector events.

5.3. **ATC Activity Model**

Activity models are the modelled sequences of tasks that the simulated operator undertakes in response to the environmental triggers. In the current simulation, the ATC activity model represents the following tasks: aircraft handoffs, transfer of control tasks, resolve conflict, meter or sequence aircraft, arrival descent clearances, departure to cruise altitude clearances, and rerouting due to weather. While the handoff task is the focus of this research effort, the controller is also concurrently responsible for the other tasks. Any deviation from the optimal response values by a controller, such as in the case of a late handoff, constitutes ‘suboptimal’ behaviour. These tasks, and the logic associated with the tasks, as coded in the Micro Saint task network model are illustrated in Appendix H. The task network models further decomposes each task into low-level units of action, or "nodes", connected together by a series of branches to represent a network or a sub-network. This task network model is linked to the operator model that possesses empirically determined response profiles (timing, workload, situational elements, etc.) and tasks become activated, or triggered, through the environment model. Tasks are serviced within this environment depending on the availability of operator “resources” and on the schedules of performance that have been programmed into the task network model.

5.4. **Task Prioritization**

It is important to note that the modelled operator is *a single channel operator who responds rationally to one task at a time*. The term “rationally” here refers to the assumption that the
human operator will decide on which task to perform at any one time by prioritising all tasks, on the basis of maximising the value of the expected outcome. The baseline model’s prioritization mechanism uses a time window to drive the performance of the controller. “Window open” is defined as the time when a task first becomes available for the controller to perform it, and “window closed” is defined as the point at which the action of the controller exceeds a specific time limit.

Referring to Figure 2 and Figure 3, the logic implemented in the ATC activity model is that the flashing begins 150s prior to the aircraft data block crossing the sector boundary, and the initiate handoff condition is supposed to occur 60s after the flashing begins. The ‘receive handoff threshold’ is defined as being 30s prior to the aircraft’s crossing the sector boundary, which marks the end of the maximum window of opportunity. The receiving controller needs to have completed his “receive aircraft” tasks, which in this case include monitoring the situation surrounding the aircraft, examining the data block information of all surrounding aircraft, and making the decision to accept the aircraft on the current flight plan (speed, heading, altitude) into the sector by the 30s receive handoff threshold. When using the DL method, accepting the aircraft into the airspace sector requires the controller to select the flashing data block with his mouse cursor and click on it. The No DL method, on the other hand, requires active communication between the ATCo and the aircraft’s flight crew.

This model uses only the actual time available as the criterion for prioritizing all tasks (i.e. not just handoff receipt tasks). That is, the modelled R-side receiving ATCo determines which aircraft handoff to process based on the time available in each window of opportunity that started when each individual aircraft icon begins flashing. At any given time, the model tabulates a list of currently queued tasks to be completed, calculates the time available for each one, and prioritizes the tasks in this queue such that items with the smallest time available of all tasks are processed first. For example, in the baseline model, when the queue size variable is set to a higher value (e.g. ShedTasksQueueSize = 10), the ATCo will essentially never shed (or drop) tasks (because the ATCo is unlikely ever to have 10 pending tasks in the ATCo queue) so the model does all the listening to ATCo readbacks, which make it slower to respond to handoffs.

To determine the task with the smallest time available in the optimal window of opportunity, the baseline model assumes perfect time estimation and therefore perfect knowledge of time available (i.e. does not use any estimates in guiding the model’s task schedule). In
particular the baseline model allows at most five critical tasks to enter into a ATCo’s queue and then, whenever a new task must be selected, strategically selects the task with the highest priority – the one critical task with the smallest time available – as the one to carry out next. This is akin to a first-in-first-out (FIFO) schedule. It is also possible that more than five critical tasks enter into the queue but this happens only with extremely large numbers of aircraft. When more than five tasks are in the queue, non-critical tasks such as pilot communications are shed.

The model of task management behaviour, depicted in Figure 6, assumes that controllers implement nominally optimal, strategic task scheduling. It uses the following queue priority level for task scheduling, from highest to lowest: receiving incoming calls from pilots with the shortest time available, resolving conflicts, and metering aircraft. The task queue measure is used to verify that the models are performing correctly.

The time available estimate, denoted as “Actual” in Figure 6 reflects the current state-of-the-art in human performance modelling, in that it assumes that the human can perfectly estimate $T_A$ with no error and with no effect of workload on these time estimates.

![Figure 6. Baseline Model of Task Management Behaviour.](image)

### 5.5. EXERCISING THE BASELINE SIMULATION MODEL

Once the simulation was developed and the environment rendered, 3 independent runs were completed: one at a low traffic load of 100% (78 handoffs) with DL communications, a second at a higher traffic level of 133% of typical traffic loads (100 handoffs) with No DL communications (relying on voice communication between ground and aircraft), and a third scenario under the higher traffic level of 133% of typical traffic loads (100 handoffs) with DL communications. Each of the simulation runs required that the simulated ATCo teams maintain
control of the airspace consistent with the rules outlined above in the FEWS experiment. The position of interest in the current simulation was the R-side ATCo of the Genera Sector 08 who was responsible for sending and receiving aircraft to and from adjacent airspace sectors, although the validation effort focussed on the receiving part of the R-side ATCo’s job.

Note that human performance modelling typically require multiple runs in order to generate reliable estimates of performance, given that each subtask is completed only once and there is variability in the modelled human performance variables. For example, consider a HPM of a pilot completing an approach and landing task. Each subtask, such as “gear down” or “flaps up,” is completed only once and the time to complete each task is usually modelled using the mean and standard deviation of some accepted distribution. In such cases, it is therefore necessary to use a large number of trials to generate reliable estimates of mean performance. In the present modelling effort, reliable estimates exist from a single simulation run, since the same human operator model is completing the same task (receive handoff) multiple times. This same human operator model is conducting a set of handoff activities in response to an environment that contains a number of aircraft. These aircraft, which fly along predetermined routes, serve as triggers for the ATCo activities (metering, handoff, etc.). The sequence and timing of the aircraft drive the ATCo’s ability to attend to the aircraft data blocks. The stochastic components of the baseline ATCo model include the modelled visual fixation times, and the communication processes and response times. The timing of the ATCo activities depends on the number of aircraft that are in the airspace sector and their distance to the sector boundary. Different ATCo traffic loads will impact the timing (the onset) of the ATCo tasks. The fact that there are 78 aircraft means that there are 78 data points related to time estimates from one simulation run.
CHAPTER 6: INITIAL VERIFICATION AND VALIDATION OF THE BASELINE MODEL

Before the baseline model (described in Chapter 5) can be used for the intended purpose of evaluating the ATCo’s response timeliness, it must first be validated. As defined in Chapter 3, model validation is the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world, from the perspective of the intended uses of the model or simulation (Balci, 1998; Law & Kelton, 2000; Sargent, 1980). A model validation approach was applied to determine how well the model represents human behaviour on the time-sensitive task of aircraft handoffs. This was the first of an iterative develop-validate approach. Further develop-validate iterations will follow in Section 3. Each model iteration phase was compared to the FEWS data set with statistical comparisons to determine whether the model differs from the FEWS (using an alpha level of .05).

6.1. ANALYSIS APPROACH

A model validation approach was applied that a) considered the specific purpose of the model, b) used an iterative develop-validate approach, c) incorporated multiple measures, and d) considered different levels of data granularities. Each is discussed below.

6.1.1. Validation Measures for the Model’s Specific Purpose

The purpose of current model was to develop a validated model of time management behaviour, specifically for the time-sensitive task of aircraft handoffs. To address this specific model purpose, the Time Correspondence measure (TC measure) was developed and applied in a time critical environment. The TC measure is a new gauge for quantitatively determining the validity of the event times in the environment and for visualizing the differences between the window open time and the handoff accept time for aircraft with identical (matched) flight paths in the model and the FEWS environment. The TC measure is considered to be a quantitative measure because it uses as its primary feature, the Time Correspondence graphs (TC graphs). Graphs are an established quantitative measure in the field of modelling (Law & Kelton, 2000). The TC graph presents FEWS data (either the window open or handoff complete times) above connected with lines to the corresponding data from the model below. Lines that are perfectly vertical indicate that the times were the same in both FEWS and model. Lines that slant down and to the right indicate that the model times were later than corresponding FEWS times. Lines that slant
down and to the left indicate that the model times were earlier than corresponding FEWS times. Lines that cross indicate different sequences in FEWS and the model, with each crossover representing two or more aircraft that were out of sequence.

To support the crossover visualizations, a calculation was completed to determine whether the aircraft orders were the same from the model as they were from the FEWS for each of the window open times and handoff complete times. This was completed by ordering all of the output according to the time that the window opened or closed and comparing whether the aircraft ID from the model was the same as the aircraft ID from the FEWS. These data are presented in tables preceding the TC graphs.

6.1.2. Iterative Approach

The iterative model develop-validate phases began with a baseline model and progressed through three subsequent develop-iterate phases in each of which a different time management submodel was introduced. Each develop-iterate model development phase used the same validation measures and techniques to determine the degree to which the time management submodel developed would bring the model performance closer to the FEWS simulation data. The iterative develop-validate process incrementally modified one parameter of the model at a time so that any changes in model performance could only be attributed only to that parameter. This approach provided a structured approach to evaluate the impact of each parameter on the baseline model’s performance in a manner consistent with Wickens et al. (2008).

6.1.3. Multiple Measures

The validation approach used multiple measures to: 1) verify the model, 2) validate the model, and 3) provide a better understanding of the likely causes of the model’s performance. The five different measures outlined next in section 6.2 were used to verify or evaluate the validity of the model. Each measure was selected given its criticality for the model developed for the specific purpose of ATCo handoff performance. In addition to the TC measure described above, measures such as workload and queue length were used to verify that the model is operating as expected, and validly represents human behaviour.
6.1.4. Aggregate Versus Fine-Grained Level of Analyses

The validation process evaluated predictions of aggregate behaviour (i.e., computing a mean value across all handoffs within a model run and comparing it to the mean value of all handoffs from a FEWS simulation trial) and predictions of fine-grained behaviour (i.e., comparing each individual handoff datapoint in the model to the FEWS data).

The aggregate measures included workload, queue length, and receive handoff duration (measures 1, 2, and 3 in section 6.2). The fine-grained measures, including the TC measure, handoff window open time and the handoff complete time, were used to explain differences in the aggregate receive handoff duration times. These validation variables are described next.

6.2. THE VERIFICATION AND VALIDATION VARIABLES

In order to compare the human data with the baseline model outputs, a series of descriptive analyses, correlation and regression analyses were conducted. Five measures were selected to assess the validity of the baseline model.

6.2.1. Measure 1: Workload

Workload is one of the multiple validation measures because evidence suggests that workload impacts the orderliness of task engagement (Corker & Verma, 2001; Corker et al., 2003; Hollnagel, 1993, 1998). Including workload as a validation variable will determine if the model represents the FEWS data source and if the environment correctly drives the operator’s workload values.

Workload in the model is the average of the visual, auditory, cognitive, and psychomotor workload estimates generated by the sequence of tasks required by the ATCo in response to the environment. In instances where the FEWS data (ATWIT ratings) possessed missing responses, the 90th percentile workload score obtained from the distribution of observed ATWIT workload values over all trials (a value of 4 on the 10 point scale) was included. A new variable called adjusted workload was created and was reported in the subsequent validation efforts. For the model, workload values were collected every 4 seconds of model time resulting in 830 workload values for each run. It was upon these values that the aggregate analyses were conducted. Workload was also used in a fine-grained analysis, which required a visual representation to be generated from the model. Subsequently, each group of 38 consecutive workload values were averaged to create 22 (830/38) workload values that mirrored the 22 values collected empirically.
in the FEWS simulation *for the fine-grained visual comparison analysis* completed between the model and the FEWS data output. Note that while the ideal was to possess an equal number of responses in each of the FEWS and the model outputs, the FEWS possessed a number of missing data points in each scenario. The experimentally specified 22-workload values were generated from the model and those cases that were missing were removed from the model comparison.

The ATWIT workload estimates from the FEWS and the averaged VACP workload output from the model both generated an overall estimate of workload. There is an extensive body of literature that suggests that subjectively rated workload can be compared to model-computed workload (Gluck & Pew, 2005; Hahler et al., 1991; Lockett, 1990), and the trends of human to model workload performance can be meaningfully compared, even though they used different scales (Corker, Gore, Fleming, & Lane, 2000; Laughery, 1999; Laughery, Plott, Engh, & Scott-Nash, 1996; Yow, 1999). Elkind et al., (1989) discuss that comparing model to human workload estimates by visual illustration and comparison of the workload trends are reasonable approaches for studying/predicting workload.

### 6.2.2. Measure 2: Queue Length

Queue length refers to the number of items in a queue waiting to be serviced. Items enter the queue and remain there when resources are not available for the model to service the task. They are sequenced based on a prioritization schedule determined by the time available to complete the task. It is important to highlight that the task queue measure relates only to the model outputs, due to the fact that there is no corresponding output from the FEWS simulation. As such, the queue length variable is used to verify that the model is operating as it would in a human operator, since queue length emerged as a function of how the operator processed the aircraft that were associated with the increase in traffic levels in the environment. In the baseline model, at each data collection stamp, which is driven by the events in the environment, the simulation outputs a number of items that are waiting to be serviced in the queue. Technically, queue length is a verification measure because there is no corresponding FEWS data; however it is included in the set of validation measures because it is used to support the validation process.

Two dependent variables associated with queue length are used to demonstrate the impact of the environment on the workload produced by the model. Average queue length is the time-weighted average number of tasks in a queue across an entire simulation. For the model, as with the workload measure, queue length values were collected every 4 seconds of model time.
resulting in 830 queue length values for each run. It was upon these values that the aggregate analyses were conducted. The maximum queue size is the maximum number of tasks that were in the queue during the model run. The maximum queue length indicates the load on the model irrespective of the amount of time that the queue length remains at a maximum value. Recall that the queue length pertains only to those tasks that can be shed/interrupted (non-critical tasks). This means that, even though the model parameter is set to a maximum of 5 in this baseline model, there may be more than 5 tasks in the queue at any given time because the model is working also on non-interruptable tasks. As a result, the variable maximum queue length can provide some insight into the reasons for the performance of the model, particularly when used along with the other validation measures. Both the average and the maximum number of tasks in the queue are variables that can be used to validate the trend associated with the model’s performance in response to the modeled environment.

6.2.3. Measure 3: Receive Handoff Duration (RHD)

Receive Handoff Duration (RHD) is defined as the time difference between the flash onset of the aircraft datablock and the receipt of the flashing datablock by the receiving ATCo. This RHD measure demonstrates the degree to which the model estimates time available and completes the receive handoff tasks in a timely manner as defined by receiving the aircraft before the aircraft crosses into the new airspace sector. The handoff condition has a clearly defined window of opportunity. The window opens when the aircraft enters the "flash" state - meaning that part of the flight data block is flashing – at which time the transferring controller allows the receiving controller to accept the handoff. The window objectively closes 120s after the data block starts flashing.

Recall that each of these simulation runs comprised either 78 or 100 handoff activities (depending on the experimental condition) by the modelled ATCos, meaning that there were 78 or 100 time estimates being completed by the modelled ATCo to form the mean sector performance. To generate the mean performance figures reported in the present thesis, a number of the analyses used only those cases that could be aligned between the FEWS and the model; 39 in the 100DL, 52 in the 133DL, 41 in 133 NoDL.

The RHD time for the handoff in the HPM was inferred from the definition of the task performance within the model and was calculated as per Equation 3. Equation 3 is different than the FEWS calculation because the model used different variables. The tasks entered into a queue
when resources were unavailable to service them and exited when resources became available. \( Q_{in} \), the time that the task entered the queue (the time that the data block began flashing) was subtracted from \( T_{comp} \), the time that the task was completed.

Equation 3. Receive Handoff Duration (RHD) Calculation for the HPM Data.

\[
RHD = T_{comp} - Q_{in};
\]

where

- \( RHD \) = receive handoff duration time;
- \( T_{comp} \) = Task completion time;
- \( Q_{in} \) = queue time in

### 6.2.4. Measure 4: Window Open Time

Window Open Time, a measure of the performance of the environment model, is the time that the aircraft becomes available for the ATCo to hand off to the adjacent sector (i.e., flash onset). The window open time was assessed to a) determine if the sequence of task window open times differed between FEWS and model, b) to compare Window Open clock time between FEWS and Model, and c) to explain differences in the aggregate RHD times.

### 6.2.5. Measure 5: Handoff Complete Time

Handoff complete time is defined as the simulation (clock) time of the receiving controller’s receipt of the flashing datablock. This is a measure of how well the model represents the actions of the ATCo, in regards to the receive hand off task. This was assessed to a) to determine if the sequence of Handoff accept times differed between FEWS and model, and b) to compare handoff complete clock time between FEWS and Model. The handoff complete time was adjusted to account for Window Open Time differences between FEWS and model, leaving a more accurate account of the effect of handoff complete time that can be attributed to the ATCo and not to the fact that the window could have opened later.

### 6.3. VERIFYING AND VALIDATING THE BASELINE MODEL

In this section, the validation measures and techniques described above are applied to the baseline model.

#### 6.3.1. Measure 1: Baseline Model Workload

##### 6.3.1.1. Mean Workload for Baseline Model

Figure 7 illustrates the average workload per experimental condition for both the FEWS adjusted data, which was made up of (a) 22 workload estimates on the 10-point ATWIT scale.
and (b) the baseline model predictions, which were computed from the original 830 VACP scores, each on a 7-point scale, by averaging groups of 38 consecutive scores, as indicated in Equation 2.

Figure 7. Mean (a) FEWS Adjusted and (b) Baseline Model Workload by Traffic Condition.

There was a significant 2x3 interaction $F(2,40)=48.66, p<.001)$. Table 4 shows significant differences in workload for each pair of traffic load conditions for both FEWS (adjusted) ($F(2,40)=57.57, p<.001$) and the baseline model ($F(2,40)=9.06, p<.001$), except for the 133DL to 133 No DL conditions for FEWS\(^8\). The lack of significance could be due to the perceived impact by the ATCo of the no datalink condition on workload estimates, which appeared to be relatively small. These data illustrate that the model adequately represented the FEWS data and that the modelled environment drove the operator’s workload values in a manner consistent with the FEWS. It is interesting to note that no significant differences existed in the FEWS data in the 133 DL to 133 No DL condition while the model does possess a significant difference. A fine-grained analysis was completed to explore the possible reasons for this difference.

Table 4. Significance Tests on Adjusted Workload for Traffic Level Increases for FEWS and Baseline Model.

<table>
<thead>
<tr>
<th></th>
<th>100 DL-133 DL</th>
<th>100 DL-133 No DL</th>
<th>133 DL – 133 No DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEWS (adjusted)</td>
<td>$t(21)=8.14^{***}$</td>
<td>$t(21)=6.86^{***}$</td>
<td>$t(21)=1$</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>$t(829)=4.36^{***}$</td>
<td>$t(829)=8.46^{***}$</td>
<td>$t(829)=3.3^{***}$</td>
</tr>
</tbody>
</table>

* $p<.05$, ** $p<.01$, *** $p<.001$

6.3.1.2. Workload Trend Output for Baseline Model

Figure 8 presents the individual workload ratings from the FEWS data (upper graphs) and the model predictions (lower graphs) for each traffic load condition. It can be seen that the FEWS and the model produced similar workload trends in the 100 DL (Figure 8a) and 133 No

---

\(^8\) As a reminder, the FEWS collected 22 workload measures while the model collected 830 instantaneous workload values.
DL (Figure 8b) conditions. As shown, workload was maintained at the low-end of each respective scale – recall that the FEWS workload was measured on a 10-point scale, and the model workload was measured on a 7-point scale. Also workload increases and decreases in the FEWS output were mirrored in the model output. Unfortunately, the 133 No DL condition (Figure 8c) in the FEWS simulation possessed a significant number of missing values, which makes drawing conclusions about the comparison of FEWS and model trends difficult in that condition.

![Figure 8](image_url)

(a)  FEWS Workload 133 DL  
(b)  FEWS Workload 133 CL  
(c)  FEWS Workload 133 NCL  
(b)  Baseline Model 133 DL Workload  
(c)  Baseline Model 133 CL Workload  
(d)  Baseline Model 133 NCL Workload  

Notes – FEWS possessed missing responses. There were 18 responses for (a), 15 for (b), 9 for (c).

6.3.2. Measure 2: Queue Length for Baseline Model

As a reminder, the variable queue length was included in the analysis as a validation measure to provide evidence that the queuing process of the HPM functions as it would in a human operator. The task queue in the baseline model was used to drive the initiation and completion of tasks and to determine which tasks should be carried out next, based on priority. The baseline model assumed that ATCo’s implement nominally optimal, strategic task scheduling. In particular, the baseline model allowed up to five tasks to enter into an ATCo’s queue and whenever a new task was to be serviced, strategically selected the task with the highest priority and with the smallest time available optimal window of opportunity (relative to the 60s optimal time as defined by the FAA’s handoff guidelines) to carry out next. When more than five tasks were in the queue, non-critical tasks such as pilot communications were shed. In determining the task with the smallest time available, the model assumed perfect time estimation and, therefore, perfect knowledge of
time available. Two related variables, maximum task queue and average task queue (see definitions above), measure the impact of the environment on the operators’ workload.

It can be seen in Figure 9 that the maximum task queue increased as the traffic load increased from 100% DL to 133% DL to 133% No DL. The increase in the maximum number of queued tasks in the 133% No DL condition suggests that the model was queuing more tasks because the ATCo’s did not complete DL handoffs and clearances due to the high workload levels. The same pattern of results was observed with mean queue length variable (see Figure 10). The queue length increased significantly when the traffic level increased from 100% to 133%, $t(829) = 9.52, p < .001$, and also when the traffic level increased from 100% to 133% No DL $t(829) = 38.95, p < .001$. Of particular interest was the large increase in queue length for the 133% No DL condition as compared to 133% DL, $t(829) = 34.166, p < .001$, indicating that the No DL condition required more tasks to be delayed than in the DL condition. This set of analyses validates that the model is operating as expected because the number of items that enter into the queue emerges based on the logic that is contained within the task management (TM) portion of the baseline model (that allows a maximum of 5 items into the queue before task shedding occurs).

![Figure 9. Maximum Number of Tasks to be Serviced in the Baseline Model Queue as a Function of the Experimental Condition.](image9)

![Figure 10. Average Number of Tasks to be Serviced in the Baseline Model Queue as a Function of the Experimental Condition.](image10)

**6.3.3. Measure 3: Receive Handoff Duration (RHD) for Baseline Model**

6.3.3.1. **RHD t-test and Results for FEWS versus Baseline Model**

Figure 11 presents the data associated with the RHD measure produced from the FEWS simulation and those RHD predictions generated by the baseline model as a function of the three experimental conditions. Figure 11 illustrates no significant differences in the mean RHD times...
for the 100% DL condition between the FEWS and the baseline model\(^9\), \(t(76)=0.85, p>.05\). A significant difference did exist, however, for the 133% DL condition between the FEWS data and the baseline model predictions, \(t(99) = 2.07, p<.05\). The same was true for the 133% No DL condition, \(t(80) = 5.51, p<.001\).

### Figure 11. Average RHD Times from the FEWS data and Baseline Model.

#### 6.3.3.2. RHD Chi-Square Test and Results for FEWS versus Baseline Model

Chi-square analyses were completed to determine if the distributions of RHD times from the baseline model and the FEWS observations came from the same population. It must be acknowledged that the chi-square tests are very sensitive to the number of bins used and how they are defined. For the current analysis, the RHD times were binned into three groups, which were defined as follows: 0 – 10s, 11-30s, and greater than 31s. These were chosen to yield bins of approximately equal size, and were applied to all traffic load conditions for both model and FEWS data.

The significant \(\chi^2\) test conducted on the RHD times for all three traffic levels suggested that the frequency distributions from the baseline model and the FEWS observations differ from each other (see Table 5).

<table>
<thead>
<tr>
<th>Traffic Levels</th>
<th>Chi-square</th>
<th>(r) value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>(\chi^2 (2, N=124) = 4.96^{**})</td>
<td>0.21</td>
</tr>
<tr>
<td>133% DL</td>
<td>(\chi^2 (2, N=165) = 3.69^{**})</td>
<td>-0.15</td>
</tr>
<tr>
<td>133% No DL</td>
<td>(\chi^2 (2, N=154) = 38.98^{***})</td>
<td>0.12</td>
</tr>
</tbody>
</table>

\(^*p<.05, **p <.001, ***p <.001\)

\(^9\) Reminder that this analysis is not limited to the matched data but did filter the inconsistent data out of the analysis.
Thus the distribution of data implied by the baseline model significantly differed from the corresponding distribution of human data as collected in the FEWS simulation. While the $\chi^2$ test is a good test for testing significance between groups, it does not provide insights into the strength of the relationship, something that correlations ($r$) do provide.

Possible reasons for this discrepancy will now be discussed. Whereas the modelled RHD times were significantly higher than the FEWS times in the 133 DL condition, the absolute difference between the means was relatively small, only about 12s. However, the baseline model overestimated RHD in the 133 No DL condition by a much greater amount, with an absolute difference of 60s, or approximately 140%. The following section will examine the histograms of the individual data points of the FEWS and the baseline model RHD, and their implication for explaining differences between the baseline model derived and FEWS data.

6.3.3.3. **RHD Histogram for FEWS and Baseline Model**

Figure 12, Figure 13, and Figure 14 present the RHD times in a histogram for both the FEWS and baseline model in each traffic load condition. For the FEWS RHD times (left panels of Figure 12, Figure 13, and Figure 14), it is apparent that the large majority of response times occurred early in the handoff window of opportunity, before 60s elapsed, and mostly within the first 30s. While the ATCos were striving to complete the handoff within 60s, it was also apparent that few outliers existed with response times above 90s when the ATCo did not correctly estimate the tasks to be completed in the window of opportunity and failed to complete the handoff in the required time. The right panels of Figures 13, 14 and 15 show the RHD times produced from the baseline model. As can be seen, the trend of RHD time was very similar to that of the FEWS simulation data for the 100% DL. For the 133 DL condition, there were more early responses in the FEWS data than the model. The RHD times produced from the model in the 133% No DL condition showed a markedly different pattern than the FEWS data. The baseline model produced RHD times that were later than the FEWS times. In the model, the large majority of response times occurred later in the handoff window of opportunity, with over 50% of the responses occurring after 120s have elapsed. Contrary to the FEWS simulation data, very few responses in the model occurred before 60s elapsed. The pattern of results shown in the histograms presented below, possess similar trends observed in the correlations provided above, in which the model accounted for more variance in the 100% condition than either of the 133% conditions. This performance difference could have occurred for one of three reasons:
1. The environment did not trigger the a/c in the same manner. For example, the model may have bunched the a/c together more than the FEWS, which could have caused an increase in the operator’s processing time.

2. The modelled operators did not process the aircraft in the same way as the FEWS operators did. The modelled ATCo could have taken longer to process aircraft than the FEWS ATCo.

3. The ATCos processed the aircraft as they arrived, rather than waiting until they were closer to the optimal time of 60s and this corresponds to more early responses in the FEWS than in the model particularly as workload increases.

Figure 12. RHD Histogram for (a) FEWS and (b) Baseline Model 100% DL.

Figure 13. RHD Histogram for (a) FEWS and (b) Baseline Model 133% DL.
6.3.4. Measure 4: Analysis of Handoff Window Open Times for Baseline Model

To explain the differences between the FEWS and the model RHD performance, an accurate understanding of the environment model was required and to assist in gaining these insights, TC graphs were generated. Figure 15, Figure 16, and Figure 17 demonstrate the TC graphs for the window open times, where the relationship between the FEWS and the baseline model’s window open times are illustrated. Recall that lines that slant down to the left indicate that the window opened earlier in the model than FEWS, and lines that slant down to the right indicate that the window opened later in the model than FEWS. Lines that crossover indicate aircraft window open times that occurred in a different order in the model than FEWS. (Note that in some cases, crossovers span multiple aircraft, so a crossover may involve more than one aircraft). The number of model window open times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the window open times were the same or different between the model and FEWS are presented in Table 6. As can be seen, approximately 52%-66% of the aircraft window open times were in a different order in the model as compared to the FEWS. This occurred because the baseline model represented an optimal ATCo, however the actual FEWS ATCo may have made decisions to alter the speed or altitude of the aircraft prior to the handoff, thus altering the window open time. This highlights the complexities involved in modelling complex tasks, such as ATC, for which there often is more than one ‘correct’ way to manage the traffic. One may be inclined to adjust the environment model such that the model’s window open times perfectly matched the FEWS ATC window open times. However, this would yield a model that is only valid for that one ATCo crew, but would not necessarily generalize to other ATCo crews. Instead, the decision was made to statistically adjust window
close times to account for differences between the model and FEWS handoff times, as will be discussed in the section below.

Table 6. Number, Relative Timing and Order for FEWS and Model Window Open Times\textsuperscript{10}.

<table>
<thead>
<tr>
<th></th>
<th>Window Opened Earlier or Same Time in Model than FEWS</th>
<th>Window Opened Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>11</td>
<td>33</td>
<td>17</td>
<td>27</td>
</tr>
<tr>
<td>133% DL</td>
<td>11</td>
<td>41</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>133% No DL</td>
<td>9</td>
<td>32</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>

Figure 15. TC Graph for FEWS and Baseline Model Window Open Times 100% DL.

Notes - Vertical Lines: FEWS time = Model time; Lines slanted down to right: FEWS time earlier than Model; Lines slanted down to left: FEWS time later than model; Crossing lines: each crossing line = 2 a/c out of sequence

\textsuperscript{10} Note only 39 values were reported here instead of the 44 reported earlier because this is all of the data that could be aligned between the FEWS and the model.
Figure 16. TC Graph for FEWS and Baseline Model Window Open Time 133% DL.
Notes - Vertical Lines: FEWS time = Model time; Lines slanted down to right: FEWS time earlier than Model; Lines slanted down to left: FEWS time later than model; Crossing lines: each crossing line = 2 a/c out of sequence

Figure 17. TC Graph for FEWS and Baseline Model Window Open Times 133% No DL.
Notes - Vertical Lines: FEWS time = Model time; Lines slanted down to right: FEWS time earlier than Model; Lines slanted down to left: FEWS time later than model; Crossing lines: each crossing line = 2 a/c out of sequence
While the overall data suggest that the 133% No DL condition produced the greatest differences in RHD times, this effect cannot be attributed primarily to the time that the window opened, since the present analysis suggests that the number of window open times that differed in order from the modelled and actual aircraft were fairly consistent across traffic conditions. The cause of the differences was therefore sought in the following analysis.

6.3.5. Measure 5: Adjusted Handoff Complete Time for Baseline Model

Whereas window open times are an indication of how the environment differed between FEWS and the model, handoff complete times indicated how representative the model was of how the FEWS operators processed the handoffs. The number of model handoff complete times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the handoff complete times were the same or different between the model and FEWS are presented in Table 7. As can be seen, approximately 44-58% of the aircraft handoff complete times were in a different order in the model as compared to the FEWS.

As seen above, some differences in the sequence of the window open times were noted. As such, the handoff complete times were adjusted to account for the differences in the window open times to eliminate any confound in handoff complete time that could be attributed to the window open time. For example, if a particular window open time was 5s later in the model than FEWS, then 5s was added to the model handoff complete clock time as compensation.

Table 7. Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times.

<table>
<thead>
<tr>
<th>Traffic Condition</th>
<th>Handoff Completed Earlier or Same Time in Model than FEWS</th>
<th>Handoff Completed Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>23</td>
<td>21</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>133% DL</td>
<td>11</td>
<td>41</td>
<td>22</td>
<td>30</td>
</tr>
<tr>
<td>133% No DL</td>
<td>8</td>
<td>33</td>
<td>23</td>
<td>18</td>
</tr>
</tbody>
</table>

The TC graphs in Figure 18, Figure 19, and Figure 20 illustrate the relationship between the FEWS and the adjusted baseline model’s handoff complete time. For all three traffic conditions, the alignment of the adjusted handoff complete time for the model and the FEWS data suggests that the model is completing the handoff in approximately the same order and time as the FEWS simulation did.
Figure 18. TC Graph for FEWS and Baseline Model Adjusted Handoff Complete Time 100% DL.

Notes - Vertical Lines: FEWS time = Model time; Lines slanted down to right: FEWS time earlier than Model; Lines slanted down to left: FEWS time later than model; Crossing lines: each crossing line = 2 a/c out of sequence

Figure 19. TC Graph for FEWS and Baseline Model Adjusted Handoff Complete Time 133% DL.

Notes - Vertical Lines: FEWS time = Model time; Lines slanted down to right: FEWS time earlier than Model; Lines slanted down to left: FEWS time later than model; Crossing lines: each crossing line = 2 a/c out of sequence
6.4. CONCLUSION OF BASELINE MODEL VALIDATION EFFORT

The goal of this effort was to demonstrate an approach with which to validate the baseline model of a complex task, for which operator responses are time sensitive. This first phase of the analysis thus far has presented a comprehensive validation approach that a) is specific to the purpose of the model (in this case the ATCo’s time-sensitive behaviours), and b) uses multiple quantitative measures that are compared to HITL data. This was the first of the iterative develop-validate approach. Further develop-validate iterations follow in Section 3. The baseline model is clearly limited in its ability to predict RHD times, presumably because the activity model is limited in the degree to which it replicates human behaviours.

Two explanations for the breakdown in the above model performance are offered. First, this could have occurred because the strategy that the ATCo used to process the handoff tasks differed between the FEWS and the model (Corker & Verma, 2001), which affected the manner in which tasks were scheduled. FAA standard operating procedures mandate that ATC must
receive the handoff 90s before sector cross, as illustrated in Figure 3, while reality dictates that actual ATCos hand aircraft off as soon as possible, reserving as much time and space as possible to resolve unforeseen events. (V. Battiste, ret. ATCo, personal communication, August 23rd, 2007). This tendency was particularly noted in the highest workload condition (133% No-DL). Specifically, it suggests that human operators reverted from *strategic* control to *opportunistic* control in the high workload condition, which affected the manner in which tasks were scheduled.

Second, it could be that the human operators simply failed to estimate the passage of time accurately due to excessive workload (Kamienski, 2006; Hart, 1987). Specifically, the model assumed that the operator always has “perfect” awareness of available time to complete a task, and, furthermore, it assumes that tasks are completed quickly, in the right order and at the right time. However, previous research undertaken in various laboratories showed that humans misestimate time, especially in high workload conditions (Block & Zakay, 1997; Hart, 1972; Hancock & Chignell, 1988; Hicks et al., 1976; Hicks, Miller, Gaes, & Bierman, 1977; Jones & Boltz, 1989; Rantanen, Levinthal, 2005; Rantanen, Levinthal, & Yeakel, 2005; Zakay, 1993). Underestimating time available results in tasks being scheduled for completion early within the window resulting in a burst of early responses. In Section 3, the baseline model will be augmented to address these explanations and the resulting models will be compared to the HITL FEWS data using the same validation approach applied here to the baseline model. These possibilities *as they relate to the RHD* will be explored in Chapter 7 and Chapter 8.
SECTION 3
DEVELOP-VALIDATE ITERATIONS

The goal of Section 3 is to outline the importance of conducting an iterative model development-model validation process and to report on the three iterations conducted in an attempt to improve the validity of the baseline model. Each model augmentation will be validated using the same validation approach and measures that were defined and illustrated in Section 2.

Chapter 7 presents background literature pertaining to time management and proposes a linking framework termed the time management model that is comprised of three elements: Task Management, Time Estimation, and Workload.

Chapter 8 augments the baseline model with a task management component and provides a validation iteration.

Chapter 9 proposes a second model augmentation, which includes a time estimation model and verifies that the model behaves as expected in a simple, generic environment.

Chapter 10 presents the formal validation effort of the time estimation augmentation of the baseline model.

Chapter 11 presents the validation of the third iteration that combines the task management and the time estimation augmentations of the baseline model.

Section 3 ends with Chapter 12, a chapter that discusses the model validation results.
CHAPTER 7: IMPROVING TIME MANAGEMENT IN COMPLEX HUMAN PERFORMANCE MODELS

7.1. THE STATE OF TIME MANAGEMENT IN COMPLEX HUMAN PERFORMANCE MODELS

The ability to accurately model critical events in a reliable manner is highly dependent on the sensitivity of the HPM software to time management, a factor that has largely been ignored by the HPM community. Hart (1991) points out that one of the primary assumptions made by operator behaviour modellers is that the human operator is a passive translator of task demands into performance output and that output is immediate, perfect and consistent. HPMs have traditionally attempted to deal with time management by forcing specific behaviours to occur in a time-scripted manner, irrespective of the environmental demands and resource availability of the simulated human operator. Characterizing behaviours in this overly simplistic fashion removes the fluidity that characterizes natural human behaviour. However, behaviours that occur in today’s military, aviation, process control, and other highly complex environments often contradict such overly simplistic operator behaviour models, in that real world human behaviour rarely occurs in such a clean and predictable fashion. In reality, humans must actively manage their time, energy and resources to carry out tasks on time, while concurrently maintaining an appropriate level of workload.

The baseline model (presented in Chapter 5) was an inadequate representation of human performance (as shown in the validation effort in Chapter 6) because it used an overly simplistic first-in, first-out (FIFO)-like approach to modelling task completion, without including elements of task management and time estimation, and without considering the effect of workload on both. The FIFO approach is nevertheless a common modelling approach, and arguably reflects the state-of-the-art of HPMs today, because no model exists that links together the existing literature and theories on the manner that humans estimate and manage time as a function of workload.

Therefore, it may be possible to improve both the current state-of-the-art model and the validity of the baseline model by defining and implementing a model that reflects the impact of workload on time estimation and time management. This will be done using a series of iterative develop-validate cycles so as to examine the impact of each change on the model performance in a structured and systematic manner using the validation approach defined in Section 2. The
following section outlines the literature on time management (i.e. task management and time estimation) and concludes with the introduction of a time management framework.

7.2. BACKGROUND LITERATURE ON TIME MANAGEMENT

7.2.1. Task Management Literature

Understanding the manner in which human operators engage in complex behaviours requires insight into strategic human behaviour, encompassing among other things task management, task scheduling, and human time estimation in response to dynamically changing environments. The term 'strategic behaviour', as compared to simpler tactical responses, implies behaviour of considerable complexity, involving recognition of global/strategic plans (Wickens et al., 1997). The contrast between global and tactical planning is that a strategic planner is strategizing, or taking a long range, overall view of the situation in deciding what to do and when to do it, whereas a tactical planner is concerned with moment to moment actions.

7.2.1.1. Task Management and Information Processing

Task management, a significant factor in determining behavioural engagement, involves task scheduling, and an effective model of task scheduling requires understanding the manner in which humans deal with environmental demands and shift their tactics in response to changes in their environment. In other words, task scheduling involves an information-processing (IP) notion of task prioritizing, task timing, and task shedding. These aspects of performance subsequently impact the task quality (output).

Research by Adams, Tenney and Pew (1991) suggested that humans were able to respond appropriately to task prioritizing, but they found that scheduling performance may break down under time pressure and other stressors. They suggested that human operators have a tendency to notice all information, but to process information related only to ‘currently-attended-to’ tasks. They also suggested that human operators were able to shift their attention to other requirements, based on context. In other words, even if information was irrelevant, it nevertheless interrupted the ongoing task to determine which queued task the information pertains to. The processing of such information tended to be neglected in most models, because no schema existed for such information entering into the cognitive store. Adams et al. also suggested that task management is an IP undertaking, which becomes most stressful when IP loads rise (requiring more tasks to
manage). Their research thus suggested that the number of tasks remaining to be completed should have an impact on the human operator’s ability to estimate time.

Hendy, Liao, and Milgram (1997) proposed an integration effort combining a number of IP models of human performance to advance the IP paradigm of human operator performance. This integration effort culminated in a dynamic IP model that is under the active control of the human operator. Their model suggested that time pressure drives a process of re-evaluating and re-investing effort, which caused the IP system to adapt by decreasing the difficulty associated with task loads. Humans adapt to excessive processing loads by changing their processing strategy so as to reduce the amount of information or increase the available time for action. This effortful process is triggered by the perceived time pressure, which in turn is driven by the motivation to achieve a certain level of performance. Performance will degrade if no effort is expended to deal with the excessive load. Time pressure will be reduced through the effortful adaptation process. Hendy, Liao, and Milgram’s research is relevant to modelling time behaviour in an integrated fashion. Additional information on IP models can be found in Appendix I.

7.2.1.2. Task Management and Control Modes

Hollnagel (1993, 1998) presented a continuum of control modes, ranging from no control at all to completely deterministic control, as illustrated on the left hand side of Figure 21. This continuum of control modes is a part of a larger contextual control model to be discussed in Section 7.3. Within this continuum, Hollnagel identified four classes of control modes, ranging from the no control end to the deterministic control end of the continuum. These are referred to as the scrambled, opportunistic, tactical, and strategic modes of operation. A summary of each mode is given to its right in Figure 21, with a list of hypothetical control parameters provided along the right side. At one end of the continuum (scrambled), the human operator has no time to plan an action, while at the other end of the continuum (strategic) the operator has a large amount of time to plan an action to an external event.
Hollnagel described *scrambled* control as being characterized by panic reactions to environmental triggers, with the choice of the next action being random. The panic reactions from the scrambled control structure can lead to *opportunistic* mode of operation, where the salient features of the environment drive the next action. This is characterized by limited planning and has been represented as chance perceptual events driven largely by the external environment. This control mode can also feed back into itself to influence human performance.

The human operator may also engage in *tactical* behaviours. Tactical performance occurs as time increases from the opportunistic control behaviours. This is characterized as a control strategy where the performer plans based on events in the present situation. This level has individual micro-oriented goal structures and methods to attain these goals operating in accordance with a known set of rules. When performance at this level fails, the individual is brought back to opportunistic control.

When there is more time available, there is greater opportunity for the human operator to plan a strategic response to the situation. This *strategic* control level is one where the operator plans at a macro level and considers system task goals based on individual response. This theory suggests that increasing the time associated with a response to an environmental trigger is
associated with an increased cognitive functioning within the human performer and that higher levels of cognition are associated with system level considerations.

Consistent with Reason (1990), Hollnagel proposed that the actions that were carried out by the human can fail to achieve their goal in carrying out an appropriate plan as a result of either accurate performance according to an inadequate plan (cognitive planning errors, or mistakes) or deficient performance (physical errors, or slips). He further argued that research surrounding human error often confuses the causes of the events surrounding human error with the internal psychological processes or cognitive mechanisms that are presumed to explain the action. In addressing this, Hollnagel outlined the inter-relationship among human internal cognitive mechanisms and control levels. The dynamics of these mechanisms demonstrated the impact that context has on the performance of the individual in the environment rather than by an inherent relation between actions.

7.2.2. Time Estimation Literature

Hart et al. (1984) proposed that one of the complexities associated with human behaviour representations involves the concept of dynamic environments and ‘windows of opportunity’ for behavioural input. They pointed out that complex dynamic environments typically involve a trade-off between missing a window of opportunity, due to indecisiveness, and making premature decisions. Accurately determining the beginning, end, duration, and transition of the window of opportunity involves correctly estimating time. Estimating time is critical for accurately scheduling required behaviours in complex environments. The impact of incorrectly estimating available time can affect task prioritization, task timing, and task shedding.

The available research on Time available (TA) to complete the task, the Time Required (TR) to complete the task, and the time to task onset (TO) are discussed below, following a review of research methods used to evaluate time estimation.11,12

7.2.3. Time Estimation Research Methods

Much of the research surrounding time estimation utilizes either prospective methods or retrospective methods. A prospective experimental model is one in which the subject is forewarned that judgments of time will be asked (experienced duration), while retrospective

---

11 Additional research considered on TA, TR and TO can be found in Appendix G.
12 The current thesis considers TA, which interacts with TO, TR as an element in time estimation but is less relevant to ATCo operations.
methods are those for which such judgments (remembered duration) are required after the relevant period, usually without any prior warning (Zakay & Block, 2004). It is asserted that ATCo’s performance is akin to a prospective reproduction task, in that ATCos are aware that they will be required to produce an estimate of the amount of time that has passed. That is, as ATCos see the datablock begin to flash, they go on to complete other tasks, when they return to the flashing data block, they then estimate whether it has been 60 seconds since it started flashing. The estimates of time available for future tasks are based on previous experience with similar tasks (using the past to guide the future task schedules).

Prospective studies (e.g. Block & Zakay, 1997; Hicks et al., 1976; Hicks, Miller, Gaes, & Bierman, 1977; Jones & Boltz, 1989; Zakay, 1993) indicate that subjective duration is typically inversely correlated with the number of stimuli processed during a time period. This means that, when warned in advance that one is going to have to estimate a time duration, one is more likely to underestimate the length of that interval if more events occur during that interval – in other words, time passes more quickly when the individual is busy.

Significant research exists that examines the effects on attention of engaging in prospective duration estimation (Michon & Jackson, 1984; Zakay, 1990; Zakay & Block, 1996, 1995; Block, Zakay & Tsal, 1997). In prospective timing tasks, attention is allocated to time estimation on an ongoing basis, due to the fact that subjects are told beforehand that their task will involve time estimation (Zakay, 1990).

Retrospective models are usually confined to memory processes (Zakay, Block, & Tsal, 1999), with retrospective estimates tending to lengthen as the amount of non-temporal information increases (Block, 1974; Ornstein, 1969; Vohs & Schmeil, 2003) leading to overestimates of time. Zakay’s model hypothesized that increasing concurrent attentional load would lengthen produced duration and shorten reproduced duration. When attention is distracted from temporal IP, prospective duration estimation should be perceived as shorter than when attention is allocated to the estimation. The impression is that, once again, time passes by

\[13\] An important distinction between produced duration and reproduced duration tasks is that with the produced duration method, a person attempts to delimit an objectively measured duration corresponding to a verbally stated time period (e.g., "Say stop after what seems like 60s to you"), while in a reproduced duration experiment, participants are asked to reproduce intervals after having been presented with stimuli (Zakay, 1993).
more quickly as workload increases.

7.2.3.1. Human Estimation of Time Available ($T_A$)

Time available, $T_A$ refers to the time remaining in the window to perform a task. Under prospective conditions of time judgment, attention plays a large role in estimating $T_A$. Assuming that attention is a limited capacity system (Gopher & Donchin, 1986; Wickens, 1992), if two tasks are processed concurrently, fewer attentional resources are available to service each task individually. Consequently, if less attention is available for estimating the time available of a particular task, the time available will be perceived as being longer. This occurs because time is perceived as passing faster. Consequently, estimated durations are shorter and more variable for more demanding tasks where subjects are required to process multiple sources of temporal information concurrently (Brown & West, 1990).

An increase in attentively demanding tasks increases the likelihood that less time available within the window of opportunity will exist to complete the sequence of required tasks. The dynamic action of recalculating the estimated time that has passed based on the currently experienced workload influences the estimated future available time. For example, if a person (correctly) estimates that 10s has passed in a 60s window, he/she would estimate (correctly) that the time available is 50s. If, however, he/she underestimates the passage of 10s (i.e. estimate that 5s has passed), he/she will overestimate the time available and be under the impression that time available is 55s when it is actually 50s. Indeed, this was found to be the case by a number of researchers studying prospective conditions of time judgement (e.g., Block & Zakay, 1997; Brown, Stubbs & West, 1992; Hicks, Miller, Gaes & Bierman, 1977; Hicks, Miller, & Killbourne, 1976; Zakay, 1993). This did not typically occur during retrospective conditions, however, when the human operator did not have a priori knowledge of needing to maintain awareness of the passage of time.

Research by Brown and Boltz (2002) suggested that there is a differential degradation factor that impacts estimates of time available, depending on the length of time being estimated. Brown and Boltz studied two methods of producing time duration, by comparing judgements of duration of prose passages under baseline, low and high workload conditions. Workload was manipulated through inclusion of distraction items in the prose. The stimuli were 36 prose passages recorded onto 3 cassette tapes, with 12 passages per tape. Each tape represented one mental workload condition. In all cases they were asked to pay attention to time passage, as well
as to detect different target stimuli. The baseline condition possessed no distraction items. The low workload condition required subjects to detect one target letter in the passage, those words beginning with the letter M. The high workload condition required that the subjects detect two types of targets, all words beginning with either the letter D or the letter L. The passage durations fell into four categories: 15–20s, 25–30s, 35–40s, and 45–50s.

Brown and Boltz presented their results in the form of two types of time estimates: absolute error and constant error. *Absolute error* indicated the absolute size of the time estimation errors but not the direction of the error, whereas *constant error* indicated the size and the direction of the error. In all cases, the ratio of average estimated to average actual times yielded values less than 1. Furthermore, their results suggest that subjects underestimate $T_A$ and that the time judgement error increased for longer task durations. They also noted a strong linear component to this relationship\(^\text{14}\) Their explanation was that competing non-temporal tasks (the letter detection task) reduce the amount of temporal information processed and stored during an interval, thus shortening a person’s perceived duration and creating the impression that proportionately less time has passed.

It is important to highlight that these findings are extendable to other competing modalities as well. For example, Boltz (2005) found that there was no significant difference in stimulus duration estimated when competing tasks involved the visual domain as compared to the auditory domain. Boltz suggested that this further supports the idea that the same underlying cognitive mechanisms mediate duration judgements regardless of event modality.

In an ATC environment, Boudes and Cellier’s (2000) research on human accuracy in time estimation also suggested that ATCos have a tendency to underestimate the time available to complete procedures, resulting in activities completed ahead of schedule. The behaviour pattern that emerged from their data suggested that humans tend to behave according to a ‘conservative bias’ (Edwards, 1982; Boudes & Cellier, 2000) in scheduling activities when performing in a strategic mode in a complex operational environment, due to the negative consequences that are associated with incorrectly projecting aircraft position and state.

\(^{14}\) Vierordt’s (1868) Law predicts a power function relationship between time judgements and task duration, not a linear relationship as concluded by Brown and Boltz.
7.2.3.2. **Human Estimation of Time Required (T_R)**

A second aspect of estimating time involves one's ability to estimate T_R, the *time required* for completing the execution of a future task, given the environmental context and the memory of previous experience with the specific projected task or task sequence. Estimating T_R is different from estimating T_A because T_A is a function of all the other activities being carried out within a particular context, whereas T_R is a function of a particular task in question. Given that T_R is less relevant for ATCo performance because the duration of the actions is so short, the concept of T_R is raised here but defined more comprehensively in Appendix K.

7.2.3.3. **Human Estimation of Time to Task Onset (T_O)**

The third component involved in human time estimation involves estimating the time to task onset (T_O), where this factor refers to the human’s estimation of when they will need to start the task. As with incorrectly estimating T_A and T_R, incorrectly estimating T_O affects the human operator's ability to undertake and prioritize tasks, schedule tasks, and drop/shed tasks. It also affects the quality of task performance. In general, humans are flawed in their ability to estimate the task T_O to fulfil the demands required by the operational environment. Research from the attention domain provides evidence for the existence of the various attentional processes that are operational when engaging in T_O estimation (Block, Zakay & Tsal, 1997; Buehler, Griffith & Ross, 1994; Dawes, 1988; Griffith, Dunning & Ross, 1990; Johnson & Sherman, Kahneman & Tversky, 1979; Kahneman & Tversky, 1982; Michon & Jackson, 1984; 1990; Zakay, 1990; Zakay & Block, 1995, 1996).

Hahler et al. (1991) developed a behaviour model to represent the manner in which human operators complete and manage a number of goals when engaging in complex military operations. These goals are highly impacted by task prioritization, operator workload and time pressure. Hahler et al.’s model utilizes workload as an intervening variable for scheduling activities and for determining the time of task onset. Hahler’s research provides insights into the manner that workload interacts with the operator’s decision to begin a task (i.e. the task onset). Hahler defines an overloaded operator as an operator whose workload exceeds a redline value of 60 on a 100-point scale they used to measure workload (based on the WinCrew workload threshold identified by Little et al., 1993; Lockett, 1997). In Hahler’s research, an overloaded operator: (1) attempts to accomplish all tasks regardless of the overload condition, and therefore will begin all tasks as they enter into the task queue; (2) does not begin a new task that will lead
Chapter 7 – Developing Time Management Models

7.2.4. **Workload and Time Management Literature**

Workload has been proposed as being one of the more critical aspects of human behaviour that impacts time estimation and task management behaviours, and vice versa (Ariely & Zakay, 2001; Boltz, 2005; Brown & Boltz, 2002; Hancock & Warm, 1989; Hart, 1975; Hart et al., 1984; Hendy, et al., 1997; Lockett, Plocher & Dahl, 1990; Mitchell, 2000). When such behaviours were forced to occur at given times and within specific time windows, pressures on the ability of the human operator to allocate resources occur (Kahneman & Tversky, 1973; Wickens, 1992). Hart (1987) suggested that time pressure is one of the end results of incorrectly estimating or scheduling the available time for a particular task to be completed. Time pressure was positively correlated with time stress (Ariely & Zakay, 2001; Hancock & Warm, 1989; Hendy, et al., 1997; Lockett, Plocher & Dahl, 1990; Mitchell, 2000).

Hart (1989) argued further that, in the real world, human operators tended to schedule task execution to reduce task loads to a “comfortable” level. In her model (validated by Raby & Wickens, 1994), Hart (1991) claimed that the overall level of workload influences the behavioural patterns undertaken by the human operator in response to the environment, emerging in a range from operator underload, to moderate load, to high load, to overload. Performance was generally "best" in the middle or even high range of workload, while performance was worst when operators were behaving at underload or overload levels. This poor performance was due largely to operators' inability to accurately schedule requisite behaviours within the allotted time.

Hart (1991) also showed that operator performance was maximized when there was an effective task performance strategy, and that this was highly influenced by estimation of $T_A$, $T_R$ and $T_O$. In order to allow enough time to complete all tasks by an imposed deadline, operators needed to estimate when it is appropriate to commence the tasks, how long it would take to perform them, and then initiate and complete the tasks in the appropriate amount of time. Hart’s (1987) research indicated that skilled operators elected to defer tasks, or even shed tasks, when workload was too high. Conversely, well-trained operators elected to perform tasks ahead of schedule when workload levels were relatively low. Combining these models, operators must
consequently engage in dynamic decision making to adjust their estimates and schedules to compensate for the various $T_A$'s, $T_R$'s, and $T_O$'s, as a function of, among other things, interruptions, distractions or fatigue, as well as subsequent changes in $T_A$. Furthermore, operators typically require re-calibration following complex behavioural engagement, so as to re-align their schedules with their contextual performance. This suggests that workload is highly influenced by the windows of opportunity that present themselves for completing goals (Hart, Battiste & Lester, 1984), and thus that workload is highly influenced by the dynamic nature of task engagement (Hancock & Chignell, 1988). See Appendix L for more background on workload and time management.

Being able to correctly represent how humans estimate and quantify time is critical for HPMs being developed today. Workload not only affects performance in executing specific task behaviours, but also impacts time estimation, task management, and scheduling of both quantitative and qualitative human operator behaviours.

7.2.5. Summary of Literature on Time Management

In a general sense, the review of literature on time management presented in the present section has shown, in relation to the topic of modelling and validation of time critical systems, the importance of accurately representing the following notions when characterizing human task performance in complex environments:

- Task management
- Time management and information processing
- Time estimation, including the requirement to properly consider people’s estimates of time available, time required and estimates of time onset in determining task schedules as a function of the projected workload.

7.3. A FRAMEWORK FOR MODELLING TIME MANAGEMENT IN COMPLEX SYSTEMS

The above literature review has served to highlight the need for a framework to link the available literature on time management into a model of human performance that is generalisable to a variety of complex domains, in which responses by human operators are time sensitive. In this section, Hollnagel’s COCOM model is proposed as a systems-level model applicable to this need, and which can be used to guide the development of a Time Management framework.
Hollnagel’s (1993, 1998) COCOM model, introduced earlier in Section 7.2.1.2, has been proposed as a method to model complex systems. COCOM asserts that human performance is determined largely by situational factors, since all human performance can be characterized by the need for the human operator to “think” and respond “correctly” given the characteristics of a specific operational environment (Hollnagel & Woods, 2005). As such, the selection of possible actions is determined by the operator’s current needs and constraints. Cognitive functioning must interact with the physical characteristics of the human operator and the environmental characteristics of the world. The result of this interaction then feeds back into the interpretive cognitive world.

The acronym COCOM was based on three main concepts: constructs, control, and competence, referred to henceforth as 3C’s (Hollnagel, 1993; 1998; Hollnagel & Woods, 2005). Constructs referred to the knowledge that the human operator possessed about the situation in which actions take place. Since constructs are the operator’s perceptions of the world, even though these perceptions may differ from the actual situation, they are artificial in the sense of being constructions (or re-constructions) of salient aspects of the situation. They are therefore usually temporary. Control characteristics were defined as the orderliness of performance and the way in which actions were carried out. Finally, competence was defined as the set of possible actions or responses that a system (including the human operator) can apply to a situation according to the recognized needs and demands. The system cannot do something that either was not available as a possible action or which could not be constructed or aggregated from the available possible actions.

Combining these three concepts led to a cyclical model of human action within a system context, as illustrated in Figure 22, where the environment (i.e., external events) modifies the human operator’s understanding of the world (construct) that then directs (controls) the operator’s actions (competence). These actions then feed back into the environment to be cycled again by the human operator. Hollnagel and Woods suggest that, in contrast to most information-processing (IP) models, which are typically characterized by the three generally sequential stages of perception, decision, and action, the cyclical model describes system performance as a mixture of feedback (error controlled) and feed-forward (cause controlled) activities (Hollnagel & Woods, 2005). This is an appropriate characterization of a joint cognitive system but appears limited as it ignores such human IP elements as the cognitive means by which the human
operator makes sense of the information in the world (that is, how the human operator perceives the constructs), and applies this (selects a control mode) to carry out his/her tasks. In most complex systems, such *inner-loop* IP functions are embedded within higher, *outer-loop* system level functions. In Section 7.4, the manner in which IP models related to time management can be interleaved within the COCOM cyclical model, is presented.

![COCOM cyclical model of human action within a system context based on the 3C’s: Constructs, Control and Competence.](image)

COCOM and the 3’Cs notion provide a framework with which to guide the development of a model of time management. As shown in Figure 23, Hollnagel and Woods’ system-level notions of constructs, control, and competence are each supported by inner-loop IP functions related to time management (shown in the inner loop). Specifically, a critical element of understanding the environment is estimating how much time is available and required to complete a task. Second, Hollnagel and Woods’ notion of control refers to the manner in which the operator’s actions are carried out. The IP component that supports this task is how the operator manages time in the face of multiple, simultaneous tasks. Finally, Hollnagel and Woods’ notion of competence refers to the operator’s actions (manual and cognitive responses) and these have a resulting impact on operator workload.
This inner loop is thus proposed as a linkage between IP models and the COCOM model. It is an approach that uses the environment to trigger the onset of common procedural models and at the same time links to two IP models of human cognition that represent how human operators engage in time estimation and task management as a function of operator workload. These IP models are presented in the following sections.

7.4. DEVELOPMENT OF A LINKING FRAMEWORK OF TIME MANAGEMENT

The Time Management framework that follows was used to guide the development of two time management IP models (time estimation and task management), which are to be called whenever the human operator engages in a time sensitive task. As shown in Figure 24, the time management framework includes a workload projecting component, a time estimating component, and a task-managing component.

The Workload Projecting component dynamically updates the operator’s actual or experienced workload based on the operator’s understanding of the environment and upcoming tasks. The workload portion of the linking framework presented in Figure 24 is where an assessment of workload by the modelled operator is completed (for example, as per Equation 2).
and the tasks are scheduled. The resulting assessment of workload influences the second component, Time Estimating, which possesses three interacting elements: $T_A$, $T_R$, and $T_O$ for completing tasks. Time estimation is the controller of the action sequences. Since these estimates are necessary for the human operator to continuously estimate current workload, the time estimating component feeds back into the first block (workload)\(^{15}\).

The same estimates also feed forward into the third block, Task Managing, in which activities are managed or scheduled to achieve the high level goals. This TM block serves to impact the rules that guide when a task enters into the schedule, when and what order the task releases from the task queue, which will be described in more detail in Section 7.4.1.

Collectively, this structure represents an approach to computationally representing the manner in which humans perceive their current workload, how this affects their estimates of time, and how these time estimates are used to schedule their required actions. The following, more detailed discussion of the mechanisms within the time management framework follows the structure of Figure 24, progressing from right to left, beginning with a discussion of task managing, and moving into the time estimating component. The workload-projecting component has been discussed in Section 7.2.4.

### 7.4.1. Task Managing

The Task Managing model at the right in Figure 24 corresponds to the inner loop IP mapping of control modes in Hollnagel’s COCOM model (see Figure 23), where changes in performance are

---

\(^{15}\) While time estimation feeds into workload in this conceptual model, it is not instantiated in the models presented in this research effort
affected by manipulating the operator’s control mode. Suitable action sequences are more likely to occur if the model has an appropriate characterization of the number of competing items that can be in the task management queue.

Task management, in turn, involves task scheduling by calling for the task that is highest in the queue in terms of priority and entry into the queue to be completed first. Two key model parameters that can be used to influence a model operating with the COCOM framework are the subjective estimation of available time and the number of simultaneous goals (included in the list of control parameters at the far right of Figure 21). These two parameters can be used to modify the manner in which a model schedules its performance. For example, in a task network model, the “number of simultaneous goals” parameter can be modelled as a variable termed “number of tasks in the queue”. When the queue gets too long, a task schedule gets invoked. The number of tasks in the queue is used as a variable to impact the “estimate of available time”, which can then be used to impact the number of and therefore the management of tasks entering and exiting the queue. The development, implementation and validation of the TM model are presented in Chapter 8.

7.4.2. Time Estimating

Referring to Figure 23, the Time Estimating structure in the centre of Figure 24 corresponds to the inner loop IP model mapping of constructs (artificial reconstructions of the salient aspects of a situation) in Hollnagel’s COCOM model. To perform the time estimation augmentation, the Time Estimating structure in Figure 24 accounts for the degradations of human estimates of time passage as a function of work load, which subsequently impacts the operator’s ability to properly schedule and sequence activities.16

The current focus is on modelling functions of upcoming tasks: estimating the \( T_A \) to complete an upcoming task. (The Time Estimating (TE) model is developed in detail and verified in Chapter 9, and then validated in Chapter 10 using the comprehensive validation approach developed in Section 2.)

The particular interest for this model is to determine if and how the estimates of mental workload link to \( T_A \), strategic task management, and the interaction among all of these factors, as

---

16 While the time estimating box is illustrated as feeding back into workload, the current instantiation did not have time estimates causing the model to modify a workload value.
operationalised by task prioritization, task timing, and task shedding. As suggested in the Time Estimating Box in the inner loop of Figure 24, the model accounted for degradations of human estimates of time passage as a function of workload. This model is a closed-loop model, in which under-estimation of time passage increases as workload increases and subsequently impacts the operator’s ability to properly schedule and sequence activities. The time estimating structure becomes involved whenever time pressure exists. The structures embedded within the time estimating model are discussed next.

Time pressure is caused by time stress in completing a series of required tasks; however the nature of the time pressures experienced will be different, depending on whether the human operator underestimates or overestimates $T_A$, before the particular task has commenced. As modelled in Figure 25, two hypothetical curves representing time available misestimates (either over- or under-estimates) prior to commencing an activity are presented. Each activity has a time threshold before which it must have been started for successful completion within the window of opportunity and after which it will be perceived as being late. If $T_A$ has been underestimated at or before the beginning of an action, then actions may be completed ahead of the nominal action’s time required for the action to take place, as in reality there turns out to be more time available to complete the goal than previously estimated by the operator. Human operators operating in this mode thus generally complete activities early. In such cases, the human operator is likely to commence the particular task with a relatively high level of perceived time pressure, which then subsides with the realisation that the time available has been underestimated.

On the other hand, if $T_A$ is overestimated prior to commencing the task, the human operator may think that he has more time than he really does, and thus does not experience a high time pressure at the beginning. As time approaches the deadline for task completion, however, the human operator, through recalibration, revises his estimate of $T_A$ to be more in line with actual $T_A$, causing time pressure to mount. The human operator will thus be required to expend energy at a continually increasing rate to complete the goal before the end of the window of opportunity for task execution (Hart, 1991).

---

17 Time stress is caused by the actual time constraints in completing a behaviour (Block, Zakay & Tsal, 1999). Time stress disrupts decision making specifically by causing people to use simple, nonlinear strategies that result in suboptimal decisions (Johnson, Payne & Bettman, 1993).
Figure 25. Relationship of Time Pressure to Elapsed Time, as a Function of Under- or Over-Estimation of Time Available prior to Beginning an Activity.

Analogous to incorrectly estimating $T_A$, incorrectly estimating $T_R$ significantly affects the human’s ability to begin subsequent tasks, prioritize and schedule future tasks, and even drop/shed tasks. It also can affect task performance quality. This is illustrated in Figure 26 for task k, where the $T_R$ for a task overlaps and exceeds the $T_A$ prediction, which is likely to result in delayed onset of subsequent tasks. Analogous to incorrectly estimating $T_A$, incorrectly estimating $T_R$ significantly affects the human’s ability to begin subsequent tasks, prioritize and schedule future tasks, and even drop/shed tasks. It also can affect task performance quality. This is illustrated in Figure 28 for task k, where the $T_R$ for a task overlaps and exceeds the $T_A$ prediction, which is likely to result in delayed onset of subsequent tasks. $T_R$ estimation is important in situations involving complex decisions where the time remaining can be quite lengthy and thus can impact the task schedule, however it is less relevant when the task is an instantaneous response to a specific situation, involving little decision making time. For example, the time it takes an ATCo to press a button is on the order of milliseconds and does not impact the task schedule because of the immediate nature of the action. That having been stated, it is assumed, for the purposes of the current model application, that $T_R$ is estimated perfectly.

$T_O$ estimation is impacted by the number of ongoing activities and the manner that the activities are being scheduled, which in turn impacts the workload experienced, and, as indicated in the closed-loop of Figure 24, subsequently affects further estimation performance.
7.4.2.1. Relationship among $T_A$, $T_R$, and $T_O$ in the Time Estimating Model

Figure 26 outlines the relationships that exist among $T_A$, $T_R$ and $T_O$. The values of $T_A$, $T_R$ and $T_O$ are functions of time (i.e. they are continually updated/revised during task execution). As represented in the figure, the operator in the general case possesses some ongoing task that was commenced at some time in the past ‘-$T$’. At the time of, or prior to, commencing that previous (ongoing) task, the operator estimated the time available ($T_A(t=-T)$) and time required ($T_R(t=-T)$) for that task. The estimated time values are indicated by the dashed arrows in the figure extending from the past, through the present and into the future, towards the end of the window of opportunity for the ongoing task.

The operator is at the current time/now ($t=0$) when he revises his estimation of $T_A'$ and $T_R'$ for the ongoing task. He is also estimating $T_O$, $T_A$ and $T_R$ for task 1 to n (with the general case denoted as task k) to factor in the revised time available of the ongoing task, given the resource loading of the ongoing activities (depending on the number of sequential activities required by the higher procedure). In the conceptual model in Figure 24, this time revision occurs all the way throughout execution of the task, but the current instantiation of the baseline model does not allow the implementation of such dynamic updating.

The time-to-onset for each task is denoted in Figure 26 by the symbol $T_O$. As suggested by the name, this parameter represents the human operator’s continuously revised estimate of the time available between the current moment ($t=0$) and the scheduled time for commencement of each remaining goal related task in the current queue, and refers to the estimate of when to start the task. In the figure, for example, we see that Task 1 is scheduled to commence at a time $T^{1O}_O$ seconds from now, at some projected time after completion of the current task. Task k, on the other hand, is projected to commence exactly upon completion of execution of Task 1, at a time $T^{kO}_O$ seconds from now. For the general case of task k, the operator estimates the $T^k_A$ and the $T^k_R$ to complete the task. In this case, we see a situation where there is projected to be inadequate time to complete task k, as the estimated time required exceeds the estimated time available. At the beginning of execution of task k, this will cause a dynamic recalculation of the time to onset for the n$^{th}$ task, which is now set to commence at the conclusion of the k$^{th}$ task. This dynamic action during time estimation demonstrates the concept of updating the time model within a modelled operator. Aspects of this relationship can change depending on the context and the
workload in the environment, as represented in the various lengths and overlap of the arrows in the figure.

![Diagram of Time Management Models](image)

Figure 26. Conceptual Representation of Key Time Estimation Factors.

The dynamic action of recalculating the $T_A$ and the $T_R$ based on the projected and experienced workloads requires successful time management by the operator (thereby emphasising the relationship between workload and time estimation of future tasks). The $T_A$ and the $T_R$ estimates, therefore, act as constraints on task management. The interaction between the $T_A$ and the $T_R$ substantiates the existence of a time estimation and management model for predicting onset, required and available time to perform behaviours.

### 7.4.2.2. Impact of $T_A$, $T_R$, and $T_O$ on Task Performance

The relationship among $T_A$, $T_R$, and $T_O$ and the anticipated effect of these estimates on task performance are summarized in Table 8. The various factors in the table are not independent. If an operator overestimates the time available, $T_A$, to complete a task, then it is likely that their $T_O$ estimate for subsequent tasks will be overestimated and will be late and the operator will be left with less time than anticipated to complete the task. Table 8 illustrates the
anticipated effect of the mis-estimation of time available, required and time to onset and the manner that these concepts are not independent. This will result in their beginning to shed tasks, which will then possibly be rescheduled, or not get done. Consistent with this, if the operator underestimates the time required, \( T_R \), then it is likely that their \( T_O \) estimate will be overestimated, causing a late task start time and s/he may have to invest increased effort to get the task done within the window of opportunity. In sum, the operator will find as she approaches the end of the task, that she still has some time left over, due to the \( T_O \) overestimate, the \( T_A \) overestimate and the \( T_R \) underestimate (Ariely & Zakay, 2001; Hancock & Warm, Hendy et al., 1997; 1989; Lockett et al., 1990; Mitchell, 2000; Rantanen, & Levinthal, 2005; Rantanen, Levinthal, & Yeakel, 2005). When \( T_R \) underestimation occurs, the operator is increasingly likely to shed tasks in order to complete some tasks within the window of opportunity, and unfulfilled tasks will build up until resources become available to service them. \( T_R \) underestimation is further heightened whenever such behaviour occurs in the overload condition (Brown & West, 1990).

### Table 8. Relationship among \( T_A \), \( T_R \) and \( T_O \)

<table>
<thead>
<tr>
<th>( T_O ) Estimate</th>
<th>( T_A ) Estimate</th>
<th>( T_R ) Estimate</th>
<th>Task Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under (early)</td>
<td>X</td>
<td>Over</td>
<td>X</td>
</tr>
<tr>
<td>Over (late)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

#### 7.4.3. Illustration of Model Interactions within the Time Management Framework

The integrated manner in which this dynamic workload structure interacts with the time estimating and the task managing component is illustrated in Table 9, which encompasses the relationship within the time management framework among the three levels of workload, shown vertically: overload, suitable (well calibrated), and underload, and the predicted performance of the human operator along the top. The latter is reflected by means of the human operator's time estimation, the effect of this on task management, the effect on the model’s behaviour, the optimal strategy in response to the situation outlined, and the effort that will be required of the operator in the specific workload level. This conceptualization is an initial linkage between the three structures of the time management framework, along with their respective effects on the HPM’s performance.

Table 9 was generated by integrating the findings from the literature reviewed and by extending existing workload-performance models.
### Table 9. Predicted (Developed Through Literature Review) Interactions Between Levels of Workload and Resulting Conception of Time, Effect on Task, and Effect on Model Behaviour.

<table>
<thead>
<tr>
<th>Workload (Competence)</th>
<th>Time Estimation (Context)</th>
<th>Task Management (Control)</th>
<th>Effect on Model Behaviour</th>
<th>Optimal Strategy*, ****</th>
<th>Effort*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underload **</td>
<td>Underestimate $T_A$</td>
<td>Overestimate $T_R$</td>
<td>Shedding – error</td>
<td>Time stress activated, Strategic behaviour/memory activation*, procedures shift in performance critical events carried out first, little regard to performance quality (no double checks performed)</td>
<td>Change: Reduce amount of information to be processed OR increase the time available before an action has to commence.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overestimate $T_A$, underestimate $T_R$</td>
<td>Perform task late Poor quality</td>
<td>No effect - correct performance</td>
<td>No effect, carry out performance in calibrated functioning mode</td>
</tr>
<tr>
<td>Suitable (well calibrated)</td>
<td>Appropriate estimate of Time</td>
<td>Perform ahead of schedule Good quality when tasks are completed</td>
<td>No time stress, procedures carried out as per initial specification, high regard for performance quality (double checks/cross checks always performed), high missed signal rate</td>
<td>Increase the amount of information to be processed OR reduce the time available before an action has to commence.</td>
<td>Low until request for increased information is provided OR time is reduced to complete goal behaviour</td>
</tr>
</tbody>
</table>

Notes: outlines the relationship that exists among the 3Cs and the time management framework. Task management maps onto COCOM’s notion of control, time estimation maps onto COCOM’s notion of context, and workload maps onto COCOM’s notion of constructs.

Notes * From Hendy et al. (1997); ** All underload algorithms and representations have been developed in this effort by Gore based on Hendy et al., (1997) and Hancock and Chignell (1988); *** Zakay (1990); **** Design-to-time Real Time Scheduling (Garvey & Lesser, 1993) algorithm.
CHAPTER 8: VERIFYING AND VALIDATING THE TASK MANAGEMENT (TM) MODEL

8.1. AUGMENTATION OF BASELINE TO “OPPORTUNISTIC-ONLY” MODE

The baseline model was augmented with a TM manipulation to account for the change from ‘strategic’ to ‘opportunistic’ control (Hollnagel, 1993, 1998) that occurs in high-workload tasks, as indicated in the literature reviewed in Chapter 7. In comparison with Figure 24, the right-most box has been augmented to include this opportunistic mode of operation, which served to impact the task ordering and time to task onset (see Figure 27). The TM model utilized the ‘conservative bias’ paradigm outlined by Edwards (1982) and Boudes and Cellier (2000) where the human operators are expected to complete all tasks in the order in which they are encountered. Conservative biases shift performance times towards “early” responses within the window of opportunity for certain high priority tasks.

The TM model used the queue length as the primary mechanism to modify the task sequence. The TM manipulation required that the number of tasks in the queue before which non-critical tasks are shed be reduced from five (in the baseline model) to zero (in the TM model). Queue length of zero was selected as the starting point for the analysis as it is the extreme value that could be used to generate opportunistic performance and could be used as a benchmark to indicate whether additional sensitivity analyses using higher load levels were required. With the-model variable ‘QueueSizeIn’ set to zero, the TM model scheduled the tasks immediately. Since, this new model did not allow multiple tasks in the queue, the operator was
Chapter 8 – Validation of Baseline Model with Task Management

forced to manage tasks *opportunistically*, rather than strategically. That is, planning was limited and only the environment drove decisions.

### 8.2. Validation of the Task Management (TM) Model

The same five measures presented in Section 6.2 to examine the performance of the baseline model were used to evaluate the performance of the TM manipulation relative to the FEWS simulation output. The treatment of the data was consistent with the explanations provided in Section 6.2.1.

#### 8.2.1. Measure 1: TM Model Workload

Figure 28 illustrates the average workload per experimental condition for (a) the FEWS adjusted data, (b) the baseline model and (c) the TM augmentation model.

**8.2.1.1. Mean Workload for TM Model**

![Figure 28](image)

Figure 28. Mean (a) FEWS Adjusted, (b) Baseline, and (c) TM Model Workload Predictions by Traffic.

There was a significant 2x3 interaction $F(2,40)=44.84, p<.001$. As with the baseline model, Table 10 shows that workload increased significantly in the FEWS ($F(2,40)=57.57, p<.001$) and the TM augmentation ($F(2,40)=7.07, p<.001$) data when the traffic level is increased in all conditions except in the DL – No DL condition of the 133% traffic case for the FEWS adjusted data. This step illustrates that the model represents the FEWS adjusted data source and that the environment correctly drove the operator’s workload values. An interesting finding was that no significant differences existed for the FEWS in the 133 DL to 133 No DL condition while the model did possess a significant difference. A fine-grained analysis was completed to explore the possible reasons for this difference.
Table 10. Significance Tests for FEWS Adjusted, BL and TM Model Workload by Traffic Level.

<table>
<thead>
<tr>
<th></th>
<th>100 DL-133DL</th>
<th>100 DL-133 No DL</th>
<th>133 DL-133 No DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEWS (Adjusted)</td>
<td>t(21) = 8.14***</td>
<td>t(21) = 6.85***</td>
<td>t(21) = 1</td>
</tr>
<tr>
<td>Baseline</td>
<td>t(829) = 4.36***</td>
<td>t(829) = 8.46***</td>
<td>t(829) = 3.3***</td>
</tr>
<tr>
<td>TM Augmentation</td>
<td>t(829) = 4.32***</td>
<td>t(829) = 25.22***</td>
<td>t(829) = 23.50***</td>
</tr>
</tbody>
</table>

*p < .05, **p < .001, ***p < .001

8.2.1.2. **TM Model Workload Trend Output**

Figure 29 presents the individual workload ratings from the FEWS data (upper graphs) and the model predictions (lower graphs) for each traffic load condition. It can be seen that the FEWS and the model produced similar workload trends in the 100 DL (Figure 29a) and 133 No DL (Figure 29b) conditions. Unfortunately, the 133 No DL condition in the FEWS simulation possessed a significant number of missing values, which makes drawing conclusions about the comparison of FEWS and model trends difficult in that condition.

![Figure 29](image)

(a) (b) (c)

Figure 29. FEWS (top) compared to TM Model (bottom) Workload Output for (a) 100 DL Condition, (b) 133 DL Condition, (c) 133 No DL Condition.

Notes - FEWS possessed missing responses. There were 18 responses for (a), 15 for (b), 9 for (c).

8.2.2. **Measure 2: Queue Length for TM Model**

The TM augmentation directly manipulated the task queue by setting the queue length for critical tasks to zero. This determined how many tasks the modelled ATCo noticed before the task of ‘listen to pilot readbacks’ were shed. During the interim period of time that was required to process the ongoing task, new tasks entered the queue. This along with any non-critical tasks
resulted in the possibility that the queue length could exceed one in both the maximum and average queue length measures. As outlined in Section 6.2.2, queue length is technically a verification measure because there is no corresponding FEWS data; however it is included in the set of validation measures because it is used to support the validation process.

It can be seen in Figure 30 that the maximum task queue increased as the traffic load increased from 100% to 133%. Furthermore, there was a large increase in the maximum number of queued tasks in the 133% No DL condition, which suggests that the model was queuing more tasks because the ATCo’s did not complete DL handoffs and clearances. The maximum number that was in the queue is however, not as high as in the baseline model (See Figure 30), which means that the TM model impacted the model by clearing some tasks out of the queue earlier than in the baseline condition.

Figure 31 presents the mean queue length data from the TM model and shows that the traffic level increased the number of items placed in the task queue on according to the increasing level of traffic. The queue length increased significantly when the traffic level increased from 100% DL to both 133% DL (t(829) = 8.29, p < .001) and to 133% No DL (t(829) = 32.64, p < .001). Of particular interest is that the queue length increased significantly in the 133% traffic condition with No DL as compared to the 133% DL condition, t(829) = 30.76, p<.001.

8.2.3. Measure 3: RHD for TM Model

Figure 32 illustrates the data associated with the RHD produced from the FEWS simulation and those RHD predictions generated by the TM model as a function of the three experimental conditions of 100% DL, 133% DL and 133% No DL.
8.2.3.1. **RHD t-tests and Results for FEWS versus TM Model**

Figure 32 illustrates that no significant differences existed between the FEWS and the TM model both within the 100% condition (t(76)=0.41, p>.05) and within the 133% DL condition (t(99) = 1.21, p>.05). A significant difference existed in the 133% No DL condition (t(80) = 3.30, p<.05). These are consistent findings with the baseline model output presented in Section 6.2.3.

![Figure 32](image.png)

**Figure 32. Average RHD Time as a Function of Traffic Level for the FEWS, the Baseline Model, and the TM Model.**

As completed for the baseline model, chi-square analyses were also conducted to determine if the distributions of RHD times from the baseline model and the FEWS observations come from the same population. As with the baseline model, data were put into three bins of approximately equal size: 0 – 10s, 11-30s, and greater than 31s.

8.2.3.2. **RHD Chi-Square Test Results for FEWS versus TM Model**

The $\chi^2$ test conducted on the RHD times for all three traffic levels suggested that the frequency distributions from the baseline model and the FEWS observations, did not come from the same population (see Table 11). These results suggest that the TM model does not faithfully represent the human RHD data in any of the three load conditions. The baseline and TM models were from the same distribution for the 100 and 133 DL conditions while they were not from the same population for the 133 No DL condition.

**Table 11. Chi-square and Correlation Analyses for FEWS, BL, and TM Model by Traffic Levels**

<table>
<thead>
<tr>
<th></th>
<th>FEWS-Baseline Chi-square</th>
<th>r value</th>
<th>FEWS –TM Model Chi-square</th>
<th>r value</th>
<th>Baseline – TM Model Chi-square</th>
<th>r value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100DL</td>
<td>$\chi^2(2,N=122)=4.96***$</td>
<td>0.21</td>
<td>$\chi^2(2,N=122)=4.39***$</td>
<td>0.26</td>
<td>$\chi^2(2,N=154)=0.76$</td>
<td></td>
</tr>
<tr>
<td>133DL</td>
<td>$\chi^2(2,N=163)=3.69***$</td>
<td>-0.15</td>
<td>$\chi^2(2,N=163)=3.27***$</td>
<td>0.01</td>
<td>$\chi^2(2,N=226)=0.76$</td>
<td></td>
</tr>
<tr>
<td>133NDL</td>
<td>$\chi^2(2,N=154)=38.98***$</td>
<td>0.12</td>
<td>$\chi^2(2,N=154)=4.85***$</td>
<td>-0.06</td>
<td>$\chi^2(2,N=226)=42.09***$</td>
<td></td>
</tr>
</tbody>
</table>

* $p<.05$, ** $p <.001$, *** $p <.001$
In summary, the TM model does not adequately represent human data as collected in the FEWS simulation. Correlations were conducted to provide insights into the strength of the relationship identified by the $\chi^2$ tests and are included in Table 11.

A more detailed examination was therefore required to explore the reasons for the breakdown in the model’s performance. Whereas the TM modelled RHD times were significantly higher than the FEWS times in the 133 DL condition, the absolute difference between the means was relatively small, only about 6s, or about 20%. However, the TM model overestimates RHD in the 133 No DL condition by a much greater amount, with an absolute difference of 26s, or approximately 60%. The reasons for the greater break down in the model’s performance can only be captured by exploring the performance of the model in all of the 100 DL, 133 DL and 133 No DL conditions in greater detail. The histogram of the individual data points of the FEWS and the TM augmentation model RHD is provided next.

8.2.3.3. **RHD Histograms of FEWS and TM Model**

Figure 33, Figure 34, and Figure 35 presents RHD times for each of the three traffic conditions for both (a) the FEWS data and (b) the TM augmentation model. It should be noted that the figures on the left side of Figure 34 (FEWS data) are the same as for the left side of Figure 14. As can be seen, the large majority of RHD times occurred early in the handoff window of opportunity, before 60s elapsed, with the greatest proportion occurring within the first 30s. While the ATCos were striving to complete the handoff within 60s, it was also apparent that few outliers existed with response times above 90s when the ATCo did not correctly estimate the tasks to be completed in the window of opportunity and failed to complete the handoff in the required time. Figure 33 (b), Figure 34 (b), and Figure 35 (b) show the RHD times produced from the TM augmentation model. As can be seen, the trend of RHD time was very similar to that of the FEWS simulation data for the 100% DL. For the 133 DL condition, there were more early responses in the FEWS data than the model. The RHD times produced from the model in the 133% No DL condition showed a markedly different pattern than the FEWS data. The TM augmentation model produced RHD times that were later than the FEWS times. In the model, the large majority of response times occurred later in the handoff window of opportunity, with over 50% of the responses occurring after 90s have elapsed. Contrary to the FEWS simulation data, very few responses in the model occurred before 60s had elapsed. The pattern of results shown in the histograms below, are similar to the trends observed in the correlations outlined above in
which the model accounted for more variance in the 100 DL condition than either of the 133 DL and No DL conditions.

![Figure 33. RHD Histogram for (a) FEWS and for (b) TM Model 100% DL.](image)

![Figure 34. RHD Histogram for (a) FEWS and for (b) TM Model 133% DL.](image)

![Figure 35. RHD Histogram for (a) FEWS and for (b) TM Model 133% No DL.](image)

### 8.2.4. Measure 4: Handoff Window Open Time for TM Model

The next step to explain the differences between the FEWS and the model RHD performance was to understand the effect of the environment on the operator model. Figure 36, Figure 37, and Figure 38 demonstrate the TC graphs for the window open times, where the relationship between the FEWS and the baseline model’s window open times are illustrated. Recall that lines that
slant down to the left indicate that the window opened earlier in the model than FEWS, and lines that slant down to the right indicate that the window opened later in the model than FEWS. Lines that crossover indicate aircraft window open times that occurred in a different order in the model than FEWS. (Note that in some cases, crossovers span multiple aircraft, so a crossover may involve more than one aircraft). The number of model window open times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the window open times were the same or different between the model and FEWS are presented in Table 12. As can be seen, approximately 52%-66% of the aircraft window open times were in a different order in the model as compared to the FEWS. This occurred because the baseline model represented an optimal ATCo, however the actual FEWS ATCo may have made decisions to alter the speed or altitude of the aircraft prior to the handoff, thus altering the window open time.

Table 12. Number, Relative Timing and Order for FEWS and TM Model Window Open Times.

<table>
<thead>
<tr>
<th></th>
<th>Window Opened Earlier or Same Time in Model than FEWS</th>
<th>Window Opened Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>11</td>
<td>33</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>133% DL</td>
<td>11</td>
<td>41</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>133% No DL</td>
<td>9</td>
<td>32</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>
Figure 36. TC Graph for FEWS and TM Model Window Open Time 100% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence

Figure 37. TC Graph for FEWS and TM Model Window Open Time 133% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
While the overall data suggest that the 133% No DL condition possesses the largest effect and that the overall RHD times were significantly different, the effect cannot be attributed solely to the time that the window opened since the present analysis suggest that the number of window open times that differed in order from the modelled to the actual aircraft were fairly consistent across traffic conditions. The case of the differences was therefore sought in the following analysis.

8.2.5. Measure 5: Adjusted Handoff Complete Time Analysis for TM Model

As with the baseline model, the window close times were adjusted to account for differences between the model and FEWS window open times. The number of model handoff complete times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the handoff complete times were the same or different between the model and FEWS are presented in Table 13. As can be seen, approximately 59-61% of the aircraft handoff complete times were in a different order in the model as compared to the FEWS.
Table 13. Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times.

<table>
<thead>
<tr>
<th></th>
<th>Handoff Completed Earlier or Same Time in Model than FEWS</th>
<th>Handoff Completed Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>20</td>
<td>24</td>
<td>18</td>
<td>26</td>
</tr>
<tr>
<td>133% DL</td>
<td>20</td>
<td>32</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>133% No DL</td>
<td>9</td>
<td>32</td>
<td>16</td>
<td>25</td>
</tr>
</tbody>
</table>

The TC graphs in Figure 39, Figure 40, and Figure 41 illustrate the relationship between the FEWS and the *adjusted* baseline model’s handoff complete time.

Figure 39. TC Graph for FEWS and TM Model Adjusted Handoff Complete Times 100% DL. Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence.
Figure 40. TC Graph for FEWS and TM Model Adjusted Handoff Complete Time 133% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence

Figure 41. TC Graph for FEWS and TM Model Adjusted Handoff Complete Time 133% No DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
8.3. **CONCLUSION OF THE TM MODEL VALIDATION EFFORT**

In this second iteration of the model development process, the validation approach previously applied to the baseline model was applied to the TM augmentation model. This model iteration resulted in conservatively biased performance, in the sense that the modelled operator’s performed immediately as resource were available to service the task. The TM model performed better than the baseline model in predicting aggregate RHD times – specifically in the 133% No DL condition. This suggests that the TM model improved on the baseline model’s deficiency in its characterization of how operators manage tasks. However, the TM model was still limited in its ability to predict RHD times, as there were significant difference between the TM model and the FEWS data. Recall that a queue size of zero was chosen as a starting point for the analysis as it was the most extreme queue size value. Since this extreme queue size did not impact the model sufficiently, further sensitivity analyses were not pursued; clearly a queue size of 1 to 4 would not improve the model validity.

The second explanation posed in Chapter 6, that the human operators failed to estimate the passage of time accurately, will be explored next. Specifically, the model assumed that the operator always had “perfect” awareness of available time to complete a task, and, furthermore, it assumed that tasks were completed quickly, in the right order, and at the right time. It can be seen that the model does not perform in this way by looking at the performance of the model in Chapter 6, particularly in the 133 No DL condition. As discussed in Chapter 7, it is known that humans do not estimate time perfectly. The resulting Time Estimation (TE) model was compared to the HITL data using the same validation approach applied to the baseline and TM models.
CHAPTER 9: DEVELOPING AND VERIFYING THE TIME ESTIMATION (TE) MODEL AUGMENTATION

9.1. MODELLING ESTIMATES OF TIME AVAILABLE (T_A)

It is important to develop accurate models of the estimates of T_A for ATC tasks because the ATCo’s primary task is to estimate the amount of time between the present time (now) and when an aircraft enters his/her sector at some point in the future. Incorrect estimates of T_A will lead to inaccurate scheduling of performance and may lead to situations where the ATCo will be forced into opportunistic kinds of behaviours involving less planning and more immediate responding rather than operating at the strategic level of operations. As previously outlined, estimating T_A is a function of all of the other tasks being carried out in a specific context. T_R is less relevant to the issues surrounding ATCo task management and time estimation because T_R for the handoff task is instantaneous, as the time it takes for the ATCo to press a button is on the order of milliseconds. It does not impact the task schedule to a large extent because of the immediate nature of the action. That having been stated, it is assumed, for the purposes of the current model application, that T_R is estimated perfectly. T_O is not being modelled here but is included to illustrate the breadth that the model will encompass in the future.

Estimating the time available to complete the activity, T_A, is akin to estimating the time remaining in the window of opportunity. In this model, it is assumed that the simulated operator has a perfect model of the end of the window of opportunity, such that, at any given time, T_A can be inferred from knowledge of the time duration that has passed thus far. That is, (s)he will always subtract her estimate of the time elapsed up to the present from the actual time in the window to get the biased estimate of time available.

This first step in the time estimation model development process was to factor in a model of estimating Time Available (T_A), based on the literature reviewed in Chapter 7, that accurately reflects the biased manner with which people (mis)estimate T_A. Recall that, when operators are under moderate to high workload, they tend to underestimate the passage of time; that is, they think less time has passed than actually has passed.\(^{18}\) With regards to the present modelling

\(^{18}\) The well-known expression “time flies when you’re having a good time” has much relevance in this context.
example, this will therefore result in an overestimate of $T_A$ – i.e. operators will think that they have more time available than they actually do. The predicted consequence of such a bias is that more non-critical tasks are expected to accumulate in the queue, with the result that more of them are expected to commence at times that are later than would otherwise be considered ‘optimal’.

### 9.2. **DEVELOPING AN ALGORITHM FOR ESTIMATED $T_A$**

Recall from Chapter 7 that Brown and Boltz (2002) showed that the factor that degrades estimates of $T_A$ depends on the length of time being estimated during tasks that required comparative judgements of duration of prose passages under baseline and high workload conditions. They showed that time judgements became shorter for longer task durations – i.e. underestimation – and that there was a strong linear component to this relationship. Table 14 presents Brown and Boltz’s (2002) data for the constant error time judgments (the ratio of estimated to actual time) for the 15-20s, 25-30s, 35-40s and 45-50s time windows, for both the baseline and difficult conditions. There was no difference in error results between the baseline and the easy task condition so only the baseline condition was used. The data presented in Table 14 represent the ratio of estimated to actual times, with values lower than 1 being time duration underestimates and values greater than 1 being time duration overestimates. Note that all values in Table 14 are less than 1.

<table>
<thead>
<tr>
<th>Duration Estimate</th>
<th>Baseline Condition</th>
<th>Difficult Condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>15s-20s</td>
<td>.95</td>
<td>.79</td>
</tr>
<tr>
<td>25s-30s</td>
<td>.90</td>
<td>.72</td>
</tr>
<tr>
<td>35s-40s</td>
<td>.88</td>
<td>.64</td>
</tr>
<tr>
<td>45s-50s</td>
<td>.85</td>
<td>.58</td>
</tr>
</tbody>
</table>

---

19 Brown & Boltz note a linear component, but the relationship between a stimulus and its perceived intensity are typically best described by a power function. While Brown and Boltz refer to this as a linear component to the relationship, it could be that they incorrectly concluded that a linear relationship exists when the small amount of data from their research comprise a portion of the longstanding power function relationship in time perception (Engen, 1971).
Although not reported by Brown and Boltz, it is useful to plot estimated time duration versus actual time to understand the relationship between them. Figure 42 and Figure 43 present plots of Brown and Boltz’s data for the baseline and high workload (difficult) conditions respectively. These figures are shown also with a power function fit to the data, together with a straight diagonal line representing ‘perfect’ time estimation (i.e., estimated time = actual time). The *power function* was selected as the appropriate regression fit to use since it is standard practice to use Stevens’ Power Law to characterize such phenomena (Engen, 1971).

The first power function trend line on the baseline condition produced an $R^2$ value of .99, yielding the power function equation presented in Equation 4. The second power function trend line on the high workload (difficult) condition produced an $R^2$ value of .99 yielding the power function equation presented in Equation 5.


$$y = 1.29x^{0.89}, \quad R^2 = 0.99$$  \hspace{1cm} (4)

**Equation 5.** Power Function Equation for Time Estimates in the High Workload Condition.

$$y = 1.94x^{0.69}, \quad R^2 = 0.99$$  \hspace{1cm} (5)

It was determined that these power function models properly represented time estimates in a manner consistent with expectations, possessing a characteristic pattern similar to that predicted by Vierordt’s Law (1868) – that is, larger durations being underestimated (Ornstein, 1968).
The power functions were further tested against a new data source, published by Boltz (2005). Those data were used to verify the performance of the power function model for the high workload condition for an actual time of 10s, with its estimate being 9.55s. The empirical research, in comparison, found that 10s durations were estimated to be 9.6 s, that is, an underestimate of 4%. For the baseline condition an actual time of 10s was estimated as 10.08s. The empirical research found that 10s duration were estimated to be 10.3s. It is important to note that the Boltz (2005) study used different stimuli containing both auditory and visual information, three different modalities (auditory, visual, and audiovisual), different durations, and possessed six different elicitation methods. For the purposes of the time estimation model, all time estimates used will be for time durations greater than 10s, because estimates lower than this are considered to be at the psychophysical level, a level too low for the consideration of ATCo performance (Grondin, 2001; Ornstein, 1968).

In summary, the power function model implementation allows for predictions of the human’s estimate of time under conditions of low workload (under Brown and Boltz’s baseline conditions) and under conditions of high workload (under Brown and Boltz’s difficult condition). These two models generalized the Brown and Boltz (2002) data into an arguably verifiable and potentially validatable power function, which allows an extrapolation to be made to predict values of ‘Y’ (a human’s estimate of time) for any given value of ‘X’ (the actual time) and for any given value of ‘Z’ (workload).

What follows is the development of a bivariate model that integrated workload on a continuous scale to influence the $T_A$ estimate.

### 9.3. WORKLOAD DEMANDS

The workload that is experienced by the operator triggered the time estimation module to simulate the human’s perception of time available to complete an activity, $T_A$. Previous research on workload outlined low, moderate, high, and overload as categories of workload that were experienced by an operator (Hancock & Chignell, 1988). Given that varying levels of workload differentially impact estimates of $T_A$, however, it was necessary to create a continuous function for estimated time as a bivariate function of both Actual Time Elapsed and Estimated Workload. Developing such a power function permitted its use for any predicted value of workload, rather than just the low (baseline) level and the high workload level.
To create this bivariate relationship, a fitted function for baseline workload from Equation 4 and the function for high workload from Equation 5 were used:

\[ y = 1.29x^{0.89}; \ y = 1.94x^{0.69} \]

where, using the same notation from earlier, ‘y’ is the estimated time elapsed and ‘x’ is the actual time elapsed.

The next step was to modify the gain and the exponent, as functions of workload, denoted henceforth by the symbol WL. A linear interpolation of the two parameters of the model – the gain and the exponent – was selected as a first approximation. In other words, using the general form of the power law functions as: \( y = Gx^p \), this first approximation bivariate model was formed by linearly interpolating between the two gains (G) and the two exponents (p), for the two extreme levels (low and high) of the independent workload variable (WL). Assuming that the baseline level of WL from Brown and Boltz’s experiment was the lowest that would be encountered, this was considered to be equal to zero Assuming that the high WL level from Brown and Boltz was the highest that would be encountered, this was considered equal to 1. The time estimation model’s gain component, G, changed according to Equation 6:


\[ G = 1.29 + (1.94 - 1.29) \times WL = 1.29 + 0.65 \times WL \]  \hspace{1cm} (6)

and the time estimation model’s exponent, p, changed according to Equation 7:

Equation 7. Linear Interpolation for Determining the Exponent, p, for the Time Estimation Model.

\[ p = 0.89 + (0.69 - 0.89) \times WL = 0.89 - 0.2 \times WL \]  \hspace{1cm} (7)

Equation 8 results when the gain component in Equation 6 and the exponent in Equation 7 were combined, to form a bivariate model:

Equation 8. The Time Estimation Model Linear Interpolation.

\[ y = (1.29 + 0.65 \times WL)x^{(0.89 - 0.2 \times WL)} \]  \hspace{1cm} (8)
Equation 9 results when the terms of Equation 8 were replaced by the Time Estimation Model’s parameters:


\[ T_{\text{estimate}} = (1.29 + 0.65 \times WL)T_{\text{elapsed}}^{(0.89 - 0.2 \times WL)} \]

Figure 44 illustrates this function with the dependant variable, the time estimate, \( T_{\text{estimate}} \), along the y-axis, and the two independent variables, actual time elapsed, \( T_{\text{elapsed}} \), along the X-axis and the workload, WL, along the z-axis. Note that this graph illustrates the behaviour of the interpolation function only for the range of 0 to 10s in time and for levels of 0 to 1 in (normalised) workload.

9.4. VERIFYING THE TIME ESTIMATION MODEL ALGORITHM

For the purposes of the time estimation model, a simplifying assumption was made that the model’s low workload (WL=0) was equivalent to Brown and Boltz’s baseline condition and that the model’s high workload (WL=1) corresponded to Brown and Boltz’s “high” workload condition\(^{20} \) provided in Equation 9. This simplifying assumption required normalizing the 0-7 workload scale of the model onto Brown and Boltz’s 0-1 workload scale by dividing all modelled values by 7, so that a modelled workload of 7 would correspond to 1 when substituted

\(^{20}\) Recall from Brown & Boltz’s high workload condition consisted of monitoring for the detection of words that began with either the letter D or the letter L (pg. 71). This is a recognized limitation.
into Equation 9. Figure 45 verifies the performance of the time estimation model by showing the relationship between the actual time available and workload, for 6 different values of time for $T_{\text{actual}} = 15, 30, 45, 60, 90$ and 120s. Note that Brown and Boltz data operated in the range up to 50s. It is most applicable for the nominal RHD time of 60s. The reader should be cautioned that the applicability is reduced as the RHD increases substantially beyond 60s. The figure illustrates the modelled misperception of *elapsed* time as a function of increasing workload given the actual times listed. The graph illustrates, in other words, that when estimates of time available are required to be made at increasing levels of workload, the elapsed time will be underestimated, as shown in the graph, causing estimates of time available to be lengthened (Ornstein, 1969) by the empirically driven time estimation model function.

Figure 45. Estimates of Time Elapsed as a Function of Changing Workload Values and Actual Time Elapsed.

In the HPM to be described below, the human operator has not experienced the "elapsed time" yet -- rather they are trying to estimate how much time is available to complete an activity. If the actual time available is 60s and the workload is high, the ATCo may underestimate the time elapsed, which results in time available estimates that are larger than the actual time available. This means that the modelled operator will schedule the tasks to be performed after the optimal time period has passed causing longer response durations during periods of high
workload. In summary, the Equation 4 and Equation 5 are specific power curves for low and high workload while the bivariate model allows interpolation to different levels of workload.

In summary, task schedules are driven by the estimates of time available based on projected workload. The TE algorithm provided in Equation 9 indicates that time estimates are influenced by the perception of elapsed time that is based on workload. The bivariate model bases projected workload on the upcoming 15 tasks, which are known to the model, to generate the modelled operator’s perception of time available to complete the tasks. This perception of time available based on the projected task is then used to schedule the upcoming sequence of tasks. As workload increases, tasks will be scheduled early because the perception of time available was known to be overestimated and therefore required earlier task onset, as represented by the function provided in Equation 9 (the bivariate model).

9.5. VERIFYING THE TIME ESTIMATION ALGORITHM MODEL USING A SIMPLIFIED ATCo TASK

A verification of the effect of the time estimation model using a simplified, generic, ATC environment was conducted, to explicitly illustrate the potential impact of the time estimation model on human performance for a hypothetical time-critical task. In other words, rather than apply the time estimation model for the first time directly within the context of modelling the activities of the actual FEWS ATCos (which is presented in later chapters), the simple generic model was first created and tested independently, in the absence of any of the confounds that occur when dealing with data resulting from real-life actions undertaken by actual ATCos. This generic task model was a simple queuing model simulating an ATCo giving descent clearances to an aircraft. It was based on the premise that the busier the ATCo was, the less time he has to do a full visual scan of a radar display to detect proximity of aircraft to the sector boundary, because he is busy dealing with other tasks. This makes the ATCo unable to update his time estimate, therefore adding to the variable of interest, the time estimation error. The longer that the modelled operator spent on tasks, the less he will update his internal clock, and the more biased will be his perception of time.

9.5.1. Generic Model Development

The modelled ATCo was required to take actions necessary to clear an aircraft for descent in a fictional airspace sector, before handing off to a downstream tower controller. The optimal time
to provide this clearance corresponds to 60s after the aircraft enters the ATCo’s sector. The time available, $T_A$, therefore varied from $t=60s$, when the aircraft enters the sector, to $t=0s$, when handoff occurred on time, to $t<0s$ if handoff was late and occurred after 60s had elapsed. If the ATCo underestimated the $T_A$, which can result in descending the aircraft too early, the aircraft will be penalized by needing to fly at a lower altitude, thereby using more fuel. If the ATCo overestimated the $T_A$, which resulted in descending the aircraft too late, the downstream ATCo would have to work harder to merge and sequence the aircraft safely into the arrival stream. As a result of these consequences, the ATCo attempted to provide the clearance without any error in his estimate of the 60s time window.

The model possessed two interacting loops to represent the use of time for managing the task schedule – either by using the actual available time, $T_A$, or the estimate of $T_A$ using the time estimation model algorithm. As each increment of time passed, the model re-evaluated the time available to complete the series of tasks in its queue. The model always descended the aircraft at 60s when it used the actual time available (which is independent of the ongoing workload). This was analogous to an operator working in the “well calibrated” condition outlined in Section 7.2. For the time estimation case, the model descended the aircraft at the time estimate driven by the time estimation model – i.e. 60s minus estimate of time passage, which in turn is dependent on the operator’s estimated workload on the next 15 tasks (as defined in Equation 2). The time that the model descended the aircraft when using the time estimate model will be influenced by the predicted workload and will be either early or late.

The two interacting loops comprise an outer loop and an inner loop. The outer loop assumed that a list of aircraft entering the sector was scanned by the ATCo at regular intervals and the time that has actually passed relative to the goal times of 60s (which means that the window of opportunity ends 60s after the time at which each aircraft is added to the list – the window open time) was updated by the modelled ATCo for each aircraft. Each time the ATCo commenced a scan, he re-sampled the entire aircraft list to determine whether enough time existed to attend to other aircraft and other tasks before needing to descend the aircraft that is closest to the window close time. The modelled ATCo used either the actual $T_A$ or the estimate of $T_A$ to decide on the next aircraft to process as described in further detail in Section 9.5.2. At the end of each outer loop scan, all time estimation errors are reset to 0.
One assumption of the inner loop portion of the model was that the inner loop scan was completed more often when the ATCo gets busy. As such, the inner loop cycle was dependent on the number of aircraft in the environment. Specifically, the formula used to determine the frequency of scan updates in both the perfect time estimation and time estimation cases presented in Table 15 – i.e. the ratio of inner to outer loop scan updates – is shown in Equation 10.

Table 15. Generic Model Inner Loop Assumptions.

<table>
<thead>
<tr>
<th>Number of Aircraft</th>
<th>Number of Times Inner Loop is Performed/Outer Loop Scan</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
</tr>
</tbody>
</table>

Equation 10. Ratio of Inner-to-Outer Loop Scan Updates for the Generic Model.

\[ f_{\text{scan}} = \left( \frac{\# \text{ of AC}}{5} \right) + 1, \]

where \( f = \text{frequency of scan update} \), \( \# \text{ of AC} = \text{number of aircraft} \)

This equation resulted in a scan being completed for each level of aircraft in the generic model. For example, in the low workload condition, or 5 aircraft, the inner loop cycle was completed 2 times \((5/5+1)\) for every one outer loop update. In the low workload condition, the modelled ATCo cycles less on the inner loop and therefore updates his/her internal time clock to a greater extent than in the higher workload levels, as reflected in Table 15 and Equation 10. In the high workload condition, or 25 aircraft, the inner loop cycle was completed six times \((25/5+1)\) for every one outer loop scan update. In the high workload condition the modelled ATCo was unable to recalibrate because of being too busy working on the other tasks and therefore could not update their clock time.

Figure 46 illustrates the two interacting loops and the tasks that are involved in this model. The model assumed that when the ATCo noticed a new aircraft entering his/her sector, it exited the outer loop (Block 2) and entered the inner loop (Blocks 4,5,6,8). In Block 4 of the inner loop, termed the “Decision Point,” the ATCo estimated the time available until the descent clearance should be issued, according to one of the two scenarios outlined below in Section 9.5.2. This was the \( T_A \) to complete the task.

Block 4 was where the ATCo uses either the actual \( T_A \) or the estimated \( T_A \) to decide to descend the aircraft. For the No Time Estimation case (as per scenario 1 in Section 9.5.2), in Block 4, when the modelled ATCo read from the list of actual \( T_A \) values that there was no time
available for one of the aircraft, they executed the descent task for that aircraft (at the top of the list of aircraft to be processed). In the Time Estimation case (as per scenario 2 in Section 9.5.2), in Block 4, when the modelled ATCo estimated that there was not enough time available, $T_A$, relative to the 60s window, he decided to move to Block 6 where he descended the aircraft, which reset the time error clock. In this case, the model proceeded to Block 8.

Figure 46. Task Network Model of a Simplified ATC Environment. The ATCo Estimates Time According to Time Estimation Model Algorithm (denoted in the Block 4).

In Block 6, the ATCo descended the aircraft and moved to the outer loop. If Block 6 was not entered, the modelled ATCo decided to attend to other aircraft and other tasks (Block 5). The time allocated to this sub-task was modelled using a normal distribution with a mean of 5s, and a standard deviation of 0.5s.

In both cases (Block 5 and 6), Block 8 was entered, where the decision to update the scan – i.e., enter the outer loop – was performed. If the model determined that the scan should not be updated, it remained cycling in the inner loop. Depending on the number of times that the inner loop was traversed, the time estimation error continued to grow. If the model determined that the scan should be updated, it passed into the outer loop as it would for cases where estimation was perfect. In such cases, there was no inner loop cycling and the error term did not grow. Once in the outer loop, the list of aircraft was re-scanned (Block 2). This cycle continued until, in Block 4, the ATCo estimated that the requisite time had elapsed, at which point they descended the aircraft. A new aircraft then joined the TMA (traffic manager advisor) list.

For the purposes of the generic model, the workload was held constant by a new aircraft joining the queue as soon as one aircraft left the queue. Five values of workload were used, ranging from low (5 aircraft), to low-medium (10 aircraft), to medium (15 aircraft), to medium-
high (20 aircraft), to high workload (25 aircraft). (The selection of these particular numbers of aircraft was based on varying fractions of a current recommendation for a recommended maximal load of 25 aircraft (Willems, 2005).) As previously outlined, for the purposes of the time estimation model, it was assumed that workload values from the model (0-7) were normalized onto the Brown and Boltz workload values (0-1). This meant that low model workload value of 0 was equivalent to Brown and Boltz’s baseline workload condition and that high model workload (WL=7) corresponds to Brown and Boltz’s “high” workload condition provided in Equation 9 with the remaining traffic loads distributed equally between workload 0 to 1.

The number of aircraft was generated by an outside task that launches aircraft according to Table 16. These aircraft populate the TMA list upon which the ATCo must act.

Table 16. The Number of Aircraft Generated by Simulation Model and Entry Time Spacing, as a Function of Workload Level.

<table>
<thead>
<tr>
<th>Workload Levels</th>
<th>Number of Aircraft</th>
<th>Sector Entry Time Spacing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Low</td>
<td>5</td>
<td>200</td>
</tr>
<tr>
<td>Low-Moderate</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Medium</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>Moderate-High</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>High</td>
<td>25</td>
<td>40</td>
</tr>
</tbody>
</table>

9.5.2. Exercising the Generic Model

Two models were generated to demonstrate the effect that the time estimation model has on the performance of the model. Both models use the same environment model but a different scheduling mechanism to decide upon the tasks to complete. Model 1 uses actual $T_A$ while Model 2 uses the estimate of $T_A$ to decide on the next aircraft to process. As outlined above, these decisions are made in Block 4 of Figure 46.

- Model 1, the baseline model scenario, assumed that the operator had perfect time estimation, and used the actual $T_A$ (real time) to descend the aircraft in the simulation. In other words, this scenario did not use the time estimates that were impacted by the ATCo’s workload to schedule and perform the activities. Model 1 therefore has the error associated with the SD in the performance of the descent task because it uses the actual simulation clock time (and SD) as the time to commence the task.
- Model 2, the time estimation model, used time estimates to perform the descent task. These time estimates impacted the performance of the perception of the window open
time according to the time estimation model’s algorithm and therefore resulted in a distribution of time error values that were later than the optimal descent times revealed by the baseline model.

Ten thousand model runs were generated for the generic model at each of five workload levels for each of the models. Table 17 provides the descriptive statistics of the time error output for each Scenario (2) x Workload (5) condition. The data from this table are plotted in Figure 47. Positive time errors caused by decisions made before the 60s are indicative of the ATCo descending the aircraft earlier than the optimal descent time. Negative time errors caused by overestimates of $T_A$ were indicative of the ATCo descending the aircraft later than the optimal descent time. The results revealed increasingly negative time errors as workload increased. This meant that, as the ATCos got busier, their estimate of the time passage became increasingly delayed (late), as they felt that they have more time available than they actually did to complete the tasks.

Table 17. Descriptive Statistics of the Time Error Distributions for Models 1 (Baseline) and 2 (TE).

<table>
<thead>
<tr>
<th></th>
<th>No A/C</th>
<th>Range</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>5</td>
<td>6.25</td>
<td>-3.10</td>
<td>3.15</td>
<td>.0030</td>
<td>1.47940</td>
<td>-.020</td>
<td>-1.143</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>5</td>
<td>6.55</td>
<td>-3.10</td>
<td>3.45</td>
<td>.1576</td>
<td>1.48623</td>
<td>-.022</td>
<td>-1.114</td>
</tr>
<tr>
<td>Baseline</td>
<td>10</td>
<td>6.31</td>
<td>-3.17</td>
<td>3.14</td>
<td>.0037</td>
<td>1.45806</td>
<td>-.011</td>
<td>-1.131</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>10</td>
<td>7.35</td>
<td>-4.09</td>
<td>3.26</td>
<td>-.1700</td>
<td>1.52686</td>
<td>-.036</td>
<td>-.940</td>
</tr>
<tr>
<td>Baseline</td>
<td>15</td>
<td>6.37</td>
<td>-3.05</td>
<td>3.31</td>
<td>.0109</td>
<td>1.45519</td>
<td>-.013</td>
<td>-1.126</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>15</td>
<td>9.61</td>
<td>-6.39</td>
<td>3.22</td>
<td>-1.0297</td>
<td>1.96382</td>
<td>-.218</td>
<td>-.575</td>
</tr>
<tr>
<td>Baseline</td>
<td>20</td>
<td>6.24</td>
<td>-3.14</td>
<td>3.10</td>
<td>.0257</td>
<td>1.46279</td>
<td>-.014</td>
<td>-1.135</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>20</td>
<td>13.09</td>
<td>-9.93</td>
<td>3.16</td>
<td>-2.3326</td>
<td>2.85134</td>
<td>-.354</td>
<td>-.694</td>
</tr>
<tr>
<td>Baseline</td>
<td>25</td>
<td>6.25</td>
<td>-3.18</td>
<td>3.07</td>
<td>.0089</td>
<td>1.46632</td>
<td>-.009</td>
<td>-1.133</td>
</tr>
<tr>
<td>Time Estimation</td>
<td>25</td>
<td>18.17</td>
<td>-15.11</td>
<td>3.06</td>
<td>-4.2188</td>
<td>4.12670</td>
<td>-.342</td>
<td>-.939</td>
</tr>
</tbody>
</table>

Table 17 also reveals that there is little effect of workload on the baseline model’s range of time errors (very close to zero as expected), skewness or kurtosis across workload levels. On the other hand, Table 17 reveals that there is an apparently large effect of workload on the time estimation model’s range of time errors, with the range increasing as workload increases. In addition, there appears to be a more pronounced effect on the skewness and kurtosis of the shapes of the time estimation model’s time error distributions across workload levels.
Figure 47. Mean Time Error of 10000 Generic Model Runs to Verify TE Model Performance.

The data from Table 17 are plotted in Figure 47, which revealed that the baseline model performs very close to zero mean error, as expected. The time estimation model time error output, on the other hand, revealed that the aircraft are being descended at increasingly late times as workload increased.

Histograms of the baseline model distributions were compared with the time estimation model distributions at each workload condition in the figures that follow, which, together with Table 17 illustrate that, as workload increased, the non-linear nature of the time estimation model increased. Figure 48 illustrates the time error distributions of the 10000 model runs of the baseline and time estimation models at the low workload level of 5 aircraft. It can be seen that little difference existed in the pattern of responses in terms of the skewness and kurtosis of the graphs. The error distributions revealed a symmetric, platykurtic distribution shape. Both of the baseline, low workload model and the time estimation, low workload model time error distributions were quite symmetric (see Figure 49).

Figure 48. Histogram for (a) Baseline and (b) TE Model Time Errors for Low Workload Runs (5 Aircraft).
Figure 49 illustrates the time error distributions of the 10000 model runs of the low-medium workload level (10 a/c) of the baseline and the time estimation models. It is apparent that the time estimation condition possessed more negative time error values (that represent slightly later descent times) as evidenced by the tail to the left side of the distribution. In addition, time error became negatively skewed as the range increased from -4.09 to +3.26, while there was little change for the baseline model. This means that when the number of aircraft / traffic was increased to a low-moderate level, the time estimation model was having an increased effect on the model’s performance. The kurtosis of the distribution also started to flatten with the time estimation model.

Figure 49. Histogram for (a) Baseline and (b) TE Model Time Errors for Low-Moderate Workload (10 Aircraft).

Figure 50 illustrates the time error distributions of the 10000 model runs of the moderate workload level (15 a/c) for the baseline and time estimation models. It is very apparent that the time estimation condition possessed more negative time error values, representing later descent times, as evidenced by the tail to the left of the distribution. In addition, time error became more negatively skewed in the time estimation model from -6.39 to +3.22 while there was little change for the baseline model. This means that when the number of aircraft / traffic was increased to a moderate level, the time estimation model was having an increased effect on the model’s performance over the previous two workload levels as expected. The kurtosis of the distribution also started to flatten with the time estimation model.
Figure 50. Histogram for (a) Baseline and (b) TE Model Time Errors for Moderate Workload (15 Aircraft).

Figure 51 illustrates the time error distributions of the 10000 model runs of the moderate-high workload level (20 a/c) for the baseline and time estimation models. It is apparent that the time estimation condition possessed even more negative time error values, representing later descent times, than the baseline model, as evidenced by the tail to the left side of the distribution. In addition, time error became even more negatively skewed in the time estimation model from -9.93 to +3.16, while there was little change for the baseline model. This means that when the number of aircraft / traffic was increased to a moderate-high level, the time estimation model was having an increased effect on the model’s performance over the moderate traffic level.

Finally, Figure 52 illustrates the time error distributions for the 10000 model runs of the medium-high workload level (25 a/c) of the baseline and time estimation models. It is apparent that the time estimation condition possessed even more negative time error values than the
baseline model, as evidenced by the tail to the left side of the distribution. In addition, the time error range became even more negatively skewed in the time estimation model from -15.11 to +3.06, while there was little change in the range for the baseline model. This means that when the number of aircraft / traffic was increased to a high level, the time estimation model was having an increased effect on the model’s performance over all of the previous workload levels. The kurtosis of the distribution also started to flatten with the time estimation model.

![Histogram](image)

**Figure 52.** Histogram for (a) Baseline and (b) TE Model Time Errors for High Workload Runs (25 Aircraft).

**9.6. SUMMARY OF TIME ESTIMATION MODEL DEVELOPMENT**

In summary, a time estimation model was developed that used a power function to model the relationship between workload and operators’ estimates of time available based on empirical literature. A model verification exercise was then conducted using a simplified, generic ATC environment to show that the time estimation model impacted time estimation in the manner that was expected and thus should be expected to have a similar effect when integrated within more complex HPMs. Specifically, it was shown through distributions of time errors that were created over 10000 model runs that the time estimation model impacted the mean, skewness and kurtosis of the distributions of time errors in the following ways:

- The mean time errors became increasingly negative relative to the baseline model, which remained positive.

- The skewness became increasingly negative (later onset times) over the baseline model.
• The kurtosis became increasingly flat over the baseline model, which remains centred between +/-3, and the range of the distribution of time errors was greatly increased under the time estimation model.

• The non-linear nature of the time estimation model as workload was increased has been verified.

This quantitative verification effort of the time estimation model provided evidence that a simple but non-trivial task network model constructed to simulate a limited aspect of air traffic control was sensitive to the parameters inside of the time estimation algorithm. As a result, as outlined in the next chapter, a formal validation effort of the time estimation model in a specific ATCo environmental context was conducted to determine the generalisability of the time estimation model to the ATCo domain.
CHAPTER 10: VERIFYING AND VALIDATING THE TIME ESTIMATION (TE) MODEL

10.1. AUGMENTATION OF BASELINE MODEL WITH TIME ESTIMATION

To augment the baseline model with the time estimation (TE) model, the Time Estimating box (centre box) in Figure 6 was modified to account for imperfections of human estimates of time passage as a function of workload, consistent with the empirical literature (Brown & Boltz, 2002; Boltz, 2005). This change is shown in Figure 53, where the times in that central block are now “Estimated” rather than “Actual”. The misestimate of time passage in the TE augmentation is according to the algorithms developed in Chapter 9 and impacts the operator’s ability to properly schedule and sequence activities.

The TE model was implemented in a closed-loop fashion, in which under-estimation of time passage increased as workload increased. The TE algorithm allowed for predictions of the human’s (under)estimate of time under conditions ranging from low workload (baseline conditions) to high workload (difficult conditions). The TE algorithm generalized the Brown and Boltz (2002) data into a power function that was verified with another data set (Boltz, 2005). This allowed an extrapolation to be made to predict a human’s estimate of time for any given value of the actual time and for any given value of workload in the range of 0 to 1.

![Figure 53. Baseline Model with TE Augmentation.](image)

The baseline model’s strategic task scheduling was combined with the TE underestimates of time passage as a function of increasing workload. In particular, this model allowed up to five tasks to enter into an ATCo’s queue and whenever a new task was to be serviced, it strategically selected the task with the highest priority and with the estimated smallest time available around...
the mean of $t=60s$ to carry out next. When more than five tasks were in the queue, non-critical tasks such as pilot communications were shed.

### 10.1.1. Measure 1: TE Model Workload

In this section, the validation measures and techniques described in Section 6.2 was applied to the TE model.

#### 10.1.1.1. Mean Workload For TE Model

Figure 54 illustrates the average workload for the three traffic load conditions for the FEWS adjusted data (a), the baseline model (b), and the TE model (c).

![Figure 54. Mean (a) FEWS Adjusted, (b) Baseline, and (c) TE Model Workload by Traffic Conditions.](image)

There was a significant 2x3 interaction $F(2,40)=49.91, p<.001$. As with the baseline model, there were significant differences in workload between each traffic load condition for the FEWS adjusted ($F(2,40)=57.07, p<.001$), and the TE model ($F(2,40)=8.91, p<.001$), except for the 133 DL to No DL comparison of the FEWS adjusted data (presented in Table 18). Again, the lack of significance in the FEWS 133 DL to 133 No-DL could be due to the high number of missing data points in the FEWS data (as shown in Figure 55). These data illustrate that the model adequately represented the FEWS adjusted data and that the modelled environment drove the operator’s workload values in a manner consistent with the FEWS. No significant differences existed for the FEWS in the 133 DL to 133 No DL condition while the model did possess a significant difference. A fine-grained analysis was completed to explore the possible reasons for this difference.
Table 18. Significance Tests for Traffic Level Increases on Workload for FEWS and the TE Model.

<table>
<thead>
<tr>
<th></th>
<th>100 DL – 133 DL</th>
<th>100DL-133 No DL</th>
<th>133 DL-133 No DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEWS (Adjusted)</td>
<td>(t(21) = 8.14^{***})</td>
<td>(t(21) = 6.85^{***})</td>
<td>(t(21) = 1)</td>
</tr>
<tr>
<td>Baseline</td>
<td>(t(829)=4.36^{***})</td>
<td>(t(829)=8.46^{***})</td>
<td>(t(829)=3.30^{***})</td>
</tr>
<tr>
<td>TE Augmentation</td>
<td>(t(829)=3.12^{***})</td>
<td>(t(829)=3.12^{***})</td>
<td>(t(829)=8.26^{***})</td>
</tr>
</tbody>
</table>

\*p<.05, **p<.01, ***p<.001

10.1.1.2. Workload Trend Output For TE Model

Figure 55 presents the individual workload ratings from the FEWS data (upper graphs) and the model predictions (lower graphs) for each traffic load condition. It can be seen that the FEWS and the model produced similar workload trends in the 100 DL (Figure 55a) and 133 No DL (Figure 55b) conditions. Unfortunately, the 133 No DL condition in the FEWS simulation possessed a significant number of missing values, limiting the conclusions that could be drawn about the comparison of FEWS and model trends.

![Figure 55. FEWS (top) compared to the TE Augmentation (bottom) Workload Output over Simulation, for (a) 100 DL Condition, (b) 133 DL Condition, (c) 133 No DL Condition. Notes - FEWS possessed missing responses. There were 18 responses for (a), 15 for (b), 9 for (c).](image)

10.1.2. Measure 2: TE Model Queue Length

Recall from Section 6.2.2 that queue length is technically a verification measure because there is no corresponding FEWS data; however it is included in the set of validation measures because it is used to support the validation process undertaken in this thesis.
It can be seen in Figure 56 that the maximum task queue increased as the traffic load increased from 100% DL to 133% DL to 133% No-DL. The increase in the maximum number of queued tasks in the 133% No DL condition suggests that the model was queuing more tasks because the ATCo’s were not completing DL handoffs and clearances due to the high workload levels. The same pattern of results was observed with mean queue length variable (see Figure 57). The queue length increased significantly when the traffic level increased from 100% DL to 133% DL \((t(829) = 4.05, p < .001)\), and also when the traffic level increased from 100% to 133% No DL \(t(829) = 30.31, p < .001\). Of particular interest is the large increase in queue length for the 133% No DL condition as compared to 133% DL, \(t(829) = 28.39, p<.001\), which indicates that the No DL condition required more tasks to be delayed than in the DL condition.

Figure 56. Maximum Number of Tasks to be Serviced in TE Model Queue as a Function of the Experimental Condition.

Figure 57. Average Number of Tasks to be Serviced in the TE Model Queue as a Function of the Experimental Condition.

10.1.3. Measure 3: TE Model RHD

10.1.3.1. RHD t-test and Results for FEWS versus TE Model

Figure 58 illustrates the data associated with the RHD measure produced from the FEWS simulation and those RHD predictions generated by the baseline model, and the TE augmentation model as a function of the three experimental conditions. Table 19 provides the FEWS, baseline model, and TE model t-test results. As can be seen, the TE model did not improve RHD predictions as compared to the baseline model. Relative to the FEWS data, the RHD times produced by the TE model were not significantly different than those produced from FEWS in the 100% condition, but they were significantly higher in the 133% DL and 133 No DL conditions. There was no significant difference between the baseline and the TE model RHD times suggesting that the RHDs come from the same population.
Figure 58. Comparison of the Average RHD in the FEWS, Baseline, and TE Models.

Table 19. Comparison of FEWS, Baseline, TE Model t-tests.

<table>
<thead>
<tr>
<th></th>
<th>FEWS – Baseline</th>
<th>FEWS – TE Model</th>
<th>Baseline – TE Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>100 DL</strong></td>
<td>t(76)=0.85</td>
<td>t(99)=2.07*</td>
<td>t(80)=5.51***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>133 DL</strong></td>
<td>t(76)=0.85</td>
<td>t(99)=2.08*</td>
<td>t(80)=5.85***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>133 No DL</strong></td>
<td>t(88)=0</td>
<td>t(99)=.01</td>
<td>t(80)=.75</td>
</tr>
</tbody>
</table>

*p<.05, **p<.01, ***p<.001

As completed for the baseline and the TM models, chi-square analyses were conducted to determine if the distributions of RHD times from the model and the FEWS observations come from the same population. As with the baseline and the TM models, data were put into three bins of approximately equal size: 0 – 10 s, 11-30 s, and greater than 31 s.

10.1.3.2. **RHD Chi-Square Test and Results for FEWS versus TE Model**

The $\chi^2$ test conducted on the aggregate RHD times for the three traffic levels suggested that the frequency distributions from the baseline model, TE model and the FEWS observations, did not come from the same population (see Table 20). These results suggest that both models did not faithfully represent the human RHD data in any of the three load conditions. The baseline and TE models were from the same distribution for all conditions.

Table 20. Chi-Square and Correlation Analyses of FEWS, Baseline and TE Models.

<table>
<thead>
<tr>
<th></th>
<th>FEWS-Baseline</th>
<th>FEWS –TE Model</th>
<th>Baseline – TE Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chi-square</strong></td>
<td>r value</td>
<td>Chi-square</td>
<td>r value</td>
</tr>
<tr>
<td><strong>100DL</strong></td>
<td>$\chi^2(2,N=122)=4.96***$</td>
<td>0.21</td>
<td>$\chi^2(2,N=122)=4.96***$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\chi^2(2,N=154)=0$</td>
</tr>
<tr>
<td><strong>133DL</strong></td>
<td>$\chi^2(2,N=163)=3.69***$</td>
<td>-0.15</td>
<td>$\chi^2(2,N=163)=55.69***$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\chi^2(2,N=226)=0$</td>
</tr>
<tr>
<td><strong>133NDL</strong></td>
<td>$\chi^2(2,N=154)=38.98***$</td>
<td>0.12</td>
<td>$\chi^2(2,N=154)=43.29***$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\chi^2(2,N=226)=1.74$</td>
</tr>
</tbody>
</table>

*p<.05, **p<.01, ***p<.001
In summary, the TE model did not adequately represent human data as collected in the FEWS simulation. Correlations were conducted to provide insights into the strength of the relationship identified by the $\chi^2$ test and are provided in Table 20.

A more detailed examination was therefore required to explore the reasons for the breakdown in the model’s performance. Whereas the modelled RHD times were significantly higher than the FEWS times in the 133 DL condition, the absolute difference between the means is relatively small, only about 6s, or approximately 28%. However, the TE model overestimated RHD in the 133 No DL condition by a much greater amount, with an absolute difference of 67s, or approximately 160%. The reasons for the greater breakdown in the model’s performance can be captured only by exploring the performance of the model in all of the 100 DL, 133 DL and 133 No DL conditions in greater detail. The histogram of the individual data points of the FEWS and the TE augmentation model RHD is provided next.

10.1.3.3. RHD Histogram of FEWS, TE Model

Figure 59, Figure 60, and Figure 61 show the RHD times produced from the TE model. As can be seen, the trend of RHD time was very similar to that of the FEWS simulation data for the 100% DL. For the 133 DL condition, there were more early responses in the FEWS data (85%) than the model (72%). The RHD times produced from the model in the 133% No DL condition showed a markedly different pattern than the FEWS data. The TE model produced RHD times that were later than the FEWS times. In the model, the large majority of response times occurred later in the handoff window of opportunity, with over 88% of the responses occurring after 60s have elapsed. Contrary to the FEWS simulation data, fewer responses in the model (12% vs. 61%) occurred before 60s elapsed. This pattern is consistent with the TM model’s output. The results illustrated in the histograms below, are similar to the trends observed in the correlations outlined above in which the model accounted for more variance in the 100% DL condition than either of the 133% DL or No DL conditions.
Chapter 10 – Validation of Baseline Model with Time Estimation

10.1.4. Measure 4: TE Model Handoff Window Open Times

The next step to explain the differences between the FEWS and the model RHD performance was to understand the effect of the environment on the operator model. Figure 62, Figure 63, and Figure 64 demonstrates the TC graphs for the window open times, where the relationship between the FEWS and the TE model’s window open times are illustrated. Recall that lines that slant down to the left indicated that the window opened earlier in the model than FEWS, and lines that slant down to the right indicated that the window opened later in the model than
FEWS. Lines that crossover indicated aircraft window open times that occurred in a different order in the model than FEWS. (Note that in some cases, crossovers span multiple aircraft, so a crossover may involve more than one aircraft). The number of model window open times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the window open times were the same or different between the model and FEWS are presented in Table 21. As can be seen, approximately 56%-79% of the aircraft window open times were in a different order in the model as compared to the FEWS. This occurred because the TE model represented an optimal ATCo; however the actual FEWS ATCo may have made decisions to alter the speed or altitude of the aircraft prior to the handoff, thus altering the window open time.

Table 21. Number, Relative Timing and Order for FEWS and Model Window Open Times.

<table>
<thead>
<tr>
<th></th>
<th>Window Opened Earlier or Same Time in Model than FEWS</th>
<th>Window Opened Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>14</td>
<td>30</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>133% DL</td>
<td>19</td>
<td>33</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>133% No DL</td>
<td>9</td>
<td>32</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>

Figure 62. TC Graph for FEWS and TE Model Window Open Times 100% DL.

Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
Figure 63. TC Graph for FEWS and TE Model Window Open Times 133% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence

Figure 64. TC Graph for FEWS and TE Model Window Open Time 133% No DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
While the overall data suggest that the 133% No DL condition possessed the largest effect and that the overall RHD times were significantly different, the effect cannot be attributed solely to the time that the window opened since the present analysis suggest that the number of window open times that differed in order from the modelled to the actual aircraft were fairly consistent with the 133% DL traffic condition. The case of the differences was therefore sought in the following analysis.

10.1.5. Measure 5: TE Model Adjusted Handoff Complete Time

As with the previous model iterations, the window close times were adjusted to account for differences between the model and FEWS window open times. The number of model handoff complete times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the handoff complete times were the same or different between the model and FEWS are presented in Table 22. As can be seen, approximately 56-68% of the aircraft handoff complete times were in a different order in the model as compared to the FEWS. The TC graphs in Figure 65, Figure 66, and Figure 67 illustrate the relationship between the FEWS and the *adjusted* baseline model’s handoff complete time.

**Table 22. Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times.**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Handoff Completed Earlier or Same Time in Model than FEWS</th>
<th>Handoff Completed Later in Model than FEWS</th>
<th># Aircraft with Same Window Close Orders</th>
<th># Aircraft with Different Window Close Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>23</td>
<td>21</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>133% DL</td>
<td>21</td>
<td>31</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>133% No DL</td>
<td>9</td>
<td>32</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>
Figure 65. TC Graph for FEWS and TE Model Adjusted Handoff Complete 100% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence

Figure 66. TC Graph for FEWS and TE Model Adjusted Handoff Complete Time 133% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
Chapter 10 – Validation of Baseline Model with Time Estimation

10.2. RESULTS SUMMARY FOR TE MODEL AUGMENTATION

The same validation approach that was applied to the two previous model development iterations was applied to the TE model iteration. The TE augmentation model did not perform as well as the TM augmentation model. The fact that the TM model succeeded in producing RHD times that were more in line with the FEWS data than the TE model suggests that the breakdown in the baseline model during the high workload scenario was due to the baseline model’s inability to adequately represent the ATCo’s task scheduling strategy rather than the model’s inability to account for biases in time estimation. The low workload conditions (100% DL and 133% DL) were unaffected by the TE augmentation. It is therefore necessary to examine the interaction that exists between task management and time estimation. Due to the fact that the TM model alone and the TE model alone did not succeed in bringing either model significantly closer to the baseline model, a third and final iteration was made to the baseline model, using both the TM and TE model augmentations.
CHAPTER 11: VERIFYING AND VALIDATING THE TASK MANAGEMENT + TIME ESTIMATION MODEL

11.1. FINAL ITERATION OF BASELINE MODEL AUGMENTATION: TM+TE

The complexity of the ATM environment and limited sensitivity to time estimation illustrates the difficulty associated with implementing an empirical model into a complex operational environment without fully representing the strategy that human operators may use to schedule their time critical tasks by representing both the task management strategy employed and the operator’s estimates of time available. Task management was manipulated as described in Chapter 8 where, the number of tasks in the queue before which non-critical tasks are shed was reduced from five (in the baseline model) to zero (in the TM model). Since that model does not allow multiple non-critical tasks in the queue, it caused the operator to manage tasks opportunistically, rather than strategically. That is, planning was limited and the environment solely drove decisions. Reducing the queue length in the model caused the model to first perform those tasks, such as handing off the flashing datablock, with the smallest time available in the window.

In this chapter the results of the final modelling iteration are presented – augmenting the baseline model with both TM and TE, as illustrated in Figure 68. This augmentation is referred to as the TM+TE augmentation. To perform the TM+TE augmentation, the Time Estimating Box (centre box) in Figure 6 was augmented as in the preceding chapter. The resulting closed-loop model caused the extent of under-estimation of time passage to increase as workload increased. In addition, this model iteration incorporated the same TM model as presented in Chapter 7 and Chapter 8, to better reflect the shift to opportunistic scheduling that ATCo’s perform when they are busy.
11.1.1. Measure 1: TM+TE Workload

In this section, the validation measures and techniques described above were applied to the TM+TE model augmentation.

11.1.1.1. Mean Workload For TM+TE Model

Figure 69 illustrates the average workload per experimental condition for (a) the FEWS adjusted data, (b) the baseline model, and (c) the TM+TE augmentation predictions (b).

There was a significant 2x3 interaction $F(2,40)=42.63, p<.001$. As with the other model iterations using the workload variable, Table 23 shows significant differences in workload between each traffic load condition for the FEWS adjusted ($F(2,40)=57.07, p<.001$), and the TM+TE augmentation model ($F(2,40)=7.48, p<.001$), except for the 133 DL to No DL comparison for both the FEWS and the TM+TE augmentation data. This lack of significance in the model could be due to the fact that the DL handoff tasks under nominal conditions possesses very similar workload No DL nominal handoff tasks or it could be due to the high number of...
missing data points in the FEWS data (as shown in Figure 70). These data illustrate that the model adequately represented the FEWS data and that the modelled environment drove the operator’s workload values in a manner consistent with the FEWS. Contrary to the other model iterations, no significant differences existed between the TM+TE model’s 133 DL and the 133 No DL workload output. This is consistent with the FEWS data.

| Table 23. Significance Tests for FEWS Adjusted, BL and TM+TE Model Workload by Traffic Level. |
|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| FEWS (Adjusted) | 100 DL – 133 DL t(21) = 3.82*** | 100DL-133 No DL t(21) = 4.32*** | 133 DL-133 No DL t(21) = 1 |
| Baseline | t(829)=4.36*** | t(829)=8.46*** | t(829)=3.30*** |
| Baseline with TM+TE | t(829)=4.57*** | t(829)=7.05*** | t(829)=1.90 |

* p<.05, **p<.01, ***p<.001

11.1.1.2. Workload Trend Output For TM+TE Model

Figure 70 presents the individual workload ratings from the FEWS data (upper graphs) and the model predictions (lower graphs) for each traffic load condition. It can be seen that the FEWS and the model produced similar workload trends in the 100 DL (Figure 70a) and 133 No DL (Figure 70b) conditions. As shown, workload was maintained at the low-end of each respective scale – recall that the FEWS workload was measured on a 10-point scale, and the model workload was measured on a 7-point scale. Also workload increases and decreases in the FEWS output were mirrored in the model output. Unfortunately, the 133 No DL condition (Figure 70c) in the FEWS simulation possessed a significant number of missing values, which makes drawing conclusions about the comparison of FEWS and model trends difficult in that condition.

Figure 70. FEWS (top) Compared to TM+TE Model (bottom) Workload Output, for (a) 100 DL Condition, (b) 133 DL Condition, (c) 133 No DL Condition. 
Notes - FEWS possessed missing responses. There were 18 responses for (a), 15 for (b), 9 for (c).
11.1.2. Measure 2: TM+TE Queue Length

Recall from Section 6.2.2 that queue length is technically a verification measure because there is no corresponding FEWS data; however it is included in the set of validation measures because it is used to support the validation process. It can be seen in Figure 71 that the maximum task queue increased as the traffic load increased from 100% DL to 133% DL but the queue length is not different in the 133% No DL. The lack of an increase in the 133% No DL condition suggests that the model was not queuing more tasks due to the high ATCo workload levels. Rather, the model is scheduling the critical tasks to be released immediately upon their entry into the queue (i.e. the model is not waiting for an optimal time to complete the tasks). The same pattern of results was observed with mean queue length variable (see Figure 72). The queue length increased significantly when the traffic level increased from 100% DL to 133% DL ($t(829) = 4.05, p < .001$) and to 133% No DL condition ($t(829) = 4.05, p < .001$). Contrary to the previous model iterations, the TM+TE model queue length is not increased significantly in the 133% traffic condition with No DL as compared to the 133% DL condition, $t(829) = 1, p >.05$. This suggests that there is an interaction between the TE model and the TM model that enables the tasks to be completed without delay, likely the estimation that there is no time available to complete the required handoff task. 

![Figure 71. Maximum Number of Tasks to be Services in the TM+TE Model Queue as a Function of the Experimental Condition.](image1)

![Figure 72. Average Number of Tasks to be Serviced in the TM+TE Model Queue as a Function of the Experimental Condition.](image2)
11.1.3. Measure 3: TM+TE Receive Handoff Duration (RHD)

11.1.3.1. RHD t-test and Results for FEWS versus TM+TE Model

Figure 73 shows the data associated with the RHD measure produced from the FEWS simulation and those RHD predictions generated by the baseline model, the TM, the TE and the TM+TE augmentation models as a function of the three experimental conditions. As can be seen in Figure 73, the model RHD times are much shorter than those from the FEWS. This has resulted because the model still did not represent the manner that the ATCo’s were operating in the FEWS HITL simulation. It can be seen however, that among the various model manipulations, the TM+TE model brought the RHD times closer than any of the other model manipulations, although still significantly different that the FEWS. Table 24 provides the FEWS, baseline, and TM+TE augmentation t-test results and illustrates that no significant differences for the 100% condition between the FEWS and the baseline model or the TM+TE augmentation model. A significant difference did exist in the 133% DL and 133 No DL condition RHD times depending on whether the RHD time was generated by the FEWS or by either of the models.

![Figure 73. Average RHD Times from FEWS and the Four Model Iterations.](image)

Table 24. Comparison of FEWS, Baseline, TM+TE Model t-tests.

<table>
<thead>
<tr>
<th></th>
<th>100 DL</th>
<th>133 DL</th>
<th>133 No DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FEWS – Baseline</td>
<td>t(76)=0.85</td>
<td>t(99)=2.07*</td>
<td>t(80)=5.51***</td>
</tr>
<tr>
<td>FEWS- TM+TE</td>
<td>t(76)=0.41</td>
<td>t(99)=1.21</td>
<td>t(80)=3.30***</td>
</tr>
<tr>
<td>Baseline - TM+TE</td>
<td>t(88)=0</td>
<td>t(99)=.02</td>
<td>t(80)=.76</td>
</tr>
</tbody>
</table>

* p<.05, **p<.01, ***p<.001

Plotted with +/- 1 SE
Chi-square analyses were conducted to determine if the distributions of RHD times from the TM+TE model and the FEWS observations came from the same population. As with the previous model iterations, data were put into three bins of approximately equal size: 0 – 10s, 11-30s, and greater than 31s.

11.1.3.2. RHD Chi-Square Test and Results for FEWS versus TM+TE Model

The $\chi^2$ test conducted on the RHD times for all three traffic levels suggested that the frequency distributions from the baseline model with TM+TE augmentation and the FEWS observations did not come from the same population (see Table 25). These results suggest that the TM+TE augmentation model did not faithfully represent the human RHD data in any of the three load conditions. The baseline and TM+TE models were from the same distribution in the 100 and 133 DL conditions while they were not from the same distribution in the 133 No-DL condition.

Table 25. Chi-Square and Correlation Analyses of FEWS, Baseline, and TM+TE Model RHDs.

<table>
<thead>
<tr>
<th></th>
<th>FEWS-Baseline</th>
<th>FEWS –TM+TE Model</th>
<th>Baseline – TM+TE Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>$\chi^2$</td>
<td>Chi-square</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>100DL</td>
<td>$\chi^2(2,N=122)=4.96^{***}$</td>
<td>0.21</td>
<td>$\chi^2(2,N=122)=4.96^{***}$</td>
</tr>
<tr>
<td>133DL</td>
<td>$\chi^2(2,N=163)=3.69^{***}$</td>
<td>-0.15</td>
<td>$\chi^2(2,N=163)=55.69^{***}$</td>
</tr>
<tr>
<td>133NDL</td>
<td>$\chi^2(2,N=154)=38.98^{***}$</td>
<td>0.12</td>
<td>$\chi^2(2,N=154)=28.07^{***}$</td>
</tr>
</tbody>
</table>

* $p<.05$, ** $p<.01$, *** $p<.001$

In summary, the TM+TE augmentation model did not adequately represent human data as collected in the FEWS simulation. Correlations were conducted to provide insights into the strength of the relationship identified by the $\chi^2$ test and are provided in Table 25.

A more detailed examination is therefore required to explore the reasons for the deficiency in the model’s performance. Whereas the modelled RHD times were significantly higher than the FEWS times in the 133 DL condition, the absolute difference between the means was 6s, or 28%. However, the TM+TE augmentation model overestimated RHD in the 133 No DL condition by a greater amount, with an absolute difference of 14s, or approximately 33%. The reasons for the greater break down in the model’s performance can be captured by exploring the histogram of the individual data points of the FEWS and the baseline model RHD, as described next.
11.1.3.3. RHD Histogram of FEWS, TM+TE Model

Figure 74, Figure 75, and Figure 76 present the RHD times in a histogram for the FEWS and with the TM+TE model in each traffic load condition. As can be seen, for the 100% DL condition (Figure 74), the large majority of RHD times occurred early in the handoff window of opportunity, before 90s elapsed, with the greatest proportion occurred early in the handoff window of opportunity, before 60s elapsed, and mostly within the first 30s. While the ATCos were striving to complete the handoff within 60s, it was also apparent that few outliers existed with response times above 90s when the ATCo did not correctly estimate the tasks to be completed in the window of opportunity and failed to complete the handoff in the required time. The right panels of Figures Figure 74, Figure 75, and Figure 76 show the RHD times produced from the baseline model. As can be seen, the trend of RHD time was very similar to that of the FEWS simulation data for the 100% DL. For the 133 DL condition, there were more early responses in the FEWS data than the model. The RHD times produced from the model in the 133% No-DL condition showed a markedly different pattern than the FEWS data. The baseline model produced RHD times that were later than the FEWS times. In the model, the large majority of response times occurred later in the handoff window of opportunity, with approximately 15% of the responses occurring after 120s have elapsed. Contrary to the FEWS simulation data, which had 88% responses before 60s, 44% of the responses in the model occurred before 60s elapsed. Consequently, to identify the reasons for the differences between the FEWS data and the model performance the similarity between the environment models is considered using histograms to examine the pattern associated with the data. The visual trend shown in the histograms below, are similar to those observed in the correlations outlined above in which the model accounted for more variance in the 100% condition than either of the 133% conditions.
11.1.4. Measure 4: TM+TE Handoff Window Open Time

To explain the differences between the FEWS and the model RHD performance, an accurate understanding of the environment model was required. Figure 77, Figure 78, and Figure 79 illustrate the TC graphs for the window open times, where the relationship between the FEWS and the TM+TE model’s window open times. Recall that lines that slant down to the left indicate that the window opened earlier in the model than FEWS, and lines that slant down to the right indicate that the window opened later in the model than FEWS. Lines that crossover
indicate aircraft window open times that occurred in a different order in the model than FEWS. (Note that in some cases, crossovers span multiple aircraft, so a crossover may involve more than one aircraft). The number of model window open times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the window open times were the same or different between the model and FEWS are presented in Table 26. As can be seen, approximately 54-78% of the aircraft window open times were in a different order in the model as compared to the FEWS. This occurred because the baseline model represented an optimal ATCo, however the actual FEWS ATCo may have made decisions to alter the speed or altitude of the aircraft prior to the handoff, thus altering the window open time.

Table 26. Number, Relative Timing and Order for FEWS and Model Window Open Times.

<table>
<thead>
<tr>
<th></th>
<th>Window Opened Earlier or Same Time in Model than FEWS</th>
<th>Window Opened Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>11</td>
<td>33</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>133% DL</td>
<td>11</td>
<td>41</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td>133% No DL</td>
<td>18</td>
<td>23</td>
<td>9</td>
<td>32</td>
</tr>
</tbody>
</table>

Figure 77. TC Graph for FEWS and TM+TE Model Window Open Time 100% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence.
Figure 78. TC Graph for FEWS and TM+TE Model Window Open Time 133% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence

Figure 79. TC Graph for FEWS and TM+TE Model Window Open Time 133% No DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
The TC graphs illustrate that the order (as reflected by the crossovers) and the onset times between the FEWS and the model are not precisely the same. While the overall data suggest that the 133% No DL condition produced the greatest differences in RHD times, this effect cannot be attributed primarily to the time that the window opened, since the present analysis suggests that the number of window open times that differed in order from the modelled and actual aircraft were fairly consistent across traffic conditions. The cause of the differences was therefore sought in the following analysis.

11.1.5. Measure 5: TM+TE Adjusted Handoff Complete Time

As with the previous model iterations, the handoff complete times were adjusted to account for differences between the model and FEWS window open times. The number of model handoff complete times that were earlier than FEWS, later than FEWS, as well as the number of aircraft for which the handoff complete times were the same or different between the model and FEWS are presented in Table 27. As can be seen, approximately 46-61% of the aircraft handoff complete times were in a different order in the model as compared to the FEWS. The TC graphs in Figure 80, Figure 81, and Figure 82 illustrate the relationship between the FEWS and the adjusted baseline model’s handoff complete time.

Table 27. Number, Relative Timing for FEWS and Model Adjusted Handoff Complete Times.

<table>
<thead>
<tr>
<th></th>
<th>Handoff Completed Earlier or Same Time in Model than FEWS</th>
<th>Handoff Completed Later in Model than FEWS</th>
<th># Aircraft with Same Window Open Orders</th>
<th># Aircraft with Different Window Open Orders</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% DL</td>
<td>11</td>
<td>33</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>133% DL</td>
<td>21</td>
<td>31</td>
<td>20</td>
<td>32</td>
</tr>
<tr>
<td>133% No DL</td>
<td>9</td>
<td>32</td>
<td>22</td>
<td>19</td>
</tr>
</tbody>
</table>
Figure 80. TC Graph for FEWS and TM+TE Model Adjusted Handoff Complete Time 100% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence

Figure 81. TC Graph for FEWS and TM+TE Model Adjusted Handoff Complete Time 133% DL.
Notes - Vertical Lines: FEWS time = Model time, Lines slanted down right: FEWS time earlier than Model, Lines slanted down left: FEWS time later than model, Crossing lines: each crossing line = 2 a/c out of sequence
It appears that the TM+TE model did not succeed in bringing the model’s RHD performance closer to the FEWS than the baseline model alone. Both the baseline and the TM+TE model predictions of RHD remain significantly different than those produced by the ATCos in the FEWS experiment for the 133 No DL condition. There did not appear to be any added benefit to the RHD prediction by both TM and TE together, although the TM+TE manipulation did bring the mean RHD times closer than any of the other model manipulations alone.

11.2. CONCLUSION OF TM+TE AUGMENTATION MODEL VALIDATION EFFORT
CHAPTER 12: A DISCUSSION OF MODEL PERFORMANCE

12.1. MODEL PERFORMANCE REVIEW

An iterative model development – model validation approach was undertaken on a baseline HPM of complex human behaviour. This iterative validation process was done to identify the model component that impacted the performance of the model. Was it the TM or the TE or were both required to cause the model to perform closer to the HITL data? Only by fully describing the model and the results in detail, can the reader ascertain the value of the contributing factors and of their success.

The baseline model of ATCo behaviour was validated by comparing its outputs to data from actual controllers operating in a complex HITL ATC simulation, on the basis of multiple measures: workload, RHD, handoff window open time, and handoff complete time. As a consequence of the unsatisfactory outcome of the effort to validate that model, the baseline model was augmented to include a task management model (based on the COCOM model) and its performance relative to the ATCo HITL performance was evaluated. The next iteration involved the proposal and development of a time estimation model. The final iteration was the combination of the task management and the time estimation model.

A model analyst that simply performed the last iteration without having undertaken the iterative validation phases as part of the current effort may have incorrectly concluded that the TE portion contained within the TM+TE augmentation possessed the desired effect, when in fact it appears that the TM seems to have had a much larger impact on the model’s performance in this complex operational environment. In addition, relying on a single aggregate measure such as the RHD in this case might cause an analyst to reject their model outright, when there are in fact portions of it that are valid (e.g. the environment – the window open and handoff complete times). Including the other variables of window open times, handoff complete times, workload, and queue length gives the model analyst more confidence that the model is operating correctly. The TM, TE, and TM+TE model iterations will now be discussed.

12.2. THE TASK MANAGEMENT (TM) AUGMENTATION

The TM augmentation included a task management allocation mode of operation that replicated
the ‘conservative bias’ (Edwards, 1982; Boudes & Cellier, 2000) performance of real, experienced human operators, which appeared to bring the RHD times closer to the performance of the ATCo in the FEWS simulation. The TM augmentation brought the performance of the baseline model closer to the performance of the FEWS human performance. When the ATCo operate in the opportunistic model they are behaving more conservatively because they are clearing off the schedule as quickly as possible in order to reserve as much time as possible to face unforeseen events that could compromise the safety of the system’s operation.

12.3. **THE TIME ESTIMATION (TE) AUGMENTATION**

The TE model was formally verified using a “generic” model of an aircraft “descent” scenario. The generic model illustrated that when the ATCo’s were tasked to estimate the opening of a window of opportunity at 60s into the future, their estimate of when this window was to open and therefore, when the ATCo was required to begin the descent task, was degraded when accounting for the human’s underestimate of $T_A$ – a common behavioural occurrence when humans face increased workload. This verification exercise showed that when the model considered operators’ non-linear bias in time estimation, the mean response error shifts from zero to a negative value. This exercise verified that the TE model operated as intended (in the sense that it degraded the estimate of time according to the Brown and Boltz (2002) and Boltz (2005) data). As such, it could be useful to consider in a complex HPM when the modelled human is required to generate a task schedule and carry out this task schedule in an operational environment. In the model of the FEWS environment, however, the adjustment of the baseline model to include perceived passage of time instead of clock time, especially during high workload periods, did not improve the model within statistical significance. Presumably, the improvement was not seen, because TM had a larger impact than TE.

While the ATCo environment did not prove to be a sufficient test case for the TE model, most likely because of the strategy shift exhibited by the controllers, the TE model should nevertheless be useful in other environments where time estimation and workload interact to drive a schedule. For example, a dispatch call centre would need accurate estimates of time available, required times and onset times to properly sequence calls for the current levels of staffing. Other examples include future time-based operations in surface transportation, where the driver of a car will be told to arrive at a pre-specified time to cross an intersection (e.g. an
event that will be common with automated highway systems). Time-based operations such as these are anticipated to require the use of the TE model augmentation.

Domains that could benefit from the TE model in its current form include tasks that utilize time sequences as opposed to space estimates, as well as domains that possess equally severe consequences associated with one of the responses of the operator. While the receiving ATCo does not necessarily benefit from an early handoff because of the possibility of having to relinquish control of the aircraft that they have accepted, the cost function of additional work is weighed against the safety buffer. ATCo’s are trained to weigh the safety margin more heavily than the added work associated with having to regain control of the aircraft. Domains that do not possess this kind of weighing (e.g. flight deck 4 D trajectory/navigation) will be more receptive to a TE model.

12.4. **THE TM+TE AUGMENTATION**

The TM+TE model did not produce a significant increase in workload with the change from 133 DL to 133 No DL conditions, similar to the actual FEWS data. The other model iterations (baseline, TM, and TE) produced results that were counter to the human data, suggesting that they did not represent workload correctly. It appears that the TM+TE is a more accurate representation of how operators use anticipated workload and $T_A$ overestimates to schedule their tasks in complex environments. By adopting this strategy, the operators maintain a more stable level of workload because they are scheduling their tasks early (because they anticipate that more time has passed than has actually passed), and not allowing as many critical tasks into the queue.

The fact that the TM augmentation model succeeded in producing RHD times that were more in line with the FEWS data than the TE augmentation model suggests that the breakdown in the baseline model during the high workload scenario was due to the baseline model’s inability to adequately represent the ATCo’s task scheduling strategy rather than the model’s inability to account for biases in time estimation. The low workload conditions (100% DL and 133% DL) were largely unaffected by the TE augmentation because these conditions did not possess sufficient workload to drive the TE augmentation algorithm in terms of the scheduling sequence, particularly when one considers the burst of early responding that occurred. When the baseline model was augmented to include both TM and TE, the model’s overestimate of mean RHD for the 133% No DL condition was reduced from that seen with the baseline model;
however, it was still significantly different from the FEWS model. A possible reason for this increase in performance was that the baseline model did not represent the human operator in enough detail. While the present TM+TE did a better job at representing the FEWS data, significant differences remain. There was no evidence of an interaction between TM and TE models that improved the performance of the model.

The fact that this could not be statistically validated could be due to the fact that in the real-world application of air traffic control, there is less emphasis on the importance of the time estimation task. Rather, the ATCo was required to complete the operations that are characteristic of the ATCo in the real world – managing their airspace, communicating (between ATCo’s and ATCo to aircraft), maintaining separation standards between aircraft, metering aircraft, handing aircraft off to adjacent sectors, resolving conflicts, and all of the tasks that are completed by an ATCo. In reality, the ATCo is required to maintain a safe airspace at all costs. In the handoff task, the conservative bias results in the ATCo handing the aircraft off as soon as it begins to flash, a mechanism that was included in the TM+TE augmentation. The ATCo engaged in this behaviour because, even though the FAA rules state that the ATCo needs to estimate the passage of 60s-120s once they notice the flashing aircraft datablock, they prefer to have a buffer of time to deal with unexpected events. In fact, while the FAA standard operating procedures mandate 60-120s before sector cross while reality dictates that the ATCos hand aircraft off as soon as possible, reserving as much time and space as possible to resolve unforeseen events - V. Battiste, ret. ATCo, personal communication, August 23rd, 2007). The true cost function that needs to be modelled is the degree that the ATCo balances this expectancy. The expectancy cost function could be implemented as a ratio of estimated $T_A$ to estimated $T_R$ and future research building off of the current research should examine the feasibility of creating this relationship.
The goal of Section 4 is to provide a discussion and interpretation of the model results and to highlight contributions to the field of both model validation and human performance modelling of complex systems.

Specifically, Chapter 13 reviews the multiple measures used to determine “validity”, the degree to which validation can be improved by increasing the resolution of the data upon which a model is validated, the model’s hidden parameters that may influence the model’s performance, and the fact that there is limited data to judge real world task validity.

Chapter 14 concludes with the significant contributions of the present effort validating complex HPMs, and the areas that can be improved upon in future research on modelling task management and time estimation.
CHAPTER 13: DISCUSSION

As outlined in Chapter 3, at the conclusion of the AMBR project, a project that used some of the world’s pre-eminent cognitive modellers, “human behaviour representation validation is a difficult and costly process [and] most in the community would probably agree that validation is rarely, if ever done (Campbell & Bolton, 2005, p. 365).” Campbell goes on to point out that there is not a general agreement on exactly what constitutes an appropriate validation of a cognitive architecture, because cognitive models are developed for a wide variety of reasons and that there is a correspondingly wide set of validation (and evaluation) objectives and metrics and associated methods associated with this multitude of reasons for the model’s development. A lack of established benchmarks and criteria exacerbates this problem.

The recent proliferation of human-system models has resulted in highly complex human behaviour models being used to generate predictions of operator performance within increasingly complex operational domain (e.g. process control, C2 / C3 operations, aircraft, ATC, and NAS operations, etc.). This proliferation is certain to continue along its growth path in the foreseeable future as computer technologies increase and the software implements more accurate representations of the human-system relationship. Many of the models that are developed for system predictions have undergone some degree of verification and validation. However, creating valid behavioural models of a human is a challenging endeavour, particularly because of the complexity of human behaviours, which are further heightened when integrating multiple models that comprise the system. Assumptions made for one submodel may interact with other submodels and may invalidate the system prediction. As a result, it is vital that the complex human models that are used to generate predictions of human-system performance be designed and validated in accordance with a principled approach.

The discussion in this chapter focuses on validation issues in general, as exemplified by the particular approaches undertaken in the present complex model development program, and raises a number of the difficulties in validating complex behavioural models. Insights into the requirements for fully validating a generalisable closed loop model of a complex system, with the ATC example being the obvious case of interest here, are highlighted, including the necessity of using multiple human performance measures for validation, the resolution of the variables used for validation, the degree to which a model should be developed for model generalisability,
recognizing hidden parameters, and the problem associated with limited availability of human data upon which to compare the model output.

13.1. **MULTIPLE HUMAN PERFORMANCE MEASURES**

Many modelling efforts have claimed to produce validated models, but have made this claim using a single measure of performance, such as only workload, only eye tracking performance, small problem solving performance, etc. (DMSO, 2001; Foyle & Hooey, 2008\(^\text{21}\); Pew, Gluck & Deutsch, 2005). Validating the model using a limited number of validation measures (often only one measure) allows model developers flexibility with respect to the manipulation that will be made to the model to get the model to perform consistently with the input data. It is often quite easy to tweak a model to perform well on one measure, while sacrificing the validity of other measures. When the model analysts change a model’s parameter, they do not examine the performance of the integrated representation of the model, rather they look at the effect of the individual parameter that they tweaked. While this is an appropriate validation process to follow for some small, non-integrated models, it is advisable that the more integrated and closed-loop HPMs conduct validation efforts use multiple human performance measures.

Prior to the commencement of this research effort, the model taken here as the ‘baseline’ had been determined to be a valid representation of the operations of ATCo’s interacting with aircraft using the FEWS display technologies and other pieces of advanced automation. However, those earlier validation efforts were based on a single measure - specifically, the number of aircraft data blocks travelling along given flight paths in the generic airspace. While this was an appropriate first step, it is arguably a verification step rather than a validation step, and in any case should be considered an incomplete validation effort because it failed to assess how the model represented the behaviours of the ATCos. The measures used to judge these behaviours and subsequent model validity were expanded to encompass measures of workload, queue length (used as a verification measure to support validation), RHD time, handoff window open time, and adjusted handoff complete time, all of which collectively provided a more complete picture of the validity of the model.

\(^{21}\) Foyle & Hooey, 2008 reported on the modelling and validation effort undertaken by 5 modelling teams. The modelling teams were left to determine a “suitable” validation for the models they developed.
Workload and queue length were included to demonstrate the importance of the closed-loop representation that exists between the environment and the human operator behavioural models. The analysis of workload revealed a significantly increasing workload trend due to increasing traffic loads both in the FEWS study and in the outputs of each of the models. The second measure, queue length\(^{22}\), was used to demonstrate the importance of the workload on the task schedule. The task queue length was used to determine the tasks yet to be completed. For each model iteration the analysis revealed that the three experimental traffic load conditions appeared to have a significant effect on performance of the models. These measures were quite useful to determine whether the model was using “experiences” gained from the simulation environment to guide future behaviours.

The third measure, the RHD, defined as the time difference between the flash onset of the aircraft datablock and the receipt of the flashing datablock by the receiving ATCo, was selected as a measure to demonstrate the degree of similarity at the aggregate level between the mean RHD of the FEWS and the models. No significant differences in the RHD values existed for the 100% traffic condition for any of the model iterations, but the model-predicted RHD values began to diverge from the FEWS for the increased traffic level condition (133%), particularly for the No DL case. As part of the process to identify the underlying reason for the discrepancy, two more measures, window open and window closed time, were included.

The window open time was defined as the simulation (clock) time of the flash onset of the aircraft datablock. Comparing the model window open time to the FEWS open time provided a measure of how well the model represented the environment – that is the movement of aircraft through the sector. Because the window open times shifted based on the activities that the operator undertook, this measure was used to assess if the sequence of tasks differed between FEWS and the model. Only a subset of the entire simulation was used for the comparison because only those aircraft that could be matched between the FEWS and the model could be used. Once the aircraft were matched, the TC measure was used to analyze the timeline and performance onset times. This method for evaluating the measure provided a clear approach to examine the onset of the various environmental variables and subsequent human actions. The

---

\(^{22}\) Recall from Section 6.2.2 that queue length is technically a verification measure because there is no corresponding FEWS data; however it is included in the set of validation measures because it is used to support the validation process.
analysis of the window open times revealed that the four models diverged from the FEWS simulation, with the majority of the models’ window open times occurring later than the FEWS.

The window open measure is one that does possess some degree of error associated with it because the window open times occur based on other ongoing tasks. Again, if this measure was the only measure used to determine the validity of the model, it is likely that the model validation effort would have concluded that the models were not validly representing the FEWS simulation. In other words, although this is an important measure, it does not tell the entire story within this integrated modelling effort. However, when taken together with the third and the fifth measure (to follow) a picture of the manner that the model is performing begins to emerge.

The fifth measure, the adjusted handoff complete time, was a measure of how well the model represented the actions of the ATCo in regards to the receive handoff task. This was assessed to determine if the sequences of handoff accept actions differed between FEWS and the model. In order to remove any differences due to window open times, and thus any bias in assessing the handoff complete time measure, an adjustment was made to each handoff complete time measure. Improving the rigour of the suite of tests through this adjustment demonstrates that creating an additional variable (and thereby using multiple measures) can add greatly to the value of the validation process of a model.

In summary, a systematic validation of the models against the FEWS HITL data has been provided using multiple measures to determine model validity.

13.2. IMPACT OF RESOLUTION OF ANALYSIS ON VALIDATION

A range of validation resolutions, from the aggregate level (i.e., comparing mean hand off times to a more fine-grained level (i.e., TC graphs and scatterplots of individual handoffs) was used in the current research program. The relative merits and the appropriateness of each approach are discussed next.

13.2.1. Use of Aggregate Data

In the domain of human performance modelling, it is common practice to validate models using aggregate analyses, as was discussed in Chapter 3. Using aggregate data provides an analyst a
very rapid way of obtaining an answer from a HPM, in much the same way as is done with empirical data (Campbell & Bolton, 2005; DMSO, 2001).

13.2.2. Use of Fine-Grained Data for Validation Efforts

Validating a model at a finer-grained level of analysis requires modelling human behaviour at a much lower level. Given that this adds substantially to the complexity of the model development effort, especially when modelling complex systems such as the ATM, it is useful to consider the added benefit from this level of analyses to the validation process, and for what purposes and research questions this increased level of fidelity will bring to the modelling effort. The following six issues that advocate for the inclusion of fine-grained analyses were identified throughout the current research project.

Models developed for the purpose of evaluating operator procedures and tasks often benefit from fine-grained analyses. In the present modelling effort, the finer-grained TC measure analyses for the window open time revealed that the model and the human ATCos processed the aircraft in different sequences. This finding would not have emerged from the aggregate analyses alone. The present model validation exercise used the finer-grained analyses (TC measure) to understand and interpret the results obtained from the aggregate RHD analyses. For example, the original purpose of the models developed for the present research was to explore the effect of ATCo technology (DL) as a function of different traffic conditions. The aggregate measures would have be sufficient to compare mean RHDs with and without DL, but were unable to uncover the human behaviour associated with task management and time estimation that contributed to those differences, as was accomplished with the finer-grained TC measure analyses. The TC measure was used to tease out the effects on performance within the airspace system that were due to the code underlying the model, due to the environment, or due to the modelled human operator’s performance.

Fine-grained analyses may also help identify trends or underlying functions in the data that may be overlooked in the aggregate analyses. The fine-grained analyses can be used to identify individual spikes in workload that would otherwise be washed-out when averaged across an entire model. Workload spikes are often more informative in identifying causes of operator workload overload than can be obtained from the aggregate measures, this is particularly true for models that use workload to determine task schedules.
Estes (2002) reported while aggregate data can help bring out trends in the data by reducing unbiased statistical error, it can also distort the data. For example, averaging data may give the impression of a process or function that is only seen in the group data, while the individual data imply different processes. In Estes’ classic example, each individual subject’s data may be in the form of a step function, but with each individual having a different value where that step occurs. When averaging across subjects, an ogive, or S-shaped psychometric function, appears in the group data; however no individual subject’s data reflected such a function (Foyle & Hooey, 2008).

The present validation effort has extended an approach that uses aggregates in combination with the fine-grained task level performance and as such, presents a refined approach to validating the implementation of an empirically-based model. This research: a) demonstrated the advantages of fine-grained analyses over aggregate measures alone, b) offered examples of how fine-grained measures were complementary to aggregate measures and c) developed a new analysis techniques (TC measure) to analyze fine-grained data.

13.3. **MODEL GENERALISABILITY**

As defined earlier, model generalisability refers to the degree to which the model can be applied outside of the specific application domain for which it has been created. Generalisability is vastly increased when models are theoretically grounded. The precision achieved when accurately modelling every aspect of an environment and triggering exact behaviours for a specific context usually reduces the generalisability of the model. This is known as overfitting the model to the data. Empirical data always contains an element of error. Models, and those that develop models, seek to reduce all forms of error (systematic and error variance). Models that have been overfit to the data are simply not generalisable to other domains, and are not scalable to larger or smaller operations when models. As Roberts and Pashler (2000) point out, a theory can “fit too much”; it can be closely fit by a similarly flexible theory making very different assumptions, and could be using an incorrect assumption while correctly fitting the data.

The present model development effort has created a generalisable time estimation model as evidenced by the power function model implementation, which allowed for predictions of the human’s estimate of time under conditions of low workload (baseline conditions) and under conditions of high workload (difficult condition). As a reminder, this model generalized the
Brown and Boltz (2002) data into an arguably verifiable and potentially valid power function which allows an extrapolation to be made to predict values of ‘Y’ (a human’s estimate of time) for any given value of ‘X’ (the actual time) and for any given value of ‘Z’ (workload).

13.4. **MODEL TRANSPARENCY**

*Model transparency* refers to the ability of the model developer/user to comprehend the relationship that exists among various models being used, the performance of the various models, which models are triggering in the model architecture, and whether a model is behaving as intended. Other researchers use different terminology, including *model traceability, model behaviour visibility, model verifiability, model validity* and *model interpretability* (Elkind et al., 1989; Foyle & Hooey, 2008; Gore, Hooey, Foyle & Scott-Nash, 2008; Gluck & Pew, 2005; Napierky, Young, & Harper, 2004).

In complex models of human performance, there are literally hundreds or thousands of parameters that impact performance. Furthermore, model “transparency” is a very large challenge for models that include representations of human cognition, since cognition is an internal mechanism that is difficult to observe directly. As such, it is often impossible to examine the effect of just one sub-model output, such as time estimation, on a particular measure. For example, with respect to the present example, since the modelled human operator didn't begin estimating time until he visually fixated on the aircraft datablock, visual scan and time estimation are intricately linked and cannot be separated.

When models are developed that integrate together a number of submodels, as exemplified by the current model development effort, the transparency of the model becomes paramount because it becomes very difficult to determine which model is active at which time in a scenario (Diller et al., 2005; Elkind et al., 1989). Model transparency also assists in clarifying the contextual triggers for the behaviours that are produced by the model, thereby increasing model interpretability. Without this level of insight into the model, and without an accurate understanding of the assumptions embedded in the model, results may be overstated. For the field to advance, it is critical that a comprehensive understanding of the mechanisms operating within a model be formalized, and therefore that there be sufficient transparency in the model’s operation. This formalization and transparency will increase the likelihood that assumptions are properly identified and noted in the model’s performance and that the correct model will be
chosen and used for the specific application. When models are transparent, the user can comprehend the model performance, thereby increasing their trust and confidence that the output from the model is in line with actual human performance. Conversely, when models are not transparent, the user cannot comprehend the model’s performance, will not have confidence, and will not trust that the model performed according to their expectation.

13.5. DATA AVAILABILITY FOR REAL-WORLD COMPLEX TASKS

The most definitive test of a model’s validity is to compare the output of the model to HITL simulation data (Law & Kelton, 2000). While this is the “holy grail” in terms of quantitative validation, there are many practical difficulties associated with acquiring the required output data from a HITL simulation or actual operational environment.

It is often difficult to conduct an extensive validation on complex real-world tasks for which only limited human data exist. Despite the fact that the data set for the current model development was very rich, it was originally collected for another purpose and as such, not all the behavioural measures that were collected had corresponding data in the HPM. For example, queue length, while a valuable measure for the validation process, had no corresponding data in the FEWS environment.

Validating a model based on a HITL simulation conducted for another purpose also posed a difficulty in the present research environment given the amount of missing data. The original premise was that the baseline model with TE augmentation might excel in very high workload conditions, as exemplified in Figure 47, but unfortunately the FEWS data set did not allow for this comparison. That is, upon analysis of the highest workload condition of 166% traffic level in the FEWS data set, it became apparent that the data set was missing excessive data points and therefore could not be used for any meaningful comparisons (and was not included in the current effort). Furthermore, even for the 133% No DL condition, there were substantial missing data points in the FEWS workload ratings given the demands of the task.

Following the discussion above regarding the value of fine-grained analyses it is important to note that, often, actual human-in-the-loop data may exist to validate at an aggregate level but not to support validation at the fine-grained level. For example, in conducting the fine-grained analyses of window open and handoff compete times in the present modelling effort,
only a subset of data that could be aligned between the model and FEWS could be included. This occurred because, once a model starts its series of required actions, it changed the occurrence of the individual handoff performance/behaviours, the measure of interest in the current examination. For example, the modelled controller changes the speed of the aircraft, which changed the time that the aircraft entered the new/next sector and therefore impacted the time that the aircraft was handed off, which subsequently impacts other ongoing tasks, e.g. possibly coinciding with conflict or metering tasks. This demonstrates the difficulty in comparing model output to HITL data.

Fully validating a complex HPM is a complex endeavour that is made increasingly difficult by the typical lack of a suitable data set with which to validate the model. Resultantly, validation efforts usually rely on small datasets, as exemplified in the present thesis. The challenge with the paucity of data is that a model analyst will undertake a large number of statistical tests on the small dataset. This increases the likelihood that the analyst might violate statistical robustness assumptions and dataset generalizability. The present thesis relied on a small number of workload scores and analysable receive handoff duration times from the FEWS to generate conclusions about the validity of the model’s predictions. The reliance on one ATCo Team might not have been representative of the real world, which in turn impacted the generalizability of the dataset and its validation effort.

Another problem with the HITL data is that it can be taken out of the HITL simulation’s context, thereby reducing the validity of the output, even though it provides a high correlation value. Roberts and Pashler (2000) suggest that good fits can be essentially meaningless because of the flexibility of many theories (the same data can be closely fit by a similarly flexible theory making quite different assumptions). This was found on a number of occasions throughout the NASA HPM simulation (Foyle & Hooey, 2008). This large scale HPM development effort enlisted human performance modellers at the beginning of the HITL simulation to identify the data and the variables necessary to conduct the modelling effort. Once the gaps in the research data were identified, a HITL simulation was conducted by NASA Ames Research Center (Foyle & Hooey, 2008) to obtain the missing data. These data were provided to the modellers, upon which they were to validate their respective model’s predictions of operator performance. It was determined throughout this effort that misperceptions and misinterpretations (and hence the potential for incorrect data manipulations) occurred with regards to the task that the pilots
completed in the HITL simulation, in spite of the considerable effort that went into documenting the HITL simulation (Foyle & Hooey, 2008). It was only with constant, iterative feedback among the modellers and the NASA HITL simulation team that it was possible to arrive at a successful quantitative validation output result (Foyle & Hooey, 2008).

The present thesis on model validation also experienced some challenges relating the self-reported workload measures collected in the FEWS simulation with those workload measures output from the model. The challenge surrounded whether the FEWS scale and the model workload scales were parallel scales which could be compared with each other, and how the data should be treated given the different granularity associated with the output. The FEWS generated workload queries every 2 minutes over 40 to 45 minutes of simulation time (for approximately 22 workload values). The model generated workload every event (for 830 values). In order to generate datasets for statistical comparisons, blocks of 38 data were averaged in the model to parallel the FEWS data. Even when this was done, it was discovered that excessive missing data existed in the FEWS as workload levels increased, which required additional dataset manipulations to maintain the 22 workload values to compare to the model. The current thesis used a very conservative treatment of the missing data by replacing the missing data with the 90th percentile value of the available observations taken over all conditions. (Another approach would have been to replace the missing data with the mean of the available observations.) An additional consideration was whether the average workload should be the average across the 4-workload channels, as used in the present thesis, or whether the average workload should be the average of the cognitive workload channel alone (Corker et al., 2003). These challenges will need to be addressed in future coordinated research among HITL simulations and HPM efforts.

---

23 Dependent on whether the FEWS simulation was allowed to run to completion.
CHAPTER 14: CONCLUSION

14.1. SIGNIFICANT CONTRIBUTIONS

A number of significant contributions to the field of human performance modelling have been provided by the research conducted in this thesis effort. These include the validation methodological approach, the TC measure, the TM model that integrates the three C’s and the COCOM structures to guide the model’s performance, and the TE model. The collection of elements that make up the validation process outlined in the current thesis is unique to the field of human performance modelling as they are applied to validate a complex time-sensitive task in an HPM.

14.1.1. The Validation Approach

This thesis has introduced and demonstrated a comprehensive iterative develop-validate approach for validating a complex, closed-loop model of ATC using multiple measures at varying levels of fidelity designed to provide a validation approach for time-sensitive tasks. A series of objective and quantitative validation measures were applied to assess the validity of a baseline model that was then carried through as model iterations were completed. The iterative approach enabled the assessment of the impact of each model manipulation to determine whether the model developed operated verifiably and validly. This comprehensive iterative develop-validate effort has culminated in a number of recommendations for validating complex models considered in section 14.2.

14.1.2. The TC Measure

A novel quantitative validation measure coined the TC measure was applied. The TC measure consists of TC graphs, which were used as a primary feature in each of a comprehensive develop-validate iteration of the model development effort to demonstrate whether the aircraft were sequenced in the same order in the model as in reality. The TC measure was used to tease out the cause of the differences in RHDs and served as a measure to address the challenge associated with validating non-observable complex human cognitive processes and their subsequent impact on human task performance in the time domain.
14.1.3. The Task Management Model

The TM model incorporated the Opportunistic control mode, which was based on Hollnagel and Woods’ (2005) and Hollnagel’s (1993, 1998) research, and served to drive the performance of the operator within the model according to the “conservative” bias (Boudes & Cellier, 2000; Edwards, 1982). The baseline model, augmented with the task management element, was shown to better represent ATCo performance particularly when workload was increased. The contribution therefore is that the opportunistic mode was shown to better represent the ATCo behaviour than the strategic mode. Representing the model in this way will allow the ability to test new procedures, technologies, and roles and responsibilities for future aviation concepts, providing the analyst with estimates of task timing, task management/scheduling, and workload.

14.1.4. The Time Estimation Model

The TE component, based on empirical data, created a power function relating time estimation to actual time and (low and high) workload conditions. This algorithm was then verified within the generic model and a validation effort was undertaken with the FEWS data set. For the field of human performance modelling to continue along its development path, a need exists to further the present research by continuing to develop and validate closed-loop models of the human operator as they engage in highly complex, time-sensitive environments.

14.1.5. The Normative Model Results

A discrepancy was illustrated between the normative model and the actual performance profiles of the FEWS operators as they performed their tasks in reality. This illustrated that either the normative model was operating incorrectly or the operators were not performing their tasks as specified by the normative model. Whereas the current thesis has focussed primarily on the former, it is interesting to speculate upon the possibility that the latter may be the source of the differences between the normative model and the human operators. Such a normative model could be used to develop future ATC operations and procedures and for ATC training purposes.

14.2. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

The present research effort has been fruitful, not only in terms of the process undertaken for developing a validated model of task scheduling and sequencing through the inclusion of time estimates that were based on workload, but also for laying the groundwork for additional, future
research. While most agree that validation remains one of the most challenging aspects of cognitive architecture research and development, the current thesis has accepted this challenge and attempted to extend a validation approach using multiple measures for cognitive model development.

Six recommendations for future research stemming from the limitations of the current research include: (1) workload projecting component of the TM model, (2) augment the TM model, (3) augment empirical data behind the time estimation model, (4) enhance the bivariate relationship, (5) expand the TE model to include $T_R$ and $T_O$, and (6) assess cross domain applicability. Each will be considered in turn.

14.2.1. Workload Projecting Component of the Time Management Model

The workload-projecting component was such that the model averaged workload over the next fifteen tasks that were scheduled to occur. Averaging workload values from a HPM results in a very large number of workload samples and therefore a very small overall workload value. This causes a closed loop model (that uses workload to drive behaviours) difficulties because workload may never reach the workload levels required to underestimate the passage of time. Future research should consider the maximum workload values in a single channel over a specific time segment in the future to drive the task schedule. For example, when workload in the cognitive channel spikes to 5.5 or greater then underestimate time passage and schedule tasks early.

Additionally, future research should build off of the current attempt at drawing a parallel between the workload output from the HPM and the HITL workload values collected empirically. HITL workload is collected at predetermined simulation events. Future research should consider taking a workload sample for only the period of time immediately preceding a modelled event, during the modelled event, and immediately following the modelled event. Any more workload values collected and used in the workload value will cause the workload parallel to be difficult to draw.

A third area for future research includes the interaction that workload projecting has with TM and TE and the combination of TM with TE. Workload has been a topic of interest for many years but there still appears to be lack of consensus on how workload impacts task management,
task scheduling and time management. Clearly, additional research is needed to better understand the workload concept and the way that it impacts task performance.

14.2.2. Augment Time Management Model

The TM model instantiated in the current modelling effort looked at the effect of opportunistic versus strategic control. The model implementation served largely as a proof-of-concept that the TM approach should be considered in the model and to show that the operator’s control mode does have a large impact on model performance. This research did not consider the scrambled and the tactical control modes, which may be more suitable to drive the performance of the TE model so future research should consider incorporating such notions. In addition, the TM model should be augmented to be more dynamically responsive to the task environment through a refinement of a time degradation function that interacts with time estimation and task schedule. One possibility is to include a ratio of estimated $T_A$ to actual $T_A$, which could then be used to drive the perception of when to schedule the commencement of the task and the subsequent schedule of task performance.

14.2.3. Augment Empirical Data of Time Estimation Model

The TE model was based on empirical data from a fairly simple laboratory time estimation task. To improve the model, time estimation data from more complex tasks characterized by higher workload such as those faced by ATCos or other C2 and C3 operators is required. To date, no such data was identified in the literature. Future HITL studies aimed at collecting these data are required and further modelling efforts need to use the data to develop empirically based HPMs.

The TE model in the current thesis mapped low and high workload from a very complex ATC environment onto the low and high workload from a simple laboratory experiment task. The low and the high workload conditions from the simple laboratory experiment may not be analogous to the low and the high workload conditions from the ATCo environment. Additional air traffic controller empirical research examining low and high workload conditions is needed to accurately populate the algorithms for the ATCo in the HPM.

Empirical research is needed to augment the model to account for the consequences associated with early or late task performance such that the TE model dynamically weighs the consequences of early versus late performance and strategically schedules the $T_O$ for tasks to occur within the environmental context and the window of opportunity (see also Section 14.2.6).
14.2.4. Enhance Time Estimation Bivariate Model Relationship

As outlined in Chapter 7, the bivariate relationship created in this thesis effort is limited, particularly for the higher workload levels as the current instantiation of the bivariate relationship becomes arguably less valid during periods of high workload. It could be that the mapping that was created between the model generated workload and the Brown and Boltz workload data was not appropriate for the task of ATM. Future research should refine this bivariate model, possibly by including a lower limit or threshold beyond which the bivariate model will not cross. This lower limit will need to be determined through the empirical research as outlined in section 14.2.3.

14.2.5. Expand Time Estimation Model to Account for \( T_R \) and \( T_O \)

The estimates of \( T_R \) and \( T_O \) were held constant in the present modelling effort because of the nature of the ATCo task, however they are likely more applicable in other application domains. The ATCo task did not require an estimate to be made of the \( T_R \) and \( T_O \) because the types of behaviours involved in ATCo operations. Expanding the TE model with the notion of \( T_R \) and \( T_O \) within an appropriate application domain will further exercise the model.

14.2.6. Assess Cross-domain Applicability

While the TM and TE models did not improve the validity for the current ATCo application, future research should assess the degree to which they may be valid for other domains, particularly those that utilize time sequences. In the ATC domain, the ATCos must weigh the option of handing an aircraft off early and potentially needing to regain control of the aircraft while it is still in their sector (but now under the control of the subsequent ATCo), against the safety margin of maintaining a safe flying environment. Domains that do not possess this kind of strategic weighting may be better suited to the Time Management Framework. An environment that may be considered is the space environment where astronaut’s behaviours and tasks require much more precise manual handling and deliberate movements. This could be used to exercise the TM and TE models and enable additional validation efforts that could utilize a multiple measure approach as exemplified in the current thesis effort.
CHAPTER 15: REFERENCES


Smith, and B. Peters (eds.) Proceeding of the winter simulation conference - Volume 1, pp. 1533-1540.


### APPENDIX A – COGNITIVE MODELS

Table 28. Summary of Cognitive Model Architectures, Common Uses and References.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Common Uses</th>
<th>Key References</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMPLE (Situation Awareness Model for Pilot in the Loop Evaluations)</td>
<td>Information processing, agent-based models, that began with the procedure oriented crew (PROCRU) model</td>
<td>Highly structured, standard procedures, driven by detected events and assessed situations</td>
<td>Baron, Muralidharan, Lancraft &amp; Zacharias, 1980</td>
</tr>
<tr>
<td>PROCRU (Procedure oriented Crew Model)</td>
<td>Derivative of the OCM that incorporates the execution of procedures in the context of manual control</td>
<td>Introduces the concept of expected net gain, a generalization of the performance index, as a means of predicting priorities among procedures to be executed</td>
<td>Baron, Muralidharan, Lancraft &amp; Zacharias, 1980</td>
</tr>
<tr>
<td>COGNET (Cognition as a Network of Tasks)</td>
<td>Theoretically based tools and techniques for performing cognitive task analyses (CTA). Designed to represent a range of individual operators' procedural knowledge and declarative knowledge.</td>
<td>Real-time, multi-tasking environments, including ATC and intelligent tutoring</td>
<td>Zachary, Ryder, &amp; Hicinbothom, 2000; Zachary, Ryder, Ross, &amp; Weiland, 1992</td>
</tr>
<tr>
<td>SOAR (State, Operator And Results)</td>
<td>A parallel matching, parallel firing rule-based system that represents both procedural and declarative knowledge.</td>
<td>Human central processing capabilities, such as learning, problem solving, planning, search, natural language and other HCI tasks, such as ATC, NASA test director, job-shop scheduling, car driving.</td>
<td>Newell, 1990; Lewis, 1997a; 1997b; 1996; Polk &amp; Newell, 1995; Miller &amp; Laird, 1996; Chong &amp; Wray, 2005; Wray &amp; Laird, 1997</td>
</tr>
<tr>
<td>EPIC (Executive Process Control)</td>
<td>Production-rule cognitive processor that contains parallel perceptual and motor processors; models system performance but lacks a theory of human performance</td>
<td>Psychological refractory period (PRP), dual tracking + stimulus response task, tracking + decision making task, verbal working memory tasks, computer interface menu search, and a telephone operator call-completion task</td>
<td>Kieras &amp; Meyers, 1997; Meyers &amp; Kieras, 1997; Kieras, Woods, and Meyer, 1997; Chong &amp; Wray, 2005</td>
</tr>
<tr>
<td>Attention / Situation Awareness Model (A/SA) – Salience, Effort, Expectancy, Value (SEEV) Attention Model</td>
<td>SEEV model of human attention, comprised of four parameters, Salience, Expectancy, Effort, and Value was modified to include noticing (N) and, thus, to create N-SEEV.</td>
<td>Model’s distribution of attention within the flight deck data, event detection latency, and duration of attentional neglect to illustrate that the model was a good fit to the empirical data. Model then applied to predict pilot responses to off-nominal events in future NextGen scenarios using a full cockpit layout with expected NextGen operational technologies and automation</td>
<td>Wickens, Goh, Helleberg, Horrey, &amp; Talleur, 2003; Wickens &amp; McCarley, 2008; Wickens, McCarley, Alexander, Thomas, Ambinder, &amp; Zheng, 2008; Gore et al, (2009); McCarley et al., (2009).</td>
</tr>
</tbody>
</table>
## APPENDIX B – INTEGRATED ARCHITECTURES

Table 29. Summary of Integrated Architectures, Common Uses and Key References.

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Uses</th>
<th>Key References</th>
</tr>
</thead>
<tbody>
<tr>
<td>General Purpose Discrete Event Simulation such as MicroSaint Sharp</td>
<td>An integrated platform that allows rapid model development through the use of flow charts and task networks. Sharp uses plug-in interfaces and object-oriented model development, allowing easy integration with other external software applications. Sharp uses a comprehensive visualization environment.</td>
<td>Aviation related tasks – Air Traffic Control task performance, aircraft numbers and flight paths flown.</td>
<td>Leiden, &amp; Kamienski, 2006; Leiden, Kamienski, &amp; Kopardekar, 2008</td>
</tr>
<tr>
<td>Cognet/iGEN</td>
<td>An integrated architecture between COGNET and iGEN, a visualization tool that assists in model development</td>
<td>Used in a submarine task to predict if a human operator would perform a certain task and the amount of time that the predicted task instance preceded the actual task.</td>
<td>Zachary, Jones, &amp; Taylor, 2002</td>
</tr>
<tr>
<td>Soar/iGEN</td>
<td>An integrated architecture between Soar and the iGEN visualization capability</td>
<td>Aviation flight deck navigation</td>
<td>Glenn et al, (2004)</td>
</tr>
<tr>
<td>Soar/EPIC</td>
<td>An integrated architecture between the Soar cognitive architecture and EPIC’s perceptual and motor processors</td>
<td>Perceptual – cognitive motor processing</td>
<td>Chong &amp; Laird, 1997</td>
</tr>
<tr>
<td>D-OMAR: (Distributed - Operator Model Architecture)</td>
<td>Represents goal behaviours (hierarchies), task instances, automatic behaviours completed in parallel, and multiple communicating agents</td>
<td>Equipment and procedure designs as part of human-machine system development in Command and Control (C2) and Command, Control and Communication (C3) systems, and aircraft and ATC operations</td>
<td>Deutsch, 1998; Deutsch &amp; Adams, 1995; IPME 1998; Laughery, 1999</td>
</tr>
<tr>
<td>DCOG</td>
<td>A modelling structure that represents local memory for actions. Applies D-OMAR architecture and assumption to agents representing skilful human behaviour</td>
<td>Represent environmental actions (e.g. aircraft actions such as velocity, altitude, etc).</td>
<td>Pew &amp; Mavor, 1998</td>
</tr>
<tr>
<td>IMPRINT (Improved Performance Research Integration Tool) / WinCrew</td>
<td>A task network modelling software tool that has been augmented to include models of operator workload</td>
<td>Military systems, including an Army tank crew size evaluation and a workload driver behaviour model.</td>
<td>Mitchell, 2003; Archer &amp; Lockett, 1997; Yow, 1999; Mitchell, Samms, Henthorn &amp; Wojciechowski,2003; Wojciechowski,2004</td>
</tr>
<tr>
<td>IPME (Integrated Performance Modelling Environment)</td>
<td>Incorporates psychological processes with physical task completion times for single and multiple operators. Uses micro-models of operator performance, including sensation and perception, cognition, task scheduling and motor outputs.</td>
<td>Team operations, team decision-making and auditory perception decrements</td>
<td>IPME 1998; Laughery, 1999</td>
</tr>
<tr>
<td>MIDAS (Man-machine Integration Design and Analysis System)</td>
<td>A 3-D rapid prototyping HPM, that links models of human anthropometry, biomechanics, and human cognition within a single environment</td>
<td>Facilitate the design, visualization, and evaluation of complex man-machine system concepts, crew stations and operating procedures.</td>
<td>Corker &amp; Smith, 1993; Gore &amp; Smith, 2006</td>
</tr>
</tbody>
</table>
APPENDIX C - HUMAN IN THE LOOP DATA FOR VALIDATION

Thirty two scenarios were conducted in the FEWS comprised of:

1. Training (11 scenarios)
2. FEWS concept x Task Load (3x3). ATCos controlled traffic at three experimental traffic levels for each of the concept designs.
3. R-side Conflict Probe Presence (2x2). ATCos controlled four additional scenarios at current and current plus 33% traffic levels with the R-side Conflict Probe Absent as a baseline for the Enhanced R-side FEWS concept.

For each of the task load levels, four simulation scenarios were created. These scenarios rotated under the automation and team configuration conditions to ensure that effects were due to the conditions and not due to differences between scenarios. Participants trained on 11 scenarios that included integrated use of the airspace and the Display System Replacement (DSR) emulation. At the end of training, participants had mastered the airspace and all of the equipment used in the experiment. Each training and experimental scenario lasted 45 minutes. Paper flight strips were not available. The operational environment triggering the behaviours of the human performance model were generated from a study conducted at the FAA Tech Center termed the Future Enroute Workstation Study (FEWS). This study utilized a generic representation of an airspace sector termed the Genera Center Air Route Traffic Control Center (ARTCC). This was a fictional airspace (ZGN sector) developed at the FAA Technical Centre that closely replicates the en-route environment. ZGN consisted of easily remembered fix names and simplified operating procedures (controllers control a high altitude feeder sector into GEN, there were reduced separation standards of 3nm and 1000ft vertical separation) and it was divided into two separate centre configurations to simulate an inter-facility operation. The Genera Center was based on projection of CONUS locations onto the ZGN airspace map. For example, if an aircraft was flying to the North-West, they found a VOR named WAS (Washington) and to the South-East they found a VOR named FLA (Florida). Controllers were required to complete a sequence of ATC activities such as situation monitoring, handoff, transferring communication, metering violations, weather, and altitude monitoring. A simplified example of the sector can be found in Figure 83.

![Figure 83](image-url)
Developing the Model Rationale – the Air Traffic Controller Environment

The familiarization with the airspace and the LOAs and SOPs used two controller stations equipped with a radarscope, a DSR keyboard, and either a trackball or an alternative input device. One high-resolution (2,048 by 2,048 pixel) monitor displayed the radarscope while another displayed either a D-side CRD or a second radarscope. The simulation pilots manoeuvred the aircraft and issued ghost controller commands. An Air Traffic Workload Index (WAK) device was mounted next to the displays within easy reach of the participant for input of workload ratings. A landline allowed inter- and intra-facility communications. A Keyboard Selection Device (KSD) and a CRD was available for use. A more detailed Figure of the controlled environment can be found in Figure 84.

Figure 84. Detailed view of ZGN airspace sector with fix, ascent, and descent waypoints.

This figure is included to provide the reader with an understanding of the FEWS environment and to enable to reader to see the two primary flight paths used in this simulation (the bolded lines both running North to South - one above R22 and the one above R18), the waypoints, the location that aircraft join/depart the flight path (illustrated by the ovals), and the sectors that surrounded the ZGN sector of interest. The outputs provided in Appendix B utilized the simulation environment illustrated in Figure 85. This simulation environment was also incorporated in the human performance model illustrated in Appendix E.
### APPENDIX D - EXAMPLE/SNAPSHOT OF ENVIRONMENT MODEL CODE

#### Table 30. Low Taskload (This Code Continues Through Line 2000 and Column AQ)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B772</td>
<td>272</td>
<td>MD80/P</td>
<td>272</td>
<td>MD80/P</td>
<td>272</td>
<td>MD80/P</td>
<td>272</td>
<td>MD80/P</td>
<td>272</td>
<td>MD80/P</td>
<td>272</td>
<td>MD80/P</td>
<td>272</td>
<td>MD80/P</td>
</tr>
<tr>
<td>B772</td>
<td>272</td>
<td>CRJ2/P</td>
<td>272</td>
<td>CRJ2/P</td>
<td>272</td>
<td>CRJ2/P</td>
<td>272</td>
<td>CRJ2/P</td>
<td>272</td>
<td>CRJ2/P</td>
<td>272</td>
<td>CRJ2/P</td>
<td>272</td>
<td>CRJ2/P</td>
</tr>
<tr>
<td>B772</td>
<td>272</td>
<td>E145/P</td>
<td>272</td>
<td>E145/P</td>
<td>272</td>
<td>E145/P</td>
<td>272</td>
<td>E145/P</td>
<td>272</td>
<td>E145/P</td>
<td>272</td>
<td>E145/P</td>
<td>272</td>
<td>E145/P</td>
</tr>
<tr>
<td>B772</td>
<td>272</td>
<td>A320/P</td>
<td>272</td>
<td>A320/P</td>
<td>272</td>
<td>A320/P</td>
<td>272</td>
<td>A320/P</td>
<td>272</td>
<td>A320/P</td>
<td>272</td>
<td>A320/P</td>
<td>272</td>
<td>A320/P</td>
</tr>
<tr>
<td>B772</td>
<td>272</td>
<td>B752/P</td>
<td>272</td>
<td>B752/P</td>
<td>272</td>
<td>B752/P</td>
<td>272</td>
<td>B752/P</td>
<td>272</td>
<td>B752/P</td>
<td>272</td>
<td>B752/P</td>
<td>272</td>
<td>B752/P</td>
</tr>
<tr>
<td>B772</td>
<td>272</td>
<td>AcType</td>
<td>272</td>
<td>AcType</td>
<td>272</td>
<td>AcType</td>
<td>272</td>
<td>AcType</td>
<td>272</td>
<td>AcType</td>
<td>272</td>
<td>AcType</td>
<td>272</td>
<td>AcType</td>
</tr>
<tr>
<td>B772</td>
<td>272</td>
<td>GEN/36L</td>
<td>272</td>
<td>GEN/36L</td>
<td>272</td>
<td>GEN/36L</td>
<td>272</td>
<td>GEN/36L</td>
<td>272</td>
<td>GEN/36L</td>
<td>272</td>
<td>GEN/36L</td>
<td>272</td>
<td>GEN/36L</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>MD80/P</td>
<td>711</td>
<td>MD80/P</td>
<td>711</td>
<td>MD80/P</td>
<td>711</td>
<td>MD80/P</td>
<td>711</td>
<td>MD80/P</td>
<td>711</td>
<td>MD80/P</td>
<td>711</td>
<td>MD80/P</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>CRJ2/P</td>
<td>711</td>
<td>CRJ2/P</td>
<td>711</td>
<td>CRJ2/P</td>
<td>711</td>
<td>CRJ2/P</td>
<td>711</td>
<td>CRJ2/P</td>
<td>711</td>
<td>CRJ2/P</td>
<td>711</td>
<td>CRJ2/P</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>E145/P</td>
<td>711</td>
<td>E145/P</td>
<td>711</td>
<td>E145/P</td>
<td>711</td>
<td>E145/P</td>
<td>711</td>
<td>E145/P</td>
<td>711</td>
<td>E145/P</td>
<td>711</td>
<td>E145/P</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>A320/P</td>
<td>711</td>
<td>A320/P</td>
<td>711</td>
<td>A320/P</td>
<td>711</td>
<td>A320/P</td>
<td>711</td>
<td>A320/P</td>
<td>711</td>
<td>A320/P</td>
<td>711</td>
<td>A320/P</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>B752/P</td>
<td>711</td>
<td>B752/P</td>
<td>711</td>
<td>B752/P</td>
<td>711</td>
<td>B752/P</td>
<td>711</td>
<td>B752/P</td>
<td>711</td>
<td>B752/P</td>
<td>711</td>
<td>B752/P</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>AcType</td>
<td>711</td>
<td>AcType</td>
<td>711</td>
<td>AcType</td>
<td>711</td>
<td>AcType</td>
<td>711</td>
<td>AcType</td>
<td>711</td>
<td>AcType</td>
<td>711</td>
<td>AcType</td>
</tr>
<tr>
<td>A711</td>
<td>711</td>
<td>GEN/36L</td>
<td>711</td>
<td>GEN/36L</td>
<td>711</td>
<td>GEN/36L</td>
<td>711</td>
<td>GEN/36L</td>
<td>711</td>
<td>GEN/36L</td>
<td>711</td>
<td>GEN/36L</td>
<td>711</td>
<td>GEN/36L</td>
</tr>
</tbody>
</table>

#### Table 31. High Taskload (This Code Continues Through Line 2100 and Column AQ)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E301</td>
<td>301</td>
<td>MD80/P</td>
<td>301</td>
<td>MD80/P</td>
<td>301</td>
<td>MD80/P</td>
<td>301</td>
<td>MD80/P</td>
<td>301</td>
<td>MD80/P</td>
<td>301</td>
<td>MD80/P</td>
<td>301</td>
<td>MD80/P</td>
</tr>
<tr>
<td>E301</td>
<td>301</td>
<td>CRJ2/P</td>
<td>301</td>
<td>CRJ2/P</td>
<td>301</td>
<td>CRJ2/P</td>
<td>301</td>
<td>CRJ2/P</td>
<td>301</td>
<td>CRJ2/P</td>
<td>301</td>
<td>CRJ2/P</td>
<td>301</td>
<td>CRJ2/P</td>
</tr>
<tr>
<td>E301</td>
<td>301</td>
<td>E145/P</td>
<td>301</td>
<td>E145/P</td>
<td>301</td>
<td>E145/P</td>
<td>301</td>
<td>E145/P</td>
<td>301</td>
<td>E145/P</td>
<td>301</td>
<td>E145/P</td>
<td>301</td>
<td>E145/P</td>
</tr>
<tr>
<td>E301</td>
<td>301</td>
<td>A320/P</td>
<td>301</td>
<td>A320/P</td>
<td>301</td>
<td>A320/P</td>
<td>301</td>
<td>A320/P</td>
<td>301</td>
<td>A320/P</td>
<td>301</td>
<td>A320/P</td>
<td>301</td>
<td>A320/P</td>
</tr>
<tr>
<td>E301</td>
<td>301</td>
<td>B752/P</td>
<td>301</td>
<td>B752/P</td>
<td>301</td>
<td>B752/P</td>
<td>301</td>
<td>B752/P</td>
<td>301</td>
<td>B752/P</td>
<td>301</td>
<td>B752/P</td>
<td>301</td>
<td>B752/P</td>
</tr>
<tr>
<td>E301</td>
<td>301</td>
<td>AcType</td>
<td>301</td>
<td>AcType</td>
<td>301</td>
<td>AcType</td>
<td>301</td>
<td>AcType</td>
<td>301</td>
<td>AcType</td>
<td>301</td>
<td>AcType</td>
<td>301</td>
<td>AcType</td>
</tr>
<tr>
<td>E301</td>
<td>301</td>
<td>GEN/36L</td>
<td>301</td>
<td>GEN/36L</td>
<td>301</td>
<td>GEN/36L</td>
<td>301</td>
<td>GEN/36L</td>
<td>301</td>
<td>GEN/36L</td>
<td>301</td>
<td>GEN/36L</td>
<td>301</td>
<td>GEN/36L</td>
</tr>
</tbody>
</table>

The tables highlight the data that is output from the environment model of the FEWS. The reason for illustrating this is to convey the data that was collected and filtered to arrive at the FEWS output and to illustrate the complexity of the environment that was coded into the numerical simulation (which was verified by me as part of the current thesis effort – see appendix D). The columns from left to right are the simulation time, the aircraft identification number (acid), the beacon code (bcncode), the sector it is in (Sector), the communication frequency the aircraft is on (freq), the aircraft type (actype), the departure runway, the arrival runway, the flying status (enroute, approach, etc.), the start time, the latitude in degrees, the longitude in degrees, the altitude in feet, the true airspeed (TAS) airspeed in knots, the indicated airspeed in knots, the heading in degrees, the altitude rate in feet per minute, the rate in degrees, the relative roll in degrees, the roll rate in degrees, the aircraft total weight, coefficient of lift, drag weight in lbs.
Figure 85. Environment model snapshot.

Figure 85 illustrates the environment model that was coded in the human performance model in MicroSaint Sharp. This figure illustrates a baseline model low taskload condition of aircraft travelling through the modelled airspace sector ZGN. The letters highlight specific variables in the simulation. [A] The communications associated with the aircraft travelling along the CHIGO flight path, [B] the VACP workload graph, [C] the communications associated with the aircraft travelling along the DARIO flight path, [D] the red point along the northern sector boundary indicates the visual fixation point of the simulated ATCo, [E] the datablock and their associated call sign information is shown in the centre of the dashed lines, [F] aircraft task listing for the CHIGO flight path, [G] the aircraft datalist for the DARIO flight path, [H] the ongoing tasks being completed by the modelled ATCo, [I] the MicroSaint Sharp code output.
APPENDIX F - BASELINE MODEL VERIFICATION PROCESS

The Model Development process has required an extensive model verification and model calibration effort of the existing baseline model that was created on the FEWS data.

First Verification Process
Model is going to be compared to the optimal human data from the FEWS simulation. This is Team #3. Team #3 has a pattern of performance that would be expected in human performance with increasing traffic levels and load levels. Given that the focus in this research program is on workload, then having a correct workload measure is critical. Obtained Model #53. Ran model, noticed discrepancy in environmental triggers in that the aircraft were not behaving as expected.
Model #53 – noticed that the model was not producing expected values in terms of the time required to do the task.
Model #53 – noticed that the model possessed some “suspect” environmental behaviours

Model #53 - The following list relates to the strange aircraft handoff times discovered in the 100% traffic scenario.

- EGF 3698 - data block became visible (accepted this a/c) at 1840 s, a/c 3698 crossed into sector 1 at 1840 s
- DAL 692 - data block became visible (accepted this a/c) at 471 s, a/c 692 crossed into sector 1 at 471 s
- DAL 414 - data block became visible (accepted this a/c) at 1648 s, a/c 414 crossed into sector 1 at 1648 s
- BTA 6918 - while this aircraft exists in the data, it does not appear that it exist as a data block in the scenario as it never appeared. Data file suggests that this data block should be around 3219.
- ASH 4276 - data block became visible (accepted this a/c) at 1914 s, a/c 4276 crossed into sector 1 at 1914 s
- AAL 6188 - data block became visible (accepted this a/c) at 2347.6499 s, a/c 6188 crossed into sector 1 at 2347.6499 s
- ASH 3577 - data block became visible (accepted this a/c) at 714.85 s, a/c 3577 crossed into sector 1 at 714.85 s
- AAL 431 - data block became visible (accepted this a/c) at 2512.099 s, a/c 431 crossed into sector 1 at 2512.099 s
- AAL 409 - data block became visible (accepted this a/c) at 2020.6 s, a/c 409 crossed into sector 1 at 2020.6 s
- NWA 788 - data block became visible (accepted this a/c) at 1240s, a/c 788 crossed into sector 1 at 1240s. 2177 s (from the data file output) - seems to be the point that the a/c has exited sector 1 and is in the new sector below. Looks like there was never an “accept” of this a/c into sector 1?

Descriptive Data Analyses Completed on Model 53 (includes all of the data with the ghost sectors included)
100% condition -
Mean - 109.78, SD 242.02, Min 3.75, Max - 1018.15 and 79 handoff receipts occurred;
133% condition -
Mean - 74.68, SD 185.73, Min 2.55, Max - 1052.15 and 90 handoff receipts occurred;
166% condition -
Mean - 110.03, SD 231.60, Min 2.45, Max - 1060.3 and 132 handoff receipts).

Identified the need to remove the ghost sectors because they were not supposed to be included (Upon observation of the activities included for the ghost sector, it became apparent that it does not make sense to include the ghost sectors in calculating anything related to the activities because there is a reduced procedural set for these operators.

Another set of descriptive analyses were required following the removal of the ghost sectors. The handoff receipt time required data with ghost sectors removed and only looking at the RHO (this is the actual time that the task took to controller to do a button press).

100% condition -
Mean - 2.44, SD 0.84, Min 1.6, Max - 6.4 and 79 handoff receipts occurred;
133% condition -
Mean - 2.51, SD 0.88, Min 1.6, Max - 6.4 and 90 handoff receipts occurred (but 100 handoffs actually occurred in the simulation);
166% condition -
Mean - 2.59, SD 0.90, Min 1.6, Max - 6.7 and 132 handoff receipts).

These model values and timings do not line up with the FEWS data. These timing problems were worked on and the model was recalibrated into new versions (and much of the iteration is the same as the above through to model 73).

As a result, a more detailed exploration of the handoff condition was undertaken to verify the operation of the handoff and the window of opportunity for the handoff to occur. The Handoff Window currently opens 90s before the sector pierce (the time that the aircraft actually crosses into the new sector). This is seen in the CreateMessagesSP() function, where it creates a MsgSP_XXX message when timeToNextS is less than 90. The window closes at the time set in MonitorTrafficTC() as TempEntity.WindowClose under the Initiate Handoff section. The model currently has this time set at “PierceTime-30”, which means the total window is 90-30, or 60s.

Model 73 – ran model – noticed that the model did not collect workload data consistently with other models. Completed a series of descriptive analyses to determine the location of the discrepancy in the 3422 lines per condition of workload output. Noted this, provided input to the model developers, and collaborated with them to develop another baseline simulation.

Model 80
Workload output is questioned in terms of 133% condition. Values seem incorrect but upon further evaluation, it became clear that the model is predicting the values verifiably. Timing information on the handoff task is suspect as the 166% condition has a lower time than the 133% condition. This effect possesses credibility because controllers begin to operate in a scrambled mode over the strategic mode when performing at the limits of their capability.
Model 80 was frozen and it is within this model that TE will be added.

Model 81
Model 81 was developed because Model 80 possessed a bug that caused metering aircraft to be generating incorrect tracks – all previous model data runs need to be deleted and rerun.

Model 82
TE algorithm created and implemented into Micro Saint Sharp. Baseline Model re-run with TE on or off depending on experimental condition.

Model 83
Task management modification added into the thesis effort. Baseline Model re-run with TM on or off depending on experimental condition.

Model 84
Generic model developed to test the TE algorithm independently of the complex operational environment of the FEWS simulation. Ten thousand model runs completed to verify the performance of the time estimation model.
### APPENDIX G - TASK ANALYSIS AND WORKLOAD TABLE

Table 32. Task Analysis and Workload (Modified) Scale Values (Miller, 2000).

<table>
<thead>
<tr>
<th>Scale value</th>
<th>Description</th>
<th>New Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>Visually Unaided (naked eye)</td>
<td>3.0</td>
</tr>
<tr>
<td>3.7</td>
<td>Visually register or detect (detect occurrence of image)</td>
<td>5.0</td>
</tr>
<tr>
<td>4.0</td>
<td>Visually discriminate (detect visual difference)</td>
<td>3.0</td>
</tr>
<tr>
<td>5.0</td>
<td>Visually inspect or check</td>
<td>4.0</td>
</tr>
<tr>
<td>5.4</td>
<td>Visually track of follow (maintain orientation)</td>
<td>4.4</td>
</tr>
<tr>
<td>5.9</td>
<td>Visually read (symbol)</td>
<td>5.0</td>
</tr>
<tr>
<td>7.0</td>
<td>Visually scan or search monitor (continuous or serial inspection multiple conditions)</td>
<td>6.0</td>
</tr>
<tr>
<td>4.0</td>
<td>Visually register or detect (detect occurrence of image) with NVGs</td>
<td>5.0</td>
</tr>
<tr>
<td>4.8</td>
<td>Visually inspect or check (discrete inspection or static condition (with NVGs)</td>
<td>5.0</td>
</tr>
<tr>
<td>5.0</td>
<td>Visually discriminate (detect visual differences) with NVGs</td>
<td>7.0</td>
</tr>
<tr>
<td>5.6</td>
<td>Visually locate/align (selective orientation) with NVGs</td>
<td>5.0</td>
</tr>
<tr>
<td>6.4</td>
<td>Visually track of follow (maintain orientation) with NVGs</td>
<td>5.4</td>
</tr>
<tr>
<td>7.0</td>
<td>Visually scan, search or monitor (continuous or serial multiple conditions with NVGs)</td>
<td>7.0</td>
</tr>
<tr>
<td>1.0</td>
<td>Auditory</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>Detect or register sound (detect occurrence of sound)</td>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
<td>Orient to sound (general orientation or attention)</td>
<td>2.0</td>
</tr>
<tr>
<td>4.2</td>
<td>Orient to sound (selective orientation or attention)</td>
<td>4.2</td>
</tr>
<tr>
<td>4.3</td>
<td>Verify auditory feedback (detect occurrence of anticipated sound)</td>
<td>4.3</td>
</tr>
<tr>
<td>4.9</td>
<td>Interpret semantic content (speech) simple (1-2 words)</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>Interpret semantic content (speech) complex sentences</td>
<td>6.0</td>
</tr>
<tr>
<td>6.6</td>
<td>Discriminate sound characteristics (detect auditory difference)</td>
<td>6.6</td>
</tr>
<tr>
<td>7.0</td>
<td>Interpret sound patterns (pulse rates, etc.)</td>
<td>7.0</td>
</tr>
<tr>
<td>1.0</td>
<td>Cognitive</td>
<td></td>
</tr>
<tr>
<td>1.2</td>
<td>Automatic</td>
<td>1.0</td>
</tr>
<tr>
<td>3.7</td>
<td>Alternative selection</td>
<td>1.2</td>
</tr>
<tr>
<td>4.6</td>
<td>Sign or signal recognition</td>
<td>3.7</td>
</tr>
<tr>
<td>5.3</td>
<td>Evaluation or judgement (consider single aspect)</td>
<td>4.6</td>
</tr>
<tr>
<td>5.6</td>
<td>Encoding or decoding, recall</td>
<td>5.3</td>
</tr>
<tr>
<td>6.8</td>
<td>Evaluation or judgement</td>
<td>6.8</td>
</tr>
<tr>
<td>7.0</td>
<td>Estimation, calculation, conversion</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td>Rehearsal</td>
<td>5.0</td>
</tr>
<tr>
<td>1.0</td>
<td>Speech</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>Simple speech (1 to 2 words)</td>
<td>2.0</td>
</tr>
<tr>
<td>4.0</td>
<td>Complex (sentence)</td>
<td>4.0</td>
</tr>
<tr>
<td>2.2</td>
<td>Motor</td>
<td></td>
</tr>
<tr>
<td>2.6</td>
<td>Discrete actuation (button, toggle, trigger)</td>
<td>2.2</td>
</tr>
<tr>
<td>4.6</td>
<td>Continuous adjutive (flight control, sensor control)</td>
<td>2.6</td>
</tr>
<tr>
<td>4.6</td>
<td>Manipulative</td>
<td>4.6</td>
</tr>
<tr>
<td>5.8</td>
<td>Continuous adjustment (rotary, vertical thumb wheel, lever position)</td>
<td>5.5</td>
</tr>
<tr>
<td>6.5</td>
<td>Symbolic production (writing)</td>
<td>6.5</td>
</tr>
<tr>
<td>7.0</td>
<td>Serial discrete manipulation (keyboard entries)</td>
<td>7.0</td>
</tr>
</tbody>
</table>
APPENDIX H - TASK NETWORK MODEL

Figure 86. Task Network Representation of the Baseline Model.

Figure 86 is an illustration of the task network model coded within Micro Saint Sharp. In this model, the receiving R-side (ATCo) undertakes a series of common tasks, which are indeterminate (repeating), in order to ensure safe operation of their airspace sector. The ATCo:

- Scans the display seeking high priority data blocks by gathering situation awareness (SA) data of the environment
- Monitors operational environment, looks at the next aircraft, and determines the action required
  - i) the next object is an aircraft,
  - ii) an aircraft action required.
  
  If the object is an aircraft then start tasks of move data block, display route, display halo, display vectors, read object, read CRD (display), read URET (display), look at keyboard, and look at neighbouring sector.
- Looks for events (conflicts, metering, handoffs, etc.) and prioritizes tasks according to the following schedule:
  - (i) initiate handoff,
  - (ii) receive handoff,
Appendix H

- (iii) initiate transfer of communication,
- (iv) receive transfer of communication,
- (v) resolve conflict,
- vi) metering conformance,
- (vii) descend aircraft,
- (viii) altitude clearance,
- (ix) complete aircraft manoeuvre (verify handoff receipt by receiving controller or other aircraft action)

- Scans the display seeking high priority data blocks to gather situation awareness (SA) data of the environment
- Manage other aircraft in the airspace sector to maintain safe separation and handle other handoffs
  - Deals with other aircraft in the airspace sector in terms of maintaining safe separation and handling other handoffs
- Start data block in flash state, perform other activities, and return to the flashing data block as close to 60s as possible (Sending R-side ATCo)
- Detect flashing data block and return to the flashing data block ideally after 60s has passed, 120s maximum. The ATCo fixates on specific aircraft requiring handoff/transfer of control (Receiving R-side ATCo)
  - Within the transfer of control boundary, the R-side ATCo undertakes handoff and transfer of control activities that are comprised of (i) R-D side communication and data block mouse clicks, (ii) handoff receipt, communication, data block mouse clicks, (iii) point out initiation, (iv) point out receipt.
- Within transfer of control boundary, undertakes handoff and transfer of control activities (R-side ATCo)
- Return to radar scope and monitor for other handoffs and conflicts

The environment (an external model) driving the task network of the behaviours outlined above is explained in Appendix E and the verification of the environment model is provided in Appendix F.
APPENDIX I – ADDITIONAL IP INFORMATION AND TASK SCHEDULES

Moray et al. (1991) suggested that, when human operators are facing several tasks, all of which must be accomplished within a fixed time span, and when the human operator was free to choose the order in which the tasks are completed, strategic behaviour was best approached from a theoretically normative perspective, in particular by using scheduling theory. Moray et al. also suggested that the most potent effects that impact strategic behaviour are from time pressure, impacting workload management and human operator performance. Furthermore, their research suggested that humans are suboptimal at scheduling tasks, in the sense that their own subjects did not follow the prescribed rules guiding behaviour, particularly when time pressure was involved. In particular, their subjects tended to switch among logically ordered jobs in a manner that deviated from the optimal strategy. Suboptimal behaviour such as this results in suboptimal task engagement, and is further reflected in estimation of task times and workload. These behavioural shifts away from the optimal strategy imply that the prescribed sequence of behaviours is highly influenced by factors internal to the human operator, as well as external factors (such as time pressure or other environmental contexts).

---

24 The concept of *time pressure* is defined as not necessarily caused by the time constraints, but instead by the time that is subjectively perceived by the operator as available to complete an activity (MacGregor, 1993; Zakay, Block & Tsal, 1999).
## APPENDIX J – OVERVIEW OF TIME ESTIMATION LITERATURE

Table 33. Overview of Time Estimation Literature including the Effects of Over/Under estimation on Timing Behaviour.

<table>
<thead>
<tr>
<th>Researcher</th>
<th>Yr</th>
<th>Empirical/Model</th>
<th>Guiding model</th>
<th>Domain</th>
<th>Time span duration</th>
<th>Estimate Measure</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rantanen, &amp; Levinthal</td>
<td>2005</td>
<td>1 Empirical Study (11 FAA ATC and supervisors)</td>
<td>Temporal Awareness (good = fewer errors, prioritized work more effectively, better rest period management)</td>
<td>ATC – Flight operations, quadrants, handoff</td>
<td>Minutes, 1 minute epochs</td>
<td>Effect of air traffic controller taskload and temporal awareness on task prioritization</td>
<td>Workload, task prioritization (which tasks were done before others)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>T\textsubscript{A}, T\textsubscript{R}, T to start, window of opportunity</td>
<td>CR = Conflict Resolution, DL = Downlink request (climb/descent), FR = Frequency Change, IH = Initiate Handoff, MV = Metering Violation, RH = Receive Handoff.</td>
<td></td>
<td>Manipulated T\textsubscript{A}, T\textsubscript{R} and time to first action (TFA)</td>
<td>Task prioritization may be driven by task characteristics that are categorical rather than continuous and quantifiable.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Examined whether a participants’ temporal awareness, that is, awareness of the TFA or TRm of each task at hand (i.e., tasks with simultaneously open WOs) played a role in their decisions to prioritize one task over another.</td>
<td></td>
</tr>
</tbody>
</table>

<p>| Rantanen, &amp; Levinthal | 2005 | 1 Empirical Study (9 Students) | T\textsubscript{A}, T\textsubscript{R}, T to start, window of opportunity | ATC timeliness in resetting a timer and attending to the next open window of opportunity. | 3.6, or 12 minutes window of opportunity | Time-based modelling of human performance | A probabilistic approach to modelling human performance. The effects of taskload on the distributions of performance variables are examined |
| | | | T\textsubscript{R}/T\textsubscript{A} to complete a behaviour Proactive to reactive performance using COCOM | | | Latency in starting the task or time to first action | |
| | | | | | Carry out each task’s instructions within a specific window of opportunity (WO), which was represented visually as a portion of the progress bar that varied from left to right at one of three speeds, and was marked at one of three locations to indicate the start of the tasks’ WO. After the participant entered the correct number and pressed enter, a new task immediately replaced the completed task. | As taskload increased, the participants were less likely to act on the experimental tasks at an earliest opportunity than under low taskload conditions, resulting in increase of ‘too late’ errors |</p>
<table>
<thead>
<tr>
<th>Year</th>
<th>Author(s)</th>
<th>Study Type</th>
<th>Task Prioritization, Workload and Time Estimation</th>
<th>Temporal Awareness</th>
<th>ATC Flight Operations - En route controllers</th>
<th>Minutes but did take summary data</th>
<th>En Route Controller Task Prioritization Research to Support CE-6 Human Performance Modeling</th>
<th>Cumulative Dwell Time NASA TLX</th>
<th>Window of opportunity Mean Time to first action Standard deviation of time to first action Objective measures of number of separation violations</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>Rantanen, Levinthal, &amp; Yeakel</td>
<td>1 Empirical Study (11 FAA ATC and supervisors)</td>
<td>Task prioritization, workload and time estimation</td>
<td>Temporal Awareness</td>
<td>ATC Flight Operations – En route controllers</td>
<td>Minutes but did take summary data</td>
<td>En Route Controller Task Prioritization Research to Support CE-6 Human Performance Modeling</td>
<td>Cumulative Dwell Time NASA TLX</td>
<td>Window of opportunity Mean Time to first action Standard deviation of time to first action Objective measures of number of separation violations</td>
</tr>
<tr>
<td>2004</td>
<td>Averty, et al.</td>
<td>One study</td>
<td>Mental workload model</td>
<td>NASA TLX, TLI</td>
<td>Workload increase, time estimate</td>
<td>Workload a function of traffic load in ATC behaviour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Koppa X</td>
<td>Review of surface transportation perceptual events</td>
<td>Car-driver roadway (Weir, 1976)</td>
<td>Measures obj and subj</td>
<td>1 to 2 minute timeframe</td>
<td>BRT, SRT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>Miro, E, Cano, M.C., Lourdes E-F, Buela-Casal, G.</td>
<td>Studies to examine the time estimation during prolonged sleep deprivation and its relation to attention activation measures Circadian Effects Linked to Fatigue</td>
<td>Stanford Sleepiness scale Objective measures – skin resistance, ability to measure 10 second intervals</td>
<td>Sleep’s effect, or lack thereof, on the internal pacemaker Prospective and retrospective duration judgements are underestimated (support Block and Zakay)</td>
<td>60 Hours sleep deprivation 10 second time intervals Circadian concept to the experience of time</td>
<td>Attentional resources are shared among all tasks that people perform. As more attention is allocated to monitor the pulse, less attention is available elsewhere. Increase in attention to time results in longer Time Estimates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td>Vohs &amp; Schmeichel*</td>
<td>Empirical 4 Studies**</td>
<td>Resource depletion model (emotion regulation)</td>
<td>Retrospective subjective time estimation</td>
<td>Overestimate</td>
<td>Actively engaged in time keeping results in longer duration judgement (factor of 2) Overestimate duration when actively engaged (factor of 1.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Year</td>
<td>General Theoretical framework: Rhythm Model comprised of four Cardinal Models</td>
<td>Internal clock model (oscillator) Combines Interval and Entrainment Models (sim to circadian clock)</td>
<td>Auditory detection/judgment Duration codes maintained in WM and compared with expectancies</td>
<td>Milliseconds Short interval timing (200-2000ms)</td>
<td>Phase mismatch causes over/underestimates of duration Judgements</td>
<td>Early ending standards were overestimated while late ending standards were underestimated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>------</td>
<td>--------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McAuley &amp; Riess-Jones X</td>
<td>2003</td>
<td>General Theoretical framework: Rhythm Model comprised of four Cardinal Models</td>
<td>Internal clock model (oscillator) Combines Interval and Entrainment Models (sim to circadian clock)</td>
<td>Auditory detection/judgment Duration codes maintained in WM and compared with expectancies</td>
<td>Milliseconds Short interval timing (200-2000ms)</td>
<td>Phase mismatch causes over/underestimates of duration Judgements</td>
<td>Early ending standards were overestimated while late ending standards were underestimated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Engen McAuley &amp; Riess-Jones X</td>
<td>1971 2003</td>
<td>Empirical Research** Webber’s Law Interval model</td>
<td></td>
<td></td>
<td></td>
<td>Focus on absolute judgement and not accurate for current effort</td>
<td>Focus on absolute judgement and not accurate for current effort</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown &amp; Boltz</td>
<td>2002</td>
<td>Series of studies** examining attention, workload and timing</td>
<td>Studying the temporal processing model (Interference model) Vierordt’s Law</td>
<td>Auditory judgement of time passage</td>
<td>Study 1 a) 13.2 s, b) 17.25 s, c) 21.30s</td>
<td>Underestimate research Attending to non temporal events = more variable and shorter time judgments</td>
<td>↑ judgement error in time est. by 4.3% from baseline to detection condition</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brown and Boltz</td>
<td>2002</td>
<td>Same as above</td>
<td>Same as above</td>
<td>Auditory judgement of time passage</td>
<td>Study 2 a) 15-20s b) 25-30s c) 35-40s d) 45-50s</td>
<td>Underestimate research Main effect for event structure</td>
<td>↑ judgement error in time est. baseline 17.1% easy 34.1%, difficult (35%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grondin, Meilleur- Wells, &amp; Lachance X</td>
<td>1999</td>
<td>Empirical research 4 studies** to determine threshold of when counting is needed to benefit interval</td>
<td>Clock model Visual + auditory stimulus Length of time of interval before clock counting become valuable</td>
<td>Psychometric measures: Millisecond measurement S1–1000 ms S2-2500 ms S3-0.7, 1.9 s</td>
<td>Estimation Benefits Difference Threshold S1 – not sig S2 – Sig. S3 – Sig.</td>
<td>Counting becomes useful at 1123 ms shorter duration 70-105 ms not benefited longer duration benefited</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Block &amp; Zakay</td>
<td>1997</td>
<td>Attentional Gate model of prospective duration estimation</td>
<td>Attentional Gate Theory Prospective time est (verbal task)</td>
<td>Overestimation Conditions</td>
<td>Prospective time est. consistently produce longer time est. than retro. est.</td>
<td></td>
<td>Prospective time est. consistently produce longer time est. than retro. est.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Block & Zakay | 1996 | Attentional Gate model of prospective duration estimation | Attentional Resource Allocation | Prospective time est (verbal task) | Pros. time ests. get trans'd to STM 2 factor model - Attending to ext. events (pulse rate is influenced by arousal level – Kahneman, 1973) and to time (am’t of att’n allocated to time). | Overestimation Conditions | Learning and subsequent learning – conditioning will lead to more attentional resources available to timing tasks  
Prospective time est. consistently produce longer time est. than retro. est. |
| St. Jean, McInnis, Campbell-Mayne, Swainson | 1994 | Empirical 2 Studies attentional processes mediate underest of time | Study of hypnosis on time under-estimation | Interval estimates Verbal time estimates of a story being read | Minutes | S1 - High Cog Load underest time by 57%  
Low Cog Load underest time by 37%  
S2 - HCL underest by 46%  
LCL underest by 14% | Underest of time interval when hypnotised under varying cog load |
| Zakay | 1993 | Attentional Gate Theory of Prospective Duration Est  
4 Studies conducted**  
Evidence of an internal mental counter | Attentional Gate Theory Temporal Relevance | Prospective time est (verbal task) | 12 s duration  
† non temporal IP load = † pulses to WM = shorter reprod than that of lower non temp IP | 4 Non temp IP load base 10s  
cond 1 - 11 s  
cond 2 - 15.5 s  
cond 3 - 18 s | Producing duration  
† clock time under high IP than low IP  
Reproducing duration |
| Killeen & Fetterman | 1988 | Math model-based theory that describes rat and pigeon behaviour | Behavioural Theory of timing – Reinforcement (signal-response) | Psychophysics measures of: Immediate timing Retrospective timing Prospective timing | Responses/ second | Behaviour switches between short stimulus duration and long stimulus duration state at 6 s |
| Triesman | 1963 | Timing with a timer model | Scalar expectancy /timing theory | Internal clock (counts the total # of pulses in a time period) guides human duration judgement | Pulses get transferred into a store and into a comparator mechanism | Perceived duration is a monotonic function of the total number of pulses transferred into the accumulator Compares WM with LTM | Does not deal with attentional processes in the conception and estimation of time |

* Vohs and Schmeichel use ratios (his version of $T_A$ and $T_R$) to indicate time estimation in study 1.
** Multiple studies (generally the third one in the set of studies) suggest that the findings are generalisable outside of the empirical research domain studied.
*** Empirical data collected by CENA on ATC workload as a function of time to airspace collision between aircraft. This data set could be used a validation data source although data did not utilize NASA TLX, nor did it use the multiple resource model for scheduling behaviours.
√ - relevant research as time frame seems appropriate  
X – not relevant research as time frame of time estimation is too micro.
APPENDIX K - HUMAN’S ESTIMATION OF TIME REQUIRED

Buehler et al. (1994) conducted a series of five studies to examine how people estimate the time required for a variety of academic and non-academic tasks. They revealed that humans had a tendency to underestimate their own time required but not the time required for tasks to be completed by others. When humans underestimated their own completion times, they demonstrated that they assumed that they were able to complete their tasks faster than they were actually able. (The equivalent implied overestimation of how much one can accomplish in a given time period is termed the 'over optimistic bias'.) When people underestimate their own completion times but do not underestimate others’, this is evidence that they are attributing external influences to others' performance rather than to their own. Kahneman and Tversky’s (1982) research suggested that peoples' underestimation of time occurs because people tend to neglect their own past experience, as they are unable to identify the similarities between current and past experiences. Also, even if people were able to identify similarities between current and past experiences, they tended not to apply this information to the current predictions being generated (Buehler et al., 1994; Kahneman & Tversky, 1982).

Buehler et al.’s (1994) findings suggested that similar experiences may not be applied to the current situation, as humans seemed to neglect past experiences when failing on a current situation, unless they were able to draw causal relations between events. This was a similar finding to that of Kahneman and Tversky (1982), who found that people tended to neglect base-rates by not taking them into account when they possessed case-based information (objective data) on which to base their judgment. Additionally, a number of researchers found that people focus on narratives (stories) when they generated inferences and forecasts (Dawes, 1988; Griffin, Dunning & Ross, 1990; Johnson & Sherman, 1990; Kahneman & Tversky, 1982). When individuals focused on narratives, there were a number of obstacles that prevented them from incorporating past experience into their current context: namely the forward nature of prediction, difficulty in defining similar experiences, and peoples' tendency to diminish the relevance of the past to the present.

A further difficulty experienced by humans engaging in task performance, as noted by Kahneman and Tversky (1979), is that the past influences one's prediction of future events, but prediction by its very nature may prevent individuals from looking back in time. This ironic
finding suggested the existence of a mechanism that actively compared the current to the past when resources were available for the human to make the comparison. It is entirely possible that, when the resources are not available for the human to enact this comparison, inaccurate task predictions are made and operators complete their actions in unexpected/unanticipated times, since their resources are taxed by experiencing varying levels of workload.
APPENDIX L - WORKLOAD AND TIME MANAGEMENT LITERATURE

Hancock and Chignell (1988) presented a model of adaptive interface design that illustrated a view of mental workload that includes both goal achievement and temporal constraints. Their model represented a 'perceived distance' between the operator’s current state and the goal state. The model described the level of effort required to achieve the time-constrained goal. In most cases the human operator possessed multileveled time-varying goals that were completed within a window of opportunity. Hancock and Chignell suggested that task-related performance varies around a set-point of workload and that, when one exceeded this set point boundary level, one was unable to complete a task. This workload exceedance in turn may trigger a behavioural load or time penalty, or both. Such a penalty may have little consequence, or, depending on its severity, it may have serious consequences associated with its violation, such as subsequently increased workload.

Hancock and Chignell’s model presented an admissible workload level defined by the boundaries of (individually determined) thresholds of discontinuity, effective floors for response times, and an effective upper limit for effective action along the temporal axis. Within the space thus defined is an area of inadmissible workload, which results from an excess of distance, in combination with the brief effective time for action. This model suggested once again the existence of a window of opportunity, as outlined above by a number of other researchers, but now the window of opportunity has been placed within a computational framework and is related to the concept of instantaneous workload. Hancock's and Chignell’s model also related workload, effort, $T_A$ for action, and skill level of the human operator.

Being able to correctly estimate and quantify times, along both quantitative and qualitative dimensions, is critical for HPMs being developed today. Workload not only affects performance in executing specific task behaviours, but also has both upstream and downstream impacts on estimating time, managing tasks, and scheduling both quantitative and qualitative human operator behaviours.
# APPENDIX M - LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>3Cs</td>
<td>Competence, Control, Constructs</td>
</tr>
<tr>
<td>A/C</td>
<td>Aircraft</td>
</tr>
<tr>
<td>ACT-R</td>
<td>Atomic Component of Thought – Rationale</td>
</tr>
<tr>
<td>AFRL</td>
<td>Air Force Research Laboratory</td>
</tr>
<tr>
<td></td>
<td>Air Man-machine Integration Design and Analysis</td>
</tr>
<tr>
<td>Air MIDAS</td>
<td>System</td>
</tr>
<tr>
<td>AMBR</td>
<td>Agent-Based Modeling and Behavior Representation</td>
</tr>
<tr>
<td>A-SA or A/SA</td>
<td>Attention-Situation Awareness</td>
</tr>
<tr>
<td>ATC</td>
<td>Air Traffic Control</td>
</tr>
<tr>
<td>ATCo</td>
<td>Air Traffic Controller</td>
</tr>
<tr>
<td>ATM</td>
<td>Air Traffic Management</td>
</tr>
<tr>
<td>ATWIT</td>
<td>Air Traffic Workload Input Technique</td>
</tr>
<tr>
<td>C2</td>
<td>Command and Control</td>
</tr>
<tr>
<td>C3</td>
<td>Command, Control, and Communication</td>
</tr>
<tr>
<td>COCOM</td>
<td>Contextual Control Model</td>
</tr>
<tr>
<td>COGNET/iGEN</td>
<td>Cognition as a NETwork of tasks</td>
</tr>
<tr>
<td>CTA</td>
<td>Cognitive Task Analysis</td>
</tr>
<tr>
<td>DCOG</td>
<td>Distributed Cognition</td>
</tr>
<tr>
<td>DL</td>
<td>Datalink</td>
</tr>
<tr>
<td>DMSO</td>
<td>Defense, Modeling and Simulation Organization</td>
</tr>
<tr>
<td>D-OMAR</td>
<td>Distributed-Operator Model ARchitecture</td>
</tr>
<tr>
<td>D-Side</td>
<td>Data Side Controller</td>
</tr>
<tr>
<td>DSR</td>
<td>Display Suite Replacement</td>
</tr>
<tr>
<td>EASE</td>
<td>Integration of ACT-R, SOAR, and EPIC</td>
</tr>
<tr>
<td>EPIC</td>
<td>executive Process-Interactive Control</td>
</tr>
<tr>
<td>FAA</td>
<td>Federal Aviation Authority</td>
</tr>
<tr>
<td>FEWS</td>
<td>Future En-route Workstation Study</td>
</tr>
<tr>
<td>G^2</td>
<td>Statistical test</td>
</tr>
<tr>
<td>HITL</td>
<td>Human in the loop</td>
</tr>
<tr>
<td>HO</td>
<td>Human Operator</td>
</tr>
<tr>
<td>HPM</td>
<td>Human Performance Model</td>
</tr>
<tr>
<td>IFR</td>
<td>Instrument Flight Rules</td>
</tr>
<tr>
<td>iGEN</td>
<td>Cognitive agent software toolkit</td>
</tr>
<tr>
<td>IHO</td>
<td>Initiate Handoff</td>
</tr>
<tr>
<td>IMC</td>
<td>Instrument Meteorological Conditions</td>
</tr>
</tbody>
</table>
Appendix H

IMPRINT
Improved Performance Research Integration Tool

IP
Information Processing

LOA
Letters of Agreement

Max
Maximum

MHP
Model Human Processor

Micro SAINT
Systems Analysis of Integrated Network of Tasks (PC)

MIDAS
Man-machine Integration Design and Analysis System

Min
Minimum

NAS
National Airspace System

NASA
National Aeronautics and Space Administration

NEXTGEN
Next Generation Air Transportation System

No DL
No Datalink

NVG
Night Vision Glasses

OCM
Optimal Control Model

RHD
Receive Hand-Off Duration

RHO
Receive handoff

R-Side Controller
Radar Controller (also known as a R-side controller)

R-Side
Radar Side Controller

SAINT
Systems Analysis of Integrated Network of Tasks

SD
Standard Deviation

SME
Subject Matter Experts

SOAR
Architecture for General Intelligence

SOPs
Standard Operator Procedures

SSE
Sum of Squared Error

SVS
Synthetic Vision System

T<sub>A</sub>
Time Available (to complete the task)

TC
Time Correspondence

TE
Time Estimation

TM
Task Management

T<sub>O</sub>
The Time to Task Onset

T<sub>R</sub>
Time Required (to complete the task)

V&V
Model Verification and Validation

VACP
Visual, Auditory, Cognitive, Psychomotor

WAK
Workload Assessment Keypad