Identifying Latent Cognitive Constructs in a Comprehensive Model of
Aviation Outcomes: The Role of the Dynamic Mental Model for Pilots

by

Kathleen Van Benthem

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy

Institute of Cognitive Science
Carleton University,
Ottawa, Ontario

© Copyright
Kathleen Van Benthem 2015

Abstract

Background: Older private pilots are at greater risk for general aviation accidents when compared with younger pilots. Predicting which older pilots are vulnerable to accidents has proven elusive, despite the inclusion of domain-independent cognitive factors, such as working memory and processing speed, in outcome models. Thus, it has been suggested that domain-dependent cognitive factors should feature in models of pilot performance. The mental model for pilots was highlighted as a domain-dependent variable particularly relevant to explanations of pilot performance (Hardy & Parasuraman, 1997). The present study specified a mental model for pilots and examined its capacity to predict aviation outcomes.

Methods: A flight simulation program of research collected performance data for 108 pilots who “flew” a Cessna 172. The findings from the flight simulation research produced a robust set of data structures utilized in the modeling phases of this work. A dynamic mental model for pilots (DMM) was developed with guidance from a framework first proposed by Herdman and LeFevre (2004). Structural equation modeling was used to validate the DMM as a central element in a cognition-oriented model of pilot performance for aviation (COMP-A). The COMP-A, based on work by Hardy and Parasuraman (1997), was used as a platform for quantifying the effects of pilot attributes and cognition on aviation outcomes.

Findings: Confirmatory factor analysis demonstrated the DMM convergence as a formative-reflective construct composed of domain-relevant components. The COMP-A included an expansive selection of predictor variables and outcomes of relevance to aviation: critical incidents and runway incursion management. The DMM scores provided the best explanation of critical incidents. The DMM also mediated the majority of effects from upstream variables, such as pilot age, on aviation outcomes

and was the only variable to predict risk of a critical incident with reasonable sensitivity and specificity.

Implications: Findings challenge current strategies for assessing the aging aviator. First, visual-spatial attention, and not working memory or processing speed, was the domain-independent variable with the largest explanatory power in the COMP-A. Second, domain-dependent, not domain-independent, constructs were the most efficacious predictors of risk.

Acknowledgements

I wish to acknowledge the pilots who volunteered their time and considerable talents flying the simulator. I learned so much about aviation, and about the strong culture of safety that exists within this community of pilots. In particular, I thank Simon Garrett for his many hours piloting our simulation protocols and providing guidance to this work. To the general aviation community in Ottawa, thank-you for encouraging me to share the results of this work with you in various venues. Every opportunity to speak was, in reality, a learning experience for me as well.

Anne Barr and Andrew Staples, your expertise in designing the simulation environment and your willingness to support me in solving any technical mishap, at a moment's notice, kept me from losing my nerve and inspired me to keep scheduling participants! Anne and Andrew, over time I grew in appreciation of your skills, but also in appreciation of your friendship. And Anne- you let me fly- wow, what an experience! I also thank Jon Wade for your generosity of time and talent in assisting me with writing programs so that the myriad of data could be handled so efficiently. This helped me more than you might have realized.

To my family and friends, thank-you is just not enough. Without your support I am not sure I would have embarked on this journey, or been able to complete it. First to my husband, Albert, to say that you were supportive and encouraging is an understatement. Your belief in me knew no end, through my fear of ever finishing, to your patient assistance in helping me understand the logic of computer programming, you were always there for me. Your encouragement is what kept me going and I look forward to everything the future has in store for us. To my children: when I started there was Josina and Gregory, and now our family has grown to include Burton and Céline. Thank-you for believing in me. I had so much fun sharing it all with you. I

hope that you will always be committed to life-long learning- it will keep you young. Thank-you to my sisters, Nicole, April, and Elisabeth, for sending me encouraging texts, and for calling me to make sure I wasn't giving up! My thanks also go to my parents, John and Tina. You both have encouraged me to study; and your own educational pursuits and attitude of life-long learning have left indelible marks on my beliefs about education. To mom, you are truly my role model and I only hope I can emulate your passion for excellence.

I greatly appreciate the commitment to excellence in my dissertation supervisor, Dr. Chris M. Herdman. Thank-you Chris for going out on a limb and allowing me to run with this project. Thank-you for letting me share ideas with you, and for listening, even when I didn't really know what I was talking about! I have never taken for granted the opportunities I had at the Herdman lab. Thank-you also for investing time towards my writing. I also thank my doctoral committee. Dr. Jo-Anne LeFevre, your own high standards have strongly influenced me and I find myself often thinking, "what would Jo-Anne do in this situation?" I am grateful for your encouragement and for allowing me to develop skills as a teacher, and student. To Dr. Jocelyn Keillor, thank-you for your encouragement and your advice. I have learned much from you regarding the transition from student to scientist. To Dr.s Taylor and Langlois, thank-you for taking the time to provide me with such valuable feedback. To the wonderful people I first met at the Institute of Cognitive Science six years ago, Dr. Brook, May, Colleen, and Liane, not only are you great encouragers to the students, you are the backbone of the program. And finally, as I do each day, I thank God for all my blessings in life. I never dreamed that this would be one of them, but I am so glad it was.

Table of Contents

Abstract	ii
Acknowledgements	iv
Table of Contents	vi
List of Tables	ix
List of Illustrations	xi
List of Appendices	xiv
Chapter 1	1
Introduction	1
1.1 A Model of Pilot Performance	3
1.2 A Brief History of General Aviation Performance Research	4
1.3 Overview	8
Chapter 2	11
Literature Review	11
2.1 Pilot Age and Flight Performance	11
2.2 Pilot Cognition in General Aviation	16
2.3 Pilot Expertise and Flight Performance	29
2.4 Mental Task Load	33
2.5 Summary of the Literature Review	37
Chapter 3	39
Methods	39
3.1 Participants	39
3.2 Simulation Environment	40
3.3 Procedure	42
3.4 Pilot Demographics and Expertise	44
3.5 Cognitive Measures	45
3.6 Flight Simulation Measures	52
3.7 Mission Difficulty Conditions	54
3.8 Mental Task Load	55
3.9 Flight path maintenance	60
3.10 Situation Awareness	61
3.11 Radio Call Communication (Prospective Memory Measure)	67
3.12 Outcome Measures	68
3.13 Data Preparation	70
3.14 Power to Detect Significant Effects	72
Chapter 4	74
Results of the Experimental Program of Research	74
4.1 Overview	74
4.2 Flight Path Maintenance	76
4.3 Situation Awareness	85
4.4 Prospective Memory for Radio Communication Tasks	100
4.5 Mental Task Load	106
4.6 Aviation Outcomes	121

4.7	Summary of the Results of the Experimental Protocol	134
4.8	Remaining Analyses in the Present Work.....	135
Chapter 5		137
Theoretical Development of the Dynamic Mental Model for Pilots		137
5.1	A Cognitive Systems Engineering Framework	139
5.2	A Mental Model for Pilots.....	139
5.3	Specifying a Mental Model for Pilots	150
5.4	The DMM in the Larger Context of Pilot Performance	169
5.5	Specifying the DMM in a Structural Equation Model	171
5.6	A Working Definition and Structure of a Mental Model for Pilots	179
Chapter 6		181
Empirical Validation of the Dynamic Mental Model for Pilots		181
6.1	Overview of the Hypothesized DMM Reflective-Formative Hierarchical Structure	181
6.2	Characteristics of the Datasets used to Model the DMM.....	183
6.3	Two-Stage Process for Testing the DMM Multidimensional Structure	188
6.4	The DMM Validation Process	192
6.5	Measurement Model Results	195
6.6	Convergence of the DMM in a Covariance-based SEM	208
6.7	Validating the DMM within a Nomological Net	218
6.8	Summary of the DMM Validation.....	221
Chapter 7		224
Specification and Empirical Validation of a Cognition-Oriented Model of Pilot Performance.....		224
7.1	Conception of the COMP-A.....	226
7.2	Overview of the Two Main Analyses	226
7.3	Domain-Dependent Cognitive Variables in the Model of Pilot Performance	227
7.4	Description of Exogenous Causal Variables.....	230
7.5	Description of Control Variables.....	232
7.6	Description of Additional Predictive but Non-causal Variables.....	233
7.7	Procedure for the SEM Analysis of the COMP-A.....	234
7.8	COMP-A Analysis One	238
7.9	Specification of Outer Measurement Models in Analysis One	239
7.10	Domain-independent Cognitive Latent Variables.....	241
7.11	Measurement Models for Self-Rating Scales	248
7.12	The Expertise Formative Measurement Model Development.....	249
7.13	Discriminant Validity of Latent Variables in the COMP-A Analysis One.....	250
7.14	Testing the Inner Path Structure of the COMP-A.....	252
7.15	Curvilinear relationships between Latent Variables	253
7.16	Distribution and Normality of Measures in the COMP-A	254
7.17	Model Fit and Quality Indices	255
7.18	Testing Causal Assumptions in the COMP-A	256
7.19	Results of COMP-A Test: Analysis One	258
7.20	Model fit and quality indices for the COMP-A (Analysis One).....	259
7.21	Results of Analysis One: Pilot Attributes	260
7.22	Results of Analysis One: Domain-independent Cognitive Measures.....	263

7.23	Results of Analysis One: Domain-dependent Cognitive Measures	264
7.24	Results of Analysis One: Non-causal Measures Predicting Critical Incidents	266
7.25	Cognitive Factors Mediating the Effects of Age on the DMM and Critical Incidents: Cognitive Markers for Age	268
7.26	Structural Path Testing in the COMP-A: Analysis Two with Domain-Independent Measures from the DCAT and Pilots 40+ Years	270
7.27	Measurement Models for Domain-independent Cognitive Latent Variables	271
7.28	Measurement Models for Self-rating Scales	272
7.29	COMP-A with the DCAT and Mental Task Load (PDT) as Domain-dependent Predictors of Pilot Performance	274
7.30	Model Quality Indices for COMP-A Analysis Two: Critical Incidents.....	275
7.31	Model Quality Indices for COMP-A Analysis Two: Incursion	276
7.32	Results of Analysis Two: Pilot Attributes.....	277
7.33	Results of Analysis Two: Domain-independent Cognitive Measures	279
7.34	Results of Analysis Two: Domain-dependent Cognitive Measures	280
7.35	Results of Analysis Two: Non-causal Measures Predicting Critical Incidents and Incursion Management	280
7.36	Control Variables.....	287
7.37	Endogeneity in the COMP-A	285
7.38	Interpreting the Role of the DMM in the COMP-A	288
7.39	Utility of the COMP-A in a Sample of Pilots Aged 60+	292
7.40	A Threshold for DMM performance as a Function of Age	295
7.41	Discriminant Utility of the CogScreen-AE Measures and the DMM in Predicting High-Risk Pilots.....	296
7.42	Summary of COMP-A Validation	300
Chapter 8		304
Discussion.....		304
8.1	Major Hypothesis One: Dynamic Mental Model Validity.....	305
8.2	Major Hypothesis Two: Modeling the DMM in the COMP-A	315
8.3	Major Hypothesis Three: The DMM as a Mediator of Effects in the COMP-A	330
8.4	Summary of the DMM and COMP-A Validation.....	332
Chapter 9		337
Implications and Recommendations.....		337
9.1	Current Protocols for Assessing Risk for Older GA Pilots.....	338
9.2	Learning from the Domain of Older Driver Assessment.....	344
9.3	Alternatives to Domain-Independent Cognitive Predictors of Pilot Performance	345
9.4	Recommendations for Identifying Older At-Risk Pilots	349
9.5	Final Note.....	352
References		353

List of Tables

Table 3.1 <i>Pilot Demographics and Expertise</i>	45
Table 3.2 <i>The DCAT Subtests Descriptions, Sample Means and Standard Deviations of Subtest Raw Scores for N=108</i>	48
Table 3.3 <i>CogScreen-AE Subtests Descriptions, Sample Means and Standard Deviations of Selected Subtest Raw Scores</i>	51
Table 3.4 <i>Total Sample Means and Standard Errors for Flight Simulation Measures</i>	53
Table 3.5 <i>Pilot objective and subjective ratings of workload confirm two mission difficulty conditions</i>	59
Table 3.6 <i>SAGAT Probe Scoring System</i>	65
Table 3.7 <i>Self-Rated SA Performance for Low- and High-Difficulty Conditions</i>	66
Table 3.8 <i>Critical Incident Count by Pilot Licensure and Rating group</i>	69
Table 3.9 <i>Mean Incursion Management Scores per Pilot Licensure and Rating Group</i>	
Table 4.1 <i>Flight Performance Factor Score Indicator Loadings</i>	77
Table 4.2 <i>Correlations between Flight Performance and Cognitive, Age, and Expertise factors</i>	79
Table 4.3 <i>Age and Expertise Predictors for Simple Model of Flight Performance in the Low-Difficulty Condition</i>	81
Table 4.4 <i>Predictors for the Final Model of Flight Performance in the Low-Difficulty Condition</i>	82
Table 4.5 <i>Predictors for the Final Model of Flight Performance in the High-Difficulty Condition</i>	84
Table 4.6 <i>Final Cognition and Expertise Regression Model of Flight Performance in the High-Difficulty Condition</i>	84
Table 4.7 <i>Pilot Characteristics by Age Group</i>	88
Table 4.8 <i>Relationship of Situation Awareness Performance to Pilot Age, Experience and Cognitive Health</i>	89
Table 4.9 <i>Regression Models for SA levels 2 and 3</i>	90
Table 4.10 <i>Relationship of SA Levels 1, 2, and 3 to Prospective Memory, Flight Path Maintenance, PDT, and Episodic Memory in Low- and High-Difficulty Conditions</i> ..	96
Table 4.11 <i>Inter-Correlations between Items of the Self-Rated Situation Awareness Scale</i>	97
Table 4.12 <i>Two-Component Matrix for Subjective SA Scale Items (Low-Difficulty Condition)</i>	99
Table 4.13 <i>Mission Difficulty Effects for Complex Simulator Tasks for Older(50+) and Younger(<50) Pilots</i>	119
Table 4.14 <i>Critical Incidents Regressed on SA level 1 (iRES), 2, and 3</i>	131
Table 4.15 <i>Critical Incidents Regressed on Prospective Memory Score for Older and Younger Pilot Groups</i>	133
Table 6.1 <i>Correlations between Observed Variables (Standardized and Ranked Dataset)</i>	187
Table 6.2 <i>DMM Hypotheses</i>	192

Table 6.3 <i>Loadings and Cross-loadings of Indicators for the Final First Order Reflective Measurement Models in the DMM (No. obs.=103)</i>	197
Table 6.4 <i>Reliability Coefficients for the Final First Order Reflective Measurement Models in the DMM (No. obs.=103)</i>	198
Table 6.5 <i>Correlation of Latent Variables with Square Root of AVEs</i>	199
Table 6.6 <i>DMM Composite Variable Indicator Weights</i>	200
Table 6.7 <i>Validation of the DMM First Order Measurement Model Structure using the Second Dataset.</i>	203
Table 6.8 <i>Measurement Model Reliability and Validity Coefficients for the Second Dataset.</i>	203
Table 6.9 <i>DMM Second Order Formative Structure for the Second Dataset.</i>	204
Table 6.10 <i>Factor Loadings for Measurement Model Factors for SA, TRP, and TM</i> 205	
Table 6.11 <i>Quality Coefficients for SA, TRP, and TM with No. Obs. = 216</i>	206
Table 6.12 <i>Quality Coefficients for the DMM with No. Obs. = 216</i>	206
Table 6.13 <i>DMM Measurement Model Standardized Estimates from CB-SEM</i>	216
Table 6.14 <i>Results of DMM Hypothesis Testing</i>	223
Table 7.1 <i>Hypothesized COMP-A Path Estimate Hypotheses</i>	236
Table 7.2 <i>Description of Subsample of Pilots for COMP-A Data Modeling</i>	239
Table 7.3 <i>Unrotated Factor Loadings and Oblique-rotated Cross-loadings for Indicators Representing Latent Variables for Working Memory, Processing Speed, Visual-Spatial Attention, and Inhibition.</i>	245
Table 7.4 <i>Quality Coefficients for Domain-independent Cognitive Latent Variables</i> 246	
Table 7.6 <i>Discriminant Validity for PROSPD, WM, INHIB, and VIS_SPA.</i>	248
Table 7.7 <i>Quality Indices for Formative Model for Pilot Expertise</i>	250
Table 7. 8 <i>Discriminant Validity of Constructs in the COMP-A Analysis One shown by Correlation among Latent Variables with Square-roots of AVEs.</i>	251
Table 7.9 <i>Latent Variable Quality Coefficients</i>	260
Table 7.10 <i>Results of Hypothesis Testing for the COMP-A Analysis One</i>	267
Table 7.11 <i>COMP-A with Pilots Aged 40 and older using DCAT as Domain-independent Predictors of Pilot Performance (n =72).</i>	271
Table 7.12 <i>Unrotated Loadings and Oblique-rotated Cross-loadings for Measurement Models in the COMP-A Analysis Two.</i>	273
Table 7.13 <i>Measurement Model Quality Indices for the COMP-A Analysis Two</i>	273
Table 7.14 <i>Results of Hypothesis Testing for the COMP-A Analysis Two</i>	284
Table 7.15 <i>Correlations Among Residual Terms with VIFs in the COMP-A</i>	287
Table 7.16 <i>Correlations Among First-Order Factors in the DMM and COMP-A variables</i>	289
Table 7.17 <i>Correlations Among Latent Variables in the COMP-A</i>	291
Table 7.18 <i>Pearson Correlations for the DMM and Predictor Variables in the Reduced COMP-A Model</i>	287
Table 7.19 <i>AUC Results for the ROC Classifier of Pilot Risk Group</i>	300
Table 8.1 <i>Comparison of the DMM Components to Literature-Based Descriptions of Mental Models</i>	311
Table 9.1 <i>Suggested Elements of a Cognitive Health Screening Tool</i>	350

List of Illustrations

<i>Figure 1.1.</i> Hardy and Parasuraman's causal model of pilot performance and aviation outcomes..	4
<i>Figure 2.1.</i> Longitudinal trend in average simulated flight summary scores by age group..	13
<i>Figure 2.3.</i> Age trends in cognitive abilities..	15
<i>Figure 2.3.</i> Age effects on five aspects of working memory..	21
<i>Figure 3.1.</i> Cessna 172 simulator cockpit running Microsoft® Flight Simulator.....	10
software and graphics	41
<i>Figure 3.2.</i> Thumb switch for auditory peripheral detection task.....	57
<i>Figure 3.3.</i> Low- and high-difficulty aerodromes.....	60
<i>Figure 3.4.</i> Circuit parameters in low- and high-difficulty conditions.....	61
<i>Figure 3.5.</i> Estimates of statistical power for a repeated measures ANOVA..	72
<i>Figure 4.1</i> An adapted model of pilot performance with the DMM as the key domain-dependent construct	75
<i>Figure 4.2</i> Average altitude deviations by age and pilot licensure and rating group...	78
<i>Figure 4.3.</i> Relation of flight path maintenance to pilot age in low (A) and high (B) difficulty conditions.....	80
<i>Figure 4.4</i> Mean situation awareness performance by age groups and situation awareness level..	94
<i>Figure 4.5</i> Modifications to SA level 1 in Endsley's model of SA in dynamic decision making.....	94
<i>Figure 4.6.</i> Interaction of workload and cue salience and age and cue salience.	102
<i>Figure 4.7</i> Structural model illustrating possible predictive links between subjective and objective measures of workload and an array of domain-independent and domain-dependent cognitive measures.	108
<i>Figure 4.8</i> Standardized loadings for the full measurement model for cognitive workload including indices for the objective and subjective workload constructs in the low- and high-difficulty conditions.....	111
<i>Figure 4.9.</i> Relationship of cognitive workload indices to an array of pilot measures for younger (<49 years) and older (49+) years pilot groups.....	115
<i>Figure 4.10.</i> Total critical incident counts for low- and high-difficulty conditions...	124
<i>Figure 4.11.</i> Distribution of incursion management scores for pilot rating group....	145
<i>Figure 5.2.</i> Histon and Hansman (2008) working mental model central to ATC cognitive processes.	148
<i>Figure 5.3.</i> Relationship of mental models to situation awareness.....	149
<i>Figure 5.4.</i> The revised dynamic mental model (DMM) and its composite elements of situation awareness (SA), task-relevant performance (TRP), task management (TM), and prospective memory (PM).....	153
<i>Figure 5.5.</i> Task-relevant performance as a function of age and mission difficulty. Higher scores on the Y-axis represent greater deviations from the ideal altitude....	158
<i>Figure 5.6.</i> The DMM as a mediator of effects from upstream predictors on downstream outcomes.....	167

<i>Figure 5.7.</i> The DMM in a simplified model of pilot performance.....	170
<i>Figure 5.8.</i> Hypothetical specification of the DMM as a multidimensional reflective-formative latent construct.	173
<i>Figure 6.1.</i> Original hypothesized first-order measurement models and second-order formative structure of the DMM.....	191
<i>Figure 6.2.</i> Three steps for structure and construct validation of the DMM.	195
<i>Figure 6.3</i> Final DMM structure with loadings, weights, and measurement model quality coefficients.....	207
<i>Figure 6.4.</i> Distribution of DMM factor scores estimated as exact linear combinations of SA, TRP, TM, and PM.....	208
<i>Figure 6.5.</i> Final standardized measurement model for the DMM using CB-SEM estimation.....	212
<i>Figure 6.6.</i> Distribution of DMM factor scores estimated by maximum likelihood.	217
<i>Figure 6.7.</i> Nomological net for the DMM with simultaneous modeling of the effect of iRES on the individual components of the DMM.	219
<i>Figure 6.8</i> Alternative model with individual prediction of SA, TRP, TM, and PM of critical incidents.	220
<i>Figure 7.1</i> Order of the main analyses and associated datasets used to test the COMP-A structure.....	225
<i>Figure 7.2.</i> Hypothesized cognition-oriented model of performance-aviation (COMP-A) with CogScreen-AE Domain-Independent Cognitive Measures.....	237
<i>Figure 7.3.</i> Curvilinear relationship between AGE and PROSPD. Shown is the best-fitting curve with segment slope coefficients and associated <i>p</i> -values. Plot produced by WarpPLS 4.0 (Kock, 2013).....	254
<i>Figure 7.4.</i> Results of measurement model and structural path testing for COMP-A with critical incidents as the outcome variable.	258
<i>Figure 7.5.</i> Relative total effect sizes for AGE, EXPERT, and domain-independent cognitive variables.	262
<i>Figure 7.6.</i> Relative total effect sizes (direct and indirect) of predictor variables in the COMP-A associated with the DMM (top) and critical incidents (bottom).....	266
<i>Figure 7.7.</i> COMP-A analysis two: critical incident outcome.....	274
<i>Figure 7.8.</i> COMP-A with the DCAT as predictors and incursion management as the outcome variable.....	275
<i>Figure 7.9.</i> Relative total effect sizes for AGE, EXPERTISE, and MENTAL TASK LOAD (auditory peripheral detection task) on domain-independent cognitive measures in the COMP-A with 144 observations.	278
<i>Figure 7.10.</i> Relative total effect sizes for predictors of the DMM (blue) and critical incidents (red) in the COMP-A with 144 observations.....	281
<i>Figure 7.11.</i> Relative effects sizes for predictors of the DMM (blue) and incursion management (purple) in the COMP-A.....	282
<i>Figure 7.12.</i> Smooth 3-D plot of moderating effect of age on relationship between incursion management scores and DCAT latent variable scores.	283
<i>Figure 7.13.</i> Reduced COMP-A for pilots aged 60 to 81 years.....	293

<i>Figure 7.14.</i> Relative total effects of predictors of critical incidents in the COMP-A model with an older pilot sub-sample from analysis two.	294
<i>Figure 7.15.</i> Curvilinear relationship of pilot age to DMM latent variable scores....	296
<i>Figure 7.16.</i> Receiver operating characteristics (ROC) curve for the DMM (blue curve) and the LRPV (green dashed curve) from the CogScreen-AE predicting low- versus high-risk pilots.	299
<i>Figure 8.1.</i> A generic “COMP” structure with hypothesized inner paths.....	305
<i>Figure 9.1.</i> Recommended neurocognitive battery for any suspected neurological impairment.	339
<i>Figure 9.2.</i> Relationship of CogScreen-AE LRPV.	340
<i>Figure 9.3.</i> Relationship of DMM factor scores to critical incidents and incursion management as a function of pilot age.....	341
<i>Figure 9.4.</i> Receiver operating characteristics (ROC) curve for the DMM (blue curve) and Taylor Speed and Working Memory (WM) factor (green curve) predicting critical incidents vs no critical incident pilots.....	343
<i>Figure 9.5.</i> Plot of DMM scores from the present work superimposed on histogram of COPA pilot ages.	349

List of Appendices

A: Psychology Ethics Review Board Approval	376
B: Participant Informed Consent Form	377
C: Participant Recruitment Poster	378
D: Sample Data Capture from Flight Simulation Protocol	379
E: Pilot Debriefing Form	380
F: Pilot Demographic and Expertise Form	381
G: NASA TLX Mental Task Load Self-Rating Scale	383
H: SAGAT Probe Answer Form	385
I: Situation Awareness Self-Rating Scale	386
J: Surprise Runway Incursion Scoring System	387
K: CogScreen-AE Subtest Abbreviations	388
L: Abbreviated Output from MPlus (v.7.11)	391
M: Comparison of Correlations of PLS-SEM and CB-SEM DMM Component Scores with Pilot Attributes and Simulator Performance	394

Chapter 1

Introduction

The purpose of the present work was to develop a comprehensive model of pilot performance and advance the identification of cognitive factors related to risk in general aviation. The development of a model of pilot performance was guided by a cognitive systems engineering (CSE) framework that acknowledged both domain-independent and domain-dependent cognitive factors relevant to pilot performance. This work also developed the *dynamic mental model* for pilots (Herdman & LeFevre, 2004), a key structure of domain-dependent cognition, and tested its utility as a predictor of aviation outcomes in the comprehensive model of pilot performance. Model results challenge current practices and inform novel methods for identifying at-risk pilots.

Identifying factors associated with risk addresses a serious problem in general aviation, where reports reveal an accident rate 40 times that found in commercial aviation (per 100,000 flight hours, General Accounting Office, 2012). Despite achieving target decreases in the accident rate for commercial aviation, the US Federal Aviation Administration (FAA) has not met targets for reducing the fatal accident rate in general aviation. To investigate why general aviation operations have such a comparatively high accident rate, stakeholders in aviation safety annually review variables associated with accidents, such as make of aircraft and individual pilot characteristics. Sixty percent of fatal accidents occur in single-engine piston aircraft that are primarily piloted by private pilots for personal flight (GAO, 2012). The single-engine piston aircraft is considered a safe aircraft with few accidents caused by mechanical failure, thus leaving pilot actions as the main cause of accidents (GAO,

2012). The relative risk of a general aviation accident also increases markedly beyond age 60, thus making older private pilots, particularly low-rated pilots, a concern (Li, Baker, Qiang, Grabowski, & McCarthy, 2005; Shao, Guindani, & Boyd, 2014). The high frequency of pilot-rated error, and the observation that older private pilots are at greater risk for critical incidents, as compared to younger pilots or pilots with more expertise, has prompted research investigating the effects of pilot cognition, age, and experience on pilot performance.

Research examining the association between cognition and general aviation has a relatively short history in the broad realm of psychological interests. In 1997, Hardy and Parasuraman conducted a literature review of predictors of flight performance and found that individual pilot factors such as cognition, age, and expertise were associated with both aviator performance and accident rates. However, at that point in time, much of the research examining pilot cognition and performance was aimed at investigating possible decrements in commercial pilot performance between the ages of 60 and 65 years, as the FAA was considering increasing the mandatory retirement age of pilots from 60 to 65 years. Little work had been undertaken regarding age and general aviation pilots. Hardy and Parasuraman (1997) noted that when age effects were reported they were most evident in communication task performance. A second observation was that most research had only examined effects of domain-independent cognition, such as working memory or cognitive speed, on pilot performance. Domain-independent cognitive variables are “general purpose” mechanisms that support a broad array of behaviour. In sum, the 1997 review found that domain-independent variables had weak associations with pilot performance criterion variables. Domain-independent cognition was also found to decline with age, for pilots and non-pilots alike, although further investigation of the protective feature of

expertise was recommended. Hardy and Parasuraman urged that future work include variables to control for experience when investigating performance in older and younger pilot groups.

1.1 A Model of Pilot Performance

To address shortcomings of earlier investigations and guide future research, Hardy and Parasuraman (1997) developed a causal model of aviation outcomes that accommodated a wide array of predictors. As shown in Figure 1.1, the model illustrates a cascading effect of pilot attributes and stressors upon both domain-independent and domain-dependent cognitive factors that, in turn, impact pilot flying performance. In Figure 1.1 solid arrows between “boxes” indicated strong causal assumptions, and dashed arrows indicated weak causal assumptions.

As shown in Figure 1.1., some cognitive effects on performance are viewed as *Domain-Independent Skills* (left centre box). Domain-independent skills are foundational cognitive abilities that presumably support a host of complex behaviours besides those associated with piloting an aircraft. The dashed arrow from domain-independent skills to pilot flying performance predicts a weak association between these two variables.

“Domain-Dependent Knowledge” pertains to aviation-related cognitive proficiencies and includes concepts such as pilot mental models, situation awareness, and decision-making (see the centre right box in Figure 1.1). Hardy and Parasuraman (1997) postulated that explanations of aviation outcomes are rooted in domain-dependent variables.

Pilot performance (represented by the lower circle) entails a variety of aviation outcomes. Hardy and Parasuraman (1997) suggested that there was a strong relationship of domain-dependent cognition to aviation outcomes, and, in particular,

between “mental models” and outcomes. On this view, the development of a mental model for pilots is pivotal to extending our ability to predict pilot performance.

Figure 1.1. Hardy and Parasuraman's causal model of pilot performance and aviation outcomes. From Hardy, D., & Parasuraman, R. (1997). Cognition and flight performance in older pilots. *Journal of Experimental Psychology: Applied*, 3(4), p.339.

1.2 A Brief History of General Aviation Performance Research

Despite Hardy and Parasuraman's (1997) exemplar, the application of psychological theory to studies of pilot performance in general aviation has been approached in a somewhat piecemeal fashion. While Hardy and Parasuraman emphasized domain-

dependent causes, most research investigating pilot performance and cognition has focused on the impact of the domain-independent variables on the first two items in the “Pilot Flying Performance” section of the model: flight simulation and ATC (air traffic control) communication. For example, it was theorized that the effect of age on pilot communication performance was mediated by domain-general working memory ability. In turn, working memory effects on performance were associated with slower processing speed and poor interference control (Morrow et al., 2001; 2003; Taylor, O’Hara, Mumenthaler, Rosen and Yesavage, 2005). Thus, previous research has delineated some of the finer relationships that exist within domain-independent mechanisms, but has not developed models that fully account for the effects of age or pilot expertise on aviation outcomes. A comprehensive account of the transmission of cognition, age, expertise, and workload effects through complex human-machine systems remains to be identified.

Some authors have investigated domain-independent cognitive effects on domain-dependent cognition. Such research is represented by the solid horizontal arrow between the centre boxes of the Hardy and Parasuraman model. Kennedy, Q., Taylor, Reade, and Yesavage (2010) examined the role of age, cognition, and expertise on domain-specific decision-making and found that older pilots (40+ years) and pilots with faster reaction times for domain-independent measures were more likely to make incorrect simulated landing decisions in poor weather conditions than younger pilots.

Other researchers have shifted from an interest in age-effects to examining a wider array of domain-independent factors that might predict pilot performance. For example, Causse, Dehais, and Pastor (2011) found that reasoning skills were strongly associated with simulated flight path deviations. Another stream of research has investigated biomarkers of age that are associated with departures from the typical

aging process. Examples of such biomarkers include brain size (Adamson et al., 2010b) and intraindividual variability (IIV) in reaction time (Kennedy, Q. et al, 2013). Pilot brain size was not associated with performance, while IIV did correlate with flight performance and was found to mediate some, but not all, performance variance accounted for by age alone. Adamson et al. (2010a) examined genetic risk factors for cognitive decline and found that pilots with the Apolipoprotein E epsilon 4 allele had lower performance on a test of visual-paired associates (immediate and delayed recall of symbol-digit pairs) than pilots without this genetic marker. The association between this genetic risk factor and flight performance was not measured. Clearly more work remains to be done before conclusions regarding the predictive utility of biological proxies for age on pilot performance may be drawn.

Minimal work has investigated effects of domain-dependent abilities on aviation outcomes. A retrospective database analysis of aviation incidents associated with pilot error found that situation awareness inaccuracies accounted for 88% of the pilot errors (Endsley, 1999). A survey of New Zealand pilots found that many GA accidents were associated with errors in procedural knowledge and selection of an appropriate strategy (O'Hare, 2006). The failure to detect pertinent information was the most common reason for incidents and the second most frequently cited reason for accidents (O'Hare, 2006). O'Hare also found that older age and increased cockpit distractions significantly predicted strategy selection errors for accidents and incidents. A survey of Canadian pilot examiners' and flight instructors' attitudes regarding younger and older pilots found there was greater concern with procedural knowledge for older pilots, particularly during flight training and currency flight examinations, as compared with concerns for younger pilots (Van Benthem, Herdman, & Garrett, 2011b). A study of airline pilots examined pilot attributes, domain-

independent and domain-dependent variables on simulated B727 performance (Hyland, Kay, & Deimler, 1994). Results demonstrated an effect of pilot age on both domain-dependent and -independent cognitive performance. Older age was associated with lower subjective ratings by “raters” for simulator measures, but was not associated with objective flight performance in the B727 simulator.

Database analyses of accident causes have also prompted an interest in the effects of prospective memory for aviation tasks on performance outcomes. Prospective memory is a domain-dependent cognitive construct associated with remembering to perform critical tasks in the future and is a skill relevant to successful daily activities (Uttl, 2008) and to aviation outcomes (Dismukes & Nowinski, 2007). A retrospective look at aviation incidents and accidents found that 74 of the 75 reports that listed memory error as a key causal factor were actually prospective memory failures (Nowinski, Holbrook, & Dismukes, 2003). Failures of prospective memory in the cockpit, such as forgetting to adjust the flaps or to lower the landing gear, can have disastrous results and can occur with pilots at any level of expertise (Dismukes & Nowinski, 2007). Prospective memory is a relevant factor in the present work because of its sensitivity to the effects of older age (Uttl, 2008; 2011). The aforementioned focus on domain-independent aspects of cognition and pilot performance has left the development of domain-dependent cognitive measures, and their effects on general aviation outcomes virtually unexplored in experimental research.

In summary, Hardy and Parasuraman (1997) claimed that the development of a mental model of pilot behaviours was necessary in order to explain the effects of individual pilot factors on performance and aviation outcomes. The present work builds on the *dynamic mental model* (DMM), a mental construct referencing the pilot mental model as formulated by Herdman and LeFevre (2004). The DMM for pilots

was originally described by Herdman and LeFevre as a latent construct indexed by pilot situation awareness, task-relevant behaviour, and cognitive workload. The present work further developed the DMM as a multi-dimensional structure indexing key domain-dependent cognitive mechanisms. To facilitate development of a comprehensive model of aviation outcomes, the present work posited the DMM as a central feature in a CSE framework.

1.3 Overview

The principal outcomes of the present work lie in the results of modeling a plausible structure for a mental model for pilots, and the investigation of the mental model as a predictor of key criterion variables highly-relevant to aviation safety. The approach taken in the present work was to develop and test a comprehensive model of pilot performance guided by a CSE framework. Latent cognitive constructs representing pilot attributes, domain-independent, and domain-dependent cognitive variables were generated and tested within this model of pilot performance. A dominant theme of this dissertation was the development of the dynamic mental model (DMM) for pilots. The DMM is a latent cognitive construct posited as a central variable in the comprehensive model of performance. Three main hypotheses regarding the DMM structured the present work.

1. First, it was hypothesized that four domain-dependent cognitive factors converge in a multilevel structure to represent the dynamic mental model (DMM) for pilots.
2. Second, it was hypothesized that the DMM, as a domain-dependent variable, accounts for significantly more variance in the aviation outcomes measures than the variance accounted for by domain-independent cognitive factors.

3. Third, the DMM was hypothesized to fully mediate the effects of cognition, age, expertise, and workload factors on aviation outcomes.

Testing the three hypotheses required a large flight simulation study where pilot performance data were collected during simulated general aviation flight missions and from two cognitive health screening tools. An outline of this dissertation and the phases of the program of research are provided below:

- Chapters 1 and 2 provide an introduction to the issues concerning the prediction of pilot performance. The literature review in Chapter 2 was guided by the variables presented in the model of pilot performance by Hardy and Parasuraman and summarizes key findings pertaining to the effects of pilot age, cognition, expertise, and workload on pilot performance. The need for the development of a mental model for pilots and the application of a CSE framework for elucidating the predictors of pilot performance and aviation outcomes is presented.
- Chapter 3 contains the methods and procedure used to capture flight simulation data with a sample of 108 pilots. Chapter 4 describes the results of various published and unpublished analyses that examined predictors of pilot performance. The relationships of pilot attributes, workload factors, and flight performance measures to the aviation outcomes of critical incidents and management of a surprise runway incursion were also investigated.
- The results of the experimental phase of the current work informed the conceptual development of a mental model for pilots: a mental model hypothesized to support basic aviator competencies. Chapter 5 traces the conceptual development of the DMM (revised from the original Herdman and LeFevre [2004] dynamic mental model structure). In Chapter 6 the structure of

the DMM was validated using both partial least squares, and covariance-based structural equation modeling (SEM).

- Chapter 7 describes predictors of pilot performance using a CSE framework based on Hardy and Parasuraman's (1997) model of pilot performance. The DMM was evaluated as a central mediating variable in two main analyses utilizing subsets from the sample of 108 pilots. In the present research the DMM and a comprehensive model of pilot performance (named the COMP-A in this work) was generated using a SEM platform. The SEM approach provided an opportunity to simultaneously model effects from pilot attributes, domain-independent and domain-dependent cognitive variables, and mental task load constructs on aviation outcomes. The SEM methodology also permitted the inclusion of latent (unobserved) variables in the model.
- A discussion of the DMM and the COMP-A validation is provided in Chapter 8. In Chapter 9 the implications of the findings involving the COMP-A are discussed. A primary issue concerns current practices for assessment of the aging aviator. While presently domain-independent variables are exclusively used to predict at-risk pilots, results from the present research indicate that domain-dependent measures are essential for identifying pilots, particularly older pilots, who might be at-risk for a critical incident.

This dissertation is unique in its development and application of the pilot mental model as a guide for generating hypotheses regarding relations between variables within a complex model of flight performance and uncovering the transmission of age, cognition, expertise, and workload effects on aviation outcomes.

Chapter 2

Literature Review

This synopsis provides an overview of the current knowledge base pertaining to individual differences and general aviation pilot performance. The results of this review serve as a rationale for the focus on pilot age, cognition, expertise, and workload as key predictor variables in a model of pilot performance.

2.1 Pilot Age and Flight Performance

The issue of age and pilot performance has been assessed using flight simulation. Longitudinal and cross-sectional analyses of a cohort of pilots, followed between one to thirteen years has consistently demonstrated superior performance by younger pilots (e.g., Kennedy, Q., et al., 2010; Taylor, Yesavage, Morrow, Dolhert, Brooks, & Poon, 1994; Taylor, Kennedy, Noda, & Yesavage, 2007; Yesavage, Taylor, Mumenthaler, Noda, & O'Hara, 1999; Yesavage et al., 2011). In an early report, researchers investigated pilots' read-back accuracy and execution of air traffic control messages and found that older pilots demonstrated less accuracy when executing air traffic commands and committed more errors in reading back messages that were composed of longer strings of digits (Taylor et al., 1994). While backward digit span ability was correlated with performance, it did not mediate the older pilots' inferior feedback accuracy. Yesavage, Taylor, Mumenthaler, Noda and O'Hara (1999) found that simulated flight performance was lower for older pilots, with age predicting 22% of the overall variance in performance. Taylor, O'Hara, Mumenthaler and Yesavage (2000) reported that four cognitive factors (working memory/processing speed, visual memory, motor coordination, and tracking) accounted for 45% of the variance in flight performance with age contributing significant variance in addition to the cognitive

factors. Taylor, O'Hara, Mumenthaler, Rosen and Yesavage (2005) and Taylor, Kennedy, Noda, and Yesavage (2007) reported lower performance by older pilots for following air traffic control messages, traffic avoidance, cockpit instrument scanning, and approach and landing ability (see Figure 2.1). In the Taylor et al. (2007) study, the effects of the passage of one year of age are demonstrated in Figure 2.1, with all age groups showing declines in performance on a flight summary score. An exception to the age-related decline was observed for the oldest pilot group (aged 60 to 69 years), when higher rating appeared to moderate the effects of age. The lower-rated (visual flight rules) oldest pilots did show the typical age-related decline in performance. Taylor et al. (2007) suggested that the phenomenon of older, higher-rated pilot performance appearing more stable over time was related to an exception with traffic avoidance scores, where more highly-rated pilots showed less degradation in abilities over time than the younger pilots.

Figure 2.1. Longitudinal trend in average simulated flight summary scores by age group. Adapted from Taylor, J. L., Kennedy, Noda, A., & Yesavage, J. A. (2007). Pilot age and expertise predict flight simulator performance - A 3-year longitudinal study. *Neurology*, 68(9), 648-654. VFR = Visual Flight Rules (lowest expertise group), IFR = Instrument Flight Rules (medium expertise group), and CFII/ATP = Certified Flight Instructor/Airline Transport Pilot (highest expertise group). Note that rates of performance declined over a three-year span for the each age and expertise group. Red line at $y=0.000$ added.

Kennedy, Q., Taylor, Heraldez, Noda, Lazzeroni, and Yesavage (2013) have further accounted for age-effects by adding a basic and complex intraindividual variability (IIV) index to an age and reaction time-based model of pilot summary performance. Performance on cognitive criterion measures has been shown to associate with basic and complex measures of IIV (Bielak, Hultsch, Strauss, Macdonald, & Hunter, 2010). The basic processing IIV index used by Kennedy, Q. et al. (2013) reflects individual deviations in scores across trials. Kennedy, Q. et al. (2013) reported reductions of 15% to 22% of the age-related effect in summary flight scores with the addition of the IIV to the model.

The literature demonstrates a persistent trend toward speculating on mediating cognitive factors when age-related declines are observed in pilot performance (Coffey, Herdman, Brown, & Wade, 2007; Morrow, Leirer, Altieri, & Fitzsimmons, 1994). In a desktop simulator study of flight performance, Coffey et al. (2007) found that older pilots missed significantly more critical events, such as cockpit malfunctions, than younger pilots. It was suggested that age-effects in “change detection” might have been responsible for the older pilots missing more events both inside and outside the cockpit. In a study comparing high-expertise older and younger pilots and highly-educated non-pilot controls, Morrow et al. (1994) found that recall of auditory material was most likely to be forgotten by older participants and that expertise (pilot versus non-pilot) mitigated age-related declines only for highly domain-specific tasks. It was suggested that the poorer performance in communication tasks by older participants was associated with declines in working memory and in particular, older subjects’ reduced capacity to store recent auditory information. In a second experiment, Morrow et al. found that recall of newly learned map knowledge was also poorer for older participants, regardless of pilot or non-pilot status. In a subsequent map-learning study by Morrow et al. (2003), note-taking was permitted during the simulated flight. With note-taking (thereby lessening demands on working memory), older pilots performed as well as younger pilots.

The well-researched phenomenon of age-related cognitive decline in the general population (Johnson, Logie, & Brockmole, 2010; Salthouse, 2000; Verhaeghen & Salthouse, 1997) and for pilots (Hardy, Satz, D’Elia, & Uchiyama, 2007; Taylor, O’Hara, Mumenthaler, Rosen, & Yesavage, 2005) supports the concern that older pilots might be at risk for poorer flight performance and greater involvement in aviation accidents when compared to younger pilots. As shown in Figure 2.3, cross-

sectional analyses suggest that the slowing of cognitive processes may begin as early as the third decade of life, and for most “fluid intelligence” tasks, there is continued decline across the lifespan (for a review of age and cognition see Salthouse, 2010). Thus, most investigations of age and pilot performance have included concomitant assessment of a range of cognitive measures.

Figure 2.3. Age trends in cognitive abilities. Adapted from Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, 30(4), 507-514

In sum, hierarchical regression models of aviation performance have shown that age predicts unique variance in many flight measures beyond that predicted by cognition and expertise factors (Taylor et al., 2005). Models of pilot performance have shown age to be a significant predictor of discrete piloting skills such as flight path deviations (Causse, Dehais, and Pastor., 2011; Taylor, O'Hara, Mumenthaler, & Yesavage, 2000), reading back air traffic commands (Morrow, Menard, Stine-Morrow,

Teller, & Bryant, 2001), or dialling in communication frequencies (Taylor et al., 2000). Causse et al. (2011) found that age effects on flight path deviations were fully mediated by executive cognitive functions and processing speed. Other analyses have taken a different approach and created criterion variables that are composites of twenty or more dependent variables collected during simulated flight (Taylor et al., 2007; Yesavage et al., 2011). Taylor et al. (2005) found that working memory, processing speed, and visual tracking accounted for significant portions of summary flight score variance, with age effects not fully mediated by the cognitive factors.

Thus far, investigators have not successfully accounted for all the effects of age on pilot performance, although there is evidence that domain-independent cognitive variables can explain portions of age-related variance for some aspects of simulated flight performance.

2.2 Pilot Cognition in General Aviation

Concern regarding the association of pilot cognition and safety is a longstanding issue. In 1998 the Human Factors and Aviation Medicine publication warned of the increased risk of dementia caused by tobacco smoking (Mohler, 1998). In 2005, the Flight Safety Foundation (FSF) also published an issue of their Human Factors and Aviation Medicine dedicated to dementia and piloting. While suggesting that most experiences of forgetfulness are not likely signs of disease, the FSF highlighted dementia, with its declines in cognitive abilities, as a concern for the aviation industry. The cognitive deficits associated with dementia were also directly related to skills associated with pilot safety (memory, learning, thinking, and reasoning) (FSF, 2005). While it has been suggested that pilots have higher than average cognitive abilities, a study of cognitive performance in commercial airline pilots found that most fluid aspects of cognition showed predictable negative relationships with age (Hardy, Satz,

D'Elia, & Uchiyama, 2007). Hardy et al. found that immediate and delayed recall of verbal material and the immediate recall of visual-spatial material remained relatively stable in the sample of pilots aged 28 to 62 years. Measures reflecting executive ability, verbal and visual learning, attention, and psychomotor speed declined gradually and steadily with older pilot age. The presence of outliers in the data was strongly associated with older age, such that the authors recommended neurocognitive screening at routine physical examinations for all pilots starting at age 40 years.

The association between pilot cognition and performance is supported by annual reports linking pilot-related error to 70% of aviation accidents (Air Safety Institute, 2012; National Transportation Safety Board [NTSB], 2012). Pilot-related error is defined as accidents caused by the action or inaction of the pilot (Air Safety Institute, 2012). Rebok, Qiang, Baker and Li (2009) investigated air taxi crash reports from the NTSB in the United States between 1983 and 2002 and reported that, of the identified errors, 27% were flawed decisions and 26% were inattentiveness. Similarly, Oster, Strong, and Zorn (1992) investigated the human factors associated with fatal and non-fatal pilot-related general aviation accidents and found that flying skills could explain 34% of all accidents, whereas pilot judgement was associated with the remaining 66% of crashes. Moreover, most pilot-related accidents occurred during takeoff or landing, phases of flight when demands on pilot cognition are greatest (Air Safety Institute, 2012). The reliance upon multiple cognitive mechanisms, such as task-sequencing, divided attention, and decision-making abilities, during these high task load phases of flight has been linked to the high accident rate during these phases. The association of pilot error with aviation accidents was similar between pilot age groups (Bazargan & Guzhva, 2011).

The close association of pilot cognition and performance has prompted experimental research using flight simulation to investigate specific cognitive factors and their association with performance in the cockpit. These studies have linked domain-independent cognitive factors, such as inhibition, visual-spatial attention, working memory, and processing speed to general flight performance. The domain-dependent cognitive factors of prospective memory and situation awareness have also been investigated regarding their relationship with flight performance.

2.2.1 Domain-Independent Cognition and Flight Performance

Inhibition. Cognitive inhibition, or interference control, is an executive cognitive function that supports attention to relevant information, discourages the intrusion of irrelevant information into working memory, and inhibits activity known as “prepotent responses” when that response is no longer appropriate (Hasher & Zacks, 1988). Studies of pilot decision-making have examined whether accuracy could be explained in terms of underlying cognitive inhibition. Causse et al. (2011, p.14) used a “spatial Stroop” measure where pilots responded to congruent or incongruent WORD–LOCATION pairings (e.g., the word “LEFT” either on the right or left side of the screen). A small correlation was observed between inhibition, and flight path deviations and accuracy for a landing decision task. However, inhibition was not a significant predictor in a regression model of performance for either criterion variable. Taylor et al. (2005) investigated whether inhibition measures from the CogScreen-AE (Kay, 1995) predicted performance on accuracy for ATC communication tasks. Pilots with better inhibitory mechanisms were found to perform significantly better on the communication tasks. Cognitive inhibition, along with processing speed, was also shown to explain almost all the relationship that existed between pilot age and a working memory span measure. The working memory measure was the strongest

predictor of pilot performance. Thus cognitive inhibition was further implicated in pilot performance via its mediation of working memory effects (Taylor et al., 2005).

Visual-spatial attention. Efficient visual-spatial attention directs mental resources to relevant stimuli and spatial locations. The need for optimal visual-spatial attention in aviation is obvious, as the limited attention possessed by the pilot must be allocated to various stimuli and locations while keeping in mind the current goals and context. For example, periodically tracking altitude, heading, and airspeed indicators in the cockpit require appropriate visual-spatial attention. Taylor et al. (2000) found that a measure of visual-spatial attention predicted significant unique variance in a multiple linear regression model explaining summary flight simulation performance. Specifically, a visual-tracking error measure was significantly correlated with the course deviation, traffic avoidance, and approach components of the flight summary score. Kennedy, Q. et al. (2010) assessed visual-spatial measures using the CogScreen-AE, where errors in monitoring a moving stimulus were recorded. Pilots kept a moving target from drifting away from centre, while they were concurrently engaged in a 2-back working memory task. In this study, visual-spatial attention was not examined directly as a predictor of flight performance, but was explored for its capacity to account for age differences in flight control measures in simulated flight. Visual-spatial attention did not significantly interact with pilot age, such that visual-spatial attention did not explain the reduced performance in flight control shown for older pilots (Kennedy, Q. et al., 2010).

Working memory. Working memory is a cognitive function described by Baddeley and Hitch (1994) as responsible for organization, maintenance, and manipulation of relevant stimuli in the short term. Traditionally, working memory has been considered a central pillar of cognition and has been featured in studies of pilot

performance. In a longitudinal general aviation study, Taylor et al. (2000; 2005) examined flight performance of pilots with a broad range of expertise ($N=276$), flying a Frasca 141 simulator. Using composite measures from the Cogscreen-AE pilot assessment battery (Kay, 1995), Taylor et al. developed cognitive factors regression models to predict pilot performance, and found that working memory, speed of processing, cognitive inhibition, and visual memory were strongly associated with summary measures of simulated flight performance. Causse et al. (2011) studied 32 general aviation pilots' performance using a desktop flight simulator that produced measures pertaining to flight path deviations and pilot decision-making at the landing phase of flight. Cognitive and psychomotor tests were examined for their association with flight performance. After controlling for pilot age and experience, a 2-back visual working memory and a reasoning test were the strongest correlates of flight path deviation scores. Psychomotor reaction time, working memory, and reasoning were found to offer a stronger account of pilot performance than pilot age. The 2-back working memory measure and error rates from a set-shifting task also predicted performance for a weather-related decision-making task.

Disentangling age from cognitive factors is a difficult undertaking as there is a clear trend for working memory to decline with age. Johnson et al. (2010) found consistent declines in performance as age increased. In Figure 2.3 a clear downward drift in several working memory mechanisms is evident across the lifespan. The standard deviation lines at the top of the graph also demonstrate the increased variability in scores as the cross-sectional sample "ages" (Johnson et al., 2010). Working memory mechanisms can include the binding of multiple aspect of a stimulus (such as colour and shape), and recognizing and recalling visual or auditory patterns. Arguably aviators must make use of these mechanisms in order to fly the aircraft,

navigate on the ground and in the air, and communicate as needed. A decline in the abilities of one or more of the aforementioned cognitive mechanisms could give rise to reductions in flying ability, perhaps mediated by declines in domain-dependent cognitive factors. Taylor et al. (2005) predicted that working memory, processing speed, and inhibition are the domain-independent cognitive variables most associated with pilot performance, and the primary mediators of the effects of age on pilot performance. Specifically, older age was thought to negatively impact processing speed and inhibition. Slower processing speed and poorer inhibition lead to working memory declines. Finally, working memory reductions are directly associated with decrements in flight performance. Thus, it was proposed that the effects of older age, slower processing speed and poor inhibition indirectly influenced pilot performance via working memory.

Figure 2.3. Age effects on five aspects of working memory. Top grouped lines represent size of standard deviations in working memory measures across age groups. Bottom grouped lines represent standardized mean scores for the working memory measures. Reprinted from Johnson, W., Logie, R. H., & Brockmole, J. R. (2010). Working memory tasks differ in factor structure across age cohorts: Implications for dedifferentiation. *Intelligence*, 38(5), 513-528.

Processing speed. Cognitive processing speed is a domain-general cognitive variable that is often used to explain variance in other cognitive measures, (e.g.,

working memory, Salthouse, 1992) and in more complex aspects of cognition, such as reasoning skills (Salthouse, Mitchell, Skovronek, & Babcock, 1989). Similarly, processing speed has been suggested as a cognitive factor influencing pilot performance (Taylor et al., 2005). Researchers of pilot performance, particularly those investigating effects of age on performance, typically regress their criterion variables on some type of processing speed measure. Results have been inconsistent with respect to the role processing speed might play in pilot performance. Taylor et al. (2005) found that processing speed was significantly correlated with accuracy on an ATC communication task, and was a plausible explanation for the significant correlation found between working memory and the communication task. In contrast, processing speed was not directly predictive of other ATC communication tasks (Morrow et al., 2003). The type of task used to measure processing speed could be a cause of conflicting results in the literature. Causse et al. (2011) utilized a simple psychomotor task requiring pilots to quickly tap a screen in response to the presentation of a target stimulus. This basic measure of speed was significantly correlated with other cognitive tests such as a two-back working memory task and inhibition (perseveration errors for the Wisconsin Card Sort [Berg, 1948]), as well as with pilot age. However, in this study the response speed measure was not predictive of flight path deviations or accuracy for a weather-related landing decision. Taylor et al. (2005, p. 123) used a more comprehensive measure of processing speed, derived from working memory-related tasks: “pattern comparison and digit copying” (from Salthouse, 1992), perhaps explaining the relationship found between the variables of read-back accuracy of the ATC messages and processing speed.

Processing speed has been posited as an isolable cognitive mechanism, which acts as a “limiter” of the speed that cognitive processes may be carried out (Kail &

Salthouse, 1994). However, it could also be argued that processing speed is not a cognitive mechanism in the same sense that working memory, for example, is a mechanism. Working memory has neural correlates; with many studies theorizing specific brain areas that mediate working memory function (see Rottschy et al., 2012 for a meta-analysis). In contrast, after many decades of probing the brain for its structural and functional purposes, processing speed has not been shown to be a mechanism that can be attributed to a specific brain area. The speed at which cognitive processes are carried out appear to result from the integrity of the material responsible for these processes, and from the structure of the processes (e.g., the larger and more complex a brain network required to carry out a process the more time it could theoretically require to produce the final action) (Kennedy, K. et al., 2015). The relationship between processing speed and brain morphology is supported by the finding that young adults with autism have slowed cognitive processing and changes to white matter microstructure, when compared to age-matched controls (Travers et al., 2014). While studies attest to a causal relationship between brain morphology and processing speed, Salthouse (2011) maintains that other causal explanations for individual differences in processing speed are plausible, such as a global speed limiting mechanism. In contrast to this view, working memory capacity and processing speed are correlated because they are both directly, but separately, influenced morphological state of the brain. This view is supported by Eckert (2011) who found that reductions in the older participants' response times was accounted for by differences in attentional networks involving the frontal cortex. Similarly, Kennedy, K. et al. (2015) observed slower response times for lexical decision tasks for older subjects; responses that were also associated with bilateral, rather than unilateral hemispheric networks of activity.

In summary, slower processing speed was associated with poorer communication task performance in the cockpit. Other tasks, such as decision-making during landing, were not associated with processing speed. The measures used to define cognitive processing speed tend to vary, and are frequently derived as a secondary measure of other tasks designed to isolate cognitive mechanisms. For example, processing speed metrics have been obtained from the reaction times associated with accurate responses for working memory tasks. Thus disentangling processing speed from working memory (Taylor et al., 2005), as a predictor of pilot performance, presents a challenge for investigators of individual pilot differences and performance. Furthermore, the argument that slower cognitive processing associated with older age is essentially a by-product of functional connectivity and morphological changes associated with the older brain also questions the relevance of processing speed as a predictor of pilot performance.

2.2.2 Domain-Dependent Cognition and Flight Performance

That domain-dependent cognition might be associated with pilot age was supported by results from a survey of pilot examiners and flight instructors. A Canadian nation-wide survey of flight instructors and pilot examiners was conducted in conjunction with the present research to determine if these frontline stakeholders noted concerns of a cognitive nature when teaching and reviewing skills of younger and older pilots (Van Benthem et al., 2011b). Almost 70% of respondents indicated that they observed differences between younger and older pilots when considering competency and safety issues related to the training or currency aspects of flight instruction. Flight instructors and pilot examiners highlighted decision-making and judgement as areas of concern for younger pilots (less than 25 years old) but reported error recognition, procedural knowledge, and physical ability as primary concerns for older pilots (age

60 years and older). Prospective memory and situation awareness during flight are two other domain-dependent cognitive factors that have been examined in relation to pilot performance.

Prospective memory. While pilots rarely forget important facts or crystallized knowledge associated with retrospective memory, there are documented instances of prospective memory failures (Nowinski et al., 2003). Prospective memory is a cognitive function that supports remembering to complete some thought or action in the future. The ability to remember to complete important cockpit tasks relies on pilot prospective memory. Utzl (2008) described three types of prospective memory tasks. Holding an intention in memory over a period of time is called prospective memory proper. For example, remembering to file a flight plan utilizes prospective memory proper. Memory for vigilance relies on maintaining an intention in consciousness while engaged in other tasks. Vigilance events usually take place over shorter time periods. The requirement to check altitude at regular intervals could be described as vigilance prospective memory. The third type, habitual prospective memory, is similar to the proper form, except that the tasks are performed routinely, such as remembering to complete radio calls at regular locations in a pattern. There are also two types of memory cues that signal the appropriately-timed execution of an intended act. Memory cues can be time-based, such as when a task should be performed at a specific future time (e.g., three o'clock). Memory cues can also be event-based, such that the execution of the task is associated with some future event (e.g., on approach). Dismukes (2008) found that pilots performed all three types of prospective memory tasks. However, habitual tasks associated with event-based memory cues were particularly common and were linked with prospective memory failures.

Prospective memory failures in aviation can have catastrophic results (for a review see Dismukes & Nowinski, 2007). In one study of the Aviation Safety Reporting System, 74 of 75 reports identifying a memory failure involved prospective memory errors (Nowinski et al, 2003). Pilots' failures to complete prospective memory tasks occurred as a result of interruptions, missed cues, changes to routine, and interleaving/time-sharing of concurrent tasks. Other work has shown that prospective memory failures were exacerbated by (a) high working memory load (Dismukes & Nowinski, 2007), (b) the lack of or inefficient encoding of the intent to remember, and (c) a mismatch between the original task cues and the cues available after an interruption (Dodhia & Dismukes, 2009). The pervasive use of checklists in the cockpit apparently does not prevent prospective memory failures, even for highly experienced pilots (Nowinski et al., 2003).

The effects of age on prospective memory have been inconsistent. Researchers report conflicting findings depending on whether data was collected in controlled laboratory settings, or in participants' home environments (see Uttl, 2008; 2011 for reviews). In some natural contexts, prospective memory was not selectively worse in older participants. However, in laboratory settings older participants have demonstrated poorer prospective memory performance (Uttl, 2008). This phenomenon is called the age paradox in prospective memory. A meta-analysis of prospective memory and aging studies found that older and younger adults performed similarly in natural contexts only when the prospective memory was of the "proper" type, and when it was associated with clear time-based memory cues. Other types of prospective memory, and tasks linked with event-based memory cues demonstrated age-related declines in performance (Uttl, 2008).

In summary, prospective memory, as a cognitive construct associated with remembering to perform critical tasks in the future, is a skill relevant to successful daily activities (Uttl, 2008) and has been associated with aviation outcomes (Dismukes & Nowinski, 2007). Prospective memory for cockpit tasks has implications in the current work as it is also a cognitive function subject to age-related declines (Uttl, 2008).

Situation awareness. Parasuraman, Sheridan and Wickens (2008, p.144) describe situation awareness as the “continuous diagnosis of the state of a dynamic world”. Situation awareness is thought to build on three levels of cognitive processing (Endsley, 1988; 1995). The first level is the perception and processing of multi-modal stimuli into meaningful units of information; at the second level this information is assembled into a comprehensive model of the aviation environment; the third level projects current information into a proposed future model that anticipates a near future state of the environment (Endsley, 1988; 1995). Endsley’s definition of situation awareness as the “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” (Endsley, 1995, p. 36) offers a comprehensive picture of situation awareness. The maintenance of an accurate, timely, and meaningful model of the environment with its relevant stimuli has obvious links to general aviation pilot performance.

Pilot loss of situation awareness has been implicated in numerous accident reports published by the NTSB (2012). Bolstad, Endsley, Costello and Howell (2010) noted that situation awareness is a key factor in general aviation safety and that poor situation awareness is a factor in poor aviation outcomes. Bolstad et al. (2010) evaluated a computer-based situation awareness training system for general aviation

pilots and found that lower situation awareness scores contributed to poorer simulated flight outcomes. Endsley and Bolstad (1994) investigated individual differences in pilot situation awareness and found that differences pertaining to perception and spatial skills were most associated with situation awareness scores. The study sample consisted of fighter pilots; therefore, generalization of the findings to general aviation should be made with caution. Situation awareness, as a theoretical construct, has proven useful in categorizing piloting errors. Jones and Endsley (1996) examined the Aviation Safety Reporting System database to identify which level of situation awareness was most associated with pilot incidents and found that over 75% of the errors identified were associated with level one situation awareness, 20% were level two and the remaining few percent were associated with level three errors.

Studies of situation awareness, navigation, and mental resources have revealed the importance of considering multiple factors when predicting performance in complex cognitive tasks. For example, a recent well-publicized study has shown that situation awareness might be linked to specific cortical areas. Maguire, Woollett and Spiers (2006) suggested that creating and maintaining mental schemas for purposes of navigation resulted in London taxi drivers having greater than normal grey matter volumes in the mid regions of the hippocampus. Adamson et al. (2012) found that hippocampal volumes in general aviation pilots were not associated with simulated flight performance at baseline. However, in subsequent measurements higher-rated pilots with greater hippocampal volumes were associated with superior flight performance. Adamson et al. reported that situation awareness capacity alone did not necessarily give rise to better performance, but that pilots with higher levels of expertise seemed better able to capitalize on the hypothesized increased situation awareness resources available. The proposed influence of age on situation awareness

and the lack of better flight performance by pilots with larger hippocampal volumes but lower expertise support a multi-factorial approach to modeling general aviation pilot performance.

Situation awareness in older general aviation pilots has not been studied specifically. In a review of the pertinent aging literature, Bolstad and Hess (2000) speculated that situation awareness performance might subject to age-effects in general aviation pilots, for all three levels of situation awareness. Bolstad and Hess suggested that age-effects in situation awareness could be associated with reductions in information processing speed, working memory capacity, inhibition, and attentional processes, that is, processes that typically show age-related declines.

In sum, the ability to form accurate, comprehensive mental representations of their current and future flight situation is a critical cognitive task for all pilots. Speculation regarding age-effects on situation awareness indicates that situation awareness, particularly level one, could be a relevant factor in causal models of aviation outcomes.

2.3 Pilot Expertise and Flight Performance

Expertise is routinely examined in studies of pilot performance. Factors associated with expertise include pilot license (student, private, commercial etc.), additional ratings (night, float etc.), years licensed, total hours flown, and recent pilot-in-command hours. Higher levels of pilot rating have often been associated with better pilot performance in flight simulators (Taylor et al., 2005; Yesavage et al., 2011). In contrast, when expertise is measured by total hours flown, the relationship between expertise and simulated flight performance is not significant (Dehais et al., 2014). In an analysis of recorded aviation accidents, a GAO (2012) investigation of pilot expertise and accident rates found that private pilots with more than 100 hours were

just as likely to be the cause of accidents than those with less than 100 flight hours (76 and 73% respectively). Nevertheless, common sense dictates that higher expertise, as determined by hours spent engaged in the task, should be accompanied by greater proficiency for any occupation (see discussions by Salthouse, 1992, 2006, 2010).

Visual Flight Rules (VFR) flight into instrument meteorological conditions is one of the most common pilot violations associated with crashes in general aviation and, as such, expertise, as well as other possible predictors of performance, have been studied in simulation studies involving weather-related tasks. Wiegmann, Goh and O'Hare (2002) studied 36 private pilots (mostly instrument rated) and found that higher expertise, as measured by total and recent flight hours, was associated with better performance when deciding to change routes due to weather conditions. However, only recent pilot-in-command hours were significantly correlated with weather-related performance. Causse et al. (2011) found that more total hours flown was associated with less deviation from the ideal flight path and with greater likelihood of selecting the correct course of action during risky weather-related landing conditions.

Time-sharing is considered an executive function task that is utilized frequently, and at times intently, during flight. Tsang and Shaner (1998) used domain-general tasks to explore whether expertise (pilot versus non-pilot) attenuated dual-task decrements when both tasks competed for the same cognitive resources. More than 20 planned statistical tests of age and expertise interactions resulted in just one moderate significant moderating effect of expertise. This analysis found that younger pilots showed less of a time-sharing decrement for a tracking/spatial ability task than did the older non-pilots. Results of the Tsang and Shaner (1998) and Morrow et al. (2003) study indicated that the moderating effect of expertise might be found for age-related

declines in some ecologically valid laboratory tasks. However, the bulk of evidence suggests that within the pilot population, greater levels of expertise do not significantly reduce the declines in performance experienced by older pilots. Interestingly, in Tsang and Shaner, an effect of age was observed for most tasks in all conditions, yet an effect of expertise was much less evident during the single task condition in comparison to any of the dual task conditions. This suggests that expertise could be associated with a more general cognitive workload factor.

In various occupational domains expertise has been posited as a variable that “protects” skills that might otherwise deteriorate due to age or cognitive decline (see Salthouse, 1991 for a brief historical review). Findings in the field of aviation have been mixed with respect to expertise compensating for older pilot age. Graded levels of pilot rating did not attenuate age-effects for air traffic control communication tasks (Taylor et al., 2005). In contrast, some support was found for the beneficial effect of increased flight hours for older pilots on air traffic control (ATC) task read-back accuracy and for expertise (as defined by pilots versus non-pilots) for read-back accuracy (with note-taking) (Morrow et al., 2003). Expertise was not found to moderate the influence of age on executing ATC instructions, or on flight summary scores (Taylor et al., 2007). Figure 2.1 demonstrates the results from the Taylor et al. (2007) study where pilots of various ratings flew a Frasca 141 flight simulator. All pilot expertise groups (other than the highest-rated group) showed below average flight summary scores by 50 years of age. After age 60 even the highest-rated group of pilots performed below the mean for the flight summary score. As shown in Figure 2.1 flight summary scores also declined from T-0 to T-1 for pilots in the youngest group (ages 40 to 49) and that performance decline was not a phenomenon only occurring with pilots over the age of 60 years. For pilots with low levels of expertise, age effects

can be seen as young 40 to 49 years. Taylor et al. pointed to the unexpected finding that the oldest highest-rated pilots improved their summary flight scores over time. That older higher-rated pilots improved over time was attributed to just one aspect of flight performance, the traffic avoidance measure, where regression to the mean and practice effects were potential causes for this phenomenon. Figure 2.1 illustrates that, despite higher levels of expertise and younger age predicting better baseline performance and lower rate of change scores, expertise did not significantly attenuate the poorer performance seen in older pilots. The Taylor et al. results were supported by an update published four years later with similar results (Yesavage et al., 2011). In a comparable study from the same research centre as Yesavage et al., flight path deviations were recorded during approach and holding patterns (Kennedy, Q. et al., 2010). Larger bank angle deviations were found to be predicted by older age, less expertise, and an age by experience interaction term, (total $r^2=0.24$). In this analysis expertise attenuated the effects of age on the bank angle deviation scores.

A caveat regarding expertise effects is necessary. Disentangling the unique effects of age from expertise (and vice versa) has been difficult to accomplish (Hardy & Parasuraman, 1997). As pilots progress through the life span there may be advancement of pilot licensure, ratings, and total flight hours. However, as suggested by Figure 2.3, there is the possibility of mental capacity reductions occurring concurrently to the accumulation of expertise through the years. In an effort to adjust for the confound between age and expertise, some researchers have controlled for hours flown when conducting correlational analyses of age and flight performance. Yet hours flown have not always predicted performance (Causse et al., 2011; Wiggins, 2011). Taylor and colleagues (2000; 2005; 2007) concluded that pilot licensure and ratings (recreational pilot, private with VFR, private with IFR, commercial etc.) is the

preferred index for expertise. However, switching to licensure as the primary measure of expertise does not eliminate the age confound. Pilots with higher ratings may also be the pilots with greater capital (time and financial), and greater cognitive aptitude, drive and ambition relating to aviation. Not unexpectedly, Taylor et al. (2005; 2007) reported that expertise was most strongly associated with flight performance when represented by pilot rating, as opposed to overall flight time or recent pilot-in-command hours.

2.4 Mental Task Load

Mental task load is a feature of the human-machine system that influences the subjective experience of the pilot and the assignment of cognitive resources needed to achieve a goal. Mental workload has been described using terms borrowed from economics. For example, mental task load references the relationship between the demands of the task and the mental capacity, or supply, available to carry out the task (Parasuraman et al., 2008). Mental task load is considered a *latent* factor because it must be measured indirectly by way of physiological markers, subjective experience, and objective measures of performance on primary tasks, or by monitoring performance in a secondary task. Mental task load can be influenced by the workload conditions of the simulated flight, and importantly to the present study, mental task load may also be a significant predictor of pilot performance and aviation outcomes. While physiological indicators of workload have been investigated in the field of aviation they were not examined in the present program of research.

Subjective measures of workload include self-rating scales such as derivations of the SWAT (Subjective Workload Assessment Technique) (Reid, Shingledecker, & Eggemeier, 1981), the Bedford Workload Rating Scale (Roscoe & Ellis, 1990) and the NASA TLX Index (Hart & Staveland, 1988). Typically at pre-determined periods

subjects are requested to rate their perceived workload regarding various aspects of the task. Subjective measures of workload frequently match manipulated workload conditions in experimental work. Dehais et al. (2014) investigated inattentional “deafness” of pilots during a simulated landing scenario. An increase in task workload, by adding wind shear during approach, resulted in missed auditory warning tones for a landing gear failure. Pilots completed subjective ratings of performance, stress, mental workload, and task difficulty for the wind shear and no wind shear conditions. Objective measures of workload were ascertained via heart rate monitoring. Both subjective and objective workload metrics were sensitive to the change in task load brought about by the wind shear factor. The high task-load condition resulted in missed auditory and visual cues for failed landing gear, and subsequent faulty decisions to land. No differences in total hours flown, the study’s measure of expertise, were found between the inattentional “deaf” group and the group who did hear the critical alarms.

Researchers have manipulated mission difficulty conditions using off-nominal events to evaluate the effects of heightened workload on simulated performance and on self-report of mental task load. The effect of small increases to approach angles increased subjective workload in pilots flying a Cessna 172 simulator (Boehm-Davis et al., 2007). When the standard descent angle was increased from 5% to 7% pilots were still able to land safely. However, as workload ratings increased significantly, approaches also became more unstable, and touchdowns occurred further down the runway. In addition, the pilots rated conditions with steep descent angles as approaching the maximum scores for cognitive workload. This is important information for stakeholders considering changes to standard operating procedures and runway designs (Boehm-Davis et al., 2007).

A straightforward method for measuring workload effects on pilot performance is via the manipulation of primary task complexity (Cain, 2007). Primary task performance can be influenced through alterations of features of the environment, tasks, and cockpit. For example, Coffey et al. (2007) increased pilot workload in a flight simulation study by adding turbulence, reducing visibility, and decreasing navigational aids. A significant workload effect on pilots' ability to maintain desired altitude was found. The effect of weather conditions is echoed by NTSB accident statistics that found that deterioration of weather conditions, with the associated increases in task workload, is a frequent cause of fatal accidents amongst private pilots (Goh & Wiegman, 2002). Numerous fatal accidents involve situations where pilots, not rated to fly in poor weather conditions, continue flight into "meteorological" conditions. The primary causes of such accidents include pilots' inability to simultaneously manage the poor weather and terrain conditions, and maintain aircraft control and spatial orientation (Goh & Wiegman, 2002).

Mental task load that exceeds a critical threshold for relevant cognitive mechanisms can have serious consequences. Wiggins (2011) hypothesized that distance flights that culminated in landings at high-event aerodromes produced workload effects that differentially targeted performance measures supported by long-term memory, as compared to performance on other tasks not reliant on long-term memory. Simulated performance was measured by deviations from the ideal altitude, pitch, and heading on the last leg of a long flight. Only altitude deviations, the performance that relied on information stored in long-term memory for use at the appropriate segment of the flight, were subject to the effects of high workload experienced on the final leg of the flight. Multiple regression models revealed that self-assessments of mental workload and capacity to control the aircraft significantly

predicted altitude deviations. Total hours flown as pilot-in-command (PIC) and fatigue management did not significantly predict altitude maintenance scores. These findings, and those of Boehm-Davis et al. (2007), illustrate that pilot subjective ratings of workload can provide important information regarding safety during flight.

Finally, another objective measure of workload in an experimental setting involves the introduction of a vigilance task performed concurrently with the primary task. Examples of secondary tasks include randomly presented auditory or visual stimuli for detection and response (ongoing during the primary task). Response times, accuracy, and d-prime response rates to the stimuli are calculated for periods of interest. Typically, longer response times and decreased detection of these secondary stimuli reflect states of higher cognitive workload. The assumption is that performance on the secondary (less important task) will vary in accord with the operators "unused capacity" (Vidulich & Tsang, 2012). Performance on the primary task might or might not be similarly affected by experimenter manipulations of workload. In Lassiter, Morrow, Hinson, Miller, and Hambrick (1996), pilots flying a desktop simulator had greater deviations from the ideal altitude and airspeed when flying during the more difficult condition of the secondary task (as compared to deviations associated with the easier condition of the secondary task). Investigators also found that when comparing extreme expertise groups (as measured by total hours flown) a high level of expertise was beneficial for younger pilots in regards to the accuracy on the secondary task during the high-workload condition. Older pilots with low expertise demonstrated the largest reductions in accuracy on the secondary task in the more difficult condition. These findings indicated that individual pilot differences interacted with level of difficulty for secondary tasks, and that older pilots with less

experience were at a greater disadvantage than younger pilots, or the older pilots with more experience, when secondary task workload was increased.

In sum, pilot mental task load is frequently assessed during simulated and actual flight to determine why changes to the flight environment (e.g., avionics, cockpit displays) influence performance and aviation outcomes. While typically congruent, subjective and objective measures of mental task load may not always correlate significantly with each other. However, both subjective reports and objective metrics have shown sensitivity to fluctuations in mental task load. The implications of older age, lower expertise, and more difficult flight conditions on *primary* measures of flight performance were not clear from the literature discussed (Lassiter et al., 1996). Although one might speculate that--in regards to pilot performance--interactions between age, expertise, and task difficulty are likely, given that secondary task performance has been affected by these interactions. Thus, in the present work the effects of flight environment manipulation on flight performance, perceived mental task load, and the potential interactions of these factors with pilot individual differences were investigated in models of pilot performance.

2.5 Summary of the Literature Review

This review of the literature has demonstrated that easily ascertained pilot attributes, such as age and expertise are correlated with pilot performance in predictable ways. Older age and higher levels of expertise have opposite effects on pilot performance such that younger and more experienced pilots routinely show better simulated flight performance than older and less experienced pilots. In contrast to some earlier expectations, pilot age and expertise did not always interact such that higher expertise “protected” older pilots from the declines in performance frequently noted for older pilots. Consequently, research turned to investigations of domain-independent

cognitive mechanisms as predictors of pilot performance. A deleterious effect of low working memory was found for a variety of pilot tasks. Effects of visual-spatial attention, cognitive inhibition, and processing speed are less consistent with regard to their associations with pilot performance. As might be expected, the executive function of cognitive inhibition was associated with pilot communication, whereas visual-spatial attention was associated with deviations from ideal trajectories. Few studies have examined the effects of domain-dependent cognition on aviation outcomes. Prospective memory for cockpit tasks is a domain-dependent cognitive variable associated with aviation outcomes and with age. Situation awareness has also been associated with critical incidents in aviation. The relationship of the three levels of situation awareness with pilot age remains to be clearly delineated. That domain-dependent cognition might be associated with pilot age was supported by results from a survey of pilot examiners and flight instructors, with respondents indicating divergent concerns regarding performance on domain-dependent cognitive variables for younger as compared to older pilots. Younger pilots were considered more likely to make errors of decision-making and judgement, whereas there was greater concern with procedural knowledge and error recognition for older pilots. The modulating impact of workload conditions has also been examined in relation to pilot performance. Both subjective and objective measures of mental task load show associations with flight performance, suggesting that a variety of workload measures might be useful in cognition-oriented models of aviation outcomes.

Chapter 3

Methods

This Chapter describes the methods employed in the experimental phase of the research. The results of the experimental phase of this work provided robust data structures used for data modeling in phases two and three. Support for the inclusion of a wide range of pilot attributes, flight performance measures, and cognitive factors in these data structures was indicated by the review of the literature. The present work utilized an experimental paradigm to collect flight performance data using a Cessna 172 simulator. Information regarding the cognitive health of pilots was collected via two standardized cognitive assessment batteries.

3.1 Participants

Data was collected from 110 pilots during two six-month research phases occurring between 2010 and 2012. Participants were recruited from local flying clubs and schools after permission from the school or club was obtained. Recruitment methods were primarily via posters in local flying clubs and schools, which also resulted in word of mouth recruitment amongst pilots. Inclusion criteria were having a minimum of a valid student's certification and a valid medical certification to fly, as well as being eighteen years of age or older. Pilots who had participated in a previous Cessna simulator study in the ACE lab were considered ineligible. All participants provided signed informed consent. Two participants were excluded from the study, one by request, and one due to lack of a current medical certificate, resulting in a final sample size of 108 certified pilots. See Appendix A for the Certificate of Ethics Clearance from the Carleton University Psychology Research Ethics Board. See Appendix B for the Informed Consent Form and Appendix C for the Recruitment Poster.

3.2 Simulation Environment

Pilots “flew” a Cessna 172 simulator that was assessed as a level 6 flight training device. The simulator structure was a converted Cessna 172 partial fuselage. The cockpit was outfitted with instruments and controls specialized for flight simulation linked with Microsoft Flight Simulator X® software. Projection graphics were produced by Microsoft Flight Simulator X “on the fly” and were not pre-rendered. Locations were geo specific in that they produced terrain and buildings etc. modeled on actual aerodromes in Canada. The graphics architecture incorporated three large screens for approximately 120 degrees of horizontal visual field of view and 45 degrees of vertical field of view. See Appendix D for a sample of the data captured by the flight simulation data acquisition computer. At approximately each second of flight the data application computer logged the time and the pilot’s location, airspeed, heading, bank, pitch, and altitude. Figure 3.1 illustrates the Cessna 172 simulator cockpit and its relative position to the graphics.

Figure 3.1. Cessna 172 simulator cockpit running Microsoft® Flight Simulator 10 software and graphics

Ecological validity of the simulation environment was enhanced by the inclusion of realistic sound effects. The consistent beneficial effects of pilot expertise on performance (discussed at various points in the present work) supported the validity of the simulator as a tool for measuring pilot performance. All measures were designed to investigate the relation of pilot cognition, age, expertise, and workload to general aviation performance. What follows is the description of procedure, measures collected via the intake demographic form, the flight simulator protocol, and two standardized neurocognitive test batteries. Some preliminary statistics regarding overall and expertise group means are provided for most measures.

3.3 Procedure

Participants provided informed written consent after a thorough description of the study purpose and activities. Participants were scheduled during the morning or afternoon according to their convenience and completed all activities within the same portion of the day. Pilot demographic and flight experience was recorded. Participants completed two cognitive health batteries, the DCAT (DriveABLE Testing Ltd, 1997) and the CogScreen-AE (Kay, 1995). The order of the batteries and the order of batteries versus simulated flight were counterbalanced where possible. Of the approximately 100 measures produced by the CogScreen-AE, two showed weak order effects: SATDIPER (perseveration errors in shifting attention task) was found to be lower when the CogScreen was administered after the DCAT and DATSCRTC (divided attention visual sequence comparison, reaction time) was lower when the flight simulation took place before the cognitive health screening. None of the DCAT subtests demonstrated test-order effects.

Regarding the simulated flight expectations, participants were provided with a written, verbal, and graphic presentation outlining the requirements for flying the “perfect”, or ideal, circuit. The presentation was followed by a practice session (four full circuits with experimenter feedback) to familiarize participants with the simulation environment. All participants spent approximately half an hour in a warm-up phase of simulated flight to become accustomed to the flight simulator avionics and gages, and to reduce learning effects. The flight simulation experimental protocol required pilots to fly three left-hand circuits in the low-difficulty condition and then three left-hand circuits in the high-difficulty condition. The simulated weather was clear for all circuits. Participants were briefed with written, verbal, and graphic instructions regarding the particular parameters of the flights before entering the

cockpit. Pilots were requested to follow “text book” circuit procedures that guide general trajectory and airspeed during all phases of the flight. One circuit required approximately six minutes to complete. Figure 3.4 illustrates the type of information provided at the start of each mission.

After the practice session, participants flew six left hand circuits: three at a low-workload airfield and three at a high-workload airfield (always in this order, See Figure 3.3). Detailed instructions regarding the radio communication tasks, simulated flight instructions and environment, and the Situation Awareness Global Assessment Technique (SAGAT) “freeze” technique (Endsley, 1988) were provided in both written and verbal format, where applicable. The pilots flew the simulator uninterrupted and without further instruction from the experimenter, except during the two situation awareness SAGAT episodes. At the end of the third circuit in the high-workload condition the experimenter introduced the incursion aircraft to initiate the unexpected runway incursion management protocol. At the completion of the avoidance manoeuvre or the landing, the experimenter terminated the simulator session.

At the end of the session, the experimenter debriefed the participant with verbal and written information pertaining to the objectives of the study, contact of university officials and experimenter, and where the participant could find further information. See Appendix E for the Debriefing Form. The majority of the data collection was undertaken by the author (80 cases) and the remaining 28 cases were handled by a second trained experimenter. No significant effect of experimenter was found for measures collected in the simulator, $p \geq 0.05$ (non-parametric Mann-Whitney U-test), other than the RMSE for airspeed in the high-difficulty condition. Despite identical instructions, the pilots flew, on average, 1.3 knots faster for the primary experimenter.

3.4 Pilot Demographics and Expertise

A pilot demographic (age, medical certification level, vision or other reported medical issues) and flight experience questionnaire was completed at the start of each session (See Appendix F). Expertise data included pilot licensure and rating, hours spent flying various aircraft, comfort level with a Cessna 172, date of last flight, any actual flight experience with the low- and high-workload aerodromes used in the study, total years licensed, total flight hours, and recent pilot-in-command hours (previous 12 months) that were obtained from the pilots' flight logs and self-report. The total number of simulator hours, or time spent flying any type of flight simulation system, was also collected. Hyland et al. (1994) found that simulator performance was significant correlated with hours spent in simulated flight (with the same aircraft type), where actual flight time in the same aircraft type as the simulator was not associated with simulator performance. Thus total flight simulation hours were considered an essential aspect of pilot expertise. Table 3.1 lists the demographic measures, total sample means, standard deviations, and range of pilot data. Half the pilots also rated their income and education levels on a five point Likert scale. Initial analyses showed no meaningful relation between income and education and flight performance and thus are not reported in this work.

In summary, there were 18 student pilots (16.7%), 42 recreational and private pilots without additional ratings (38.9%), 28 private pilots with additional ratings (non-instrument rated) (25.9%), and 20 instrument rated, commercial, flight examiners/instructors, airline transport, or military pilots (18.5%). Pilot age ranged from 19 to 81 years with 72 pilots aged 40 years or older (66.7%). Determining if the present sample is representative of the population of pilots in Canada is difficult to establish, as categories of pilot licensure contain pilots who hold more than one

license. For example, an airline pilot may also hold a private pilot's license. Importantly this volunteer sample of pilots contained a broad range of ages and levels of expertise. Analyses regarding gender and flight performance were not undertaken due to the low number of female pilots in this sample, $n=6$.

Table 3.1
Pilot Demographics and Expertise

Measure	Description	Mean (SD)	Range
Date of Birth	Age of pilot at testing	45.5(15.1)	19 - 81
Pilot Licensure and Rating	Level of pilot certification	2.5(1.0)	1 - 4
Pilot-in-Command Hours	Hours flown as PIC (past 12 months)	31.1(46.7)	0 - 308
Total Hours Flown*	Total logged hours as licensed pilot	634.3(1243.7)	5 - 8000
Total Years Licensed*	Total years since first licensed	13.4(14.5)	1 - 63
Total Simulator Hours*	Total hours spent on any flight simulator system	81.8(181.0)	0 - 1000

Note. Pilot licensure and rating categories are as follows: Student (1), VFR (2), Private VFR+ (3), and Private IFR/ATP/MP/Commercial (4). *= significant differences as a function of age (groups split at age 50 years and older). Older pilots had, on average, more total hours flown, more years licensed, and fewer simulator hours, $p<.01$ (Significance calculated with Mann-Whitney U Tests for 2 independent samples). Licensure and rating by age group was calculated with a *chi-square* test, $p=.06$.

3.5 Cognitive Measures

3.4.1 The DCAT

The DCAT is a computer-based cognitive health screening tool developed for the identification of “at-risk” older drivers. The DCAT consists of six subtests that evaluate cognitive abilities related to driving errors specifically associated with drivers with dementia. The DCAT assesses motor speed and control, span of attentional field, spatial judgment and decision-making, speed of attentional shifting and executive function (DriveABLE Testing Ltd, 1997). The DCAT produces age-

normed z-scores that reflected response times and accuracy of the responses. The DCAT has shown 94% sensitivity (correctly identifying drivers with dementia who failed an on-road driving test) and 98% specificity (correctly identifying drivers without dementia who passed an on-road driving test) (DriveABLE Testing Ltd, 1997). Because this tool predicts risk for older adults for the complex task of driving, the DCAT might also be a useful predictor of performance for older aviators. The current study is the first to use the DCAT in an aviation setting. The following is a description of the three subtests utilized in the study. Table 3.2 provides a brief description, z-scores, and standard deviations for three DCAT subtests for the sample of 108 pilots.

The DCAT subtest 1: motor speed and control. Participants briefly viewed visual stimuli (blue boxes) on a 12-inch computer monitor. For this task, participants depressed a button on the desktop unit and maintained button depression until a blue box appears on the computer screen. Participants responded by releasing the button as soon as a box appeared on the screen and reached up to touch the screen in the location where the blue box was shown. This was repeated as the location of the stimuli moved randomly across the screen. This task is presumed to engage the cognitive processes of sustained and divided attention, processing of visual stimuli with accompanying motor responses and maintenance of multiple goals. Scores reflected speed and accuracy of responses.

The DCAT subtest 2: span of attentional field. This test is a dual-task measure incorporating a lexical decision and periphery stimulus detection task. Participants were required to discriminate a target word from non-target words that were briefly (<1 second) presented individually in the center of the computer screen while concurrently monitoring for the presentation of a target dot that was periodically

presented in the periphery. The first portion involved single-task lexical decision task and single target stimulus detection. The second, dual-task portion incorporated the lexical decision task and the periphery detection task with both the target and distractors. Scores reflected the speed and accuracy for single and dual task conditions.

The DCAT subtest 6: the identification of driving situations. DCAT subtest 6 is an audio/video-version of a driving Situational Judgment Test (SJT) (see Hunter, 2003 for use of SJT with general aviation pilots). Typically, situational judgement tasks describe a domain-specific scenario with four or five optional “best” responses pertaining to some best action based on the situation presented. The DCAT Subtest 6 required participants to watch ten short (approximately 5 to 20 seconds) narrated video clips of actual driving footage and then answer a question associated with an aspect of scene detail, spatial judgment, or driving safety related to the just-seen video clip. Participants were required to simultaneously integrate the auditory and visual stimuli for each trial of the ten video clips and maintain the stimuli in working memory to answer each trial question correctly. After the presentation of each stimulus and a brief inter-stimulus interval, a question was presented auditorally, with four potential “best” multiple-choice answers appearing on the next screen. Participants selected one of four text options on the touch screen to answer questions such as “how should you respond in this situation?” or “what is the most dangerous aspect of this scene?” General aviation situational judgment tests scores have shown strong associations with pilot involvement with critical aviation incidents, with better scores correlated with fewer critical incidents (Hunter, 2003).

Table 3.2

The DCAT Subtests Descriptions, Sample Means and Standard Deviations of Subtest Raw Scores for N=108

DCAT™ Subtest	Description	Mean <i>z</i> - Score (<i>SD</i>)
Motor Speed and Control	Reaction time to release button and touch location of stimulus on touch screen.	.428 (.351)
Span of Attentional Field	Location recall of visual stimuli in one of four main quadrants surrounding target word in the centre of the screen (with and without distractors). Recall of centre word as a target or foil.	-.376 (.800)
Identification of Driving Situations	Situational judgement test integrating visual and auditory information from a video clip.	-.284 (.529)

Note: The derived score for each subtest is an age-normed *z*-score based on reaction time and accuracy. The derived age-normed *z*-scores are generated via a proprietary algorithm. Subtests 3, 4, and 5 of the DCAT were collected but not utilized in this work.

3.4.2 CogScreen- Aeromedical Edition (AE)

CogScreen-Aeromedical (AE) is a computerized cognitive screening tool designed to assess cognitive processes deemed relevant to the complex tasks of an aviator (Kay, 1995). The CogScreen-AE was originally developed for detection of small changes in brain function in military and commercial pilots. CogScreen-AE purports to measure attention, immediate and short-term memory, working memory, visual-perceptual functions, sequencing functions, logical problem solving, calculation skills, reaction time, and dual-task processing (Kay, 1995). Computerized results for 11 major subtests are provided for each participant. See Table 3.3 for a brief description of the 11 major subtests. More than 95 raw and derived scores are produced in the CogScreen-AE output. Scores reflect accuracy, number of errors, total span, reaction time (correct), and standard deviation of individual median reaction time, dual-task decrement ratios, or “thruput”. The “thruput” scores are derived from correct responses

divided by time per response and divided by total task time. Five “Taylor Factor Scores”, which are computed via an undisclosed algorithm, are also automatically calculated. Only the non-derived scores of percent accuracy, total span, reaction time, and errors were used in the present analyses of pilot neurocognitive performance.

CogScreen-AE was standardized and validated with 584 commercial airline pilots, aged 25 to 67 years (Kay, 1995). The Cogscreen-AE is used as a baseline test of cognitive performance for most US military pilots. Commercial airlines in North America also use the CogScreen-AE as part of their pilot recruitment procedures (personal communication with Gary Kay, May 2012). CogScreen-AE measures have demonstrated significant correlations with pilot age and pilot simulator performance (Hyland, Kay, & Deimler, 1994). However, due to a small sample size, Hyland et al. (1994) were unable to determine whether age accounted for significant amounts of variance in flight-simulator performance once pilot cognition, as per CogScreen-AE performance, was partialled out of the correlation.

CogScreen-AE testing was conducted using a Windows XP computer with Elo - Touch systems 2216 AccuTouch USB Touch monitor (Elo Touch Solutions). The 11 subtests of the CogScreen-AE were administered. The subtests are labeled Backward Digit Span, Math, Visual Sequence Comparison, Symbol Digit Coding, Matching to Sample, Manikin, Divided Attention, Auditory Sequence Comparison, Pathfinder, Shifting Attention, and Dual Task. CogScreen-AE provides raw and derived scores based on mean (or median) reaction time and accuracy of responses, thus providing a minimum of three measures for each element of the 11 subtests. Only the raw scores from the CogScreen-AE were used in the present work. Completion time of the CogScreen-AE ranges from approximately 45 to 75 minutes.

The CogScreen-AE was administered to all the pilots in the second phase of the data collection ($n = 55$). One participant was unable to complete the full CogScreen due to difficulties completing the tasks. Data was unexplainably not recorded by the computer for the Backward Digit Span (accuracy) for one other pilot. Table 3.3 contains the CogScreen-AE subtest descriptions. Sample means and standard deviations are provided for selected subtest items.

Table 3.3

CogScreen-AE Subtests Descriptions, Sample Means and Standard Deviations of Selected Subtest Raw Scores

CogScreen-AE Subtest	Description	Label (metric)	Mean (SD)
Backward Digit Span	Immediate reverse reproduction of digit series (visual stimuli): series range from 3 to 6 digits.	BDSAC (Percent Accuracy)	76.7(18.6)
Math	Arithmetic word problems presented visually: multiple choice response.	MATHAC (Percent Accuracy)	74.8(24.6)
Visual Sequence Comparison	Visual Comparison of a sequence of letters and numbers: both series presented simultaneously.	VSCRT (Reaction Time)	2.5(0.8)
Symbol Digit Coding	Immediate (~90 sec.) and delayed (20 to 30 min.).	SDCRT (Reaction Time)	1.6(0.4)
	Delayed recall of pairs of symbols and digits.	SDCDRAC (Percent Accuracy)	75.9(29.3)
Matching to Sample	Pattern recognition for checkerboard-type stimuli: two stimuli choices are available for selection after initial presentation of target.	MTSRT (Reaction Time)	1.3(0.3)
Divided Attention	Visual tracking with response to stimulus criterion with and without simultaneous (dual) presentation of the visual sequence comparison subtest.	DATDRT (Dual, Reaction Time)	0.5(0.2)
Pathfinder	Selection of digits, letters or alternating digits and letters to complete an ascending series presented visually.	PFCRT (Combined letters and digits, Reaction Time)	1.3(0.4)
		PFCAC (Combined letters and digits, Percent Accuracy)	95.4(6.6)
Dual Task (DTT)	Visual-motor tracking and 2-back working memory task each done alone or together in the dual task condition.	DTPDAC (Dual, 2-back recall, Percent Accuracy)	76.6(17.2)

Note. BDS, $N=53$. Reaction time displayed in seconds. Descriptions adapted from Kay, G. 1995. *CogScreen-Aeromedical Edition. Professional Manual*. CogScreen, LLC.

It was determined a priori to not utilize the CogScreen-AE “Thruput” measures. The thruput measures are derived scores calculated as described by Kay (1995, p. 133). The scores do not reflect a pure measure of time, or accuracy. Thruput scores are calculated as:

$$\text{Thruput} = ([\text{Accuracy}/100] \times 60 \text{ seconds}) / \text{Median Response Time Correct Trials}$$

Interpretation of thruput scores is difficult because a score can be identical for a pilot who obtains perfect accuracy but low median response time, and another pilot who obtains half the accuracy but whose median response time was twice as fast. Therefore, holding accuracy constant, longer median response times will always reduce the thruput score. Furthermore, on categorical response tasks (e.g., “same” or “different responses”) always selecting one response (obtaining a chance 50% accuracy), but responding very quickly could result in the same thruput score as someone who responds only half as fast, but responds thoughtfully and achieves a score of 100%.

3.6 Flight Simulation Measures

The present work specifically investigated aviator tasks and outcomes associated with private pilot flights in small aircraft. In small single-piloted aircraft the pilot is solely responsible for *aviation, navigation, and communication*. Table 3.4 provides the list of measures collected via the flight simulation protocol, sample means and standard deviations for performance data collected in the low- and high-difficulty conditions. The incursion management measure was only collected in the high-difficulty condition as this measure was based on a one-time unexpected event.

Table 3.4
Total Sample Means and Standard Errors for Flight Simulation Measures

Measure	Description	Total Sample Mean (SD) per Difficulty Condition	
		Low	High
Flight Path Deviations	Average deviation from ideal trajectory (RMSE)		
Altitude	Deviation from the ideal altitude in the downwind leg (feet).	99.4(85.0)	144.7(118.4)
Airspeed	Deviation from the ideal airspeed in the downwind leg (knots).	8.1(3.0)	9.3(2.9)
Heading	Deviation from the ideal heading in the downwind leg (degrees).	10.3(2.5)	9.4(2.5)
Situation Awareness	Percent accuracy relevant details of the simulated "world".		
Level 1	Other aircraft details and ownship.	53.2(22.4)	36.9(20.7)
Level 2	Relative position of ownship and circuit traffic.	84.5(24.1)	65.5(22.3)
Level 3	Projection of relative position of ownship and circuit traffic in near future.	79.8(22.3)	64.9(22.3)
Retrospective Memory	Recall of ownship during specific circuit event (episodic memory control, maximum 3 points).	1.60(.63)	1.19(.81)
Prospective Memory	Radio communication completion rates		
Radio Calls for Six Circuits	Radio calls completed (maximum = 18).	16.2(2.0)	15.0(2.6)
Outcome Measures	Indicators of pilot risk during flight		
Critical Incidents	Raw score for number of crashes, lost events, contact with terrain, or non-runway landings (range =0-5).	0.12(0.4)	0.44(0.8)
Incursion Management	Timing, quality, and outcome of incursion event (maximum = 10)		5.4(3.8)

Note. RMSE= root mean square error. RMSE calculated as average square-root of the squared difference between ideal and actual flight parameter. Other than Incursion Management, each measure was collected in the low- and high-difficulty condition. Larger flight path deviations and reduced task accuracy were generally found in the high-difficulty condition.

In brief, minimizing flight path deviations by maintaining proper altitude, airspeed, and directional heading by means of manual operation of aircraft controls are basic aviator skills. This maintenance of a correct trajectory is a key aspect of pilot training and evaluation. Pilot situation awareness supports navigation of the relevant space and knowledge of locations of other aircraft, whether on the ground or airborne. With many private pilots flying in uncontrolled airspace (without air traffic control information) and cockpits not outfitted with GPS technology, the ability to create and maintain an accurate mental model of the airspace and relevant traffic remains an essential task in general aviation. Communication with other aircraft and ground personnel in a clear and timely fashion are also aviator activities of critical importance for private pilots. Prospective memory, a cognitive function supporting remembering tasks for future implementation, also supports communication and other tasks carried out before and during flight. The consequences of critical incidents and unexpected events occurring during flight are particularly relevant outcomes for flights in small aircraft piloted by private pilots. Because private pilots in small aircraft are generally the single crew member, avoiding and handling incidents falls to them alone. Pilots in single-pilot operations must also be prepared to handle an emergency and prepare other aircraft occupants for the event. Sections 3.8 to 3.11 provide details regarding measures collected during the flight simulation.

3.7 Mission Difficulty Conditions

Two discrete mission difficulty conditions were created: low and high. The development of the mission difficulty conditions was done in consultation with pilot examiners and four highly experienced pilots. Subjective feedback and initial *t*-tests of flight performance during pilot testing were used to confirm differences in mission difficulty. In the low-difficulty condition pilots flew at an aerodrome with simple flat

terrain and no other aircraft during circuit one; one other aircraft during circuit two; and two other aircraft during circuit three. The aerodrome in the low-difficulty condition was the Microsoft FSX TM rendering of the aerodrome located at Yorkton Municipal Airport, Saskatchewan, Canada (elevation: 1,635 ft. / 498 m; coordinates 51°15' 53' N and 102°27' 41' W). In the high-difficulty condition, pilots flew in unfamiliar mountainous terrain with two other aircraft during circuit one; three other aircraft during circuit two; and four other aircraft during circuit three. The aerodrome in the high-difficulty condition was the Microsoft FSX TM rendering of the aerodrome located at Kaslo Airport, British Columbia, Canada (elevation: 2,354 ft. / 717 m; coordinates 49°54' 13' N and 116°56' 07' W). In both the low- and high-difficulty conditions pilots were required to maintain situation awareness of ownship and all other relevant aircraft. Pilots were required to monitor the radio calls from other circuit traffic. In each condition the simulated weather was clear and no critical events with respect to instrumentation or aircraft controls were introduced.

3.8 Mental Task Load

The manipulation of mission difficulty was hypothesized to create two cognitive workload conditions. Mental task load was measured via an objective set of measures using an auditory peripheral detection (PDT) task that ran concurrently with the simulator protocol. Subjective measures of workload were collected via self-rated measures of task workload.

3.8.1 Objective Measure of Workload

A novel auditory peripheral detection task (PDT) was used as an objective measure of workload. The PDT was considered a proxy measure of cognitive workload because it measures key cognitive resources that remain available to the pilot after resources

necessary for the primary task (the flight tasks) are employed. Previous work employing PDTs to gauge workload during transportation tasks have used visual stimuli because the PDT modality should be equivalent to the chief modality used in the primary task (Patten, Kircher, Östlund, & Nilsson, 2004). Piloting tasks include aviating, navigating, and communicating. All of these tasks have an auditory component; therefore the stimulus created for this work was an auditory tone. The tone was generated at 1500 Hertz and delivered randomly with 10- to 20-second inter-stimulus intervals throughout the simulated flight. The tone presentation was binaural and administered via a headset. Pilots wore a small thumb switch that they were instructed to depress when the auditory stimulus was detected (see Figure 3.2). Reaction times (ms) and accuracy rates (percent of correct hits) were recorded. The thumb switch was worn on the pilot's hand of choice and remained there for both mission difficulty conditions. Depression of the switch did not interfere with control of the yoke.

Figure 3.2. Thumb switch for auditory peripheral detection task. The white oval shape indicates the position of the thumb switch and illustrates how depression of the switch did not interfere with control of the yoke.

Pilots practiced responding to the auditory PDT during the warm-up session of the flight simulator task to ensure their understanding of the task. Within-subject differences in mean reaction time and accuracy rates between the low- and high-difficulty conditions were analyzed using *t*-tests to confirm the existence of two dichotomous workload states. The data acquisition computer recorded correct hits and false alarms. The raw data was preprocessed using Python v2.7.1 (van Rossum, 1995) code written to facilitate extraction of the derived measures for the low- and high-difficulty conditions. Measures for the PDT included average response times for correct hits, the individual standard deviation of response times for correct hits, and the hit rate. There were extremely few false alarms, thus a D-prime measure would

not have been informative and was not derived for the PDT. The individual standard deviation of response times was a measure of intra-individual variability.

3.8.2 Subjective Measure of Mental Task Load

A modified version of the NASA Task Load Index (TLX) (Hart & Staveland, 1988) was used to index pilots' perception of workload in both mission difficulty conditions (See Appendix G). The NASA TLX has been used for more than 20 years to evaluate workload in aviation and other settings requiring complex cognitive behaviours. The TLX scale provides seven categories reflecting task workload and uses a 10-point rating scale: physical demand, mental demand, temporal demand, performance, effort, frustration, and overall demand. Descriptions of the seven features of workload were available on the scale for easy reference for the pilots. Preliminary descriptive results showed that the item reflecting "performance" did not correlate as expected with other measures in the scale. This was the only item where a low-workload indication was physically reversed on the scale, which may have led to some confusion (since participants tended to complete the forms fairly quickly). The performance item was removed from further investigations due to its unreliable status as an indicator of workload. Item scores were unweighted.

Confirmation of the low- and high-difficulty conditions. Mental task load was manipulated such that pilots flew in two distinct workload conditions: labeled low- and high-difficulty conditions. Mission difficulty was experimentally manipulated via terrain complexity (prairie versus mountainous) and circuit traffic volume (maximum two versus four other aircraft). Increasing the amount of traffic requiring monitoring by the pilots, in combination with a flight environment containing a greater number of perils (mountains, valleys, and water in close proximity) resulted in pilots consistently rating the second, high-difficulty condition, as a significantly more demanding task

(see Table 3.5). The results of the subjective ratings were compared with the objective measures of workload, which also demonstrated significantly longer response times and lower hit rates for the PDT in the high-difficulty condition. These results support the inclusion of a workload factor as a predictor of performance and allow for conclusions based on mission difficulty in subsequent investigations of pilot performance.

Table 3.5

Pilot objective and subjective ratings of workload confirm two mission difficulty conditions

	Workload Condition		Standard Error	<i>t</i> (104)
	Lo w	High		
Objective Workload Measures				
Mean Response Time (ms)	0.854	0.932	0.014	-5.841
Mean Individual SD of Response Time (ms)	0.258	0.282	0.010	-2.326 (<i>p</i> =.022)
Hit Rate (% Accuracy)	83.38 %	69.27%	2.033	6.937
Mean Subjective Task Load Ratings (range=1-10)				
Mental Demand	6.70	8.18	0.244	-6.112
Physical Demand	4.13	5.43	0.205	-6.348
Temporal Demand	4.86	7.12	0.244	-9.272
Frustration	4.62	6.59	0.296	-6.649
Effort	6.20	7.78	0.233	-6.78
Overall Workload	5.58	7.49	0.276	-6.920

Note. The *p*-value associated with each t-test is *p*<.001 unless indicated.

Figure 3.3. Low- and high-difficulty aerodromes. View indicates direction of takeoff and subsequent “touch-and-go” landings in the Kaslo, BC aerodrome (right half of Figure). Bottom left of Yorkton, SK runway indicates takeoff location (Left half of Figure) Left-hand circuit required rotate speed of 80 knots, ascent to 500 feet (above ground [AG]), followed by crosswind turn, downwind turn at 1000 feet (AG), maintenance of 100 knots airspeed for majority of downwind leg, followed by base turn, decent to 500 feet (AG) for final turn, and touch-and-go at approximately 60 knots airspeed. Maps from Google Maps TM (terrain-view).

3.9 Flight path maintenance

The circuit parameters for the simulation task were designed in consultation with pilot examiners and highly-experienced pilots. Preliminary testing revealed that the task was realistic and met standard circuit operating procedures. Aircraft airspeed, altitude, and heading were logged at one-second intervals on the data acquisition computer. Flight path error was measured by calculating deviations from the “perfect” circuit altitude, airspeed, and heading for each difficulty condition for the downwind leg of the circuit. The parameters of the downwind leg required an altitude of 1000 feet above ground level; airspeed was 100 knots; and heading was 180 degrees difference from the runway heading. The perfect circuit parameters were carefully described to participants in detail and practiced for at least four full circuits in the warm-up phase of the protocol. Pilots were made aware via verbal, graphic (a colour printed Google

EarthTM overhead image of each aerodrome, a generic aerodrome parameters graphic as shown in Figure 3.4), and written instructions as to the required altitude, airspeed, and heading for the entire circuit, in particular for the downwind leg (long length of circuit parallel to the runway) of the circuit. At each data point (approximately one data point per second spent in the downwind leg of the circuit) a value was calculated, which represented the root mean square error (RMSE) between the actual airspeed, altitude, and heading of the pilot and the “perfect” airspeed, altitude, and heading. There were separate RMSE scores for the low- and high-workload conditions. See Figure 3.4 for details of the circuit flight parameters.

Figure 3.4. Circuit parameters in low- and high-difficulty conditions. RMSE calculated for points between RC3 and RC5. “Probe” site indicates location where scenario was “frozen” once per difficulty condition for situation awareness. “Incursion” site indicates where the offending aircraft was located (not indicated on participant material). The incursion scenario only took place once: at the end of the high-difficulty condition.

3.10 Situation Awareness

3.10.1 Objective Measure of Situation Awareness

Situation awareness measures utilized SAGAT methods and required participants to respond to queries in the midst of flight (Endsley, 1995, 2000). The SAGAT procedure was always as follows: at approximately halfway down the base leg of the second

circuit in each mission difficulty condition the flight scenario was “frozen”. The experimenter quickly covered all the cockpit gauges. This was always done very quickly and before pilots had an opportunity to recover from the surprise of the sudden freezing of the scenario mid-flight. A series of questions were given in a quick manner to allow speedy resumption of the flight scenario. Participants were given a clipboard and a response sheet with a circuit-shaped rectangle pre-drawn on the paper (this was also to allow for quick responses to the probes.) See Appendix H for the probe answer sheet. Probe questions were verbally asked, and were written at the bottom of the response sheet by the pilot. No explicit time limit was given; however, all participants were asked to provide an answer as quickly as possible to maximize recall.

Questions were developed to reflect three levels of situation awareness (see Table 3.6). Pilots indicated (from memory) their heading, airspeed, and altitude at the time of the probe, and the call sign and aircraft type of the other aircraft in the circuit (Level 1); their ownship in the circuit at the time of the probe and where all additional aircraft were situated (Level 2); and, where they predicted their aircraft would be in two minutes, and where the other circuit aircraft would be in three minutes (Level 3). Monitoring all traffic communication in the circuit was an essential element of situation awareness. In the low- and high-workload conditions the probes occurred at the beginning of the “base” leg of the circuit when there were one or three other aircraft in the circuit, respectively. Total raw scores for each level of situation awareness were converted to percent correct for comparison purposes. A perfect score for a probe was 2, a partial response was 1, and no correct information was 0. See Figure 3.3 for the location of the SAGAT probe site in the base leg of the circuit. Queries were designed to reflect basic information relevant to general aviation at or

near an uncontrolled aerodrome. A control question that tested episodic memory was also included so that retrospective memory ability could be later partialled out from SA performance.

Queries were vetted with four experienced GA pilots, one of whom was a pilot examiner. Once a participant had provided their responses the cockpit gauge covering was removed and the flight simulation was resumed. The entire SAGAT procedure had a duration of approximately 2 to 4 minutes. Endsley (1995; 2000) found that delays of six minutes did not result in the decay of information relevant to SA. Each probe was administered twice (once for each mission difficulty condition). The same SA queries were used for all pilots (i.e., SA probes did not vary between pilots as is often done by other researchers [Endsley, 1995; 2000]). Table 3.4 provides mean scores for 108 pilots for the three levels of SA.

For clarity it is noted that SA level 1 was later renamed iRES, “i”= information and “RES” = resolution. iRES is different from SA level 1 probes used by other researchers in that scores do not reflect binary, correct versus incorrect, responses. (Endsley, 1995; 2000). Instead, iRES, as shown in Table 3.6, is a measure reflecting the resolution of the information pilots could extract from the environment while flying the missions. The simulated “world” was replete with information such as ownship (heading, airspeed, and altitude), and the make and full call-sign of all other relevant traffic. iRES scores reflected the level of detail pilots could relay regarding this virtual world. To understand iRES consider the concept of resolution as it pertains to a computer screen. A computer monitor with a million pixels can convey more bits of information than a monitor displaying 500,000 pixels. The amount of information that can be displayed is considered the resolution of the screen. As a group, pilots also “displayed” varying resolution regarding the detail of information conveyed for SA

level 1. For example, iRES points were obtained by pilots for relaying the call-signs of relevant traffic at the airfield. As shown in Table 3.6, the point system was graded such that more points were awarded in accord with greater levels of correct specificity regarding the call-sign (0 points for no correct details, 1 point for partial correct, and 2 points for fully correct details). Thus, the iRES point system reflected a pilot's resolution of the simulated world.

In relation to the SAGAT procedures, accuracy for probes has not been shown to be impacted by the time lapse between freezing of the scenario and the time of questioning, the time taken to respond to probes, nor by the number of freezes inserted into the flight simulation protocol (Endsley, 1995). Despite this it was still possible that, within a sample with a large range of pilot age and expertise, pilots with poor long-term memory could be disadvantaged by the SAGAT-type assessment of SA. Therefore, an episodic memory query (pertaining to ownership when a particular auditory event had occurred) was included as a control for the potential retrospective memory confound. The retrospective memory question was used as a control factor permitting long-term memory effects to be *partialled out* of SA performance. The SAGAT queries for each level of SA and the retrospective memory probe are listed in Table 3.6. The average percent accuracy (aggregated for each level) for the three levels of SA and the retrospective memory control probe (range, 0-2) for each difficulty condition are shown in Table 3.4.

Table 3.6
SAGAT Probe Scoring System

	Probe	Points Awarded
Level 1 Probes		
	"Please write your response on the response sheet..."	
1.	Your current altitude?	2 - within +/-100ft 1 - within +/-200ft 0 - <200ft
2.	Your current airspeed?	2 - within +/-5kts 1 - within +/-10kts 0 - <10kts
3.	Your current heading?	2 - within +/-10deg 1 - within +/-20deg 0 - <20deg
4.	The call sign of other relevant traffic?	2 - full call sign 1 - partial call sign 0 - incorrect or none
5.	The type of aircraft of other relevant traffic?	2 - full type 1 - partial type 0 - incorrect or none
Level 2 Probes		
	"Please indicate with an X on the response sheet..."	Same scoring for probes 6-9
6.	Your current position?	1 - correct leg and 1 - correct within one-third for downwind/runway leg or 1 - correct within one-half for base or crosswind leg 0 - wrong leg of circuit
Level 3 Probes		
	"Please indicate with an X on the response sheet..."	
7.	The current position of other relevant traffic?	Same as Level 2 but 1 point maximum awarded if position incorrect as per actual future position, but correct based on response recorded from Level 2
8.	Your position in two minutes will be?	
9.	Other relevant aircraft position in three minutes will be?	
Retrospective Memory		
10.	Indicate your location in the circuit at the time when the most recent aircraft to join the circuit first communicated?	Same scoring as per Level 2 probes

3.10.2 Subjective Measure of Situation Awareness

Subjective situation awareness was measured using a 7-point scale designed for the program of research at the ACE Laboratory at Carleton University (Hou, Kobierski, & Brown, 2007). The scale reflected 11 aspects of situation awareness typically encountered during flight and referenced awareness of the environment, tasks, time, priorities, relative position of ownship and other aircraft, future events, and overall level of situation awareness (See Appendix I). The scale has shown sensitivity to levels of information provided to operators in complex simulations of uninhabited aerial vehicles (Hou et al., 2007). The scale was administered twice: at the end of each mission difficulty condition. The median response for each item on the subjective situation awareness scale for each difficulty condition is listed in Table 3.7.

Table 3.7
Median Self-Rated SA Performance for Low- and High-Difficulty Conditions

Subjective SA Scale Items	Mission Difficulty	
	Low	High
The navigation environment	5	4
Tasks to be performed	6	5
Time to complete specific tasks	6	5
Time to complete perfect circuit	5	5
Task priorities	6	5
Aircraft's position relative to other aircraft in the circuit	6	4
Impact of other aircraft on ability to fly the "perfect circuit"	6	5
Ability to predict future circuit events	5	4
Awareness of other aircraft activities	6	4
Awareness of other aircraft location(s)	6	4
Overall level of situation awareness	5	4

Note. Item series sample size range from $N= 89$ to $N=91$. The range for each rating was 1-7. Higher ratings indicated better self-rated SA performance.

3.11 Radio Call Communication (Prospective Memory Measure)

The radio call completion rate was the index of prospective memory in the present study. Participants were provided detailed instructions regarding the number and the timing of the radio calls they were to complete during each circuit. As per standard procedures, the pilots' radio calls were to convey information pertaining to their identification and current location. Each circuit required the completion of six radio calls. Figure 3.3 illustrates the expected location for each radio call. One radio call was required during each of the following six circuit events: (1) RC=1, initial rolling on runway (or subsequent runway touch-and-go or runway overshoot events), (2) RC=2, airborne, (3) RC=3, turning downwind, (4) RC=4, mid-downwind, (5) RC=5, turning base, and (6) RC=6, turning final.

Prospective memory performance is associated with the salience of the memory cue, which triggers the appropriate behaviour, with high cue salience related to better performance. The radio calls with low-salience cues were calls two (airborne) and four (mid-downwind). These calls had very limited visual cues, such as changes to the horizon, and no physical cues, such as turning the yoke. The two high-salience cue calls were calls three (turning downwind) and five (turning base). These two calls (turning downwind and turning base) had inherent high cue salience because the call to be remembered matches the activity during the event, thus providing strong congruence between the cue and the memory task. High-salience memory cues were the visual changes to the horizon, and physical cues where the pilots' physical actions were congruent with the information in the radio call. Turning base was considered the radio call with the strongest cue salience because of the physical act of turning and descending, and presence of visual changes to the horizon and approaching runway to

the left. The remaining two radio calls, one (runway) and six (final approach) were considered to have moderate-salience cues.

Participants were not penalized for making extra calls, but only received a point for each radio call completed during the prescribed circuit event. Six calls per circuit for each of three circuits in each of the two workload conditions allowed for a maximum of 36 radio communication calls to be recorded for each participant. Participants heard, via their headset, the virtual “pilots” in the circuit traffic complete similar radio calls. The virtual pilots also completed the radio calls at the same six locations in the circuit as was required of participants. The average prospective memory score (aggregated into a single score for each difficulty condition) is found in Table 3.4.

3.12 Outcome Measures

3.12.1 Critical Incidents

The experimenter recorded critical incidents during simulated flight. Critical incidents included controlled flight into terrain or water, landing at a location other than the specified runway, loss of separation with other aircraft or terrain, and occurrences of becoming lost during flight. The experimenter recorded critical events as they occurred and were coded with respect to occurrence in the low- or high-difficulty condition. Thirty-four of the 108 pilots incurred at least one critical incident. Table 3.8 provides the description of the number of critical events that were associated with each expertise group. A Pearson *chi-square* analysis demonstrated a significant effect of pilot licensure and rating group on critical incidents,

$$\chi^2 (15, N=108) = 25.92, p = .039.$$

Table 3.8
Critical Incident Count by Pilot Licensure and Rating group

Pilot Licensure and Rating	Critical Incident		Group Size
	No	Yes	
Student	7	11	18
VFR/PPL/Recreation	28	14	42
PPL with Additional Ratings	19	9	28
IFR/Military/ATP/Commercial	20	0	20
Total	74	34	108

3.12.2 Surprise Runway Incursion Management

The runway incursion scenario was the last “task” in the flight simulation protocol. Importantly, pilots were completely unaware that a runway incursion would occur. In the final high-difficulty condition pilots were requested to fly two touch-and-go circuits and at the end of the third circuit participants were to come to a full stop on the runway. While the pilot was on final approach a “rogue” simulated aircraft was introduced at the distal end of the runway. The runway incursion aircraft was traveling in the wrong direction (i.e., towards the oncoming pilot participant). To manage this runway incursion successfully, pilots must have accomplished three main tasks in quick succession. The initial requirement was to correctly perceive and interpret the environmental stimuli (i.e., there was a vehicle on the runway that the pilot was landing on and it was travelling towards the participant’s aircraft, causing imminent unsafe separation of aircraft). The second task was to use the interpreted information to make a decision regarding the required procedure to avoid the quickly approaching incursion (abort the landing procedure and initiate a “go around”). Finally, the correct decision regarding evasive action in the air must have been carried out in a timely and

efficient manner. The incursion management score was based on the quality, timing and result of the pilot response to the runway incursion. See Appendix J for the scoring details. Inter-rater consistency was checked informally to confirm consistency between the qualitative aspects of this measure (e.g., the timing of manoeuvres score). A rough method for testing for possible experimenter effects was to examine the pattern of scores for each experimenter. A one-way ANOVA revealed no significant difference in incursion management scores between raters, $F > 1$. For some participants the incursion scenario was their very last task of the study protocol, meaning that they had completed the two neurocognitive batteries before engaging in the flight simulation. To ensure that possible testing-induced fatigue did not influence incursion management a one-way ANOVA was conducted between participants who flew first, and those who completed the neurocognitive batteries first. No evidence of a test order effect was found, $F > 1$. Table 3.9 provides a description of the mean incursion management scores per pilot licensure and rating group.

Table 3.9
Mean Incursion Management Scores per Pilot Licensure and Rating Group

Pilot Licensure and Rating	Mean (SD)	N
Student	2.44(3.15)	18
VFR/PPL/Recreation	4.71(3.77)	42
PPL with Additional Ratings	7.32(3.22)	28
IFR/Military/ATP/Commercial	6.55(3.28)	20
Total	5.35(3.80)	108

3.13 Data Preparation

All data collected via pen and paper was handled according to the scoring procedures specified for that measure and entered into the electronic database (SPSS) by hand for analysis. All other data collected via electronic means was stored on a secure hard drive and transferred via electronic means to the electronic database. Some data

collected from the flight simulator was pre-processed electronically using Python v2.7.1 scripts to reduce file size and streamline analysis in SPSS. Pre-processing was undertaken to calculate RMSE scores for the downwind leg of the circuit, and to calculate PDT accuracy, reaction time, and reaction time standard deviation for correct trials only.

Missing or partially-collected data was monitored and handled as per recommendations depending on the statistical test utilized. On a few occasions data collected by the data acquisition computer during simulated flight did not record values for the PDT or RMSE measures. This resulted in very little lost data (<3% for any one measure). For these cases the missing value was replaced with the series mean. The subjective situation awareness and cognitive workload self-rating and rating scales were not administered to the first 15 subjects and were therefore “missing- not at random”. However, the analyses that included the subjective measures were undertaken using the full sample of 108 pilots (where missing values were replaced with the series mean). Current recommendations from the literature reject the summary removal of otherwise useful subjects from statistical analyses due to small amounts of missing data (Dong & Peng, 2013; Wilkinson, 1999). In contrast, various methods for replacing missing values are proposed, such as replacement with the series mean. This is discussed in greater detail in Chapter 6. As mentioned, the CogScreen-AE was administered during the second wave of data collection ($n=55$). When analyses included data from the CogScreen-AE, the 53 subjects from the first wave of data collection were not included in the analysis. The skew and kurtosis features of raw data distributions were calculated as necessary for individual statistical tests. For example, in the structural equation modeling environment all data were

ranked and standardized as a method for handling potential skew, kurtosis, and outlier biases in statistical results.

3.14 Power to Detect Significant Effects

A sample of 108 pilots met study criteria and flew three circuits in each of two mission difficulty conditions. The power of the test for most analyses in this dissertation was estimated using G*Power 3.1.7 (Faul, Erdfelder, Lang, & Buchner, 2007). Parameters for a statistical analysis, as might be used in the current study, were utilized in a sample power analysis (as shown in Figure 3.5). As illustrated in Figure 3.5, reasonable power estimated at 0.80 was produced with a sample size of 108, for a repeated measures ANOVA. Power of 0.80 is considered the minimum acceptable power for a statistical test (Cohen, 1992). The preceding estimation was based on a combination of the following parameters: 1) an F -test with three independent groups 2) two repeated measures 3), an effect size of 0.19, and 4) a type-one error rate of $\alpha=0.05$.

Figure 3.5. Estimates of statistical power for a repeated measures ANOVA. Plot created using G*Power from Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.

Statistical analysis based on fewer groups or simple one-way ANOVAs yielded power greater than 0.80. An example power analysis was also calculated for the smaller sample size of $n=55$ pilots from the second phase of data collection. The power analysis indicated that a bivariate correlation of $r(53)=0.33$ (one-tailed) could be detected with the following parameters: power = 0.80, $n=55$, and type one error rate of $\alpha=0.05$. These a priori analyses indicated that there was sufficient power in the planned statistical analyses to detect small effects in the selected samples. For the subsample of 55 pilots it was possible that very small effects might not have reached significance. In light of this, for the SEM analyses, effect sizes based on Cohen's (1988) *f-squared* estimates of effect size were also examined to ensure that all relevant predictors were identified for criterion variables. *F-squared* estimates are calculated to represent global and local relative effect sizes for multivariate regression analyses, and are the preferred parameter for determining the value of predictors in complex models (as compared to *p*-values for example) (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012).

Chapter 4

Results of the Experimental Program of Research

4.1 Overview

This chapter reviews the results of key analyses from the experimental component of this work. Once sufficient power to detect small to moderate effects was obtained, analyses using subsamples from the eventual total sample of 108 pilots were undertaken. Thus, this review describes a number of smaller published studies that explored the relationships between variables that Hardy and Parasuraman (1997) proposed as essential to a comprehensive model of pilot performance. These analyses were presented as conference papers or other published works from 2011 to 2015. The results highlight the effects of pilot age, cognition, expertise, and workload on flight performance (unpublished analyses are also described to clarify these relationships as required).

Analyses investigated pilot characteristics, cognitive workload, and domain-independent cognition predictors of the domain-dependent skills of flight path maintenance; situation awareness (Van Benthem, Herdman, Brown & Barr, 2011a; Van Benthem, Tolton, & Herdman, 2013); and prospective memory for pilot radio communication tasks (Van Benthem, Herdman, & LeFevre, 2011c; Van Benthem, Herdman, LeFevre, & Tolton, 2012a). The utility of subjective and objective measures of cognitive workload in predicting domain-dependent pilot performance was examined (Van Benthem & Herdman, 2014a). Predictors of the outcome measures of critical incident rates (Van Benthem, Herdman, Tolton, & Barr, 2012b) and runway incursion management (Van Benthem & Herdman, 2014b) were also investigated.

Figure 4.1 illustrates a model of pilot performance based on Hardy and Parasuraman (1997) and Herdman and LeFevre (2004). The individual studies discussed here represent various hypothesized relationships depicted in the Hardy and Parasuraman model. The findings from the experimental component of this work are discussed in the context of a relevant body of literature. The results of these analyses were used to guide specific hypotheses regarding the DMM structure (outlined in Chapter 5), and to inform the data modeling described in Chapters 6 and 7.

Figure 4.1 An adapted model of pilot performance with the DMM as the key domain-dependent construct (based on the models by Hardy and Parasuraman [1997] and Herdman & LeFevre, [2004]). Other than the suggested causal link *from* domain-dependent skills *to* domain independent skills (dashed arrow) all other relationships, as indicated by arrows in this figure were investigated in the experimental portion of this work.

4.2 Flight Path Maintenance

Error in flight path trajectory is a basic measure of aviator competency. Deviations from the ideal altitude, airspeed, and heading are generally the criterion or dependent variable in studies of general aviation pilots (Causse et al., 2011; Taylor et al., 2000, 2005, 2007). In light of previous work it was hypothesized that working memory, visual tracking, pilot expertise and age would significantly predict deviations from the perfect trajectory.

This analysis of flight path deviations is modeled after the report by Tolton, Van Benthem and Herdman (2012). In this study, rather than examine the RMSE for altitude, heading and airspeed separately a comprehensive measure of flight performance was derived using factor analysis. Flight performance component scores were calculated as exact linear combinations of altitude and airspeed RMSE in the low-difficulty condition, and altitude, airspeed, and heading RMSE in the high-difficulty condition. Thus, one flight path maintenance score was computed for each pilot in the low- and high-difficulty conditions (scores were reversed so that better scores were associated with higher values). The RMSE scores were regressed on pilot cognition, age, and expertise factors to build models of basic aviator performance (see Table 3.4 for the average RMSE for altitude, heading, and airspeed in each mission difficulty condition). Table 4.1 lists the indicator factor loadings for the low- and high-difficulty condition flight performance factor scores for the subset of 55 pilots.

Table 4.1
Flight Performance Factor Score Indicator Loadings

Indicator	Flight Performance (<i>p</i>)	
	Low-Difficulty Condition	High-Difficulty Condition
Altitude	0.878 (<.001)	0.808 (<.001)
Speed	0.808 (<.001)	0.639 (.003)
Heading	Negative small loading	0.729 (<.001)

Note. Factors computed as exact linear composites. Significance tested using bootstrapped standard errors (SE). Due to small negative loading Heading was not included in the low-difficulty factor.

Figure 4.2 highlights the observation that, even in the low-difficulty downwind segment of the circuit, large deviations from the ideal altitude were observed. Pilots deviated by an average of 87.5 feet or 144.5 feet, in the low- and high-difficulty condition. Canadian flight certification testing regulations note deviations in altitude beyond 200 feet (Transport Canada, 2012). As shown in Figure 4.2, higher-rated pilots showed minimal altitude deviations. Beyond 40 years of age, lower-rated older pilots showed noteworthy altitude deviations in the low-difficulty condition. In the high-difficulty condition only the highest rated pilots did not deviate beyond 200 feet. In the high-difficulty condition, starting at age 40, student, recreational, and private VFR+ pilots were observed deviating beyond 200 feet.

Figure 4.2 Average altitude deviations by age and pilot licensure and rating group. Line at 200 feet indicates a noteworthy deviation for private pilot flight tests in Canada (Transport Canada, 2012). Only the highest-rated pilots did not exceed the 200 feet deviation threshold.

To generate models of flight performance in both low- and high-difficulty conditions correlation analyses were first conducted with predictor variables representing cognition, pilot age, and composite expertise. Cognitive measures that significantly correlated ($p < .05$, two-tailed test) with flight performance were retained for subsequent model building.

As expected, flight performance in the low-difficulty condition was significantly correlated with performance in the high-difficulty condition, $r(55) = 0.483$, $p < .001$. Table 4.2 lists the variables with significant correlations with low- and high-difficulty factor scores. Appendix K contains full titles of abbreviated CogScreen-AE subtests.

Table 4.2
Correlations between Flight Performance and Cognitive, Age, and Expertise factors

	Pearson Correlations(<i>p</i>)	
	Low-Difficulty Condition	High-Difficulty Condition
LRPV		-0.331(.014)
Working Memory		
MTSAC	0.367 (.006)	0.421 (.002)
PFCAC	0.434 (.001)	0.347 (.010)
DTTPDAC	0.440 (.001)	
		-0.312 (.022)
VSCRTC	-0.299 (.028)	-0.353 (.009)
MATHRTC		-0.350 (.009)
SATINRTC		
Visual Tracking/Dual Task		0.333 (.014)
DTTDHIT	-0.537 (<.001)	
DTTDABS	-0.434 (.001)	-0.282 (.039)
Expertise	0.358 (.008)	
Age	-0.452 (.001)	
Age by Expertise	-0.364 (.007)	

Scatterplots were inspected for each flight performance and predictor variable pair. One extreme outlier case was found in the MTSAC and low flight performance plot. To moderate the effects of this outlier the MTSAC and low-difficulty flight performance score was replaced with the mean (without this case). Flight performance was significantly correlated with age in the low-difficulty condition and marginally correlated in the high-difficulty condition, $r(55)=-0.396$, $p=.003$ and $r(55)= -0.253$, $p=.065$, respectively. A scatterplot of the data was generated to explore why a moderate correlation of $r(55)= -0.253$ for age and high-difficulty condition flight performance was non-significant. Concerning the high-difficulty condition, a slope

similar to that found for the low-difficulty condition was observed (see Figure 4.3), with the exception that the outliers appeared to belong to younger pilots that in the high-difficulty condition. Bootstrap estimates (samples=1000) indicated a correlation between -0.48 to -0.01 (95% CI).

Figure 4.3. RMSE=root mean square error. Relation of flight path maintenance to pilot age in low (A) and high (B) difficulty conditions. Deviation scores were reversed so that better performance was associated with higher flight performance scores. Flight performance factor scores were exact linear combinations of altitude and airspeed RMSE in the low-difficulty condition and altitude, airspeed, and heading RMSE in the high-difficulty condition.

To determine final models of flight performance, stepwise multiple linear regressions were conducted separately for the low- and high-difficulty conditions. The stepwise method is preferred when the purpose of the model is to fine tune predictors from a larger set of predictors with known relationships to the criterion variable (Cohen, Cohen, West, & Aiken, 2002). Before examining all the relevant cognitive predictors, an initial model composed of age and expertise was examined first. In the low-difficulty condition an initial model of flight performance including age, composite expertise, and an age by expertise interaction term was generated. The pilot age by expertise interaction term was explored because expertise has frequently been

considered a moderator of age effects on flight performance. All r^2 values reflect are adjusted. A simple model of flight performance in the low-difficulty condition was significant at $F(2,50)=12.10$, $p<.001$, $r^2=0.30$. See Table 4.3 for regression model parameter estimates for pilot age and expertise in the low-difficulty condition. Table 4.4 illustrates the parameter estimates for the final model of flight performance in the low-difficulty condition, which also included the significant cognitive measures. While all potentially relevant cognitive measures were examined, only the measures with significant t-values were retained in the final model. The model including pilot age, expertise, a pathfinder working memory measure, and a dual-task ability measure was significant at $F(4, 48)= 11.27$, $p<.001$ and $r^2= 0.44$. Once the cognitive variables were entered into the equation pilot age no longer contributed significant variance. Expertise did remain as a significant predictor. The age by expertise interaction did not account for significant variance in flight performance in the low-difficulty condition.

Table 4.3

Age and Expertise Predictors for Simple Model of Flight Performance in the Low-Difficulty Condition

Model	<i>B</i>	<i>SE</i>	<i>beta</i>	<i>t</i>
(Constant)	1.360	0.362		3.758***
Pilot Age	-0.027	0.007	-0.433	-3.710***
Expertise	0.129	0.036	0.423	3.620***

Note. ***= $p< .001$. The pilot age by expertise interaction did not contribute significant variance in the simple model of flight performance in the low-difficulty condition. *SE*= Standard error.

Table 4.4
Predictors for the Final Model of Flight Performance in the Low-Difficulty Condition

Model	<i>B</i>	<i>SE</i>	<i>beta</i>	<i>t</i>
(Constant)	0.402	0.158		2.549*
Pilot Age	-0.006	0.009	-0.097	-0.695
Expertise	0.094	0.033	.0307	2.822**
PFCAC	0.124	0.048	0.291	2.599*
DTTHIT	-0.091	0.032	-0.375	-2.806**

Note. *= $p < .05$. **= $p < .01$. PFCAC= Pathfinder Accuracy and DTTHIT=Dual Task Boundary Hits (CogScreen-AE).

Because age was a factor of primary interest in this analysis and it achieved marginal significance in a two-tailed correlation test with flight performance in the high-difficulty condition, age was selected as a predictor in the initial model. In the high-difficulty condition an initial model of flight performance involving age and composite expertise was significant at $F(2,50)=6.04$, $p=.004$, $r^2=0.162$. See Table 4.5 for regression model parameter estimates for pilot age and expertise.

To determine the final model of flight performance in the high-difficulty condition the relevant cognitive variables were entered into the model in the manner described above. Table 4.6 illustrates the parameter estimates for a model of flight performance in the high-difficulty condition. A model including pilot age, expertise, and the pathfinder working memory measures was significant at $F(3, 49)= 7.04$, $p<.001$ and $r^2= 0.26$.

Further analyses were conducted with the simple model of performance in the high-difficulty condition to examine whether the CogScreen LRPV measure was predictive of performance. The LRPV is frequently used in neurocognitive assessments of older pilots in determining re-certification. The LRPV has been described as a measure sensitive to brain dysfunction (Kay, 1995). The LRPV has

shown strong association with age (Kay, 1995) and, in this sample, the LRPV was significantly correlated with age, $r(55)=-0.346$, $p=.01$. The LRPV was added to the age and expertise model LRPV to determine if the LRPV accounted for additional variance beyond that of age. The results showed that the LRPV contributed an additional 4.5% unique variance. When LRPV replaced age the final model accounted for 5.1% more variance than the age and expertise model ($r^2=0.21$).

In sum, for the initial simple models of performance, both age and expertise were significant predictors of low- and high-difficulty flight performance. In the high-difficulty condition the LRPV was a better predictor of performance than was age. While pilot expertise remained an important predictor of performance, once cognitive factors were entered into the model the age factor became insignificant. The effects of age appeared to be subsumed by working memory, and dual task tracking ability. The models of flight performance revealed that expertise, a number and letter pathfinder working memory measure, and a dual-task tracking ability measure (low-difficulty condition only) consistently predicted flight path deviation in the low- and high-difficulty conditions. Expertise did not significantly attenuate the effects of age on performance in the low- or high-difficulty condition.

Table 4.5
Predictors for the Final Model of Flight Performance in the High-Difficulty Condition

Model	B	Std. Error	<i>beta</i>	<i>t</i>
(Constant)	1.008	0.439		2.294*
Pilot Age	-0.021	0.009	-0.307	-2.400*
Expertise	0.120	0.043	0.353	2.763**

Note. *= $p < .05$. **= $p < .01$.

Table 4.6
Final Cognition and Expertise Regression Model of Flight Performance in the High-Difficulty Condition

Model	B	Std. Error	<i>beta</i>	<i>t</i> -value
(Constant)	-15.802	6.163		-2.564*
Pilot Age	-0.013	0.009	-0.184	-1.435
Expertise	0.097	0.042	0.285	2.323*
PFCAC	0.167	0.061	0.353	2.734**

Note. *= $p < .05$. **= $p < .01$. PFCAC= Pathfinder Accuracy (CogScreen-AE).

Correlational analyses between pilot age and PFCAC and DTTHIT, the two cognitive variables predicting flight path deviations, were conducted to explore the hypothesis that age effects could be subsumed by cognitive variables. Both PFCAC and DTTHIT were significantly correlated with age such that older age was associated with reduced pathfinder accuracy, $r(54) = -0.357$, $p = .008$ and substantially more tracking errors, $r(53) = 0.610$, $p < .001$. A linear regression analysis revealed that approximately 37% of the variance in tracking errors is accounted for by age, thus suggesting that a combination of working memory and visual tracking is sufficient to subsume the single effect of age (approximately 17%) on flight path deviations in low-difficulty conditions.

4.2.1 Results of the Flight Path Maintenance Analyses in the Context of Previous Findings

The finding that pilots who performed less well on working memory, processing speed, and visual tracking tasks also demonstrated poorer flight path maintenance is consistent with other general aviation flight simulation studies (Causse et al., 2011; Kennedy, Q. et al., 2010). In a comparable study, pilots flew a Frasca 141 flight simulator where flight path deviations were recorded during approach and holding patterns (Kennedy, Q. et al., 2010). Larger bank angle deviations were found to be predicted by older age, less expertise and an age by experience interaction term, ($r^2=0.24$). Expertise attenuated the effects of age on the bank angle deviation scores. Secondary analyses were completed with CogScreen-AE measures similar to those measures that correlated with flight path deviations in the present study. The measures represented visual tracking (dual task tracking error and boundary hits) and processing speed (pathfinder and symbol digit coding throughput). Unlike the present results, in the Kennedy, Q. et al. (2010) analysis with the cognitive variables entered into the regression model the age and age by expertise variables remained significant. Key differences are that the present study did not find expertise to be a moderator of age effects, with the effects of age apparently subsumed by the visual tracking and pathfinder variables. Findings from this work and other flight simulation studies show that older pilots are more likely to incur greater deviations from the ideal trajectory, but expertise and cognitive factors are better predictors of performance than pilot age (Causse et al., 2011; Taylor et al., 2005; 2007).

4.3 Situation Awareness

Situation awareness (SA) refers to the formation of accurate and comprehensive representations, or models, of the current and near future state of the environment.

Pilot SA is an example of a domain-dependent cognitive factor in the causal model of flight performance put forth by Hardy and Parasuraman (1997). According to Endsley (1988; 1995) SA relies on three general cognitive mechanisms. The first is the perception and integration of stimuli into meaningful units of information. A second mechanism binds relevant information into a comprehensive model of the environment. The third process projects the current model into a likely future model of the environment. By this definition, SA is reliant on several cognitive functions that somehow work in tandem to produce accurate and frequently updated representations of relevant aspects of the world. SA is responsive to top-down direction such as pilot attention and goals. At the same time, some aspects of SA are affected by elementary cognitive mechanisms such as working memory and processing speed, which support the production of SA in a bottom-up fashion (Bolstad, 2001; Gugerty & Tirre, 2000; Gutzwiller & Clegg, 2013). Because SA is influenced by cognitive mechanisms that are also susceptible to effects of age (Salthouse, 2010), it is likely that SA is also influenced by age (Bolstad & Hess, 2000).

Situation awareness has received considerable attention in the aviation literature, with loss of situation awareness associated with over 70% of pilot-caused general aviation accidents (Endsley, 1999). Investigations of the predictors of performance for each level of SA were undertaken for this analysis (Van Benthem et al., 2011a). The congruence between actual and self-ratings of SA performance was also examined to explore the potential advantages of including self-assessment data in models of performance (Van Benthem & Herdman, 2013).

4.3.1 Predictors of situation awareness in simulated flight.

An initial exploration of SA was conducted in 2011 and included data from 27 private pilots (Van Benthem et al., 2011). The purpose of the analysis was to determine how pilot cognition, age, and expertise influenced the three levels of SA. It was hypothesized that each variable uniquely predicted performance at each level of SA. The cognitive measures were DCAT subtest 6 (a working memory-type measure) and average response time for the mental task load measure, the PDT. SA scores from the low- and high-difficulty conditions were averaged to produce one score for each level of SA. SA probes were conducted using the SAGAT technique (Endsley, 1995). The SAGAT procedure is described in section 3.9.1.

Sample characteristics are described in Table 4.7. Older pilots ($M=59.4$ years, $n=11$) had flown for significantly more years, had similar total flight hours, but logged significantly less recent PIC hours than the younger pilots ($M=40.1$ years, $n=16$). Findings showed that age and recent pilot-in-command hours were associated with SA level 1; age, working memory, and PDT (RT) were associated with SA level 2; and, working memory and PDT (RT) were associated with SA level 3. Figure 4.4 illustrates the relative effect of age for each level of SA.

Table 4.7
Pilot Characteristics by Age Group

	Younger Pilot Group		Older Pilot Group	
	Range (Mean)	SD	Range (Mean)	SD
Age (Years)	27-50 (40.1)	7.1	52-76 (59.4)	7.6
Pilot Certification (Years)	0.5-25 (9.3)	7.2	0.5-40 (22.3)	13.7
Pilot-in-Command past 12 Months (Hours)	10-60 (31.8)	16.2	0-32 (8.2)	11.9
Total Flight Time (Hours)	99-1000(322.6)	249.3	90-1309 (397.2)	370.4

Note. SD= Standard Deviation

Figure 4.4 Mean situation awareness performance by age groups and situation awareness level. From Van Benthem, K., Herdman, C. M., Brown, M., & Barr, A. (2011a). The Relationship of age, experience and cognitive health to private pilot situation awareness performance. Presented at the 16th International Symposium on Aviation Psychology, Dayton, Ohio.

Table 4.8 outlines the Pearson correlations of each predictor at each level of SA. Individual regression analyses were also conducted to determine the relative influence of hypothesized predictors for each level of SA. A model utilizing recent PIC hours and age accounted for 28% of the variance in SA level 1, $F(2,22)=4.30$, $p=.027$. A model predicting 54% of the variance in SA level 2 scores was produced with working

memory, PDT response time, and age, $F(3,20)=7.84$, $p=.001$. Similarly, a model constructed with working memory and PDT response time accounted for 35% of the variance in SA level 3.

Table 4.8

Relationship of Situation Awareness Performance to Pilot Age, Experience and Cognitive Health

Situation Awareness Level	Peripheral Detection Task	DCAT 6	Pilot Age	Recent Pilot-in-Command Hours
Level 1: Perception	ns	ns	-0.49	0.44
Level 2: Schema	-0.50	0.46	-0.49	0.40
Level 3: Future Schema	-0.41	0.50	-0.50	ns

Note. All correlations are significant at $p<.05$ level, two-tailed. From Van Benthem, K., Herdman, C. M., Brown, M., & Barr, A. (2011a). The Relationship of age, experience and cognitive health to private pilot situation awareness performance. Presented at the 16th International Symposium on Aviation Psychology, Dayton, Ohio. ns= non-significant.

These findings demonstrated that SA level 1 was influenced by a different pattern of predictors than SA levels 2 and 3. SA level 1 was not predicted by other lower-level cognitive processes, perhaps because SA level 1 is also representative of a lower-level cognitive function. With respect to expertise, in this analysis only private pilots were included in the sample. With licensure not available as a variable in the homogenous group of private pilots, recent PIC hours and total hours were the remaining expertise factors that could potentially predict performance. For both SA levels 1 and 2 only recent PIC (Pilot-in-Command) hours was a significant expertise-related predictor.

To further delineate the fundamental differences between SA level 1 and levels 2 and 3, two multiple linear regressions were performed. The first analysis determined

the magnitude of SA level 1's effect on SA level 2 (averaging SA from the low- and high-difficulty conditions). This second analysis determined the relative contributions of level 1 and level 2, to the most comprehensive aspect of SA, which was level 3. It was hypothesized that SA level 1 strongly predicts level 2 (e.g., more than 25% of the variance in performance). The second hypothesis was that both level 1 and level 2 equally predicted SA level 3. The findings from the two multiple regressions are presented in Table 4.9. The first hypothesis was partially fulfilled in that level 1 did predict SA level 2, however to a much smaller degree than anticipated, $F(1,87)=9.00$, $p=.004$, $r^2=0.094$. The second hypothesis was also partially supported in that only SA level 2 significantly predicted level 3, $F(2,86)=20.47$, $p<.001$, $r^2=0.32$ (0.29 from SA level 2).

Table 4.9
Regression Models for SA levels 2 and 3.

Model	B	SE	<i>beta</i>	<i>t</i>
SA Level 2 (dependent)				
(Constant)	60.519	5.299		11.421***
SA Level 1	0.318	0.106	0.306	3.000**
SA Level 3 (dependent)				
(Constant)	31.619	6.967		4.538***
SA Level 1	0.001	0.093	0.001	0.013
SA Level 2	0.543	0.089	0.567	6.087***

Note. **= $p<.01$. ***= $p<.001$. SE = Standard Error.

To determine whether this trend held with the larger sample of 108 pilots the Pearson correlations between SA levels 1, 2 and 3 were compared (averaged across the mission difficulty conditions). In this larger sample SA level 1 was significantly correlated with SA level 2, $r(108)=0.307$, $p=.001$, and was not significantly correlated

with SA level 2, $r(108)=.179, p=.064$. As expected from the initial analysis, SA level 2 and 3 were moderately correlated, $r=0.550, p<.001$.

4.3.2 Results of the SA Analyses in the Context of Previous Findings

This finding that SA level 1 was not significantly associated with working memory is supported by other research that investigated the differential effects of cognitive mechanisms on the various levels of SA. Cass (2011) examined SA and Sea King helicopter pilots in the Canadian military and also noted that SA level 1 was somewhat distinctive and that accuracy for the first level of SA was reduced in comparison to accuracy for SA levels 2 and 3. That SA level 1 was not predicted by working memory mechanisms is also supported by Gutzwiller and Clegg (2013) who tested the association between a comprehensive measure of working memory and each level of SA. Contrary to their hypothesis, SA level 1 (for a complex “fire-chief” scenario-based decision-making task) was not predicted by working memory. Similar results were found in a study, this time using expert pilots, where SA level 1 was only predicted by low level visual-spatial tasks, and not tasks that tapped working memory (Sulistyawati, Wickens, & Chui, 2011). In contrast, the spatial tasks showed weak associations with SA levels 2 and 3.

4.3.3 Repositioning SA level 1

A synthesis of the findings from the investigation of the predictors of each level of SA and the relative contributions of level 1 to both levels 2 and 3, indicated that SA level 1 was functionally different from levels 2 and 3. The initial process of all models that describe an SA-like construct is consistently described as the monitoring and apprehension of information from the environment. In Endsley’s information

processing-based concept of SA, level 1 is defined as the perception of elements in the current “situation”. Neisser’s (1976) model of the perception-action cycle also begins with perception. The perception-action cycle is generally considered a precursor to more modern models of SA. In the Neisser model emphasis was put on the ecological drive to seek out information. Smith and Hancock (1995, p. 137) defined situation awareness as “adaptive, externally directed consciousness”. The beginning of this consciousness also arises from the initial registration and sense-making of stimuli from the internal or external environment. The idea that monitoring and integrating information from environments is conceptualized exclusively as part of SA is a problem for other models of cognition. In fact, it is understood that all cognitive mechanisms claim perception and integration of information as their first process (Doyle & Ford, 1998). In light of the functionally-distinct role that SA level 1 appears to play in the SA process and because of the likely dependence on information detection and integration by numerous cognitive mechanisms I proposed a modification to Endsley’s model of SA in dynamic decision-making (see Figure 4.5). The modification requires repositioning SA level 1, currently encapsulated within the SA framework, to a more central and accessible location (suggested modifications to the model are outlined in Figure 4.5). So located, “perception” is liberated to be utilized as a supporting mechanism for various other cognitive functions that rely on the ongoing integration of information from the environment. Recall that both O’Hare (2006) and Endsley (1999) reported that mis- or non-detection of information was the most commonly cited reason for incidents or accidents related to pilot-error. In addition, if the three level information processing model is correct then SA level 1 should produce accuracy scores as good as, or even better than levels 2 or 3, since the second and third levels build on the first level. However, in this work and others

(Cass, 2011; Sulistyawati et al., 2011) SA level 1 accuracy was significantly lower than SA level 2.

In sum, I am suggesting that the lower accuracy for SA level 1 might be tied to its general purpose function. It appears that the SA level 1 probes developed by researchers tap a broader array of information than pilots actually monitor for just SA purposes. It is also possible that the small amounts of information from the hypothetical SA level 1 that are subsequently integrated into mental models (represented by SA levels 2 and 3) might become exponentially more “information-laden” as they integrate with pre-existing information structures (SA levels 2 and 3). It is also out of necessity that pilots monitor a much larger assortment of stimuli than are used for SA, because SA is just one aspect of the complex network of activities required for competent flight. For example, other sorts of information to be monitored are the memory cues required for prospective memory tasks, and the details of auditory information delivered by communication channels. The opportunity to explore how SA level 1 (renamed iRES, explanation to follow) functioned as a general purpose information monitor was afforded by its new central location in the model of pilot performance. As shown in Figure 4.5, Endsley’s model of SA was modified such that SA level 1 was repositioned in a central location, as a general purpose mechanism for perceiving stimuli. Stimuli are perceived, processed, and are utilized in the manufacturing of situation models, *and* are used for other cognitive functions (e.g., prospective memory) as needed.

Figure 4.5 Modifications to SA level 1 in Endsley's model of SA in dynamic decision making. The initial schematic is adapted from Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 32-64. The new "iRES" box illustrates that SA Level 1 is removed from SA, and is available for association with other cognitive functions. "iRES" box and arrows to the from iRES are not elements of Endsley's original model of SA in dynamic decision-making.

One purpose in removing SA level 1 from the "bonds" of the SA framework was to allow direct access between information perception and integration, and other cognitive mechanisms. To test the validity of this supposition a further secondary analysis was conducted to determine if, in fact, SA level 1 was consistently related to other cognitive constructs measured in this dissertation. Mechanisms that necessarily rely on information perception and integration—but are not traditional SA tasks—and were measured in this work included prospective memory, flight path maintenance,

the PDT measures, and episodic memory. Table 4.10 provides an overview of the association of each level of SA (in low- and high-difficulty conditions) with prospective memory, flight path maintenance, the PDT, and episodic memory. It is evident from Table 4.10 that SA level 1 is associated with all other domain-dependent variables. These associations are not as apparent with SA levels 2 and 3. To avoid confusion in subsequent model building phases of this work, and to acknowledge the ubiquitous association of SA level 1 with other cognitive constructs, SA level 1 is referred to subsequently as *iRES* (information resolution). The label *iRES* is a fitting description for the measure that was originally utilized to collect the SA level 1 scores during the SAGAT technique in the present study. Each probe for SA level 1 was scored according to the level of detail that each participant could recall. If, for example, a partial call sign for another aircraft in the circuit was recalled then a partial score, reflecting the amount of information that was recalled for that particular probe, was assigned for that query. Each probe was scored as 0, 1, or 2, in accord with the level of detail provided. The probe score therefore reflects the *resolution* at which the visual and auditory information was perceived and integrated into the mental representation of the scene.

Table 4.10

Relationship of SA Levels 1, 2, and 3 to Prospective Memory, Flight Path Maintenance, PDT, and Episodic Memory in Low- and High-Difficulty Conditions.

Cognitive Construct	SA Level		
	1	2	3
Prospective Memory			
low-difficulty	L, H		L
high-difficulty	H		
Flight Path Maintenance			
Altitude			
low-difficulty	L, H		
high-difficulty	L, H		
Airspeed			
low-difficulty	H		
high-difficulty	H	H	H
Mental Task Load (PDT)			
Reaction Time			
low-difficulty	H		H
high-difficulty	H		
Hit Rate			
low-difficulty	L, H		
high-difficulty			
Episodic Memory			
low-difficulty	L, H		
high-difficulty	L, H		

Note. L= significant ($p < .05$) Spearman's ρ between SA (low-difficulty scores) and predictor variables. H= significant ($p < .05$) Spearman's ρ between SA (high-difficulty scores). As illustrated in the red oval, the L and H in the column SA level 1 and row for the low-difficulty prospective memory variable indicates that SA level 1 in low- and high-difficulty conditions were significantly correlated with prospective memory performance (low-difficulty condition). Whereas, only SA level 1 high-difficulty was significantly correlated with prospective memory (high-difficulty condition). This Table illustrates the ubiquitous association of iRES (SA level 1), and the sparse association of SA levels 2 and 3 with other flight performance parameters.

4.3.4 Congruence between Subjective and Objective Measures of Situation Awareness

The ability to predict SA performance using pilot self-ratings of SA would be beneficial considering the notable association between SA errors and critical incidents (Endsley & Bolstad, 1994; Bolstad et al., 2010). In the present analysis, subjective (or

self-rated) situation awareness was measured using a 7-point scale designed specifically for aviation tasks (Hou et al., 2007). The scale consisted of 11 aspects of situation awareness typically encountered during flight and included references to awareness of the environment, tasks, time, priorities, relative position of ownship and other aircraft, future events, and overall level of situation awareness. In light of the critical nature of SA it seems advantageous to explore reliable methods of predicting SA performance. Two analyses were conducted to explore the relations between the SA self-rating scale items (first individually and then as two factors determined by principal component analysis) and performance on the three levels of SA. In these studies the sample consisted of the 93 participants who had flown the full simulator protocol and responded to the SA self-rating scale in both low- and high-difficulty conditions.

Table 4.11
Inter-Correlations between Items of the Self-Rated Situation Awareness Scale

	Items on the SA Self-Rating Scale										
	1	2	3	4	5	6	7	8	9	10	11
1 Navigation Environment		0.438*	0.355*	0.377*	0.321*	0.368*	0.324*	0.379*	0.315*	0.365*	0.661**
2 Tasks Performed	0.344*		0.710*	0.542*	0.747*	0.464*	0.442*	0.434*	0.385*	0.393*	0.448**
3 Specific Tasks	0.495*	0.754*		0.743*	0.632*	0.370*	0.282*	0.253*	0.192*	0.263*	0.390**
4 Perfect Circuit	0.373*	0.591*	0.679*		0.533*	0.338*	0.287*	0.241*	0.225*	0.295*	0.426**
5 Task Priorities	0.335*	0.493*	0.410*	0.545*		0.433*	0.524*	0.483*	0.393*	0.413*	0.404**
6 Relative Ownship	0.270*	0.301*	0.278*	0.303*	0.396*		0.583*	0.720*	0.699*	0.723*	0.649**
7 Impact of Other Aircraft	0.244*	0.232*	0.345*	0.407*	0.357*	0.520*		0.575*	0.570*	0.468*	0.484**
8 Predict Events	0.239*	0.237*	0.284*	0.264*	0.438*	0.661*	0.544*		0.843*	0.777*	0.699**
9 Other Aircraft Activities	0.186	0.397*	0.342*	0.423*	0.529*	0.642*	0.550*	0.642*		0.792*	0.675**
10 Other Aircraft Locations	0.218*	0.342*	0.295*	0.218*	0.376*	0.653*	0.477*	0.550*	0.624*		0.691**
11 Overall SA	0.442*	0.455*	0.465*	0.493*	0.607*	0.546*	0.461*	0.469*	0.605*	0.516*	

Note. Spearman's ρ correlations above the diagonal are from the high-difficulty condition and below are from the diagonal from the low-difficulty condition. * = $p < .05$ and ** = $p < .01$.

Individual items of the SA self-rating scale and individual pilot differences were regressed on the three levels of SA (Van Benthem et al., 2013). A linear regression

model found that lower overall SA (a score averaging the three levels of SA) was predicted by older age ($\beta = -0.25$), lower working memory ($\beta = 0.18$), and a pilot's self-rating for awareness of future events ($\beta = 0.32$). A second regression model found that lower SA level 1 (high-difficulty condition) was predicted by older age ($\beta = -0.31$), lower pilot rating ($\beta = 0.25$), slower processing speed ($\beta = 0.18$), and a pilot's self-rating for awareness of the navigation environment ($\beta = 0.21$). Only two items from the scale predicted the observed SA performance. These results indicated that either the items did not fully reference the notion of SA as prescribed by the objective measurement of SA, or that pilots are poor estimators of their own performance pertaining to SA.

4.3.5 Principal Components Analysis for Situation Awareness Rating Scale Items

The weak relationship shown between the subjective and objective measures of SA might be explained by confirming the factor structure underlying the SA self-rating scale. To examine whether the SA self-rating scale may actually measure more than one factor pertaining to SA a principal components analysis was undertaken. Table 4.12 demonstrates a plausible two factor structure, where the first factor represents information pertaining to ownship and the position of other aircraft. The second factor appears to reflect information pertaining to tasks, timing of tasks, and the general flight environment.

Table 4.12
Two-Component Matrix for Subjective SA Scale Items (Low-Difficulty Condition)

Scale Items for Low Mission Difficulty Condition	Component	
	1	2
The navigation environment		.628
Tasks to be performed		.782
Time to complete specific tasks		.866
Time to complete perfect circuit		.839
Task priorities		.545
Aircraft's position relative to other aircraft in the circuit	.816	
Impact of other aircraft on ability to fly the "perfect circuit"	.698	
Ability to predict future circuit events	.859	
Awareness of other aircraft activities	.817	
Awareness of other aircraft location(s)	.843	
Overall level of situation awareness	.645	

Note. Rotation Method: Varimax with Kaiser Normalization. Loadings below .50 are not shown.

4.3.6 Summary of situation awareness performance and effects

In sum, the present findings indicated that SA is a cognitive construct with levels that were differentially influenced by pilot attributes and cognitive mechanisms. SA levels 2 and 3 were most closely linked with each other, and both significantly correlated with age, reaction time, and working memory. SA was also influenced by pilot expertise, with this relationship most evident for the youngest and oldest pilots, and with SA level 1. To permit SA level 1 to function as a central variable in subsequent complex models of aviation outcomes, SA level 1 was conceptually distanced from SA levels 2 and 3, and was relabeled as iRES. Results from two analyses

demonstrated that the items from the subjective SA self-rating scale appeared to index two distinct components of SA: awareness of tasks, and awareness of ownship and the position of other aircraft. The separation in constructs resulting from the PCA with the self-rated SA scale items was leveraged in Chapter 6, where the three items pertaining to task management from the SA self-rating scale were utilized as manifest variables for a task management (TM) latent variable (a first-order variable in the DMM). This use of self-rated items represents the only occurrence in the model of pilot performance, to be discussed in Chapters 5 to 9, where subjective items were permitted to assume the same causal role as objective indicators.

4.4 Prospective Memory for Radio Communication Tasks

Prospective memory is a cognitive function that supports remembering to complete some thought or action in the future. The ability to remember to perform tasks in the cockpit relies on pilot prospective memory. Failures of prospective memory during flight can have catastrophic results (for a review see Dismukes & Nowinski, 2007). This has led some to explore how pilot attributes might contribute to prospective memory failures. The effects of age on prospective memory have been inconsistent. To date there are no studies that specifically examine the effects of age on prospective memory for aviation tasks. Researchers have also found that the type of memory cue associated with triggering the prospective memory was also associated with remembering to perform the task (Delprado et al., 2013; see McDaniel & Einstein, 2000 for a discussion on memory cues; and Utzl, 2011 for a meta-analysis on cue-type and prospective memory). In general, it was observed that the less salient or noticeable the cue, the less likely the prospective task was recalled and completed.

In the present work, prospective memory was measured by the number of radio calls pilots remembered to complete for six specified locations in each circuit. The

turning downwind and turning base locations were two radio calls that were inherently accompanied by salient recall cues. At the turn for downwind and base legs of the circuit there are visual cues related to horizon angle changes (due to banking in the turn) and there are physical cues arising by the manipulation of the yoke to initiate and control the turn. Despite the lack of physical motion with the Cessna 172 simulator system used in this work, turns did result in a slight to moderate physical sensation of turning, presumably brought about the strong visual cues indicating that the plane was banking in a turn. Two other calls were considered to have low-salience recall cues: airborne (or “away”) and mid-downwind. At the two low-salience cue positions in the circuit there was little to change in the horizon and no physical manipulation of the yoke. It was hypothesized that mission difficulty and older age influenced performance such that older pilots (>49 years) would forget to complete radio calls more often than younger pilots (<50 years), and that fewer radio calls would be completed in the high-difficulty condition. It was also predicted that recall for the two low-salience cue radio calls would be less than the two calls with the high-salience cues. Age and cue salience were hypothesized to interact such that older adults experienced greater effects from low-salience cues than the effects experienced by the younger pilots.

Figure 4.6. Interaction of workload and cue salience (A) and age and cue salience (B). From Van Benthem, K., Herdman, C.M., Tolton, R., & LeFevre, J. (2015). Prospective memory failures: effects of cue salience, workload, and pilot individual differences. *Aerospace Medicine and Human Performance*, 86(4), 366-373.

Preliminary analyses using a sample of 50 pilots revealed that older age and higher mission difficulty resulted in significantly fewer calls being completed (Van Benthem et al., 2011c). In a later analysis, repeated-measures ANOVA with a sample of 100 pilots (Van Benthem et al., 2012a) revealed significant effects from memory cue salience, workload, and age. Mission difficulty and pilot age interacted with memory cue salience in the manner hypothesized (see Figure 4.6). The airborne and mid-downwind radio calls (identified a priori as having low-salience cues) were significantly more likely to be forgotten by older pilots, particularly in the high-difficulty condition (as compared to the radio calls for downwind and base turns). This three-way interaction between cue salience, workload, and pilot age was marginally significant ($p=.08$).

Predictors of total prospective memory completion rates (regardless of memory cue) were investigated in a subsequent analysis with the full sample of 108 pilots for each difficulty condition. Results showed that the DCAT subtest 6 (a complex measure

of working memory), pilot licensure and rating, and pilot age significantly predicted performance in the low-difficulty condition, $F(3, 104) = 10.49, p < .001, r^2 = 0.23$. In the high-difficulty condition working memory, DCAT subtest 2 (measuring span of attentional field), and pilot expertise were predictors of performance, $F(3, 104) = 6.71, p < .01, r^2 = 0.16$.

4.4.1 Results of the Prospective Memory Analyses in the Context of Previous Findings

In regard to the first hypothesis, mission-difficulty effects showed that a highly demanding background environment was associated with significantly poorer prospective memory performance for all pilots. Furthermore, there was a drop in prospective memory scores when the radio calls had low-salience cues in the high-workload scenarios. The effects of workload in this study are consistent with observations by Dismukes and Berman (2010) that high workload magnified prospective memory errors in aviation settings. High workload might impact performance by causing an increase in missed task cues (Nowinski et al., 2003), and by inefficient encoding of the intent to remember at the outset of the mission, possibly due to greater distractions in higher workload conditions (Dodhia & Dismukes, 2009).

It was also found that older pilots did not perform as well as younger pilots on a naturalistic prospective memory task. These findings are contrary to other research where in naturalistic settings normal older adults typically perform as well as, or better than, younger adults (Rendell & Craik, 2000; Rendell & Thomson, 1999). The impact of low-salience memory cues helps to explain this difference (Delprado et al., 2013). The significant age by cue salience interaction, found in this work, points to older pilots being more affected by low-salience cues than younger pilots. In other research,

naturalistic situations tended to have high-salience cues, or otherwise uncontrolled cue salience. The present research is unique in that it was able to clarify the relation between cue salience and prospective memory performance (in naturalistic aviation settings) by controlling the salience of cues. The finding that radio calls associated with less salient cues are less well-remembered is supported by Nowinski and Dismukes (2005) who suggested that, in real-world prospective memory situations, forgetting to attend to tasks could result from competing tasks in the environment and from low salience associated with any cues that might trigger the previously-intended action.

The finding that the cognitive mechanisms of working memory and attentional field of view also predicted performance give some indication as to why the older pilots were more likely to forget to complete the radio calls, particularly those that were not associated with strong visual or physical cues. The current results are supported by the suggestion from Logie, Maylor, Della Sala, & Smith (2004) and West and Bowry (2005) that working memory is a mediator of prospective memory performance. Both found that increased load on working memory demands reduced laboratory-based prospective memory performance, with older subjects experiencing more of an effect of load than the younger subjects experienced.

The result that expertise did not significantly attenuate the effect of age on radio call completion was in keeping with the findings from Sanjram and Khan (2011) who also found that the salience of the memory-cue was a significant predictor of prospective memory, whereas subject expertise was not. Sanjram and Khan (2011), using a computer programming task, found that it was multi-tasking ability, rather than expertise, that predicted prospective memory errors. In the Nowinski et al. (2003) report only airline transport flights were investigated, meaning that the prospective

memory errors were committed by highly-trained commercial pilots. Thus, in a study where prospective memory performance might be somewhat worse for very low expertise pilots as compared to high expertise pilots, this small effect of experience might not be strong enough to offset the much larger deleterious effects of age.

In summary, prospective memory for communication tasks in the cockpit was predicted by pilot age, cognition, expertise, and by workload conditions. A second key finding was the association of prospective memory performance to the features (salience) of the memory-cue associated with the to-be-remembered task. Pilots were significantly more likely to remember to complete the task when the cue possessed salient physical and visual features. The relevance of cue-salience was made clear when no effect of age on performance was observed for high-salience cue radio calls. Prospective memory was also sensitive to interactions between predictors, such that particularly poor performance was observed for older pilots and those with low working memory. The sensitivity of prospective memory performance to cue-salience and to working memory suggests that prospective memory in the cockpit relies on an “automatic” mechanism for cue detection and task execution (Einstein & McDaniel, 2005). In contrast, the “strategic” mechanism is an alternative theory to the automatic mechanism of prospective memory and proposes that intentional monitoring of the environment for the appropriate cue is the main supporting process for prospective memory successes (Einstein & McDaniel, 2005; Gynn, 2003). More recently, computational modeling of prospective memory activity suggested that a combination of top-down control and bottom-up perceptual processes work in tandem to support performance (Gilbert, Hadjipavlou, & Raoelison, 2013). The present results do not completely dissociate automatic versus strategic mechanisms, as automatic processes are only thought to be effective when the memory cue is “focal” to the ongoing activity

(Einstein & McDaniel, 2005), as they were in the present study. Regardless, the present work suggested that top-down processes, perhaps supported by expertise, may have some effect on prospective memory performance. In contrast, the constellation of effects from age, cognition, workload, and cue-salience supported an automatic retrieval mechanism: a mechanism strongly influenced by lower-level perceptual processing and working memory ability. The underpinning of prospective memory by a network of cognition-oriented predictors supports the inclusion of prospective memory in a model of pilot performance based on a cognitive systems engineering framework, and suggests that prospective memory might also be a useful predictor of outcome variables in the present work.

4.5 Mental Task Load

Mental task load can be defined as the mental resources required of an operator to undertake complex tasks. There is arguably some mental task load threshold that, once crossed, can lead to reduced performance and increased risk to the pilot. Mental task load can be influenced by environmental factors, such as manipulations of mission difficulty, as was done in this work. The following analyses investigated pilot mental workload states, measured both subjectively and objectively. For a variety of reasons, different tasks impose varying levels of cognitive load upon a pilot's cognitive system. Understanding how mental task load influences performance on domain-independent and domain-dependent skills can inform a model of pilot performance. Furthermore, increasing the precision in measuring the relationship between mental task load and pilot performance might also assist in identifying pilots most at risk of poor performance.

In the analyses that follow, the concept of mental task load, or cognitive workload, and its potential value in models of aviation outcomes was explored by

examining two measures of mental task load. First, the analyses quantified how much variance in an array of cognitive tasks, representing domain-independent and domain-dependent cognition, could be predicted by subjective measures of mental task load. The second predictor of performance made use of a peripheral detection task (PDT). The PDT is an objective measure of the cognitive resources that remain available to the pilot after other resources necessary for the primary task (the flight simulator tasks) are employed (Vidulich & Tsang, 2012). The PDT serves as a proxy of mental task load. In this analysis (Van Benthem & Herdman, 2014a) the NASA TLX subjective rating scale (Hart & Staveland, 1988) and a novel auditory peripheral detection task (PDT) measure (an objective measure of cognitive workload) were first validated and refined using factor analysis. Second, using structural equation modeling (SEM), the latent workload constructs were investigated simultaneously as predictors of performance for an array of criterion variables, ranging from domain-general cognition to complex aviator tasks (see Figure 4.7). Two models were produced, one for younger pilots (<49 years) and one for older pilots (49+ years). Though not *causal* in nature, a workload measure could improve the prediction of performance beyond that predicted by upstream factors in a comprehensive model of pilot performance. Pilots completed the NASA TLX scale immediately after each mission difficulty condition. The objective (PDT) measure of pilot workload represented the cognitive resources available to the pilot that remain after mental resources are supplied to all relevant cognitive and motor tasks required to conduct the mission. The objective measure of cognitive workload represented a potential causal factor in the model. The auditory PDT data was collected continuously while pilots completed the flight simulation protocol.

Figure 4.7 Structural model illustrating possible predictive links between subjective and objective measures of workload and an array of domain-independent and domain-dependent cognitive measures. From Van Benthem, K., & Herdman, C. M. (2014). Congruence of self-rating with pilot performance: Validation of a task load index for human-machine systems. Presented at the *Aerospace Medical Association Annual Scientific Meeting*, May 14, 2014, San Diego, CA. TLX= task load index; PDT= peripheral detection task; ProSpd= processing speed; WMem= working memory; TRP= task-related performance (flight path maintenance); and SA= situation awareness. The full structure tests the association between both workload indices, and domain-general cognition and flight task constructs (only one arrow from both measures are shown here). All performance measures are composite factors derived from CogScreen-AE and flight simulator measures. Age, expertise, and mission difficulty, not shown, were included in the model as control variables.

4.5.1 Factor analysis for objective and subjective indices of cognitive workload

It was hypothesized that the five items from the subjective NASA TLX scale and the three indicators for the objective measure of workload loaded significantly on their respective latent construct. It was also expected that objective and subjective measures of workload would show weak to moderate correlations in both mission difficulty conditions.

Confirmatory factor analysis was undertaken in the MPlus 7.11 (Muthén & Muthén, 1998) platform to develop latent factor scores for the subjective and objective

workload measures. MPlus 7.11 uses covariance-based analysis and is the preferred SEM method for theory-building when previous structures have already been identified. Once the measurement models for the objective and subjective workload constructs were developed, the outer structural modeling was undertaken in WarpPLS 4.0 (Kock, 2013), a modeling platform recommended when the goal of modeling is predominantly prediction, and theory-building to a lesser extent. The original full measurement model, including the PDT and NASA TLX indicators for both mission difficulty conditions, showed poor model fit, $\chi^2 = 212.907$ (66, $N=93$), $p < .001$, RMSEA (root mean square error of approximation) = 0.082, and CFI (comparative fit index) = 0.884. Good model fit is characterized by a non-significant *chi-square* value ($p > .05$), RMSEA below 0.06, and CFI above 0.95 (Schermelleh-Engel, Moosbrugger, & Müller, 2003). A description of the final model and the minor refinements undertaken are described as follows.

In the original, poorly fitting model, the subjective workload construct included the following indicators: mental, physical, and temporal demands, frustration, effort, and overall workload. The “frustration” and “physical” items loaded poorly (standardized *beta* values were less than 0.65) and were removed as indicators of the subjective workload construct. The item representing frustration also unexpectedly correlated negatively (in both mission difficulty conditions) with the item querying temporal demands. The negative correlation of “frustration” with “temporal demands” suggests that the question regarding how much frustration was associated with the flight did not function as intended in the scale. The remaining items of mental and temporal demands, effort, and overall workload revealed standardized loadings at 0.70 or higher (see Figure 4.8).

The three original indicators for objective workload were retained for the final measurement model as they each loaded significantly on the construct. The standardized item loadings for reaction time and hit rate were greater than 0.65. The loadings associated with the standard deviation of reaction times of 0.59 [low-difficulty condition] and 0.33 [high-difficulty condition]) were considered low, but were retained as final model fit was good and loadings were significant and were in the expected directions (consistent with the reaction time and hit rate loadings). The measurement error terms for reaction time and standard deviation of reaction time (both derived from the same original reaction time values) were permitted to correlate to improve model fit (see Figure 4.8).

Figure 4.8 Standardized loadings for the full measurement model for cognitive workload including indices for the objective and subjective workload constructs in the low- and high-difficulty conditions. From Van Benthem, K., & Herdman, C. M. (2014). Congruence of self-rating with pilot performance: Validation of a task load index for human-machine systems. Presented at *the Aerospace Medical Association Annual Scientific Meeting*, May 14, 2014, San Diego, CA. tlxlo= NASA TLX construct (low-difficulty); tlxhi= NASA TLX construct (high-difficulty); tl1 and tl7= mental demands; tl3 and tl9= temporal demands; tl5 and tl11=effort; tl6 and tl12=overall workload; objlo= Objective cognitive workload construct (low-difficulty) objhi= Objective cognitive workload construct (high-difficulty); o1 and o4=PDT average reaction time; o2 and o5=PDT individual standard deviations of reaction time; and o3 and o6=PDT hit rate. This measurement model validation was undertaken before the validation of the model shown in Figure 4.7.

The final hypothesized model showed good fit with the actual variance/covariance matrix, $\chi^2 = 86.69(68, N=93)$, $p=.063$, and good fit indices, RMSEA= 0.05, and CFI= 0.97. The first hypothesis that all the NASA TLX scale items and all the PDT measures would load significantly on their respective constructs was partially met. The final TLX model retained four of the five items from the subjective rating scale, with “physical demands” and “frustration” proving to be

unreliable indicators. Two items from the objective workload construct loaded strongly as expected: average reaction time and hit rate. The individual standard deviations of reaction time indicator loaded weakly in each mission difficulty condition. Despite their low loadings the overall model quality indices for the objective workload construct were very high and since the indicators are not scale items (generally scale items loading below 0.60 are removed) the low loading items were retained.

The second hypothesis that the subjective and objective workload constructs correlated significantly with each other in each mission difficulty condition was not supported by the results of the measurement model development. As shown in Figure 4.8 by the double-headed arrows between the workload constructs (top of figure), the two types of workload construct did not correlate significantly with each other. The low-difficulty subjective measure did correlate significantly with the same construct in the high-difficulty condition, $r(91)=0.46$. The same correlation pattern was observed for the objective measure of workload, $r(91)=0.55$. Significant cross-correlations *between* constructs were not observed. With the measurement models for subjective and objective workload constructs completed, the full structural equation analysis could be undertaken to determine if each workload construct significantly predicted performance for cognitive tasks ranging from simple domain-independent cognition to more complex domain-dependent cognitive tasks.

4.5.2 Relation of cognitive workload indices to younger and older pilot performance

The full hypothesized structural model is shown in Figure 4.8. Once the complete structural model was developed in the WarpPLS 4.0 platform (Kock, 2013), additional

measurement model reliability and validity indices were determined. Quality indices relevant to reflective measurement models included composite reliability and Cronbach's alpha, two indices that reveal the average degree that scale items are inter-related (Cronbach, 1951; Sijtsma, 2009). Model quality was also measured by the Average Variance Extracted (AVE) that reflects the discriminant validity of each construct by determining if most of the variance in the indicators is related to the latent construct. A third quality measure is the Full-collinearity Variance Inflation Factor (VIF), which tests that coefficients associated with items are not biased due to high inter-item correlations. The composite reliability and Cronbach's alpha were acceptable (ideally above 0.60 [Fornell & Larcker, 1981]) for the subjective workload construct, 0.930 and 0.900, respectively. The AVE (ideally above 0.50, [Fornell & Larcker, 1981]) was 0.769 and the full collinearity VIF (ideally below 3.3, [Kock, 2013]) was 1.013. Quality indices for the objective workload construct revealed that composite reliability and Cronbach's alpha were acceptable, 0.813 and 0.652, respectively. The AVE was 0.595 and the full collinearity VIF was 1.013. All measurement model quality indices for both subjective and objective constructs were considered acceptable according to accepted reflective measurement model standards (Kock, 2013).

The results of the structural equation modeling are shown in Figure 4.9. In this figure the metric of the y-axis is standardized *beta* values. The larger the absolute magnitude of the *beta* value the stronger the relationship between the predictor and criterion variable. The x-axis contains bars representing the predictive strength and direction of the subjective and objective workload constructs for younger (<49 years) and older pilot groups (49+ years). Age groups were split to maximize equal group sizes. The model was tested separately according to age groups because of the strong

age effects generally associated with domain-independent measures such as processing speed and working memory (Salthouse, 2000, 2010) and the effects of age associated with iRES, flight path maintenance and SA that were shown previously in this work. It was important to determine whether workload indices predicted cognitive performance similarly for older and younger pilots, or whether age moderated this relationship too. If age group appeared to differentially impact the association between measures of workload and cognitive performance then subsequent data modeling should be careful to examine possible interactions between age and workload on cognitive tasks.

Figure 4.9. Relationship of cognitive workload indices to an array of pilot measures for younger (<49 years) and older (49+) years pilot groups. From Van Benthem, K., & Herdman, C. M. (2014a). Congruence of self-rating with pilot performance: Validation of a task load index for human-machine systems. Presented at the *Aerospace Medical Association Annual Scientific Meeting*, May 14, 2014, San Diego, CA. PS= processing speed; WM= working memory; TRP= task-related performance (flight path maintenance); SA= situation awareness (factor scores for SA construct including SA level 2 and 3). Bars represent the strength (standardized *beta* values) and direction of the subjective and objective workload construct for younger and older pilot groups for five performance constructs (PS, WM, iRES, TRP, and SA). The larger the absolute magnitude of the *beta* value the stronger the relationship between the predictor and criterion variable. Darker bars represent more complex tasks. Lower PS scores represent faster reaction times. Bars that extend past 0.16 (y-axis) were significant at $p=.05$.

It was hypothesized that the objective and subjective indices of mental task load would significantly predict the tasks measured during the flight simulation protocol (iRES, TRP, and SA). It was also hypothesized that the NASA TLX, as a subjective measure, was better at predicting performance where pilots had access to conscious awareness of performance, namely domain-dependent skills of task-relevant performance (TRP) and SA. It was assumed that performances on lower level

cognitive mechanisms, such as processing speed and working memory, were not as likely to be accessible to conscious awareness, and therefore should show less of a relationship to the NASA TLX. This hypothesis was not well-supported, as shown in Figure 4.9 subjective ratings moderately predicted both the domain-independent and domain-dependent cognitive tasks. The effect of working memory was not in the expected direction, because higher subjective ratings of workload predicted better working memory scores. One possible explanation for this result was that higher working memory was associated with more-experienced pilots, who were in turn more likely to acknowledge high-workload conditions. A two-step post hoc analysis investigated this hypothesis and found that DCAT subtest 2 was significantly influenced by pilot rating such that higher rating was associated with better DCAT scores, $F(3, 104) = 2.79, p = .044$. Second, t -tests (undertaken for each rating group) for the repeated NASA TLX scale showed that the lowest-rated student pilots did not tend to rate the high-workload condition as significantly more difficult than the low-workload condition. In contrast, the highest-rated pilots indicated that the high-difficulty was associated with greater mental task load when compared to the low-difficulty condition, but this pattern was not seen for the lowest-rated pilots. This pattern was not attributed to ceiling effects, as the 95% confidence intervals for the mean TLX ratings for the lowest rated group did not reach the ceiling score (maximum score was 10/10), and the mean TLX score for the “overall workload” item was higher for the *more experienced* pilots, in comparison to the pilots with the least expertise. Thus a greater sensitivity to mission difficulty, and higher DCAT subtest 2 scores were associated with the highly-experienced pilots. It was noteworthy that the result of the post hoc analysis suggested that pilot rating was associated with this complex measure of working memory (attentional field of view). A causal relationship

between expertise and working memory cannot be speculated as a result of this relationship. However, in the modeling phase of the present work this “spillover” effect, where more experience causes better working memory, was explored. The association of higher mental task load and better working memory was not observed for the PDT (the objective measure of workload).

It was expected that there should be no differences in the overall pattern of prediction between the younger and older pilot groups. For example, it was hypothesized that the PDT, as a robust measure of mental task load, should predict performance equally well for both younger and older pilots. The introspection required to complete the NASA TLX was also not expected to be a factor impacted by age, and therefore the predictive utility of the NASA TLX should also be impervious to age group differences. For the WM construct this hypothesis was supported in that both measures of mental task load performed similarly for both age groups. Additionally, for both older and younger pilots, the NASA TLX was the strongest predictor of iRES, a measure reflecting the resolution of the information extracted from the environment by pilots during flight. SA was one of the least well-predicted measures by the NASA TLX for both age groups. SA was also poorly predicted by the PDT, but only for the younger pilots. In contrast, the NASA TLX significantly predicted TRP (the flight path maintenance measure), however, this was only found for the older pilots. The general result is similar to the Wiggins (2011) study where a self-rating of overall mental workload on a seven-point rating scale was associated with performance related to altitude deviations. Unlike the results of the current work, the Wiggins’s (2011) sample was considerably younger than the older group of pilots in this analysis.

Other than the moderate prediction of all tasks by the PDT for older pilots, all other relationships showed that the more complex the task, the less likely it was to be significantly predicted by a workload measure. In fact, the PDT was not particularly useful for predicting any of the simulator measures for the younger pilot group.

4.5.3 Results of the Workload Analyses in the Context of Previous Findings

The findings of this analysis show that lower level cognitive tasks, not collected in the simulator, tend to be similarly associated with both workload measures. The more complex flight tasks showed different patterns of association with cognitive workload depending on the type of measure and the pilot age group. The literature pertaining to workload indices and actual performance is also inconsistent regarding the predictive utility of workload measures. Different patterns of association between tasks, age, and type of measurement tool might have been observed due to the differential sensitivity of the measurement tool types. The PDT has been shown to be sensitive to peaks in workload, whereas subjective measures of workload reflect overall task workload (Jahn, Oehme, Krems, & Gelau, 2005; for a review see Ba & Zhang, 2011). The finding that the PDT showed consistent association across all simulator tasks for the older pilots, but did not significantly correlate with these tasks in the younger group suggests that older pilots might have incurred stronger mission difficulty effects than did the younger pilots. A follow-up analysis to investigate the premise that older pilots were more affected by peaks in workload was conducted by examining the *t*-tests for paired scores for simulator tasks between the low- and high-difficulty conditions, for older and younger pilots. These scores reflect the drop in performance that occurred because of the high-difficulty condition. If there was a greater drop in scores for the older pilots this could explain the strong association of the PDT with the simulator

tasks for just the older pilots. Table 4.13 shows that none of the five simulator measures associated with this analysis reveal that older pilots experienced larger difficulty-related reductions in performance. Thus, the sensitivity of the PDT to peaks in workload is not the likely explanation for the better association of the PDT with performance for older pilots. Another explanation involves the metrics of the PDT, with two of the three metrics linked to PDT response times. The ability to achieve high performance on the complex simulator tasks also requires the speedy detection of stimuli (both inside and outside of the cockpit) and the efficient and quick perceptual-motor ability to maintain the desired aircraft trajectory. The PDT and the simulator tasks both rely on similar foundational cognitive abilities (processing speed), which could be mediating performance for both tasks. Such a premise could be tested in subsequent data modeling efforts.

Table 4.13
Mission Difficulty Effects for Complex Simulator Tasks for Older(50+) and Younger(<50) Pilots.

Flight Parameter	(t-test)	
	Younger Pilots	Older Pilots
RMSE Alt Low – RMSE Alt High	-3.318**	-2.915**
RMSE Spd Low – RMSE Spd High	-3.311***	-1.166
SA1 Low - SA1 High	4.965***	4.820***
SA2 Low - SA2 High	5.181***	2.251*
SA3 Low - SA3 High	3.988***	2.367*

Note. *= $p < .05$. **= $p < .01$. ***= $p < .001$. $N=93$. Alt=altitude. Spd =airspeed.

The difficulty in locating the primary causes of workload effects has been noted in the literature. Pilot workload has been studied in the general aviation domain to determine whether pilot workload was affected by the introduction of new technology (Cain, 2007; Vidulich & Tsang, 2012). Oman, Kendra, Hayashi, Stearns, and Burke-Cohen (2001) examined the effect of various navigational devices on simulated approach and landing performance and subjective workload. Oman et al. found that subjective measures of workload using a modified Bedford workload rating scale (Roscoe & Ellis, 1990), completed after the landing task, were similar with and without new displays, and that a mapped vertical navigation display improved performance significantly. On the Bedford scale pilots indicated whether the workload allowed for attentional resources sufficient for completion of the tasks. Oman et al. concluded that the new displays improved performance without increasing workload. Others have argued that unidimensional subjective workload measures, such as the Bedford scale, are actually more indicative of a pilot's perception of the success of the task rather than an actual measure of mental or physical workload (Cain, 2007). This study illustrates the difficulty that researchers have encountered when measuring workload for general aviation tasks.

Another issue that arises with subjective measures of workload pertains to their correspondence with more objective measures of workload. In a review of workload measures in aviation, Cain (2007) suggested that the correspondence between subjective and objective physiological measures of workload has not been shown consistently in the literature. Wilson G. (2002) also found that a variety of physiological markers of workload including heart rate and spectral analysis of electrophysiological data were not consistent with subjective measures of mental workload for both easy and more difficult simulated landing scenarios. In the Wilson

G.(2002) study pilots rated familiar tasks as low in mental task load, whereas physiological markers collected simultaneously indicated higher cognitive load states. In this case, the subjective ratings were suggested to reflect familiarity rather than actual workload.

In summary, the clearest trend revealed in the present study is that iRES, a domain-dependent measure with important implications for pilot performance (see earlier section on SA), was well predicted by the NASA TLX by both older and younger pilots, and by the PDT for older pilots (see Figure 4.9). The mediation of workload effects on aviation outcomes by iRES is an effect that was explored in subsequent models of pilot performance. The significant prediction of TRP by the NASA TLX by older pilots, and the weak but consistent significant prediction of all measures by the PDT for older pilots also suggest that modeling of aviation outcomes should incorporate the non-causal but predictive relationship with workload measures. The observation that the NASA TLX predicted the basic aviator competency of flight path maintenance for older pilots informs efforts to develop screening tools for older pilots who might be at risk during flight.

4.6 Aviation Outcomes

Discussed below are previously published and new analyses regarding critical incidents and runway incursion management (Van Benthem et al., 2012a; Van Benthem & Herdman, 2013; 2014a, in press). The majority of this discussion describes analyses where critical incident counts and runway incursion management were individually regressed on the predictor variables of flight path maintenance, objective and subjective SA, and prospective memory. The results of these analyses guided the data modeling in subsequent phases of this research.

4.6.1 Critical Incidents

The accident rate in general aviation is a concern in the industry and is a frequently reported statistic by national safety boards. General aviation operations result in the most accidents per hour flown, and are responsible for 96% of all reported accidents (GAO, 2012). Critical incidents include accidents, events, or emergencies that could result in collisions, flights into terrain, fire, or engine failures. Critical incidents could also involve occasions where the pilot becomes lost or loses situation awareness, or situations where fuel supply was not managed. Examples of critical incidents in the present work include serious loss of situation awareness (requiring experimenter assistance), controlled flight into terrain, and non-runway touch downs.

Van Benthem et al. (2012b) found that incurring a critical incident was predicted by older age, lower pilot licensure and ratings, and lower working memory performance. Of these factors, expertise was the strongest predictor, followed equally by working memory and age.

Descriptive statistics and distribution of the critical incidents count variable. Although rare in the industry overall, critical incidents are arguably the most important outcome pertaining to general aviation safety. Accordingly, pilot-related critical incidents were considered a key outcome measure in this work. Examining explanatory variables for critical incidents required estimation methods designed for count variables. As shown in Figure 4.10, the probability density function representing the occurrence of critical incidents in the present study demonstrated the typical Poisson distribution with a high zero count (Ramsey & Schafer, 2012). The range of total critical incident occurrences per pilot was zero to five, with 34 out of 108 pilots incurring at least one critical incident. In the low- and high-difficulty conditions the variance, 0.19 and 0.64, was slightly larger than the mean, 0.14 and 0.45. Normally, a

variance greater than the mean indicates overdispersion in the data (McCulloch & Nelder, 1989). Overdispersion is a problem when calculating effects of predictors in Poisson regressions due to “inefficient” standard errors (Ramsey & Schafer, 2012). The overdispersion could result in biased standard errors that are either too large or too small, thus leading to type one or type two errors. The IBM SPSS 22.0 statistical software package used in the present analysis was equipped with methods for dealing with mild overdispersion such as the robust (to incorrect model specification) estimation of the covariance matrix. SPSS 22 also generated a deviance *goodness-of-fit* statistic (deviance value/*df*) and the robust scaled deviance *goodness-of-fit* statistic. Ideally, the deviance statistic, which indicates extra-variation in the Poisson distribution, should be close to the value of 1.0. Ramsey and Schafer (2012, p. 685) suggest examining the deviance *goodness-of-fit* statistic with the results of a regression using the Poisson model with a log-link function where a “rich” combination of predictors is used. Poisson regression tests in the present study examined the deviance statistic. Because the deviance *goodness-of-fit* statistic was not considered large, the negative binomial function was not selected instead of the Poisson function to estimate the effects of predictors of critical incidents. Only if the negative binomial model provided a better fit to the data (by way of a lower deviance *goodness-of-fit* statistic and lower information criteria value [Bayesian Information Criteria]), then would a negative binomial model be considered for parameter estimation (Ramsey & Schafer, 2012).

Figure 4.10. Total critical incident counts for low- and high-difficulty conditions.

4.6.2 Surprise Runway Incursion Management

Runway incursions are a priority safety issue facing the aviation industry. According to the Transportation Safety Board of Canada (TSB, 2012) the risk of collisions on runways is one of four main safety threats in Canadian aviation. Not unlike other accident statistics, general aviation pilots are more likely to be involved in a runway incursion event when compared with pilots flying in commercial operations. The FAA's Annual Runway Safety Report (2010) noted that general aviation pilots were implicated in 77% of incursions caused by pilot deviations. The nature of most general aviation flights could be contributing to the greater risk for general aviation pilots. For example, many general aviation flights are single pilot operations and often occur at uncontrolled aerodromes where the responsibility for detecting incursions, rests solely with the approaching pilot. Thus pilot factors have been studied, albeit somewhat indirectly, in research investigating factors predicting performance when encountering surprise events.

Wickens and Long (1995) introduced a surprise incursion at the end of a flight simulation study to examine the effects of conformal versus non-conformal symbology in head-up and head-down displays. Subjects in this study included 14 pilots with at least a private pilot rating and the remaining 18 pilots were rated at the certified flight instructor or commercial (including airline) rating or higher. Detection latency for the incursion (a wide-body jetliner) was most adversely affected by the head-up display. The phenomenon of “cognitive tunneling” was suggested as the mechanism that interfered with detecting the incursion stimulus (Wickens & Long, 1995). Cognitive tunneling acts to limit out-the-window registration of stimuli due to the focal location of the information on the windscreen and the associated locking of attention onto the head-up display (Jarmasz, Herdman, & Johannsdottir, 2005). Individual pilot factors have been shown to influence the number of critical events detected during simulated flight. While not exploring runway incursions per se, Coffey et al. (2007) found that when compared to younger pilots (24.4 years, $n = 7$), older pilots (64.7 years, $n = 7$) missed more unexpected events both inside (cockpit gauge “freezing”) and outside the cockpit (nearby traffic at the same altitude).

The flight simulation protocol for this dissertation included an unexpected runway incursion event that occurred on final approach at the high-difficulty aerodrome. To be considered a true surprise event, the incursion scenario occurred only once, and at the end of the last circuit (so as not to impact subsequent performance). The runway incursion management score quantified overall performance with respect to the handling the incursion, and the outcome of the surprise event. Pilots received a score (from 0 to 10) that referenced the timing, quality, and outcome of the runway incursion task.

A previously-published analysis from this work found that incursion management was predicted by pilot rating, self-rated mental workload and situation awareness, and an interaction between age and the DCAT 6. Together these factors accounted for approximately 29% of the variance in runway incursion management, $F(6, 86) = 5.87, p < .001$. The size and direction of the standardized *beta* coefficients indicated that incursion management scores increased for pilots who correctly estimated the high level of mental demands in the high-difficulty condition and who rated their own situation awareness for the impact of other aircraft on their ability to fly the “perfect” pattern as high. In addition, pilots with higher levels of pilot rating tended to have better runway incursion management scores. The age by DCAT 6 interaction was interpreted such that older, but not younger, pilots with lower DCAT 6 scores tended to demonstrate poorer incursion management.

Descriptive statistics and distribution of incursion management scores.

Figure 4.11 illustrates the distribution of incursion management scores for each pilot rating group. It appeared that scores for pilots in the two middle-rated groups followed a bimodal distribution. Two clearly different patterns of results were associated with the lowest- (student) and the highest-rated pilots (e.g., commercial pilots). The low-expertise pilots have scores stacked at the lower end of the score range, whereas high expertise pilots have scores primarily at the high score range.

Figure 4.11. Distribution of incursion management scores for each pilot rating group. VFR= visual flight rules. PPL with Addn'l Ratings= private pilot license with additional ratings. IFR= instrument flight rules. ATP= airline transport pilot. Note the bimodal distribution found for the largest pilot subgroup (VFR/PPL/Recreation).

A *chi-square* analysis of pilot rating group by incursion management score was significant, $\chi^2 (30, N=93)= 48.17, p=.019$. The significant *chi-square* appears to result from the lowest rating group where scores did not extend beyond five out of 10, and from the highest rating group where a greater number than expected perfect scores and zero scores were found. To determine whether the apparent bimodal distribution in the second lowest rating group (recreational pilots and private pilots without additional ratings) was a result of an underlying categorical explanation, such as age group or recent pilot-in-command hours, separate *chi-square* analyses were conducted. The first analysis explored the effect of older age (two groups split at age 50 years) on incursion management scores. The second analysis explored the effect of low, medium, and high recent pilot-in-command hours (by equal group size split) on incursion management scores. Both *chi-square* analyses showed that neither age nor recent pilot-in-command hours significantly predicted incursion management scores

for the second to lowest rating group, $p > 0.1$. Thus, pilot characteristics might not be responsible for the bimodal distribution of incursion management scores in the recreational and non-rated private pilot. Based on the preliminary *chi-square* analyses it was determined that further investigation of explanatory variables of incursion management would include the sample of 93 pilots, of all rating and licensure groups, who had completed the subjective rating scales and the DCAT subtests.

A final check to ensure the appropriateness of linear regression with the incursion management scores as a dependent variable involved the inspection of standardized residuals from a simple linear regression model with pilot rating as the predictor variable. The standardized residual values were between -2.5. and 1.8 and were approximately normally dispersed around the mean of zero. The normal distribution of the residual values from the linear regression suggested that the standard errors generated by the model were not be inflated, thereby providing validity of the significance of explanatory effects based on p -values.

The section below describes previously published and new analyses that explore the predictors of both critical incidents and incursion management. Earlier analyses in the present work have shown that pilot characteristics can significantly influence flight path maintenance, situation awareness, and prospective memory for communication tasks. In the following section, the effects of these domain-dependent cognitive abilities on the two criterion variables in the present work are explored.

4.6.3 Domain-Dependent Predictors of Critical Incidents and Incursion Management

Flight path maintenance as a predictor of aviation outcomes. It is not understood how a pilot's tendency to deviate from the perfect altitude and airspeed

translates meaningfully to pilot risk for a critical incident. Common sense suggests that the further a pilot flies from their ideal position the greater the likelihood of an accident. To investigate the relationship between flight path maintenance and aviation outcomes a secondary analysis was completed to investigate how deviations in altitude and airspeed influenced the number of critical incidents incurred by pilots. Because it was important to understand how a pilot's basic flight path maintenance performance under normal conditions influenced the number of critical events, performance factor scores were produced using altitude and airspeed deviation measures from the low-workload condition. Factor scores were created using principal component analysis. Scores were normalized and truncated to +/- 3 standard deviations from the mean. A reliable method of determining the effect of a continuous or ordinal predictor on a discrete dependent variable, such as the count of critical incidents incurred, is to use the Poisson regression (Cameron & Trivedi, 2013). The Poisson regression model provided exponentiated estimates of the log odds associated with each predictor. The number of critical incidents was regressed on flight path maintenance factor scores (with pilot licensure and rating as a control variable). The omnibus *chi-square* test was significant, $\chi^2 (2, N=93)=26.12, p<.001$. The Wald *chi-square* for flight path maintenance was significant, $\chi^2 (1, N=93)=5.73, p=.017$. The exponentiated *beta* coefficient for flight path maintenance was 1.45, indicating that an increase of one standard deviation in flight path maintenance resulted in a 45% increase in the likelihood of incurring an additional (or one) critical incident.

A similar procedure was followed with the surprise incursion management outcome variable using ordinary least squares regression. Because the incursion scenario always occurred during the high-difficulty condition a flight path

maintenance factor score was generated for each subject from the linear combination of RMSE values for altitude, heading, and airspeed deviations from the high-difficulty condition. The flight path maintenance factor score was not significantly correlated with incursion management. Similarly, high-difficulty flight path maintenance factor scores did not account for a significant amount of variance in a regression model with incursion management as the dependent variable, ($p>.1$).

Situation awareness as a predictor of aviation outcomes. SA might also function as a domain-dependent predictor of aviation outcomes (as posited by Hardy & Parasuraman, 1997). In this analysis, critical incidents and incursion management were regressed on iRES, and SA levels 2 and 3 (pilot licensure and rating included as a control variable). Table 4.14 shows the results for the Poisson regression model with critical incidents as the criterion variable. The model produced for critical incidents was significant, $\chi^2 = 43.414$, $df=4$, $p<.001$. Note that the SA level 3 contribution to the model was not significant, Wald *chi-square* <1.0. The exponentiated *beta* values may be interpreted such that for each 1% increase in situation awareness accuracy for iRES and SA level 2 there is a corresponding 2% decrease in the likelihood of incurring a critical incident (2.2% decrease for iRES). A regression model for incursion management as the criterion variable was tested using ordinary least squares regression and was not significant, $F<1.00$, with neither iRES, SA level 2 or 3 contributing significantly to the model.

Table 4.14
Critical Incidents Regressed on SA level 1 (iRES), 2, and 3.

	Wald <i>chi-square</i>	<i>df</i>	<i>p</i> - value	<i>Exp (B)</i>
Critical Incidents				
(Intercept)	24.590	1	.000	24.280
SA Level 1 (iRES)	5.193	1	.023	0.978
SA Level 2	8.710	1	.003	0.980
SA Level 3	0.765	1	.382	0.993

Note. *df*= degrees of freedom. *Exp(B)*= exponentiated *beta* coefficient.

In models of aviation outcomes subjective ratings can have predictive, but not causal roles. The SA self-rating scale used in this work (Hou et al., 2007) would have additional utility if it also predicted performance on the two aviation outcomes. Critical incident rates and incursion management were regressed on subjective SA scale factor scores. The standard regression model produced for incursion management was non-significant, $F < 1.10$. The Poisson model produced for critical incident rates was also not significant, Wald *chi-square* < 1.0 . The utility of a subjective measure of SA in predicting aviation performance was found by Snow and Reising (2000) who used the SA-SWORD (Vidulich & Hughes, 1991) measure and compared results to objective measures of SA. In their sample of 12 military pilots higher subjective ratings of SA were associated with particular cockpit display types. The same association was not found for objective measures of SA collected using the SAGAT technique. Vidulich and Hughes raised a concern with subjective measures in general, which have been criticized as reflecting overall performance in tasks, rather than actual SA (Endsley, 1995).

Prospective memory and aviation outcomes. In keeping with the initial results that age, cognition, and cue salience were strong predictors of prospective memory, it was predicted that the number of critical incidents incurred by a pilot would show

stronger links to prospective memory for older as compared to younger pilots. Prospective memory has been implicated in serious aviation accidents involving pilots with various levels of expertise (Dismukes & Nowinski, 2007). Earlier in this work, a significant effect of memory cue salience was shown for prospective memory performance. However, the Pearson correlation between the low-salience radio call completion rates and total critical incidents ($-0.39, p < .001$) was similar to the correlation between high-salience cue radio call completion rates and total critical incidents ($-0.34, p < .001$). Therefore, the total number of completed radio calls was used in a Poisson regression analysis where the count of total critical incidents was regressed on the combined number of calls completed in the low- and high-difficulty conditions (maximum score = 36). Poisson regression was used in place of linear regression due to the “count” distribution of the critical incident variable (Cameron & Trivedi, 2013). Because older pilot age was associated with a greater number of prospective memory failures separate analyses examining the influence of prospective memory on critical incidents were undertaken for younger (less than 50 years, $n = 66$) and older pilots (50+ years, $n = 42$). This method tested the notion that prospective memory ability significantly predicted critical incidents for older pilots to a greater degree than for younger pilots. Table 4.15 shows the total number of critical incidents regressed on prospective memory score for older and younger pilot groups. The model suggests that for an increase of one percent radio call performance there is a corresponding 15% decrease in the likelihood of a critical incident occurring for pilots in the younger group, but only an 8% decrease in likelihood for older pilots. The omnibus test was significant for younger, $\chi^2 = 9.14, df = 1, p = .003$ and older pilots, $\chi^2 = 11.00, df = 1, p = .001$. These results are contrary to the hypothesis that older pilot critical incident rates would have a stronger association to prospective memory, when

compared to younger pilots. These findings highlight the importance of prospective memory for both older and younger pilots.

Table 4.15

Critical Incidents Regressed on Prospective Memory Score for Older and Younger Pilot Groups.

Model	Wald <i>chi-square</i>	<i>df</i>	<i>p-value</i>	<i>Exp(B)</i>
Younger Pilots (<i>n</i> =66)				
(Intercept)	3.648	1	.056	79.940
Prospective Memory	5.156	1	.023	0.845
Older Pilots (<i>n</i> =42)				
(Intercept)	3.496	1	.061	8.378
Prospective Memory	4.830	1	.028	0.919

Note. *df*= degrees of freedom. *Exp(B)*= exponentiated *beta* coefficient. Poisson regression model used for critical incidents (a count variable).

The relationship between prospective memory and incursion management is not an association that might be normally hypothesized. Rather, the quality of the response of pilots to sudden events should draw upon procedures that are honed with experience (Van Benthem & Herdman, 2013; Van Benthem & Herdman., 2014b). It was possible that the detection of a runway incursion could be associated with prospective memory for radio calls that relied on highly salient memory cues. Detecting cues for both runway incursions and for completing radio calls are arguably both reliant on the salience of the cues available in the environment. Therefore, a model for incursion management was tested that included separate prospective memory scores for the low- and high-salience cues, and pilot licensure and rating. The model with three predictor variables was significant, $F(3, 104)= 8.95$, $p<.001$. As expected, pilot expertise was associated with incursion management, $t= 3.89$, $p<.001$, $r^2= 0.10$. Performance for low-salience memory cue radio call completion rates was not a significant predictor of incursion management, $r^2<0.02$. On the other hand,

performance on the high-salience memory cue radio calls did significantly predict incursion management, $t= 2.60$, $p=.01$, $r^2=0.09$. The association between prospective memory, and critical incident rates and incursion management has important implications in models of aviation outcomes.

4.7 Summary of the Results of the Experimental Protocol

The experimental phase of this dissertation explored the effects of pilot age and expertise, and domain-independent cognitive variables, on flight path maintenance, situation awareness, and prospective memory. When possible, the effect of cognitive load represented by low- or high-difficulty conditions was considered in addition to the predictors of flight performance. The aforementioned variables were also examined in regard to their association with two aviation outcomes: critical incidents and incursion management. The results were instrumental in developing the mental model for pilots and for building a comprehensive and theory-driven model of pilot performance.

Hardy and Parasuraman (1997) and Herdman and LeFevre (2004) predicted that a variety of pilot attributes and stressors (mission difficulty and high mental task load) would impact domain-dependent cognitive variables. Figure 1.1 illustrates the hypothesized causal path of pilot and environmental effects on cognitive and aviation performance (Hardy & Parasuraman, 1997). The experimental phase of this work demonstrated that pilot expertise and age were relevant explanatory variables of domain-dependent factors of pilot performance. The ability to maintain appropriate altitude, heading, and airspeed were directly influenced by pilot age and expertise. For example, greater deviations from the perfect position during the downwind leg of the circuit were associated with older age, lower levels of expertise, and with the higher-difficulty mission condition. The mission difficulty level and pilot attributes were

found to similarly influence situation awareness and measures of prospective memory. While pilot age was implicated in performance, results from multiple regression models found that cognitive factors such as working memory, visual tracking, and processing speed were, at times, better markers of pilot performance than age alone was. In sum, these findings suggested that predicting aviation outcomes requires a broad array of explanatory variables, and a statistical method that supports simultaneous determination of both direct and indirect effect of predictors on downstream variables.

4.8 Remaining Analyses in the Present Work

As per the goals of this dissertation, before specifying a comprehensive model of pilot performance, it was necessary to develop and validate a mental model for pilots. The dynamic mental model (DMM) was expected to operate as a direct influence, and a mediator of upstream variables, on aviation outcomes. Herdman and LeFevre (2004) predicted that situation awareness, cognitive workload, and task-related performance (flight path deviation in this work), were indices of a mental model for pilots. Hardy and Parasuraman (1997) suggested that further developments in predicting pilot performance were reliant on understanding the role of the mental model for pilots. Chapter 5 considers theoretical concepts relevant to the operationalization and specification of a mental model for pilots. Chapter 6 describes the process and results of indexing a mental model for pilots, and testing its structural and theoretical validity.

In regards to specifying a comprehensive model of pilot performance, the evidence from the research just discussed supports the hypothesis that domain-dependent cognitive factors predict critical incidents. Incursion management, the second aviation outcome in this work, was also predicted by expertise and prospective memory: thus confirming the relevance of upstream factors outlined in the model of

pilot performance by Hardy and Parasuraman (1997). In Chapter 7 the validated mental model for pilots is posited as the central variable in a comprehensive model of pilot performance. The utility of this model to predict critical incidents and incursion management was tested in two main analyses. The model is also used to investigate how upstream effects are transmitted via direct and indirect paths to influence outcome variables. In particular the mental model for pilots is posited as a central variable with significant potential to mediate effects from pilot attributes and domain-independent cognitive factors on critical incidents and incursion management.

Three main hypotheses guided the remainder of the present work:

1. The four domain-dependent cognitive factors of SA, TRP, TM, and PM converge in a multilevel structure to represent the dynamic mental model (DMM) for pilots.
2. The DMM, as a domain-dependent variable, accounts for significantly more variance in the aviation outcomes measures than the variance accounted for by domain-independent cognitive factors.
3. The DMM fully mediates the effects of cognition, age, expertise, and workload factors on aviation outcomes.

Chapter 5

Theoretical Development of the Dynamic Mental Model for Pilots

As described in Chapters 1 and 2, older and less experienced pilots show greater risk for critical incidents compared to younger or more experienced pilots. Investigating sources of pilot risk within complex systems requires a framework that accounts for human, environmental, task, and technological factors. This framework should also support the explicit modeling of interactions between system factors. Within the aviation domain, human performance evaluation is guided by theory that integrates facets of the system such as task load, basic aviator competencies, and pilot attributes.

Earlier work examining factors associated with pilot performance might have been disadvantaged by a lack of emphasis on domain-dependent cognitive factors. This dissertation makes explicit use of a domain-dependent cognitive construct, the dynamic mental model (Herdman & LeFevre, 2004), to generate and test hypotheses related to general aviation flight performance. The results are used to extend our understanding of how age, cognition, expertise, and workload factors are conveyed through complex human-machine structures.

Hollnagel and Woods (1983; 2005) emphasized the “joint cognitive” nature of human-machine systems and called for a cognitive systems engineering (CSE) framework to modeling performance (a description of the CSE approach is provided in the following section). One such framework, a causal model of pilot performance by Hardy and Parasuraman (1997), highlighted environmental, system, and pilot factors as predictors of flight performance and aviation outcomes. See Chapter 1 for a description of the causal model of pilot performance by Hardy and Parasuraman. While foundational cognitive functions such as working memory and processing

speed have a longstanding place within pilot human factors research, it was the *mental model for pilots* that was emphasized as the central cognitive factor in the model suggested by Hardy and Parasuraman. It was suggested that the specification of the mental model for pilots was a complex but necessary endeavour if we were to extend our capacity to predict effects of individual differences upon performance and aviation outcomes. Herdman and LeFevre (2004) took up this challenge and followed Hollnagel and Woods (1983) by employing a CSE framework to specify a cognition-focused *dynamic* mental model for pilots. The dynamic mental model (DMM) was proposed as a novel criterion variable for quantifying the effects of various display configurations for multi-operator military-like missions. As shown in Figure 5.1, the DMM reflected important components of the pilots' mental representation of the ongoing tasks and situation. While the DMM was designed with military pilots in mind, its elements are generic and appear appropriate for utilization with a sample of general aviation pilots. The DMM, as outlined by Herdman and LeFevre, represents the first attempt at detailing the indices of a mental model for pilots and testing its responsiveness to changes in the aviation system. The present work expanded the conceptual basis for the selection of variables designed to index a mental model for pilots.

To address the goal of locating sources of risk for older pilots, the present work developed a model of pilot performance based on the work of Hardy and Parasuraman (1997) and Herdman and LeFevre (2004). The overarching structure of the model of pilot performance was guided by the causal model outlined by Hardy and Parasuraman. The indexing of the mental model for pilots (the DMM) was guided by the work of Herdman and LeFevre (2004). The present work reconstructed the DMM and situated it centrally in a comprehensive model of aviation performance.

5.1 A Cognitive Systems Engineering Framework

The development of a comprehensive model of pilot performance and aviation outcomes, and the specification of a mental model for pilots was undertaken using a CSE approach. According to Hollnagel and Woods (1983), a CSE approach orients the modeling of performance in human-machine systems towards relevant cognitive mechanisms. The CSE framework provides a cognition-focused and holistic approach to understanding and designing for the human in complex and often highly automated systems (Hollnagel & Woods, 1983). One function of a CSE framework is to enable efficient and safe design of “joint cognitive systems” that reflect and respect the limitations and strengths of both the human and technical features of the system (Hollnagel & Woods, 2005; Woods & Hollnagel, 2006). In their original treatise on the man-machine system, Hollnagel and Woods emphasized that any interaction between an operator and technology was characterized by *psycho-logic* behaviour on the part of the human (1983, p. 586). Psycho-logic meaning that human response was constrained by the limits and nature of the cognitive processes involved in the tasks. On this view, efforts towards predicting pilots who might be at risk should ensure that multiple aspects of cognition are considered and, as suggested by Hardy and Parasuraman, include indices of the pilot’s mental model. The CSE framework is both a guide for constructing the DMM, and for testing the DMM’s role in a cognition-oriented model of pilot performance.

5.2 A Mental Model for Pilots

Hardy and Parasuraman (1997) and Herdman and LeFevre (2004) indicated that explanations of aviation outcomes could be expedited by the inclusion of a variable that represented a mental model for pilots. Before a mental model for pilots can be included in a comprehensive model of pilot performance, the mental model must be

defined and *specified*. However, the literature pertaining to mental models is very diverse and definitions vary greatly from one domain to the next. The following discussion reviews some of the key descriptions of mental models from various fields.

5.2.1 Broad Definitions of the Mental Model

An oft-quoted description of a mental model comes from Rouse and Morris (1985, p.7) who stated that mental models are the “mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states.” Rasmussen (1979) suggested that a general definition beyond that of Rouse and Morris was futile because each domain and system to which a mental model was linked was rife with its own taxonomy. Despite their differences, descriptions of mental models from various domains illuminate our understanding of a *mental model* concept, and explore how such a concept might play a role in predicting aviation outcomes. Characterizations of mental models are found in domains that pertain to reasoning and deductive processes, interactions with complex systems, and to the broad arena of learning (particularly in relation to reading comprehension).

The necessity of a mental model for reasoning and deduction for any given problem space has been outlined. Logician Charles Sanders Peirce (1931–1958, posthumously), and philosopher and cognitive psychologist Kenneth Craik (1943, posthumously) wrote of the importance of mental models as the mind’s tool for reasoning, and for understanding the world (Johnson-Laird, 2004). A mental model reflects what is believed about the world, although, not necessarily with a one-to-one mapping between entities and model elements (Johnson-Laird, 2004). The mental model provides the structure for information about the problem and links this

information with data structures from schema in long-term memory (Holland, Holyoak, Nisbett, & Thagard, 1986). Holland et al. (1986, p. 12) differentiate mental models from schemata (or scripts). Schemata are responsible for “chunking” together relevant bits of information. The mental model, on the other hand, is a dynamic structure that represents information pertaining to rules or procedures, guides activity in the problem space, and permits simulation of information used primarily to test the logic and accuracy of deductions (see also Wilson J. & Rutherford, 1989).

Human factors work has posited mental models as mental structures generated by users for understanding and interacting with a human-machine system. Mental models reflect the users', or teams of users', understanding of the purpose, function, and relations between elements in the system (Converse, 1993; Kieras & Bovair, 1984; Norman, 1983; Wilson J. & Rutherford, 1989). Human factors-based design explicitly acknowledges the users' mental model of *how to operate* the system, and *what the technology does*. The user interface must convey this information efficiently and accurately for an expedient mental model to be constructed by the user. Accurate and efficient mental models enable and support learning and the efficacious use of technology in complex systems (Cooper, Reimann, & Cronin, 2007). A mental model is a functional requirement for understanding cognitive systems because the human mind is constrained by working memory processes, which limit the amount of information manipulated consciously at any one time (Rabardel, 1995). This chunking of system information is posited to support three aspects of the human-machine system: Jonassen and Tessmer (1996 in Sheehan & Tessmer [1997]) suggested that “mental models contained three kinds of interconnected knowledge: knowing that (declarative), knowing how (skills), and knowing why (causal principles or functions). Conditional knowledge (knowing when) may also be a part of a mental model (p. 19).”

On this view, the ability to actually respond by using the procedural knowledge was not considered essential to the definition of a mental model.

The importance of users' mental models was shown by Kieras and Bovair (1984) in three experiments that demonstrated that overall use of a device and inferring the procedures required for use benefitted from learning a "device model" as opposed to learning rote procedures. The mental model group, who understood the device model, but had not received training of rote procedures, were faster and more efficient (using shortcuts) when handling the standard of off-nominal task requirements. Similarly, Zhang and Wickens (1987) demonstrated that device models that highlighted the functions and interrelations of key system parts, even those composed using fairly sparse graphical representations, were effective in improving performance, as compared to groups who were not exposed to the device model aid. Response times and detection of failures in the system were not influenced by access to the device model, whereas diagnostic accuracy was positively associated with the (assumed) generation of a functional mental model of the system.

The role of mental models has also been illustrated in the field of education, Lev Vygotsky and Jerome Bruner have described mental models as structures utilized by children during development and learning (see Lutz & Huitt, 2004). According to these constructivist theorists, the acquisition of information via the building of mental models about the world and the relations between contents of the world is by definition what *learning* is, and, at the same time, is the process that enables learning to occur. Some have investigated how comprehension benefits from adequate spatial mental models that are generated during reading (Brunyé & Taylor, 2008). The generation of accurate spatial mental models was hampered by secondary visual-spatial tasks (during the reading task), which lead to reduced performance (in

comparison to controls who did not complete the interference visual-spatial tasks while reading) for reporting locations related to the narrative. Brunyé and Taylor (2008) demonstrated that spatial mental models formed during reading are subject to working memory constraints: thus suggesting that mental model generation may rely upon lower level supportive cognitive functions.

A common characteristic of mental models, regardless of whether the model is applied to reasoning, interfacing with a system, or for learning, is their parsimonious reflection of the world. According to Besnard, Greathead, and Baxter (2004, p. 119), mental models are “simplified, cognitively acceptable versions of a too complex reality”. On this view, mental models are built from the procedural knowledge relevant to pursuing goals and from minimal bits of information from the current situation. Thus, the mental model is important because of its functionality, rather than its completeness (Besnard, Greathead, & Baxter, 2004; Craik, 1943; Jones, Ross, Lynam, Perez, & Leitch, 2011). Doyle, Ford, Radzicki, and Trees (2001, pp. 6–8) have studied mental models in dynamic systems and strongly put forward the notion that mental models are more than incomplete: they are “vastly oversimplified....messy... sloppy....and unstable over time”. Due to the nature of human memory systems mental models also have boundaries that are “fuzzy... and ill-defined”. According to Doyle et al. (2001) mental models are deficient in running simulations with more than a few variables, are prone to error, and are poor modelers of non-linear relationships. Despite these limitations, it is unlikely that humans could operate in the world as masterfully as we do without mental models. An early descriptor of mental models comes from Rasmussen (1979, p. 8) who stated that “The efficiency of humans in coping with complexity is largely due to the availability of a large repertoire of different internal models of the environment from which rules to control behaviour

can be generated ad hoc. An analysis of the form of these internal mental models is important to the study of human interaction with complex man-made systems.”

5.2.2 Aviation-Related Definitions of Mental Models

Previous work has explored features of mental models for the aviation domain. The mental model for pilots (the DMM) espoused by Herdman and LeFevre (2004) operationalized three indices of performance sensitive to environmental and pilot factors. According to Herdman and LeFevre the DMM is a *latent* cognitive construct, which entails that the construct must be measured indirectly through its indices. As shown in Figure 5.1, the dynamic mental model was designed as a latent construct indexed by situation awareness (SA), task-related performance (TRP), and cognitive workload measures. Situation awareness, TRP, and workload are also latent constructs that are indexed via observed variables (discussed in the section pertaining to indexing the revised DMM). The indexing of the DMM with SA, TRP, and cognitive workload measures was anticipated to “capture a large portion of the variance in human-machine systems” (Herdman & LeFevre, 2004, p. 25). Specifically, the DMM served as a structure for evaluating the effects of new technology in the aviation system. Herdman and LeFevre characterized the DMM as multi-dimensional, incorporating mechanisms such as selective attention, long-term memory, and perceptual-motor coordination.

Individual components of the DMM have shown sensitivity to the manipulation of conditions within a simulated flight environment. Herdman and LeFevre reported on three experiments where metrics of situation awareness and workload were sensitive to experimental manipulations regarding heads-up versus head down displays; for advanced optical sensors and digital moving maps; and for conditions that included additional crew with advanced optical sensors.

Figure 5.1. Cognitive systems engineering (CSE) framework. From Herdman, C. M., & LeFevre, J. (2004). *Tactical Aviation Mission System Simulation Situational Awareness Project*. (No. DRDC-CR-2007-002). Defence Research and Development Canada. Carleton University, Ottawa (Ontario). DTIC Document.

Hardy and Parasuraman (1997) also described features of a mental model for pilots. The pilot mental model represented procedural knowledge and the capacity to implement cognitive strategies at appropriate times. On this vein, the mental model for aviation developed slowly and was fine-tuned in accord with expertise. Hardy and Parasuraman suggested that SA, another model-like mechanism, was separate from the pilot mental model because SA reflected a dynamic aspect of aviation not considered (by those authors) as part of a pilot mental model. That mental models change slowly and do not directly include the dynamic aspect of SA is also echoed by Endsley (1995), who has extensively studied SA in aviation settings. The difference between pilot mental models that are associated with long-term memory facilities, and dynamic features of situation models produced by SA mechanisms has been stressed (Wickens, 2008). The present work took a different approach and incorporated both

the dynamism and knowledge structures associated with SA as integral to a pilot mental model that supports basic aviator competencies.

In the aviation domain, pilot mental models have been examined with respect to the impact of glass cockpits (Sarter & Woods, 1991, 1992, 1994), the automation of information systems (Sarter, Mumaw, & Wickens, 2007), and regarding scanning behaviour in the cockpit (Bellenkes, Wickens, & Kramer, 1997). Sarter and Woods (1992, 1994) and Sarter et al. (2007) depict the pilot mental model from the *user interface* perspective, where indexing of the mental model was accomplished by asking pilots knowledge-related questions about the flight management system after the experimental protocol. Simulated flight performance was associated with this operationalization of the pilot mental model: specifically, gaps in knowledge pertaining to the system were associated with poorer flight parameters and more errors during the flight protocol. On the other hand, Bellenkes et al. (1997, p. 557) indexed the “pilot mental model of the flight dynamics” using cockpit scanning data gathered from a head-mounted eye-tracker during simulated flight. The dwell-time and frequency of “visits” to cockpit indicators were considered in relation to deviation from the ideal position of the aircraft. Results demonstrated that more expert pilots flew with greater precision with regards to altitude and heading, with corresponding shorter dwell-times but increased frequency of scans to cockpit indicators. A more “refined mental model” was attributed to the expert pilots who were assumed to gather more information in a shorter amount of time, and “cross-couple” information such that a quick gaze on one indicator provided expert pilots with information relating to another indicator (without the gaze to the second indicator) (Bellenkes et al., 1997).

A simulated cognitive architecture implemented a mental model for pilots called the *internal situation model* (Karikawa, Takahashi, Ishibashi, Wakabayashi, &

Kitamura, 2006). In accord with Karikawa et al., the simulated mental model is composed primarily of information from the flight environment and includes details regarding terrain, other aircraft, ownship, and operations by the “pilot”. The internal situation model was influenced by cognitive agents related to the perception of visual and auditory information from the environment. Enhanced outcomes were observed with more informative cockpit displays, which were also thought to enhance the pilot mental model (Karikawa et al., 2006).

Histon and Hansman (2008) describe “working mental models” for air traffic controllers (ATC) that are a central to ATC cognitive processes. Working mental models are dynamic, task, and situation specific, and play a role in predicting future states of the system. According to Histon and Hansman (2008, p.71), mental models support the “generation and maintenance of situation awareness as well as various decision-making and implementation processes.” Figure 5.2 locates the working mental model in the larger ATC cognitive system (Histon & Hansman, 2008). The mental model is considered the “working” one, when it has been selected from a “library” of working models that the user has access to. Histon and Hansman (2008, p. 72) focus on the “snapshot” in time aspect of mental models and report on experiments (mainly non-ATC participants) designed to operationalize mental models. Outcomes included the time to detect conflicts and false alarms. The complexity associated with the ATC scenarios, described as the “degrees of freedom” were associated with poorer outcomes, and, as per their definition of working mental models, outcomes appeared to closely reflect a situation awareness construct (Histon & Hansman, 2008). As indicated in Figure 5.2, the working mental model as outlined by Histon and Hansman (2008) directly influences decision processes. Other than “long-term memory store” all

other potential predictors of the outcome variable (decision processes) could be mediated by the mental model scores.

Figure 5.2. Histon and Hansman (2008) working mental model central to ATC cognitive processes. From Histon, J. M., & Hansman, R. J. (2008). Mitigating Complexity in Air Traffic Control: The Role of Structure-Based Abstractions. Retrieved from <http://18.7.29.232/handle/1721.1/42006>

Some aspects of the Histon and Hansman (2008) model are similar to a mental model portrayed by Endsley (1995). As shown in Figure 5.3, Endsley also envisions a pilot having access to many mental models. These mental models for aviators are structures on which information pertaining to situation models, goals, procedural scripts, and plans are organized. A mental model is updated by SA and coordinates plans, tasks, and procedures. The activity of the mental model is guided by the overall goals of the mission. The mental model as described by Endsley is discussed in more detail in the section below on SA.

Figure 5.3. Relationship of mental models to situation awareness. Adapted from Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 32-64.

In summary, regardless of the domain, a mental model has consistently been described as the representation of relations between entities in the real world, and includes knowledge structures and situation models. The mental model has been shown to be dynamic in nature, as it appears to be influenced by pilot expertise and features of the flight environment (e.g., cockpit displays). The present work assumes that the evolution from a simple to a complex mental model is imperative for the advancement of novice to expert operator in aviation systems.

5.3 Specifying a Mental Model for Pilots

While the kinds of tasks pilot engage in vary according to expertise and flight purpose, it was a *prima facie* understanding of the present work that all pilots require a basic mental model for aviation to support competent performance. The focus of the present work was to index and test a mental model for pilots that supported basic aviator competencies for general aviation flight environments.

In accord with the description by Rouse and Morris (1985), a mental model for pilots might reflect a description of the purpose and form of the aviation system, and provide explanations regarding the functions and state of the system. In line with a more human-factors' description of "device" or system models, a mental model for pilots might also incorporate a mechanism for predicting the state of the system in the near future and allow for mental simulations of potential future states. From higher to lower level features, a mental model for pilots represents *a knowledge and capacity network* that integrates:

- mission goals;
- knowledge and skills regarding how and when to engage in tasks and maneuvers;
- current and predicted future status of ownship, cockpit controls, and gauges;
- and a dynamic situation model that includes (but is not limited to) terrain or other aircraft positions relative to ownship.

The DMM is a mental model implemented in the mind of the human operator, which supports aviator activities. In the present work, the conceptual architecture of the DMM was based on current theories regarding neurocognitive networks. These theories speculate that complex cognitive functions result from distributed yet interconnected and mutually-constraining brain areas. Likewise, the idea that we fully

apprehend cognition by focusing on singular mechanisms is giving way to the understanding that exploring networks of activity yields a better appreciation of how cognition is realized in the mind (Tognoli & Kelso, 2014). In the same way that different brain regions might network functionally to produce other complex behaviour, the networks associated with the DMM integrate and support each other to mediate the dynamic mental model for pilots.

Herdman and LeFevre (2004) formulated a pilot mental model as a latent structure that was indexed solely by cognitive factors. The present work builds upon the original configuration of the DMM (see Figure 5.1). In this work, SA and TRP were retained as elements of the DMM. A brief discussion regarding the removal of the workload construct follows here.

5.3.1 Workload Constructs in the Comprehensive Model of Pilot Performance

Going forward, situation awareness (SA) and task-related performance (TRP) were retained as components of the DMM, whereas workload was not retained as an index of the DMM. Although the construct of workload remained a predictor of performance in the model of aviation performance, workload was not conceptualized as a construct that contributed to a network of mechanisms comprising a mental model. The DMM is a cognitive construct, therefore each element of the DMM must also be a cognitive mechanism, but workload, itself is not a mechanism. Despite this, the construct of workload is relevant to aviation performance (Cain, 2007), and therefore was incorporated in the comprehensive model of pilot performance. First, mission difficulty was a predictor of flight performance and a “control” variable. This conception of workload is similar to the model by Hardy and Parasuraman (1997), where workload effects were classified as pilot stressors. The control aspect of

workload was relevant to the validation of the model of pilot performance because relationships between variables in the model should be stable, in either low- or high-workload conditions. Second, mental task load, as measured by an auditory perception task (PDT) performance was included as a domain-independent cognitive variable. The latent variable reflecting mental task load was conceptualized in a manner similar to Tsang and Shaner (1998), where PDT performance represented “reserve” mental resources, available for managing the secondary PDT task while engaged in the primary task of flying the aircraft. The third opportunity for a workload construct in the model of pilot performance was as a non-causal predictor of aviation outcomes. The non-causal aspect of workload was operationalized via the self-rated NASA TLX scale (task load index, Hart & Staveland, 1988). The use of the TLX provided an opportunity to determine whether pilot self-ratings of task workload explained variance in outcome measures and thereby augment the model’s ability to predict the criterion variables.

5.3.2 New Constructs in the Revised DMM

Two new constructs, task management (TM) and prospective memory (PM), were added as new components of the revised mental model for pilots, and replaced the original workload construct. Therefore, it was proposed that the architecture of a mental model for pilots capable of supporting basic aviator activity is mediated by four primary cognitive mechanisms. Figure 5.4 illustrates the latent structure of the DMM. The four elements of the DMM include:

- i. a dynamic situation model of the present and future physical environment (SA);
- ii. knowledge and skills needed for task-relevant performance (TRP);

- iii. task management mechanisms that incorporate knowledge of tasks and task priorities relevant to the overall goals of the mission (TM); and,
- iv. prospective memory mechanisms that support the timely execution of mission-related tasks (PM).

Figure 5.4. The revised dynamic mental model (DMM) and its composite elements of situation awareness (SA), task-relevant performance (TRP), task management (TM), and prospective memory (PM).

Thus formulated, the DMM was modified from the original concept by Herdman and LeFevre (2004). The SA and TRP features are consistent with the original description of the DMM. Task management and PM are two cognitive constructs new to the DMM that are hypothesized to be key components of a mental model for pilots. Task management refers to knowledge that enables engagement in the right task at the right time. Task management was considered the most “executive” of the DMM composite elements as it maintains overarching and sub-goals, and contains information regarding how or when tasks can be eliminated, suspended, or switched for other tasks. Task management should be strongly associated with pilot expertise. Hardy and Parasuraman (1997) highlighted TM as an important aspect of a pilot’s

mental model. Prospective memory mechanisms support the maintenance of plans to commit future acts or thoughts, and pilot awareness of when tasks must be performed. As outlined in Chapter 1, PM is a domain-dependent cognitive variable with links to aviation outcomes (Nowinski et al., 2003). Prospective memory mechanisms also aid in the recognition of memory cues that are available from the environment or from internal representations in the mind of the pilot (Uttl, 2008; Van Benthem et al., in press).

5.3.3 Distinction between the Mental Model for Pilots and Individual Mechanisms of SA, TRP, TM, and PM

Stating that a mental model for pilots represents a cognitive architecture mediated by SA, TRP, TM, and PM is not the same as stating that the entire repertoire of relevant information from the environment is actually contained within the mental model for pilots. For example, the pilot mental model must have access to a broad array of information from the situation model, information generated via SA processes. The SA information is used to update the DMM. The DMM integrates the relevant SA information with its other TRP, TM, and PM features. Determining the quality of that SA mechanism within the pilot mental model is therefore accomplished by indexing the situation model. The capacity of the pilot mental model regarding its SA function is inferred from the quality of the information the pilot has integrated into the various SA levels of processing (these levels are discussed below). By modeling indices of SA, TRP, TM, and PM simultaneously *as one construct* I have constructed a hypothetical, or conceptual, model of a mental model for pilots. While one could argue that there are similarities, the DMM should not be confused with a user's generation of a device model used to interface with a human-machine system. A

rationale for the inclusion of SA, TRP, TM, and PM as the four mechanisms supporting a mental model for pilots is provided below.

Pilot situation awareness. The ability to navigate accurately in the world supposes an accurate model of the world that is also updated frequently and accurately at regular intervals. Parasuraman et al. (2008, p.144) described situation awareness as the “continuous diagnosis of the state of a dynamic world”. Situation awareness is thought to build upon three levels of cognitive processing (Endsley, 1988; 1995). The first level is the perception and processing of multi-modal stimuli into meaningful units of information; at the second level this information is assembled into a comprehensive schema of the aviation environment; the third level projects current information into a future schema that anticipates a near future state of the environment (Endsley; 1988; 1995). Endsley’s definition of situation awareness as the “the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future” (Endsley, 1995, p. 36) offers a comprehensive picture of situation awareness. The maintenance of an accurate, timely, and meaningful model of the environment with its relevant stimuli has obvious links to pilot performance. Similar to Hardy and Parasuraman, Endsley positions situation awareness as a factor outside the mental model. As shown in Figure 5.3, goals direct the selection of an appropriate mental model. The mental model enables understanding and projection of the current situation model. In turn, situation awareness provides feedback to the mental model regarding necessary changes to the model. While the boxological methods used to express these concepts required they be physically separated from one another, the interdependency of mental models with situation awareness is clear. It is also clear from Figure 5.3 that the content of the mental model is not specified. The functions of

mental models are illustrated, yet the “stuff” that make up the models is not described (not even in very generic terms). Endsley (1995, p.43) quotes Rouse and Morris’ (1985) definition of a mental model (as also quoted here earlier) and suggests that mental models are “complex schemata that are used to model the behavior of systems.” Complex schemata have a close association with situation awareness. Endsley states that a mental model has an attention focusing role. For example, if a goal consists of employing a missile, attention is directed by the mental model toward information regarding altitude, heading, and airspeed for own and threat aircraft. The information pertinent to the goal, in this case of missile deployment, is integrated into the situation awareness model of the pilot. The mental model consists of mechanisms for directing attention when goals are known. If the mental model directs the human operator to detect and assimilate information into a three-dimensional model of the physical world (or SA level 2) then some kind of situation awareness feature must be part of the mental model for pilots. It is not logical that the mental model directs the seeking and use of information in situation models of the world, uses that information to select scripts for action, or even modify goals, if situation awareness mechanisms were not also inherently part of the mental model. Therefore, the present work advocates that the mental model for pilots is partially mediated by a mechanism pertaining to situation awareness. While the actual situation model might reside conceptually outside of the mental model proper (as in Figure 5.3) the mental model functionality is contingent upon information from situation awareness and must be equipped to somehow appropriately use this information. It entails that situation awareness measures also serve as reasonable indices of relevant aspects of a pilot mental model.

Task-relevant performance. Task-relevant performance represents the various skills and competencies of the pilot that support a wide range of aviation

tasks. While some have suggested that mental models do not contain a performance element e.g., possessing the competency to actually perform some task (Jonassen & Tessmer, 1996), the TRP component of the DMM does incorporate perceptual-motor scripts for carrying out tasks in a competent manner. Descriptions of mental models routinely include components that refer to knowledge and skills pertaining to relevant procedures (Hardy & Parasuraman, 1997; Rouse & Morris, 1985). Figure 5.3 also illustrates that procedural scripts are associated with mental models. The “running” of scripts, such as the servo-mechanistic (Craik, 1943) responses of the operator to cockpit controls (e.g., small physical adjustments that ensure maintenance of desired altitude, heading, and airspeed), is controlled in part by the mental model. Using the same logic as was employed with the issue of situation awareness, it was surmised that a mental model for pilots manage procedures, even procedures of a very automatic nature such as operating aircraft controls to maintain a desired airspeed and trajectory. Rasmussen (1979) also referenced this sensory-related monitoring of the environment related to device control (or simulated control), and suggested that mental models of human-machine systems were responsible for the automatic detection of relevant sensory information, the detection of information mismatch, the detection of missing information in the environment.

Selection of the task-relevant performance measure should be guided by the experimental manipulation being tested (Herdman & LeFevre, 2004). The ability to maintain ideal altitude, airspeed, and heading is a key skill that has substantial precedence in the literature as a basic index of flight performance (Causse et al., 2011; Taylor et al., 2000; 2005; Yesavage et al., 2011). Flight path deviations, as measured by the root mean square error from ideal airspeed, altitude, and heading, represent task-relevant performance that reflects foundational skills for pilots. As shown in

Figure 5.5, task-relevant skills illuminated the effects of experimental manipulations (in this case, altitude deviation in low- and high-difficulty conditions). The more difficult mission resulted in greater deviations from the ideal altitude. Workload effects were significantly more pronounced for the two older pilot groups as compared to the younger group. Figure 5.5 illustrates the variability of TRP that occurred within heterogeneous sample of pilots with varied expertise and age. In the present, work task-relevant performance was the facet of the mental model for pilots indexing the model's ability to monitor and direct critical procedural skills. As shown in Figure 5.5, TRP varied in a predictable way in accord with pilot age and mission difficulty (both deleterious effects).

Figure 5.5. Task-relevant performance as a function of age and mission difficulty. Higher scores on the Y-axis represent greater deviations from the ideal altitude. Note that the youngest pilots demonstrate less of an effect of higher difficulty than the two older pilot groups.

Task management. Task management (TM) is a new element of the revised DMM and refers to knowledge possessed by the pilot of standard operating procedures for the aircraft being flown and of knowledge relevant to the goals of the mission. For example, managing tasks with acumen requires knowing what to do and when to do it to fulfil the flight mission. This aspect of the mental model references Jonassen and Tessmer (1996) who highlighted the knowledge structures pertaining to tasks (e.g., knowing when to do something). When handling unexpected events, TM also requires knowledge regarding what activities can be abandoned and replaced by new actions designed to manage the unexpected event. Task management, as an element of the DMM, reflects the proposal by Hardy and Parasuraman (1997) that procedural knowledge regarding tasks is an integral feature of a mental model for pilots.

Prospective memory. Prospective memory is also a new component of the revised DMM and was hypothesized as integral to a mental model for pilots. Competent pilot performance is reliant upon the detection of cues from the environment to trigger the execution of tasks that were planned at some time in the past. The procedural information utilized by the DMM must include a cognitive mechanism that enables the appropriately timed execution of tasks in the future. For example, PM supports recall of planned physical or mental events to be carried out at timely points during the mission, such as lowering flaps or making radio calls at specific time points (or positions in a circuit). A study of the Aviation Safety Reporting System found 74 of 75 reports that identified a memory failure involved PM errors (Nowinski et al., 2003). Researchers have also found that the type of memory cue (or PM trigger) associated with the task was associated with remembering to perform the task (Delprado et al., 2013; see McDaniel & Einstein,

2000 for a discussion on memory cues; and Uttl, 2011 for a meta-analysis on cue-type and prospective memory). Generally, it was noted that the less noticeable (less salient) the memory cue was, the less likely the prospective task was recalled and acted on. *The addition of prospective memory to the DMM illustrates the synergistic nature of each DMM element: competent task management and execution of tasks rely upon prospective memory, and prospective memory is supported by cues and situation models generated by situation awareness.*

Indexing the latent constructs of SA, TRP, TM, and PM. The following discussion briefly describes the measures used to index the latent constructs of SA, TRP, TM, and PM. Researchers in pilot performance have developed a number of methods for studying situation awareness. The Situation Awareness Global Assessment Technique (SAGAT) is frequently reported in studies of situation awareness (Endsley, 1995). SAGAT involves engaging the pilot in simulated flying scenarios, freezing the scenario at random points unknown to the pilot, and questioning the pilot on particular aspects of the scenario that are associated with the three levels of situation awareness (Endsley, 1995). For example, queries might include the pilot reporting on the location of ownship or other aircraft. The SAGAT method has shown good validity in indexing a pilot's ability to create and maintain an accurate model of the environment (Wickens, 2008). In the present work, SA was indexed via performance on direct measures of SA using the SAGAT method.

Task-relevant performance is also an original feature of the DMM, and is retained in the revised DMM specifically as perceptual-motor responses that support basic piloting ability. For aviators, TRP refers to the (usually) small adjustments to controls primarily by hands and feet to attain and maintain a desired trajectory. Inputs to the system are a direct reflection of the perceived discrepancy between the actual

and desired trajectory. Larger movements include reaching for trim or flap adjustments, or sharp turns on the yoke when responding to unanticipated situations. In general, however, TRP has been described as servo-mechanistic in nature (Craik, 1943). In this work, TRP was measured by flight path maintenance. Figure 5.5 illustrates the variability in TRP that can occur in a sample of pilots with varying expertise and age, thus solidifying its importance in any model of pilot performance highlighting individual factors.

Task management is the only component of the DMM that was indexed via a self-report measure in the present work. Subjective ratings regarding a pilot's awareness of tasks to be performed, and how and when the tasks should be undertaken were used to gauge the task management aspect of the pilot mental model.

Prospective memory was indexed via the rates at which pilots correctly remembered to make radio calls at particular points during flight. Radio calls were required at six different positions in the circuit. Therefore, 18 radio calls were associated with the low-difficulty and 18 with the high-difficulty conditions. Chapter 3 provides a full description of the method and procedures utilized in the present work.

5.3.4 Assumptions of the DMM

Five assumptions regarding the DMM are discussed below. The DMM is formulated first as an integrated network of functions, second as an automatic mechanism, third as one model among potentially many models, and fourth, as the prime mediator of upstream effects upon pilot performance and aviation outcomes positioned “downstream” in the model (see page 152 for a description of upstream and

downstream factors in models of pilot performance). Finally, the dynamic nature of the DMM is assumed.

The DMM supported by integrated neural networks. The DMM, and necessarily its four composites, follow the same precepts as other complex cognitive functions of the mind in that the DMM is not envisioned as “housed” in one localized region of the brain. The DMM components are likely supported by complex interconnected neural circuitry, each component further operationalized by its own network of brain regions. Thus, each of SA, TRP, PM, and TM are mediated by circuits unique to their own processes, and, most likely, by some overlapping circuitry. It is beyond the scope and methodology of this dissertation to determine the complex circuits within the brain that enable each component of the DMM. Yet, it is important to state that the DMM relies upon an integrated network of brain regions, where alterations in function in one or more nodes within the network may result in substantially reduced DMM quality. Converging evidence from brain imaging research suggests that cognitive functions are mediated via multiple brain centres (Kennedy, K. et al., 2015; Tognoli & Kelso, 2014). The reliance of cognitive performance upon integrated brain networks is supported by findings pointing to the integrity of the connective white matter tracts as a key determinant in cognitive performance, in addition to effects from localized grey matter. See Bressler and Menon (2010) for a review of structural, functional, and effective connectivity studies pertaining to cognition. Interestingly, before the domain of modern neuroscience had even emerged, in 1943 Craik (1943, p. 53) postulated that mental models were supported physically in the brain via “neural patterns by stimulation of the sense organs”. The current work makes no claim on actual neural patterns or brain networks involved in the production of a mental model. Yet, there was *prima facie*

understanding that various brain networks must and do support the cognitive functions of relevance to a mental model for pilots.

Automaticity and the DMM. A mental model is necessarily complex when applied to situations involving aviation, although, when a mission is well-rehearsed such as routine circuit flight, the DMM might seem to be produced effortlessly. The pilot mental model should be updated and used in an automatic fashion. For example, when new relevant information is encountered the information should be used and stored by the model automatically. The DMM could be characterized as the mechanism that explains how one might operate an aircraft on mental “autopilot”. The mental autopilot phenomenon may be detected fleetingly by pilots after realizing that *somehow* they have competently flown a significant distance without awareness of attention to the task. In fact, there might appear to be no residual memory of actions for the previous few minutes while the mind was seemingly engaged in other tasks (e.g., thinking of plans for later or previous conversations). A more universal example is driving on “autopilot”. It is likely that most drivers have experienced that same phenomenon where the task of driving has proceeded competently despite the mind was seemingly engaged in thoughts unrelated to operating the vehicle. A task-specific DMM automatically controlling behaviour (TRP) or monitoring the environment for information (SA) must exist for the mental autopilot phenomenon to occur successfully. Because the DMM was hypothesized to be continually updated in the mind, chiefly by automatic processes, there is a dilemma concerning how upstream factors are able to influence the DMM.

Automaticity of mental processes have been described for perceptual-motor tracking (Craik, 1943), visual search (Shiffrin & Schneider, 1977; 1984; Schneider & Shiffrin, 1977), and even the automatic nature of automaticity itself (i.e., with

repetition, sequences become automatic without a deliberate intent for automaticity) (Bargh & Chartrand, 1999). It is generally presumed that automatic mental processes, perhaps due to their practiced and repetitive characteristics, operate without the constraints that limit effortful or controlled processes. A strong version of this premise states that “execution of an automatic process consumes no attentional resources; nor does it interfere with other ongoing cognitive processes, or leave any trace of itself in memory” (Kihlstrom, 2010 p. 586). Paradoxically, as early as Broadbent (1954) it has been understood that both automatic and controlled behaviours are implemented by brain processes, which are inherently limited in capacity. Regardless, there is general consensus that even the automatic processes that support complex tasks, such as driving, are entertained effortlessly with little taxing of brain resources (Shiffrin & Schneider, 1984). The motor acts accompanying the mental autopilot also have keen automaticity associated with them. Craik (1943) strongly influenced current thought pertaining to human factors in complex systems and described the servo-mechanistic behaviours of the human operator undertaken during tracking behaviours. Trajectory corrections occurring in tandem with regular sensory “sweeps” of the environment are described as ballistic, with the automatic triggering of a complete sequence of perceptual-motor tasks upon discernment of misalignment (Craik, 1943). This servo-mechanistic activity is akin to the TRP of pilots in the current study where tracking of altitude, airspeed, and heading deviations was coupled with seemingly automatic perceptual-motor responses resulting in small adjustments to flight controls. The current research was not concerned with the extent of automaticity involved in the construction of the DMM. Rather, automaticity in DMM construction and maintenance is ostensibly understood.

Previous understanding of automatic mental processes and resultant behaviour claims that these processes are not limited by other ongoing mental activities or by finite capacity mechanisms such as working memory (Kihlstrom, 2010; Shiffrin & Schneider, 1984). In contrast, it was the intent of this research to establish that “automatically” produced mental models—regardless of how frequently or well-formed the model is—are indeed constrained by the same limited capacity mechanisms supporting controlled (non-automatic) processes. There is evidence that automatic cognitive phenomena are affected by working memory. A structural equation model with a causal path from working memory to reading comprehension (skilled adult readers) was partially mediated by the construct represented by mind wandering scores (a proxy for executive attention ability) (McVay & Kane, 2012). The significant relationship between working memory and the automatic (uncontrolled) mental process of mind wandering reasonably suggests that working memory, processing speed, or other cognitive mechanisms known to predict controlled cognitive behaviours might also predict the DMM. By extension, factors such as age, with known relationships to working memory or processing speed might also predict DMM function (for an extensive meta-analysis regarding age effects upon cognitive function see Verhaeghen & Salthouse, 1997). Pilot expertise is another factor that is likely to support the automatic nature of the DMM, although opposite to pilot age in its effects upon performance.

With repeated exposure to a task the network of SA, TRP, PM, and TM mechanisms learn to self-assemble in a manner that enables competent mission performance. The DMM is sensitive to experience as practice improves the efficient integration of information from the environment for SA and TRP. Experience also strengthens TRP by improving precision of perceptual motor responses needed for

trajectory control. The automaticity, or self-assembling, which occurs with learning is critical for any DMM, as conscious and purposeful coordination of SA, TRP, PM, and TM is time-consuming, mentally fatiguing, and cumbersome. An example of the automaticity that develops with the DMM is demonstrated by considering the difference between flying an aircraft for the first time and flying the same aircraft after five years of piloting. In fact, automaticity within the DMM likely exists as a continuum and not as an “all or none” phenomenon.

The Mental model for aviation versus access to multiple “mental models”.

As shown in Figure 5.3, Endsley (1995) portrays the mind as having access to many potential mental models, each associated with a different aspect of the system. Similarly, the DMM is assumed to be one model among many that are accessed by the pilot during various aviation activities. The current work posits the DMM as a model supporting the “basics” of flying an aircraft. The DMM is a general mental model for aviation reflecting a broad base of system information including the pilot’s own aircraft, a known volume of airspace, and other aircraft within this airspace. Also inherent in the DMM is an awareness and capacity to perform the activities required for competent aviation within this system. On this view, the DMM represents the mental model that operates during basic and highly automated flight situations, such as the phenomenon that supports flying on mental “autopilot”.

The DMM as a mediator of upstream effects upon pilot performance. The DMM was also assumed to mediate effects from elemental cognitive mechanisms, such as working memory or processing speed, upon criterion performance variables, such as critical incidents or management of surprise events. The difficulty encountered by previous research in accounting for major portions of variance in outcome measures might be due in part by conceptual and theoretical gaps that exist when

explanations of high level performance outcomes (such as critical incidents) rely upon domain-general cognitive mechanisms (such as processing speed). In the present work, the DMM functions as an intermediary construct that bridges the relations between upstream predictors and downstream performance factors. As shown in Figure 5.6, upstream factors influence aviator performance both directly and indirectly, as mediated by the DMM. The DMM was not the only mediating factor that was investigated in this work, for example, the influence of pilot age on SA or PM might also be mediated by processing speed or working memory ability. The concept of constructs acting as mediators of upstream effects is important in complex models of aviation performance. In the present work, direct and indirect effects from upstream variables upon downstream outcomes were quantified, allowing for the comparison of the relative effects from various predictors upon performance criteria.

Figure 5.6. The DMM as a mediator of effects from upstream predictors on downstream outcomes. In this figure the effect of the predictor upon the outcome is partially mediated by the DMM, as represented by the dashed arrow that demonstrates remaining effects that exist between *predictors* and *outcomes* after mediation by the DMM.

Dynamic features of a mental model for pilots. The DMM was considered a dynamic construct in the present work for several reasons. The DMM has dynamic properties because its supporting mechanisms are continuously acting in concert with each other, but at various times one or more elements of the DMM dominate the landscape of the model. While SA is a forefront process while navigating through new airspace, TRP may represent a primary function for maintaining a flight path during poor weather conditions. Dynamism is also represented by the continuous integration of information from the environment and from long-term memory with the current schema. Herdman and LeFevre (2004) defined the DMM as a support for being aware of what was happening during flight. This definition entails the idea that a mental model for pilots is anything but static. A final, less temporal, aspect of the DMM's dynamic trait pertains to the idea that the DMM is responsive to underlying support mechanisms, which oscillate based on operator experience, brain health (or states such as fatigue), and maturation across the lifespan. A human factors notion of mental models also emphasizes their dynamic features (Besnard et al., 2004). Mental models are updated and maintained by pilots, but this dynamism is constrained by cognitive factors and environmental factors such as confirmation bias (monitoring for information that is consistent with the current mental model and ignoring other non-consistent information) and task workload, respectively (Besnard et al., 2004). The emphasis on the dynamic nature of a mental model for pilots is not entirely consistent with the model described by Hardy and Parasuraman (1997) who reported that mental models are slow-to-change entities. This is likely because Hardy and Parasuraman did not include situation awareness as an element of the mental model for pilots, but instead envisioned it as a body of skills and procedural knowledge that might wax and wane mainly in accord with pilot experience.

5.4 The DMM in the Larger Context of Pilot Performance

Hardy and Parasuraman (1997) specified a comprehensive model of pilot performance that accounted for pilot attributes and stressors, cognitive abilities, cognitive skills specific to the aviation domain, and outcomes of relevance to aviation. Figure 5.7 depicts the DMM within this larger context of pilot performance. This model of pilot performance, based on the model by Hardy and Parasuraman, was used to guide the development of hypotheses in present work. While the model by Hardy and Parasuraman did not depict the human-machine system explicitly, it was understood that changes to the aircraft or the environment were considered pilot “stressors”. The effect of stressors were transmitted through the model and eventually reflected in pilot performance. Hardy and Parasuraman explicitly positioned the mental model for pilots centrally as a domain-dependent cognitive variable. In the current model of pilot performance the DMM is positioned as a domain-dependent cognitive construct that receives and transmits effects from upstream predictors, and also acted as a direct predictor of outcome variables.

Figure 5.7. The DMM in a simplified model of pilot performance. Adapted from Hardy, D., & Parasuraman, R. (1997). Cognition and flight performance in older pilots. *Journal of Experimental Psychology: Applied*, 3(4), p.339. The full structural equation model tested in this dissertation was based on this model of pilot performance.

Data from this research agenda was utilized to statistically model the DMM and determine if the four elements converged as one construct to represent the mental model for pilots. Specifically, it was the premise of the present work that SA, TRP, TM, and PM comprehensively index a mental model for pilots. SA provides the structure for understanding current and near future position of others and ownship. Automatic responses by TRP allow for safe navigation through the “situation”. Knowing what tasks to perform and how to perform each one specifically and with correct prioritization is represented by TM. Last, being aware of environmental or internal cues related to the execution of future acts is provided by PM. The four

composites together build one formative construct known in the present work as the DMM.

5.5 Specifying the DMM in a Structural Equation Model

Factor analysis and structural equation modeling (SEM) were the primary methodologies used to test the three main hypotheses associated with this work. These hypotheses were:

1. The four domain-dependent cognitive factors of SA, TRP, TM, and PM converge in a multilevel structure to represent the DMM for pilots.
2. The DMM, as a domain-dependent variable, accounts for significantly more variance in the aviation outcomes measures than the variance accounted for by domain-independent cognitive factors.
3. The DMM fully mediates the effects of cognition, age, expertise, and workload factors on aviation outcomes.

SEM is a “second generation” statistical technique for generating and testing statistical models. In SEM, factors or variables in the model can be latent (unobserved) or observed variables. The inner model of a SEM diagram is the path model. It illustrates the hypothesized direction of causation between variables. The outer models constitute the measurement models for the latent variables. SEM makes use of factor analysis to generate these latent variable structures. Inner path estimates are often determined via covariance-based modeling methods (e.g., maximum likelihood, Jöreskog, 1967) or variance-based techniques (e.g., partial least squares, Wold, Ruhe, Wold, & Dunn, 1984). Thus, SEM is known as a hybrid technique that may include confirmatory factor analysis, path analysis and regression. The DMM was one of a number of measurement models in the larger model of pilot performance tested in the present work.

SEM structures start with hypotheses regarding predictor and criterion (dependent) variables. Latent factors are specified according to their reflective or formative (definitions to follow) nature. The latent structures in SEM are known as measurement models. Measurement models are tested and refined using factor analysis to ensure that the latent structure is represented by a reliable and valid structure of indicators (also called observed or manifest indicators). Well-constructed measurement models, along with directly observed variables, are posited in the larger theory-driven SEM structure. After an initial test of the SEM structure, a model generating process refines the model so that good “fit” is achieved. Figure 5.7 illustrates the basic structure for the comprehensive model of pilot performance that the SEM modeling in the current work was based upon. Figure 5.8 details the specification of the DMM measurement model that was posited as a central variable in the model of pilot performance.

5.5.1 The Multidimensional Features of the DMM

The multidimensional aspect of the DMM has been alluded to in this chapter. The DMM was hypothesized as formed by four first-order constructs, that themselves are latent constructs. The multilevel aspect of the specified the DMM as a multidimensional construct with first and second-order latent constructs (see Figure 5.8).

Figure 5.8. Hypothetical specification of the DMM as a multidimensional reflective-formative latent construct. Arrows pointing *from* the first-order factors to manifest variables indicate reflective construction. The four first-order constructs show arrows pointing *to* the DMM (the second-order factor), indicating formative construction. The specification of first-order manifest variables is discussed in Chapter 6. DMM= dynamic mental model; SA= situation awareness; TRP= task-related performance; TM= task management; PM= prospective memory.

When specifying the structure of a latent construct the number of composite dimensions must first be determined. A latent construct with two dimensions has first and second-order levels. The second task when specifying latent constructs is determining the causal relationship (or direction of influence) between observed variables and their associated latent constructs. The direction of influence indicates whether a latent construct is a formative or a reflective construct (Bollen & Bauldry, 2011). A formative construct is formed, or caused, by its composite variables (arrows show direction *from* indicator *to* the latent construct). In a formative design, the latent construct coalesces from linear combinations of elements rather than from their overlapping variance (Lee, Cadogan, & Chamberlain, 2014; Petter, Straub, & Rai, 2007). A reflective latent variable “causes” the association between the latent construct and its indicators (arrows show direction *from* latent construct *to* the observed variable) (Lee et al., 2014). Presumably, for appropriately selected reflective indicators, a large portion of the variance from each measure overlaps with the variance of the remaining reflective indicators. This common variance is thought to be caused by the reflective latent construct.

The DMM is hypothesized as a formative construct because it is caused or *formed* by four other constructs (Petter et al., 2007). This is necessarily so, as a mental model for pilots cannot be fully understood by just one underlying entity. This view is supported by the descriptions of mental models discussed earlier, that a mental model is a composite-type construct composed of various unique, but essential elements. A mental model with the complexity to support aviator tasks should access multiple cognitive mechanisms to enable the attainment of mission goals (Endsley, 1995; Hardy & Parasuraman, 1997). The formative nature of the DMM is also supported by the consideration that if one component were removed from the model, then the entire

understanding of the DMM changes (Petter et al., 2007). The meaning of a formative construct relies upon the combination of *all* its composite elements. This composite definition is the hallmark of a formative construct (Lee et al., 2014). Each composite contributes some unique aspect of the construct to help “form” the underlying construct; hence this kind of model denotes a formative structure (Diamantopoulos, Riefler, & Roth, 2008). Without SA, for example, the notion of a mental model for pilots changes considerably. The accuracy of this hypothesis, that the DMM is a formative construct, was evaluated after the DMM modeling process. Generally, if the composite elements are very highly correlated with each other then it is more likely that a construct is a reflective rather than a formative construct (Petter et al., 2007). Formative constructs are usually associated with composite variables that have weak, or non-significant intercorrelations. Structural equation modeling tested the theory that four composite elements converged empirically as one construct. Specific hypotheses regarding the nature of the DMM as a formative construct were tested and discussed in Chapter 6.

In contrast, each of the four constructs that represented the first-order variables of the DMM are hypothesized as reflective latent constructs. Reflective latent constructs are not formed by their indicators; instead the reflective latent construct is considered the “cause” of an indicator (Petter et al., 2007). SA, TRP, TM, and PM are considered unitary concepts and have been described in this manner in this Chapter. For example, each reflective indicator associated with the SA construct is designed to index just that one construct. A construct is considered reflective if it is associated with indicators that could be removed (individually) from the model without altering the underlying concept behind the construct (Petter et al., 2007). For example, each indicator for SA should be equally capable of reflecting the concept of situation

awareness. If the SA construct was indexed by SA levels 1 and 2, instead of levels 1, 2, and 3 the underlying idea behind the construct should not change. It was hypothesized that the four measurement models, one for each SA, TRP, TM, and PM construct, converged as a reflective latent constructs. The results of the data modeling were evidence for the hypothesis that each composite of the DMM was best specified as a reflective construct. After data modeling, the indicator variables associated with each reflective construct were examined with regard to the strength of their association with the other indicators in their measurement model. Indicators within one reflective measurement model should correlate well with the other indicators in that same model. For example, if the indicators from one measurement model (e.g., SA) have strong intercorrelations this supports the selection of manifest indicators and the specification of the latent variable as reflective in nature.

In SEM, the distinction between formative and reflective latent constructs is formalized by the direction of the arrows either to or from the indicator to the construct. A formative model shows the arrows *from* manifest variables *to* the latent construct (indicating direction of cause). For reflective constructs the arrow is in the direction from the latent construct to the indicator. Figure 5.8 illustrates the direction of arrows to and from indicators in the DMM as just described here. Whether SA, TRP, TM, or PM is measured as reflective constructs (as hypothesized here) or are composed of formative constructs does not change the overarching form of the DMM. For example, if one very informative indicator were found for each of SA, TRP, TM, or PM then reflective measurement models would not be required and each composite (as measured by one indicator) would still be combined in a formative manner to build the DMM. In fact, the constructs that build the DMM were composed of a combination of reflective and formative latent constructs. However the present work

made use of first-order latent constructs that were reflective in construction. According to model typology, the DMM with a reflective first-order and formative second-order structure, is a “type II” hierarchical latent construct (Becker, Klein and Wetzels, 2012, p. 364).

Value of a Multilevel DMM. An advantage to modeling the DMM as a formative construct pertains to model parsimony. It was possible to model SA, TRP, TM, and PM individually, where each variable was related to antecedents and downstream variables in unique ways. However, in a comprehensive model of pilot performance, modeling SA, TRP, TM, and PM as separate constructs would quickly lead to a very complex web of relationships. Simple models are generally preferred to complex structural equation models because interpreting parameters is more straightforward, and the overfitting of variables to the variance covariance matrices and the capitalization of chance correlations are less likely when the model is less complex (Hair Jr, Hult, Ringle, & Sarstedt, 2013). Modeling the DMM as a formative construct contributed to simplicity in investigating upstream effects on downstream outcomes, and to the primary goal of this work, which was to determine whether there was empirical support for a *functional coalition* of cognitive mechanisms that support aviator tasks. In sum, the development of a DMM for pilots, situated centrally as a domain-dependent predictor of aviation outcomes (as illustrated in Figure 5.7), permitted an account of the transmission of cognition, age, expertise, and workload effects through a comprehensive model of pilot performance.

5.5.2 Material Causation and the DMM

An important concept in the present work was the specification of the DMM as a formative latent construct where each element contributed some unique and relevant

aspect of a mental model for pilots. The individual elements of the DMM did not simply direct influence on the DMM, they were used in a material sense to build the DMM. This conception of the DMM relied on the *material causation* explanation of a formative latent variable (Lee et al., 2014). Hypothetically speaking, if “bits” of information were measurement units for SA, TRP, TM, and PM then the combination of these bits of information (as theorized in Figure 5.8) produces a new conceptual entity called the DMM. The latter feature of the DMM, its description as a new entity, differs from Lee et al.’s (2014) proposition that material causation implies that the formative construct is not a separate conceptual entity from its constituent elements. In this work, the DMM was both reliant upon the material comprising SA, TRP, TM, and PM, and, at the same time, representative of a new structure. The following analogy explains the formative nature of the DMM: the single concept of a *house*, clearly relies upon four distinct elements composed of roof, walls, door, and floor. Together the roof, walls, door, and floor combine to form a house. The elements are joined together structurally, to physically and conceptually create a house. Each element used to build the house is distinct, and the final construct--the house--is also distinct from its parts. On this view, the combination of elements is possible regardless of whether the constituent elements were made of the same “stuff”. Analogous to the roof, walls, door, and floor combining to form a house, so also do SA, TRP, TM, and PM combine as the building blocks of the DMM. The primary purpose in empirically testing the DMM as a formative construct was to determine if the four distinct elements of SA, TRP, TM, and PM converged algebraically as one construct, where each part contributed significantly to the final structure. Finally, the initial metaphor that bits of information are the material “stuff” of cognitive processes follows a position widely attributed to John von Neumann (1958) who proposed that, like a

digital computer, neurons in the brain, by firing or not firing, provide bits of information. Similarly, George Miller (1956) suggested that the mind processed bits of information: for example, one bit of information was required by the mind in order to make a decision between two equally-likely options. However, the present work did not propose any type of informational unit for cognition, and the reference to the informational bit is used solely to explore the material causation explanation of the DMM. The DMM was described earlier as enabled by a structured and goal-driven network of brain areas. The material causation assumption is best thought of as the DMM being mediated by the integration of neural networks. Individual segments of the DMM network can operate independently to support SA, TRP, TM, and PM functions, but segments can also work in concert to generate a mental model for aviation, a model that supports goal-driven aviator tasks.

5.6 A Working Definition and Structure of a Mental Model for Pilots

To conclude, the DMM is a conceptual model that was operationalized to reflect a “base” mental model for pilots. The DMM is a basic mental model because it was assumed to support the fundamental competencies of a pilot required for GA flight. Such a mental model for pilots was indexed by SA, TRP, TM, and PM. The mental and physical capacities supported by SA, TRP, TM, and PM were assumed to index the full complement of cognitive functions required for basic aviator activities. At the same time, the four components of the DMM are cognitive mechanisms that can also exist independently of the DMM.

A principal goal of the present work was to determine whether a single construct, composed of four different cognitive functions, was a plausible structure for representing a mental model for pilots. The DMM was specified as a multilevel formative-reflective construct signifying that the independent mechanisms of SA, TRP, TM, and PM were combined to build a single unifying construct. As described in the following Chapter, the DMM structure was validated using factor analysis to test the hypothesis that the measurement models for SA, TRP, TM, and PM were well-specified as a single formative construct. The DMM was also validated, using SEM, within a larger model of pilot performance, where the DMM's direct and mediating effects on aviation outcomes were evaluated. The DMM, as well as pilot attributes and other cognitive factors, were assessed in the larger model of pilot performance in regards to their value as predictors of risk for two outcomes relevant to GA: critical incidents and the outcome associated with a surprise runway incursion during landing.

Chapter 6

Empirical Validation of the Dynamic Mental Model for Pilots

A significant portion of the present work involved modeling a plausible structure to represent a mental model for pilots. Rigorous reliability and validity testing of the content and structure of the mental model for pilots was necessary before examining its utility as a predictor of key criterion variables. This chapter describes the methods and results of testing the first of three main hypotheses in this program of research:

- 1. The four domain-dependent cognitive factors of SA, TRP, TM, and PM converge in a multilevel structure to represent the dynamic mental model (DMM) for pilots.**
- 2. The DMM, as a domain-dependent variable, accounts for significantly more variance in the aviation outcomes measures than the variance accounted for by domain-independent cognitive factors.**
- 3. The DMM fully mediates the effects of cognition, age, expertise, and workload factors on aviation outcomes.**

6.1 Overview of the Hypothesized DMM Reflective-Formative Hierarchical Structure

The DMM was hypothesized as a latent construct with multiple levels where the first-order constructs were reflective and the second-order was formative. Figure 6.1 illustrates the formative composite model for the DMM. The first-order latent variables of SA, TRP, TM, and PM are modeled (reflected) via their associated measured variables. Carefully selected first-order constructs, and their manifest indicators, assisted in developing a structure with the capacity to support the complex activities tasks required for aviation. The selection of manifest indicators was guided

by Petter et al. (2007) who recommend that the first-order constructs in a formative latent variable ensure complete coverage of all facets of the formative latent variable, while avoiding redundancy. Each first-order construct and its initial measurement model structure are discussed below. In this work the concepts of latent variable, latent construct, and latent factor are synonymous. All latent variables are constructed, and therefore indexed, via their measurement models.

6.1.1 Situation Awareness (SA)

The SA measurement model indicators were the measures for SA levels 1, 2 and 3. As discussed in the results section of the experimental phase of the work, SA level 1 appeared to represent a unique construct as compared to the SA levels 2 and 3. For that reason SA level 1 was relabeled as iRES in Chapter 4. Before eliminating SA level 1 as a potential indicator for SA in the first-order measurement models it was decided to examine the loadings and cross-loadings of SA level 1. Only if SA level 1 did not appear to strongly and uniquely belong to the SA construct would it be removed as an indicator in the SA measurement model.

6.1.2 Task-Related Performance (TRP)

The measurement model for TRP consisted of the RMSE scores for altitude, airspeed, and heading maintenance. The measure related to heading maintenance was retained in the initial structure of the TRP construct even though it was shown to be negatively associated with the other RMSE measures in the low-difficulty condition. See page 145 in Chapter 5 for a discussion on the selection of the task-related performance indicators. If the variable representing RMSE heading continued to demonstrate a contrasting pattern of relations with the variables for altitude and airspeed, then it would be considered for removal from the TRP construct.

6.1.3 Task Management (TM)

Task management was reflected by three indicators from the self-rating scale for situation awareness. The analysis of the SA self-rating scale (see Section 4.3.5) revealed that the three task-related items appeared less related to the subjective SA measure when compared to the other items regarding ownship and position of other aircraft. Therefore the self-rating for awareness of tasks to be performed, the specific nature of the tasks, and the task priorities were selected as manifest indicators of the TM construct. This use of subjective indicators in a “causal” role was an exception to the premise that subjective measures in the larger model of pilot performance were expected to predict, but not be the cause of aviation outcomes. Future work might address this issue by designing paradigms that include objective task management indicators.

6.1.4 Prospective Memory (PM)

The PM measurement model was represented with one measure, the percentage of radio call completed in each mission difficulty condition. The single PM measure represented the sum of the 18 radio calls per difficulty condition, used to create one variable for testing within the DMM. Recent SEM publications have reported that using one best indicator is an acceptable and, at times, the preferred option as this reduces artificially high r^2 values that arise from multicollinearity effects associated with reflective indicators (Hayduk & Littvay, 2012). For this reason the 18 items from the PM task were *parceled* into one overall percent accuracy score.

6.2 Characteristics of the Datasets used to Model the DMM

Full details (e.g., ages, expertise etc.) regarding the dataset of 108 pilots is found in Chapter 3. The main dataset used in the SEM analyses contained roughly 216

observations for each observed variable. While the sample consisted of 108 pilots, the collection of observed variables used to model the DMM occurred twice, once for each mission difficulty condition in the flight simulation protocol. For example, each pilot had a score for SA Level 2 in the low-difficulty condition and another SA Level 2 score in the high-difficulty condition. This method resulted in a maximum of 216 observations in the flight simulation series. A similar method where multiple observations (per pilot) from flight simulation data for SEM analyses has been reported in the literature (Svensson, Angelborg-Thanedz, & Wilson, 1999; Svensson & Wilson, 2002).

6.2.1 Sample Size

Sample size is frequently a concern for structural equation modeling (Goodhue, Lewis, & Thompson, 2012). Goodhue et al. (2012) conducted Monte Carlo simulations for both CB-SEM and PLS-SEM estimation methods and found that even with small to moderate departures from normality, sample sizes of $N=90$ were able to achieve at least 80% power to detect medium size effects. For normally distributed measures and approximately $N=200$ the power to detect medium effect sizes was approaching 100% for both methods of estimation.

6.2.2 Missing Data

Regarding the entire dataset, there was less than 3% missing data, other than the items from the two self-rating scales. The self-rating scales for situation awareness and task workload (TLX) were not administered to the first 12 subjects and a few others had one or two responses missing from the scales at random. The missing values from the first twelve subjects represented approximately 11% of the total self-rating scale responses. It was selected to replace the missing values with the series mean. While

not ideal, the loss of approximately 12 subjects with otherwise complete response sets was weighed against some loss of series variance potentially occurring by utilizing the series mean for the missing responses. The logic for this decision to retain these subjects, as opposed to removing them from the dataset, using list- or casewise deletion, or multiple imputation methods to replace missing data is described in more detail below.

The DMM was tested using ten observed variables: SA levels 1 to 3, RMSE score for altitude, airspeed and heading, the three task-related items from the SA self-rating scale, and the PM measure of radio call completion rates. The three items from the SA self-rating scale, which represented the TM construct, had approximately 14% missing data (the first twelve subjects and a few random missing values). The remaining observed variables had less than 3% missing data (missing at random). Multiple independent sample *t*-tests were carried out to determine if the subjects with missing data from the SA self-rating scales differed significantly from the rest of the sample with respect to age and rating, or performance on the remaining seven observed variables in the DMM. Pilots with missing data from the self-rating scales had similar performance on the SA measures and the PM measure as compared to those subjects who had complete responses for the self-rating scales. The *t*-tests demonstrated a nonsignificant trend for the missing data group to be younger and have lower pilot ratings, $p > .05$. It was possible that younger pilots, eager to fly the Cessna 172 simulator, were the first to respond to the recruitment efforts, and therefore part of the first group 12 subjects who did not receive the self-rating scales to complete. A potential method for dealing with the missing data for the first 12 subjects was “multiple imputation”, where pilot age and rating were conditional variables. This method resulted in means for the imputed values for each SA scale item that were

extremely similar to the original series means (without the replaced values). In addition, a few replacement values generated by the multiple imputation algorithm exceeded the actual limitation of the scales (only responses 1 to 7 for the SA scale and 1 to 10 for the TLX scale were permitted). Therefore, the generation of implausible values, and the similarity between the mean of the imputed values and the original series mean was the rationale for using the series mean for handling missing data for the self-rating scale measures. Another important reason for retaining the subjects with missing self-rating scales related to the bias that is introduced into analyses when list- or case-wise deletion methods are selected. Retaining subjects with missing data in the dataset (where reasonable), rather than employing list- or casewise deletion is strongly supported due to the bias in mean structure and intercepts that can arise from the elimination of groups of subjects perhaps representing distinct characteristics in the sample (Dong & Peng, 2013; Wilkinson, 1999). Finally, retaining these younger and less experienced pilots was particularly important in work described in Chapter 7 where a comprehensive model of pilot performance was tested.

6.2.3 Data Normality

Understanding the distribution of each measured variable was important when modeling data, as CB- and PLS-SEM estimation algorithms are sensitive to data that do not meet the assumptions of normality. It is possible that properly specified models may be rejected because of non-normal distributions (Hooper, Coughlan, & Mullen, 2008). As mentioned, recent Monte Carlo simulations demonstrated that CB-SEM and PLS-SEM are robust to small deviations from normality (e.g. skew up to ± 1.1 and kurtosis up to ± 1.7). Histograms charting the distribution of each observed variable were examined for indications of skew or kurtosis. While minor deviations from

normal were observed for most variables, the test statistics found that only the PM measure demonstrated considerable skew (negative) and kurtosis (leptokurtic), -1.25 and 1.71, respectively. To address issues of normality and the presence of outliers in the data, in particular for the PM measure, the “jackknifing” resampling method was used to reliably estimate the standard errors for observed variables when testing the DMM (and for later comprehensive model testing as well) in the PLS-SEM software. See Kock (2013, p. 24-25) for a description of the jackknife standard error estimation method.

In sum, the dataset used to test the DMM consisted of ten observed measures collected during the simulated flight protocol, resulting in a 10 by 216 observation matrix. Responses for each measure were transformed into z-scores with a mean of zero and a variance of one. Each series was also ranked to minimize the effects of outliers. Information regarding the range and means for the raw, unstandardized and unranked data is found in Chapter 3. Correlations between the observed variables in the initial DMM structure are found in Table 6.1.

Table 6.1
Correlations between Observed Variables (Standardized and Ranked Dataset)

	SA1	SA2	SA3	RmseAlt	RmseSpd	RmseHdg	T1	T2	T3
SA1									
SA2	.361**								
SA3	.254**	.604**							
RmseAlt	-.313**	-.099	-.158*						
RmseSpd	-.207**	-.165*	-.149*	.458**					
RmseHdg	-.025	.100	.059	.021	-.033				
T1	.182**	.110	.056	-.093	-.071	.112			
T2	.151*	.058	-.002	-.169*	-.098	.135*	.697**		
T3	.228**	.103	.104	-.093	-.066	.119	.628**	.501**	
PM	.306**	.269**	.283**	-.331**	-.171*	.007	.180**	.196**	.233**

Note. * = $p < .05$ and ** = $p < 0.01$, two tailed. No. of observations = 216.

6.3 Two-Stage Process for Testing the DMM Multidimensional Structure

The basic modeling and evaluation of the DMM structure followed a two-stage process using PLS-SEM. PLS-SEM utilizes algorithms that are known to emphasize the predictive relations between variables (Hair Jr. et al., 2013). PLS-SEM was selected as the primary method for testing the DMM and the subsequent comprehensive model of pilot performance (see Chapter 7) because the purpose of this dissertation was to investigate the predictive utility of cognitive factors in a model of aviation outcomes. Additionally, the importance of the predictive relations between cause and effect variables in formative model structures makes PLS estimation particularly appropriate in this work (Petter et al., 2007). For both stage one and two, as described below, only observations from group 1 ($n=103$, as determined by randomized group selection) were used. As described in the next section, model cross-validation was used as an auxiliary method of validity testing of the DMM. The full dataset of 216 observations was randomly split into two sets: one with 103 and another with 113 observations.

In stage one, the measurement models for the reflective first-order latent variables were developed using the WarpPLS 4.0 statistical package using “PLS Mode A”. The use of this mode is standard procedure for testing reflective measurement models. Latent variables were assembled with the indicators as described above and illustrated in Figure 6.1. For example, the SA construct was reflected by SA level 1, 2 and 3. Using PLS Mode A “outer model weights were calculated through a least squares regression where the latent variable score is the predictor and the indicators are the criteria” (Kock, 2013, p.22). This analysis produced weights, component loadings, and associated p -values for each indicator. If indicators loaded inadequately (e.g. loadings lower than 0.60, or reliability coefficients less than 0.70) then the poorly

loading indicators were considered for removal. Once a measurement model was determined, the quality of that model was assessed using accepted reliability, and construct and discriminant validity criteria. After the three model structures were complete, the resulting factor scores (generated by PLS estimation) were saved and added as variables to the data file. Factor scores are used as a generic term in this work and do not reflect “factors” such as generated by *common factor analysis*. The generic use of the term factors is supported by work showing that latent variable scores from exact linear combinations (derived from weighted composites, as were created in this analysis) perform almost identically to factor scores estimated via common factor analysis algorithms (Velicer & Jackson, 1990), particularly where sample size was adequate e.g. $N=108$ (Velicer & Jackson, 1990) and for formative measurement models with few indicators (Fornell & Bookstein, 1982).

In stage two, the saved factor scores for SA, TRP, TM, and PM latent variables were used as composite indicators of the DMM latent formative construct. This formative model was evaluated using accepted methods that focused on the statistical significance of the weights, representing the relative causal effects *from* composite latent variables *to* the DMM latent formative construct (Becker et al., 2012; Diamantopoulos et al., 2008; Petter et al., 2007; Ringle, Götz, Wetzels, & Wilson, 2009). It was hypothesized that the factor scores for SA, TRP, TM, and PM, as a representation of a mental model for pilots, converged empirically and demonstrated good model fit. The fit coefficients for evaluating the formative model included the *p*-values for composite indicator weights (not loadings as evaluated for reflective latent variables), variance inflation factor values (VIF), weight-loading signs, and effect sizes associated with each composite→latent construct relationship. This two-stage method of generating and testing formative models was completed as per WarpPLS

4.0 instructions (Kock, 2013, and personal communication), and as per other PLS-SEM software packages (Becker et al., 2012). Figure 6.1 illustrates the first-order measurement models and the second-order formative structure of the DMM, which were evaluated individually. Each hypothesized parameter is labeled in Figure 6.1. Table 6.2 lists the hypotheses associated with each parameter tested in stages one and two.

Figure 6.1. Original hypothesized first-order measurement models and second-order formative structure of the DMM. Labels on each arrow refer to a specific model hypothesis in Table 6.2. C.A. = Cronbach's alpha (indicating average degree of relatedness between indicators within each latent variable [Cronbach, 1951; Sijtsma, 2009]).

Table 6.2
DMM Hypotheses

Model Element	Label in Figure 6.1	Parameter Estimate Hypotheses
Stage One		
SA Reflective Loadings	1 a-c	LVs show convergent validity in that manifest variables i) load significantly with their respective latent construct ($p<.05$), ii) have low cross-loadings on other LVs, and iii) have AVE values greater than 5.0. Discriminant validity is shown by the square root of the AVE for an LV being larger than the correlation of that LV with other LVs. LVs show good reliability by having high composite reliability ($>.70$) and high Cronbach's alpha ($>.60$).
TRP Reflective Loadings	2a-c	
TM Reflective Loadings	3a-c	
Stage Two		
SA, TRP, TM, and PM Composite Indicator Weights	4 a-d	Composite indicators in a formative model indicate i) relevance by significant weights ($p<.05$), ii) non-redundancy by low full collinearity VIF (<3.0), iii) proper direction of causality <i>from</i> indicator <i>to</i> formative LV by WLS=1, and iv) adequate causal effect sizes, ES $>.02$.
Composite Indicator Correlations	Not labeled	Evidence of a formative (rather than reflective structure) is shown by low correlations between composite elements of the DMM, $r<.5$.

Note. SA= situation awareness; TRP= task-related performance; TM= task management; PM = prospective memory; AVE = average variance extracted; LV = latent variable; VIF=variance inflation factor; DMM= dynamic mental model. PM is a single-indicator construct therefore no measurement model exists for PM. WLS= weight loading sign. ES= effect size.

6.4 The DMM Validation Process

In lieu of the standard one-step validation of SEM structures, three auxiliary validation activities were undertaken. This was considered an important aspect of validating the DMM, as there are no extant latent structures representing a mental model for aviation against which the DMM could be judged. Figure 6.2 illustrates the model validation process. First, the final model structure derived from group one model generation was cross-validated using the dataset from group two ($n=113$).

6.4.1 Model Cross-Validation

Modelers using SEM frequently stop the model evaluation process before cross-validating their model on unique data; thus, final model structures based on post hoc data-driven considerations (rather than theory determined before model generation) may be over-fitted to the data and not reflect structures that are robust across unique samples. To allow for cross-validation of the DMM, the full sample was randomly split into two similarly sized groups ($n = 103$ and $n = 116$). The Microsoft Excel™ “random number” generator was used to split the full data file of 216 observations. The DMM was tested and model generation was undertaken using the first sample of $n = 103$. The final DMM model structure generated from group 1 data was cross-validated with data from the second group ($n = 113$). Measurement invariance techniques were applied to ensure that good fit of the DMM was found for both the first and second datasets (as per MPlus 7.11, Muthén & Muthén, 1998). If no significant differences were found for key parameters between groups 1 and 2 then this was taken as evidence of model structure validity, and that the DMM structure represented sound theoretical considerations, and not peculiarities of the first dataset.

Tests of multigroup measurement invariance for second-order weights (weights relating each first-order construct to the DMM latent variable) generated from dataset one and two were undertaken. A lack of significant difference between the corresponding weights from group one and two indicated that the DMM formative structure was robust across the two unique datasets. Only after the cross-validation step was completed were the group one and two datasets combined for further validation of the DMM.

6.4.2 Validation of the Final DMM Structure Using Covariance-Based SEM Algorithms

The finalized DMM structure, as determined by PLS-SEM, was also tested within a covariance-based SEM (CB-SEM) platform. All hypotheses regarding the DMM parameters were examined using traditional CB-SEM model quality and fit indices. Ideally the DMM should generate adequate quality and fit indices for PLS-SEM and CB-SEM methods. A discussion of the different SEM estimation methods precedes the CB-SEM validation results.

6.4.3 Validating the DMM in a Nomological Net.

To test construct validity of the DMM, the complete DMM structure (with $N=216$) was tested within a nomological net. A nomological net is created by selecting antecedent and/or effect variables that theoretically have strong associations with the formative construct being evaluated, and then examining the links between the newly developed formative construct and these other variables (Diamantopoulos & Winklhofer, 2001; Götz, Liehr-Gobbers, & Krafft, 2010). Concepts in a nomological net should be strongly related to each other. In this case “iRES” was selected as the antecedent and “critical incidents” as the outcome variable. Large effect sizes from iRES to the DMM and from the DMM to critical incidents were considered evidence of the content validity of the DMM. This is because the resolution at which pilots integrate information from the environment (iRES) should strongly predict a pilot’s mental model for flying, and consequently, the mental model for flying should strongly predict the number of critical incidents incurred by a pilot.

Figure 6.2. Three auxiliary steps for structure and construct validation of the DMM.

6.5 Measurement Model Results

The results of modeling each first-order construct are described below. The selection of indicators for SA, TRP, TM, and PM were based on theoretical considerations discussed in Chapter 5 and on the findings from the experimental phase of the dissertation. The manifest indicators are illustrated in Figure 6.1. As reported below, the DMM was modeled with settings (e.g. standard error estimation) in accord with recommendations by Kock (2013, personal communication) using WarpPLS 4.0 SEM analysis software. Model quality was tested (at both first and second levels) using accepted validity and reliability techniques.

6.5.1 Stage One: Testing the Structure of the First Order Measurement Models of the DMM with Random Sample One (No. obs =103).

Developing and evaluating the measurement models required the testing and possible elimination of indicators. While a model with acceptable model fit may result from modeling efforts there is always the risk that some “overfitting” of the model has occurred. Ideally the model is tested on new data, and various levels of model invariance testing are completed to provide evidence that the model structure is robust

across various population groups or samples. In the present work, the sample of 216 observations was randomly split into two groups. Group one, with 103 observations was used to test the DMM and modify its structure according to results of the factor analysis and formative model testing. The SA, TRP, and TM measurement models were tested simultaneously using a least squares algorithm where factor scores were exact linear combinations of “outer” model indicators. The construct validity and reliability of each measurement model is discussed. Construct validity for each measurement model was supported by convergent and discriminant validity indices.

First order measurement model generation. The latent structures for SA, TRP, and TM were examined individually to confirm that each indicator loaded significantly on its respective latent variable. Satisfactory convergent validity in reflective models was evidenced by loadings of at least 0.6 and cross-loadings less than 0.40 (Kline, 2005). As expected from earlier work involving the three levels of SA, the SA1 indicator had a loading below 0.60 (0.567) and was therefore removed from the SA model. The “heading” variable for the TRP latent variable also loaded below 0.60 and was negatively related to the SPEED and ALT measures (-0.475) and was thus removed as an indicator from TRP. The task measures (T1-T3) loaded above 0.70 for the TM latent variable. The oblique-rotated cross-loadings were low for the TM items (< 0.33). Only non-normalized loadings were examined as this permitted examination of relative relationships between indicators and the latent construct when only two indicators reflected one latent construct (Hair, Sarstedt, Ringle, & Mena, 2012).

Table 6.3

Loadings and Cross-loadings of Indicators for the Final First Order Reflective Measurement Models in the DMM (No. obs.=103)

	SA	TRP	TM	PM	Type (as defined)	SE	P value
SA2	(0.896)	-0.046	-0.061	0.020	Reflective	0.070	<0.001
SA3	(0.896)	0.046	0.061	-0.020	Reflective	0.073	<0.001
RmseAlt	0.043	(-0.858)	-0.024	-0.154	Reflective	0.069	<0.001
RmseSpd	-0.043	(-0.858)	0.024	0.154	Reflective	0.073	<0.001
SATPerf	0.014	0.012	(0.921)	-0.072	Reflective	0.062	<0.001
SATSpec	-0.092	0.030	(0.833)	0.035	Reflective	0.073	<0.001
SATPr	0.089	-0.050	(0.714)	0.052	Reflective	0.105	<0.001
PM	0.000	0.000	0.000	(1.000)	Reflective	0.105	<0.001

Note. Loadings (bracketed) are unrotated and cross-loadings (unbracketed) are oblique-rotated. SE= standard error.

Loadings and cross-loadings for the modified measurement model structures are shown in Table 6.3. Loadings between indicators and their respective latent variable were above 0.7 and cross-loadings were below 0.16. The high loadings and low cross-loadings indicated good agreement between the latent variable concepts and the indicator variables chosen to represent these concepts. Standard errors were also examined as evidence of normal distribution densities and to investigate whether lower loadings might be a result of non-normal distributions. The indicator reflecting task priorities (SATPr, or T3 in Figure 6.1) and PM had higher SEs in comparison to the remaining indicators. Not surprisingly, the task priorities measure also had the lowest loading in the TM model.

The quality of each measurement model was evaluated according to coefficients that represented model reliability and validity (see Table 6.4). Standard reliability coefficients were estimated for SA, TRP, and TM. The latent variable, PM, was not examined for reliability because it was represented by just one indicator (the raw scores for the radio communication prospective memory measure). A single indicator always measures its construct perfectly. Two common reliability coefficients are

Cronbach's alpha and composite reliability (CA and CR, respectively). Both CA and CR should be greater than 0.70 (Fornell & Larcker, 1981). The SA and TM latent variables achieved acceptable CA and CR values between 0.76 and 0.89. The TRP construct demonstrated lower reliability with a CA of 0.68 and a CR of 0.85. Because the indicators for RMSE altitude and airspeed both loaded strongly on the TRP construct neither indicator was considered for removal.

Table 6.4

Reliability Coefficients for the Final First Order Reflective Measurement Models in the DMM (No. obs.=103)

	SA	TRP	TM	PM
R-squared				
Adj. R-squared				
Composite reliab.	0.891	0.849	0.865	1.000
Cronbach's alpha	0.755	0.643	0.763	1.000
Avg. var. extrac.	0.803	0.737	0.684	1.000
Full collin. VIF	1.082	1.137	1.042	1.234

Note. R-squared and Adj. R-squared are not calculated for individual reflective measurement models. AVE=Average Variance Extracted. CR= Composite Reliability. CA= Cronbach's alpha. VIF = variance inflation factor. Not Shown: The loading and reliability coefficients for PM are not shown because, as a latent variable with a single indicator, the coefficients were necessarily 1.0.

Convergent validity was evidenced by average variance extracted (AVE) values at or above 0.50 (Fornell & Larcker, 1981), meaning that, on average, most of the variance within indicators were attributed to the latent factor it represented. As shown in Table 6.4, AVEs were acceptable, with coefficients ranging from 0.68 to 0.80. Full collinearity VIFs are <1.3 for all latent variables indicating that risk of bias in loadings due to multicollinearity was low. Discriminant validity was also assessed using the AVE value for each latent variable (Fornell & Larcker, 1981). Evidence for discriminant validity was evidenced when the square roots of the AVEs for each latent

variable were greater than the correlations between latent variables. As shown in table 6.5, the square roots of the AVE (representing the degree of inter-indicator correlation) for each measurement model was greater than the correlation between the latent variables. The pattern of correlations between latent variables and the square roots of AVEs suggested that each latent variable was discriminated empirically from the other latent variables. It was expected that the latent variables associated with the DMM had weak to moderate inter-correlations because each construct represents a distinct element necessary for the DMM composition. As shown in Table 6.5, the intercorrelations between latent variables in the DMM were low, with the PM latent variable demonstrating low but significant correlations with the other three latent variables

Table 6.5
Correlation of Latent Variables with Square Root of AVEs

	SA	TRP	TM	PM
SA	(0.896)	0.146	0.093	0.265
TRP	0.146	(0.858)	0.042	0.341
TM	0.093	0.042	(0.827)	0.195
PM	0.265	0.341	0.195	(1.000)

Note. The AVE for each latent variable is on the diagonal.

Measurement model reliability and validity tests indicated that the SA, TRP, and TM constructs were adequately and uniquely reflected by their respective indicators. Therefore, the factor scores for SA, TRP, and TM were saved as new variables in the dataset for use in phase two of DMM modeling.

6.5.2 Stage Two: Testing the Structure of the Second Order Model of the DMM

As described in Chapter 5, the DMM is modeled formatively, where composites are

theorized to “cause” the DMM. Figure 6.1 illustrates this structure with the arrow directions going *from* SA, TRP, TM, and PM *to* the DMM. After confirming the quality of the measurement models for SA, TRP, and TM, the DMM was modeled using the resultant factor scores from the measurement models. Specifically, in the two-stage process of formative construct modeling the standardized and ranked factor scores for SA, TRP, TM, and PM were saved, added to the dataset, and used as indicators for the DMM. Using “PLS Regression” and “Formative” modes, and jackknifed standard errors (102 samples) DMM factor scores were generated as exact linear combinations of the SA, TRP, TM, and PM factor scores. The DMM composite variable indicator weights and quality coefficients are provided in Table 6.6.

Table 6.6
DMM Composite Variable Indicator Weights

	DMM	Type (as defined)	SE	P value	VIF	WLS	ES
Iv_SA_2	(0.389)	Reflective	0.090	<0.001	1.092	1	0.238
Iv_TRP_2	(0.386)	Reflective	0.084	<0.001	1.116	1	0.235
Iv_TM_2	(0.277)	Reflective	0.174	0.057	1.049	1	0.120
Iv_PM_2	(0.509)	Reflective	0.069	<0.001	1.237	1	0.407

Note. VIF= Variance Inflation Factor. ES=Effect Size. The scale of the TRP was reversed so that the RMSE scores related to the latent variable in the same direction as SA, TM, and PM. The *SEs* were estimated using the “jackknifing” method (resampling =102, as per Kock, 2013).

Evaluating the quality of a composite (formative) latent variable structure requires different indices than those used to evaluate reflective measurement models. Traditionally reflective measurement models are evaluated using the established reliability and validity guidelines associated with factor analysis. Due to the direction of effects that emanate *from* the composite indicator and point *to* the latent variable, and not vice versa, formative measurement models must necessarily be evaluated

using other indices. As per Table 6.2, hypotheses regarding the empirical convergence of the DMM were based on the p -values for composite indicator *weights* (not loadings as evaluated for reflective latent variables), variance inflation factor values (VIF), weight-loading signs and relative effect sizes. Table 6.6 displays the four quality coefficients for the DMM as column headers for each composite indicator (listed by rows). Results demonstrated that the weights associated with SA, TRP, and PM composite indicators were significant, ($p < .05$). TM was marginally significant at $p = .057$. Lateral collinearity was assessed by examining the variance inflation factor (VIF) for each composite indicator of the DMM. As shown in Table 6.6 all VIFs are below 2.0. A threshold of 3.3 has been recommended for formative indicator VIFs (Kock, 2013), although VIFs below 2.0 is a strict threshold. The weight-loading signs (WLS) for each indicator are also +1. Values of -1 indicate instances of a Simpson's Paradox whereby the path coefficient sign does not match the correlation sign between the predictor and criterion variable. Another sign of formative construct quality is the f^2 effect sizes for the composite indicators. As per Cohen (1988), f^2 effect sizes below 0.02 are extremely weak and not worth considering when discussing effects on criterion variables. The effect sizes for the DMM composite variables range from 0.12 to 0.41, which categorizes the effect sizes as small to large (Cohen, 1988). Although the TM weight was marginally significant, TM contributed a significant amount of variance in the DMM, its WLS was positive and its VIF was low. The overall acceptable quality of the TM suggests that it contributed relevant weight with respect to the DMM factor scores. In addition, a theoretical case was made in Chapter 5 of this work regarding the relevance of TM. Therefore, the TM latent variable was retained as a composite variable in the DMM. In regards to the hypotheses outlined in Table 6.2, hypotheses were confirmed, with two exceptions, after the model

generating phase. The two hypotheses not supported were the inclusion of iRES (SA1) and RMSE HEADING in their respective first-order measurement models.

6.5.3 Model Cross-validation Results

The three measurement models and the DMM structure, as generated in phases one and two, were validated by applying the final model structure to the untested second half of the randomly split data file. The model validation process provided confidence that the final measurement models were not overfit to peculiarities in the first half of the dataset. As shown in Table 6.7, the loadings and cross-loadings for each indicator were very similar to the loadings in the final model for the first dataset. All loadings were high and cross-loadings were low. Using the pooled standard error method multigroup measurement invariance was tested using the loadings for the first-order measurement models and the weights for the second-order DMM structure (Kock, 2014).

Table 6.7

Validation of the DMM First Order Measurement Model Structure using the Second Dataset.

	SAii	TRPii	TMii	PMii	Type (as defined)	SE	P value
SA2	(0.895)	0.008	0.109	-0.055	Reflective	0.069	<0.001
SA3	(0.895)	-0.008	-0.109	0.055	Reflective	0.073	<0.001
RmseAlt	0.109	(0.850)	0.042	-0.185	Reflective	0.069	<0.001
RmseSpd	-0.109	(0.850)	-0.042	0.185	Reflective	0.069	<0.001
SATPerf	0.042	0.033	(0.905)	-0.037	Reflective	0.061	<0.001
SATSpec	-0.061	-0.094	(0.870)	-0.033	Reflective	0.061	<0.001
SATPr	0.018	0.059	(0.879)	0.070	Reflective	0.072	<0.001
PM	0.000	0.000	0.000	(1.000)	Reflective	0.085	<0.001

Note. SA, TRP, and TM were labeled along the top row with “ii” to denote the use of the second dataset. Loadings (bracketed) are unrotated and cross-loadings (unbracketed) are oblique-rotated. SE= standard error. Indicator loadings were not significantly different ($p > 0.05$) from corresponding parameters estimated from the first sample.

Table 6.8

Measurement Model Reliability and Validity Coefficients for the Second Dataset.

	SAii	TRPii	TMii	PMii
R-squared				
Adj. R-squared				
Composite reliab.	0.890	0.839	0.915	1.000
Cronbach's alpha	0.752	0.617	0.861	1.000
Avg. var. extrac.	0.801	0.723	0.783	1.000
Full collin. VIF	1.171	1.113	1.095	1.241

Note. Acceptably high AVEs, CR and CA coefficients. *R-squared* and *adjusted R-squared* are not calculated for individual reflective measurement models in models without inner paths.

Table 6.9
DMM Second Order Formative Structure for the Second Dataset.

	DMM	Type (as defined)	SE	P value	VIF	WLS	ES
Iv_SA_2	(0.390)	Reflective	0.088	<0.001	1.159	1	0.253
Iv_TRP_2	(0.381)	Reflective	0.091	<0.001	1.108	1	0.241
Iv_TM_2	(0.319)	Reflective	0.118	0.004	1.088	1	0.169
Iv_PM_2	(0.450)	Reflective	0.064	<0.001	1.218	1	0.337

Note. Weights produced for the second dataset were not significantly different from corresponding weights estimated from the first sample, ($p > 0.05$).

The measurement models for SA, TRP, and TM also demonstrated good reliability and validity, as shown in Table 6.8. The second-order test of formative latent variables in the DMM using dataset two also revealed significant weights, low VIFs, and high effect sizes for each latent variable (see Table 6.9). Testing demonstrated a stronger (in comparison to group one estimates) and significant effect of TM on the DMM, thus confirming the decision to retain TM as a composite element of the DMM.

Multigroup measurement invariance testing was undertaken to examine differences between corresponding parameters estimated in the first and second-order model structures. Pooled standard error t -tests examining the differences in first level loadings, and second level weights between group one and two were all nonsignificant, $p > .05$. The tests of measurement invariance indicated that the first and second-order factor structures were robust for both datasets and that similar parameters were produced in each case. Therefore, the two datasets were combined for further validation testing.

The final model with parameter estimates, as shown in Figure 6.3, suggested that all first-order measurement models were acceptable. The significant weights from SA,

TRP, and TM to the DMM and their low intercorrelations (0.08 to 0.31) indicated acceptable formative specification of the DMM. The details regarding the validation of the DMM using the full dataset are described in the remaining sections.

6.5.4 Measurement Model Validation for SA, TRP, and TM with 216 Observations

The final loadings for indicators in each measurement model were significant at $p < .001$ for all indicators. Table 6.10 provides the individual loadings for each indicator.

Table 6.10
Factor Loadings for Measurement Model Factors for SA, TRP, and TM

	SA	TRP	TM	PM	Type (as defined)	SE	P value
SA2	(0.895)	0.015	0.033	-0.021	Reflective	0.049	<0.001
SA3	(0.895)	-0.015	-0.033	0.021	Reflective	0.051	<0.001
RmseAlt	0.073	(0.854)	0.003	-0.165	Reflective	0.048	<0.001
RmseSpd	-0.073	(0.854)	-0.003	0.165	Reflective	0.050	<0.001
SATPerf	0.030	0.020	(0.908)	-0.051	Reflective	0.043	<0.001
SATSpec	-0.074	-0.073	(0.854)	-0.006	Reflective	0.047	<0.001
SATPr	0.044	0.053	(0.817)	0.064	Reflective	0.061	<0.001
PM	0.000	0.000	0.000	(1.000)	Reflective	0.066	<0.001

Note. Standard errors estimated using Jackknifing methods.

The final quality coefficients for SA, TRP, and TM measurement models demonstrated acceptable model quality as indexed by reliability coefficients (Cronbach's alpha and composite reliability) and validity coefficients (AVE and full collinearity VIFs). Table 6.11 provides the individual loadings for each indicator.

Table 6.11

Quality Coefficients for SA, TRP, and TM with No. Obs. = 216

	SA	TRP	TM	PM
R-squared				
Adj. R-squared				
Composite reliab.	0.890	0.843	0.895	1.000
Cronbach's alpha	0.753	0.629	0.824	1.000
Avg. var. extrac.	0.802	0.729	0.740	1.000
Full collin. VIF	1.118	1.112	1.064	1.228

Note. *R-squared* and adjusted (adj) *r-squared* are not calculated for individual reflective measurement models. Note acceptably high AVEs, CR and CA coefficients. DMM Formative latent variable Quality Indices with No. Obs. = 216

The DMM was modeled again using the reflective-formative multilevel structure. Model quality was assessed using indices appropriate for formative structures (the reflective structures were analyzed as above). The path weights from each composite element of the DMM was significant ($p < .001$), as was the effect size for each element (ESs range from 0.154[moderate] to 0.367[large]). All weight loading signs were positive indicating appropriate direction of causation (from composite element to the DMM). VIFs were also low, demonstrating the unique contribution of each element in the DMM. Estimated quality coefficients are provided in Table 6.12 below.

Table 6.12

Quality Coefficients for the DMM with No. Obs. = 216

	DMM	Type (as defined)	SE	P value	VIF	WLS	ES
Iv_SA	(0.385)	Formative	0.062	<0.001	1.118	1	0.240
Iv_TRP	(0.383)	Formative	0.062	<0.001	1.098	1	0.238
Iv_TM	(0.309)	Formative	0.096	<0.001	1.062	1	0.154
Iv_PM	(0.475)	Formative	0.044	<0.001	1.219	1	0.367

Note. Standard errors estimated using Jackknifing methods.

Figure 6.3 Final DMM structure with loadings, weights, and measurement model quality coefficients. The average variance extracted (AVE) and composite reliability coefficients for SA, TRP, and TM are listed, respectively, under each construct. All loadings and weights were significant ($p < .01$). DMM composite weight loadings signs were positive, and effect sizes were larger than 0.10.

6.5.5 Distribution of DMM Composite latent variable Scores

The factor scores for the DMM are shown in Figure 6.4. The density of scores follows a normal distribution, with no significant skew or kurtosis. This distribution of scores represents two DMM factor scores for each pilot, one score for the low-difficulty, and one for the high-difficulty condition.

Figure 6.4. Distribution of DMM factor scores estimated as exact linear combinations of SA, TRP, TM, and PM.

6.6 Convergence of the DMM in a Covariance-based SEM

The second auxiliary test of DMM reliability and validity was undertaken by simultaneously modeling the full measurement model and the second-order DMM structure using covariance-based SEM (MPlus 7.11, Muthén & Muthén, 1998). While PLS-SEM model fit indices are sufficient for exploring the adequacy of a model, the

validation using CB-SEM provided additional evidence for the theoretical basis for the DMM structure. Thus modeling the DMM using CB-SEM provided an alternate method of confirming the validity of the DMM as a multi-dimensional reflective-formative construct.

A brief description of the different algorithms used by PLS- and CB-SEM demonstrates why the use of both methods was appropriate in the present work. PLS-SEM is often the method of choice when explanations of variance and predictability of variables are of interest (Hair et al., 2012). PLS-SEM is similar to regression analysis as it attempts to maximize the explained variance in a criterion variable by one or more predictor variables (Wold et al., 1984). For these reasons PLS-SEM was selected as the primary method for examining predictive relationships within the DMM and the full model of pilot performance (described in Chapter 7). On the other hand, CB-SEM minimizes the discrepancy between the variance-covariance matrix as specified by the modeler and the observed variance-covariance matrix (Jöreskog, 1967). CB-SEM is typically applied to extend or confirm theories. A difference between CB- and PLS-SEM pertains to factor score determinacy. CB-SEM estimates reflective models using common factor analysis; this produces factor scores that are indeterminate (Velicer & Jackson, 1990). Although with large enough samples this does not normally pose a problem when interpreting the factors (Velicer & Jackson, 1990). CB-SEM also does not utilize a “fixed point” estimation method, meaning that factor scores for latent variables are not stable across various models. Factor scores fluctuate depending on the addition or removal of either up- or downstream variables in the model (Kock, 2013). As mentioned earlier, the fixed point issue was overcome in WarpPLS 4.0 by selecting the “regression” option to “freeze” latent variable scores in the (*as per* WarpPLS 4.0, Kock, 2013). Thus, using regression estimation the inner

path model did not impact the outer model weights. Each estimation method also differed in their ability to model multidimensional constructs. An advantage of CB-SEM is the simultaneous estimation of all levels of a multidimensional model. With current software packages, a two-stage estimation of factor scores is used in PLS-SEM [Becker et al., 2012]). While the two-stage process is slightly more cumbersome, it is a simple process to model the first level of the DMM, save the factor scores for each first level latent factor, and then model the latent variable factor scores as composite indicators for the DMM.

Acknowledgement of the complementary (rather than contrary) nature of the two estimation methods has recently been stressed (Hair et al., 2012). Presumably there are instances where the purposes of a model and the nature of the dataset benefit from evaluation by both estimation methods, thereby providing a richer understanding of the model in question. Incidentally, it seems to be a rare occasion when *only* theory or *only* prediction should be of interest to a modeler. In regards to the DMM, theory development and prediction are both valuable outcomes of data modeling. Moreover, simulation studies have shown that where the data expresses multivariate normality and sample size is adequate, similar parameters are found for both estimation methods (Fornell & Bookstein, 1982; Jannoo, Yap, Auchoybur, & Lazim, 2014). Parameter estimates differ in well-known ways, as per their individual algorithms. Typically, PLS-SEM over-estimates reflective measurement model loadings and underestimates structural path coefficients in comparison to CB-SEM. In contrast, CB-SEM models tend to over-estimate structural path and underestimate measurement model coefficients. This is a well understood trade-off between methods (Hair et al., 2012).

6.6.1 Recommended Constraints on the DMM Specification Using CB-SEM

Three conditions were imposed to maximize the comparison between the CB-SEM and PLS-SEM estimation i) use of the identical standardized and ranked dataset, ii) identical specification of model path structures (with the exception that the PLS-SEM used the two stage method to first estimate first-order factor scores), and iii) the constraining of equal starting weights for each path from first-order latent variable to DMM. The specification of equal starting weights also approximated the similar weights that were estimated using the PLS-SEM method. Determining weights a priori is recommended practice when developing formative models in CB-SEM (Diamantopoulos & Temme, 2013; Lee et al., 2014). The CB-SEM goodness of fit indices determined if the multidimensional model and formative parameter estimates, structured as described, represented a good reflection of the actual variance-covariance matrix. The results of the CB-SEM are illustrated in Figure 6.5 and are discussed below.

Figure 6.5. Final standardized measurement model for the DMM using CB-SEM estimation. Significant correlations between the composite indicators of the DMM are shown as bi-directional arrows between the composite indicators. Note that the correlations between SA and TM, and TRP and TM were not significant. All shown paths are significant at $p < .05$. No. of observations = 216. Appendix L contains output from CB-SEM analysis.

6.6.2 Results of the CB-SEM DMM Estimation

Figure 6.5 illustrates the resulting structure and parameter estimations of the DMM modeled using CB-SEM algorithms. Before overall fit of the model was evaluated, the parameters from each measurement block of the DMM were examined according to

the hypotheses listed in Table 6.2. For example, hypotheses 1a-c referred to the significance of the loadings from the SA construct to each SA manifest variable. It should be noted that the manifest items for the DMM first-order factors that did not load onto their associated latent variable were also not modeled in the CB-SEM analysis (e.g. SA1 was not part of the SA latent variable measurement model). As shown in Figure 6.5, the observed variables pertaining to situation awareness in the present and future both loaded significantly with SA. In fact, as previously shown with PLS-SEM, all manifest variables demonstrated significant loadings from their respective latent variable. The composite indicators of SA, TRP, PM, and TM also demonstrated significant relationships from the latent variable to the higher order DMM. The CB-SEM concurred with the PLS-SEM finding that both the reflective and the formative paths within the DMM were specified adequately. The correlations between first-order factors were also modeled, and they revealed mostly low (but significant) correlations. Concerns regarding the formative nature of the DMM would have been noted if the correlations between the factor scores of the composite indicators were above 0.4 or 0.5. However, in each case the correlations were below 0.40.

CB-SEM Goodness of fit indices for the DMM. The resulting model converged within the maximum number of iterations and was over-identified (as required for estimation) with 29 free parameters and 15 degrees of freedom. The first step in assessing the goodness of fit of a fitted model is the rejection of the baseline model. The baseline model is specified in the analysis with all structural paths assumed to be zero, loadings from latent variables to observed indicators as one (1), and the variance of the latent variable as zero. In a baseline model, only the intercepts and residual variances of the observed endogenous variables are estimated. High *chi-*

square coefficients, with *p*-values below 0.05 support rejection of the baseline model. The baseline model for the DMM was summarily rejected, χ^2 (28, *N*=216)=478.0, *p*<.001.

CB-SEM models are traditionally evaluated using both absolute and incremental fit indices. A common absolute fit index is the *chi-square* test of model fit. The *chi-square* test assumes multivariate normality or data that has been adjusted to account for skew and/or kurtosis. MPlus 7.11 provided a correction factor (the “MLR” estimation) to adjust for possible effects of skew in the sample data. The correction factor produced for this dataset was equal to 1.0, therefore the model was run using the basic maximum likelihood algorithm and not the MLR since non-normality of the indicator distributions was not indicated. The *chi-square* fit index examined the size of the discrepancy between the researcher-specified model covariance matrices and the sample matrices. Lower *chi-square* values and an associated *p* value greater than 0.05 indicated a model with good overall fit with the sample data. The DMM had very good overall fit, χ^2 (15, *N*=216)=15.994, *p*=.382. The root mean square error of approximation (RMSEA) and the standardised root mean square residual (SRMR) measure the square root of the residuals between the sample matrices and specified matrices and good fit is indicated by a value close to zero. Both the RMSEA and SRMR values were adequate at 0.02 (CI: 0.0 – 0.068, *p*=.569) and 0.03, respectively. The incremental fit indices typically used in CB-SEM evaluation tested whether the specified model was significantly different from the baseline model. Two common fit indices are the comparative fit index (CFI) and the Tucker-Lewis Index (TLI) where values below 0.95 indicate poor fit (Schermelleh-Engel et al., 2003). The DMM demonstrated acceptable CFI and TLI, 0.99 and 0.98. Both the absolute and

incremental fit indices suggested that the DMM, specified as a reflective-formative multi-dimensional structure, was a plausible model with adequate representation of the variance-covariance structures in the observed data and significantly better fit than the baseline model.

As discussed earlier, common factor analysis for each measurement model utilized maximum likelihood algorithms to estimate loadings for each indicator. As listed in Table 6.13, each loading was significant at $p < .01$. MPlus 7.11 did not generate an index of reliability for measurement models. However, Cronbach's alpha (C.A.) was calculated by hand as:

$\text{sum}(\text{factor loadings})^2 / (\text{sum}(\text{factor loadings})^2 + \text{sum}(\text{error variance}))$ (as per Cronbach, 1951)

The reliability index (Cronbach's alpha) for the SA, TRP, and TM latent variables were all above 0.7, and are listed in Table 6.13.

Table 6.13
DMM Measurement Model Standardized Estimates from CB-SEM

Manifest Variable	R-Square	Cronbach's alpha	Standardized Loading	Composite Path to DMM
SA LV		0.719		0.449
S2	0.563		0.750	
S3	0.647		0.805	
TRP LV		0.714		0.530
AL	0.785		0.886	
SP	0.268		0.517	
TM LV		0.833		0.454
T1	0.839		0.916	
T2	0.576		0.759	
T3	0.467		0.683	
PM				
Single Indicator		0.599		

Note. All unstandardized paths are significant at $p < .01$ LV= latent variable. S2= SA Level 2. S3=SA Level 3. AL = RSME Altitude. SP= RMSE Airspeed. T1= Awareness of tasks to be performed. T2= Awareness of specific nature of tasks to be performance. T3= Awareness of task priorities. All unstandardized indicator loadings significant at $p < .01$. Cronbach's alpha calculated manually post hoc.

6.6.3 Distribution of DMM Latent Variable Scores as Estimated by Maximum Likelihood

The latent variable scores estimated for the DMM using CB-SEM maximum likelihood estimation are shown in Figure 6.6. The density of scores followed a normal distribution, with no significant skew or kurtosis. This range, spread, and density distribution of scores resembled the distribution of factor scores resulting from the PLS-SEM estimation method using ordinary least squares, and as shown in Figure 6.4.

Figure 6.6. Distribution of DMM factor scores as estimated by maximum likelihood.

6.6.4 Summary of the CB-SEM DMM Validation

The validation process using CB-SEM provided additional evidence for plausible content and structure for a mental model for pilots. Absolute and incremental fit indices demonstrated good fit, such that the manifest variables for each measurement model were appropriately assigned to each latent variable. Importantly, there was convergence of the higher order structure whereby SA, TRP, TM, and PM formed one construct, with each latent variable contributing significantly to the DMM. Low to moderate correlations between the first-order measurement models also suggested a formative, rather than a reflective structure for the DMM. This formative structure supports the theory outlined in Chapter 5 where the DMM was conceptualized as unitary construct that was formed by four domain-dependent cognitive constructs. With the structure and content of the DMM supported by reliability and validity testing, a final process was undertaken to test the notion that the DMM was indeed a construct relevant to aviation outcomes.

6.7 Validating the DMM within a Nomological Net

The meaning and relevancy of the DMM was examined by testing its predictive utility within a nomological net. “A necessary condition for a construct to be scientifically admissible is that it occurs in a nomological net, at least *some* of whose laws involve observables” (Cronbach & Meehl, 1955, p. 281-302). A nomological net was constructed by selecting antecedent and/or effect variables that should have strong associations with the formative construct being evaluated (Cronbach & Meehl, 1955; Diamantopoulos & Winklhofer, 2001; Götz et al., 2010). The direction and strength of the association between the formative construct and the antecedent and criterion variables should reflect what is already accepted knowledge regarding the constructs. Figure 6.7 illustrates the nomological net for the DMM that contains variables that were hypothesized to have logical relationships with the DMM. The DMM was preceded by the antecedent “iRES” and followed by the outcome variable of critical incidents. Hypothetically the DMM had a strong relationship with iRES and predicted a significant amount of variance in number of critical incidents incurred. The *beta* coefficient from iRES to the DMM was 0.49, which indicated that 24% of the variance in the DMM was accounted for by iRES. The magnitude of the *beta* coefficient from the DMM to critical incidents was -0.48, meaning that 23% of the variance in critical incident counts was accounted for by the DMM. The direction of the relationship is also as expected, such that as the mental model performance increased when the number of critical incidents decreased. This example of a nomological net reveals that the DMM “behaved” in the model as expected if the DMM was a relevant predictor of flight performance.

The nomological net test permitted the examination of the relationship of individual components of the DMM to iRES. Figure 6.8 illustrates the individual

relationships that iRES had with the DMM composites. The *beta* coefficients from iRES to SA, TRP, TM, and PM were 0.36, 0.31, 0.24, and 0.35, respectively. The relationship of iRES to individual elements of the DMM demonstrated that all components of the DMM were significantly related to iRES, thus validating the effect from iRES to each DMM component individually.

Figure 6.7. Nomological net for the DMM with simultaneous modeling of the effect of iRES on the individual components of the DMM, for comparison purposes only.

6.7.1 Alternate Nomological Nets

The utility of the DMM was further validated by determining whether the formative modeling of SA, TRP, TM, and PM—as a single construct—resulted in a reduction in the variance accounted for in critical incidents, as compared to a model where each latent variable individually predicted critical incidents. An alternative nomological net is illustrated in Figure 6.8, where each latent variable is individually associated with

critical incidents. The resultant r^2 value for critical incidents is also 0.23. If the variance explained in CRIT by its antecedent had decreased significantly (e.g. by more than 0.02), then the benefits of the parsimony provided by the DMM structure would be offset by a loss in predictive power associated with this multilevel structure.

Figure 6.8 Alternative model with individual prediction of SA, TRP, TM, and PM of critical incidents. Note the r^2 value for critical incidents is not reduced as compared to the r^2 resultant from the formative DMM model in Figure 6.7.

Comparing CB-SEM and PLS-SEM nomological net results. A similar test was undertaken using CB-SEM methods, again using MPlus 7.11. In this analysis, the effect of the DMM on CRIT was estimated using Poisson regression, as is appropriate for count dependent variables (see Chapter 4 for an earlier similar analysis). Results of this analysis found a significant path between the DMM and CRIT ($beta = -0.530$, $p < .001$). These results may be interpreted as, for each one unit increase in DMM there is a corresponding decrease of 0.530 log critical incident counts. To allow direct

comparison, a Poisson regression estimation of the effect of the DMM on total critical incident counts, was undertaken using the PLS-SEM scores derived as exact linear composites of SA, TRP, TM, and PM (DMM factor scores produced by WarpPLS 4.0 were imported into SPSS 22.0 and analyzed using Poisson regression). A strong association as evidenced by the *beta* coefficient of -0.631, $p < .001$, thus indicating that the use of component-based proxy scores did not result in the loss of predictive power of the DMM, as compared to CB-SEM estimation techniques. This result using component scores was also useful for supporting the decision in Chapter 7 to develop a comprehensive model of pilot performance using the PLS-SEM platform.

Validation of the DMM relationships to individual pilot differences and incursion management. A final nomological net-type test compared the Spearman rank correlations of DMM PLS-SEM component and CB-SEM factor scores with other relevant measures of pilot individual differences and the incursion management scores (*rho* coefficients shown in Appendix M). Spearman correlations confirmed that stronger relationships between the DMM and other relevant measures (e.g., pilot age, rating, and recent pilot-in-command hours) were found for the component-based scores, as compared to the CB-SEM factor scores. An exception was the relationship between the DMM and incursion management, which indicated that the CB-SEM algorithms may have performed more efficiently with the bimodal distribution observed for the incursion management scores than PLS-SEM (see Chapter 4, Figure 4.11).

6.8 Summary of the DMM Validation

As summarized in Table 6.14, hypothesis testing indicated that the DMM converged as a multilevel formative construct. First, standard reliability and validity testing revealed that the measurement models for SA, TRP, and TM had acceptable fit.

Second, the original premise by Herdman and LeFevre (2004), that the DMM as a latent variable was indexed by the cognitive constructs of SA and TRP, was validated. As theorized in the present work, TM and PM, two other aviation-related cognitive mechanisms, were also found to be acceptable indices of the DMM, such that the four components together formed one unitary construct. The full formative structure of the DMM demonstrated good empirical fit with four first-order factors each contributing significantly to the model. The validity of the structure of the DMM was evidenced by multiple model cross-validation, where key parameters were found to be invariant across two unique datasets. The CB-SEM model test also demonstrated acceptable model fit indices, such that the variance-covariance matrix as specified by the DMM structure was adequately similar to the actual data structures. Finally, nomological net tests demonstrated construct validity of the DMM, where it predicted and was predicted by relevant variables. The significant role of the DMM in the nomological net pertaining to pilot performance, in conjunction with the theoretical arguments outlined in Chapter 5, support the indexing of a DMM for pilots by SA, TRP, TM, and PM. Chapter 7 outlines the role of the DMM in the larger model of pilot performance where the effects of upstream variables on the DMM and the effects of the DMM on downstream variables were quantified. The analyses in Chapter 7 provide a more definitive understanding of the role of domain-dependent factors in diagnosing pilot risk, and inform current decision-making protocols regarding pilot fitness-to-fly.

Table 6.14
Results of DMM Hypothesis Testing

Model Element	Parameter Estimate Hypotheses	Decision
Stage One		
SA Reflective Loadings TRP Reflective Loadings TM Reflective Loadings	LVs show convergent validity in that manifest variables i) load significantly with their respective latent construct ($p < .05$), ii) have low cross-loadings on other LVs, and iii) have AVE values greater than 5.0.	Accept Caveat: CB-SEM indicated loading of 0.517 for Airspeed on TRP
	Discriminant validity is shown by the square root of the AVE for an LV being larger than the correlation of that LV with other LVs.	Accept
	LVs show good reliability by having high composite reliability ($> .70$) and high Cronbach's alpha ($> .60$).	Accept
Stage Two		
SA, TRP, TM, and PM Composite Indicator Weights		
Composite Indicator Correlations	Composite indicators in a formative model indicate i) relevance by significant weights ($p < .05$), ii) non-redundancy by low full collinearity VIF (< 3.0), iii) proper direction of causality <i>from</i> indicator <i>to</i> formative LV by WLS=1, and iv) adequate causal effect sizes with ES $> .02$.	Accept
	Evidence of a formative (rather than reflective structure) is shown by low correlations between composite elements of the DMM, $r < .5$.	Accept

Chapter 7

Specification and Empirical Validation of a Cognition-Oriented Model of Pilot Performance

The principal outcomes of the present work lie in the results of modeling a plausible structure for a mental model for pilots, and the investigation of the mental model as a predictor of criterion variables highly-relevant to aviation safety. To achieve these goals, the present work was organized according to three key hypotheses that were tested within a comprehensive model of pilot performance. This chapter describes the methods and results of testing the second and third hypotheses in the present work:

1. The four domain-dependent cognitive factors of SA, TRP, TM, and PM converge in a multilevel structure to represent the dynamic mental model (DMM) for pilots.
2. **The DMM, as a domain-dependent variable, accounts for significantly more variance in the aviation outcomes measures than the variance accounted for by domain-independent cognitive factors.**
3. **The DMM fully mediates the effects of cognition, age, expertise, and workload factors on aviation outcomes.**

Results from two main analyses are provided. In the first analysis, model testing was based on 110 observations where measures from the CogScreen-AE (a cognitive health screening tool for pilots) were used to generate a broad set of domain-

independent cognitive constructs. The second major analysis included data from 144 observations aged 40 years and older, where domain-independent measures were generated from the DCAT¹, a cognitive health screening tool designed for older drivers, and from a custom auditory peripheral detection task (PDT).

Figure 7.1 illustrates the procedure for testing the comprehensive model structures. The pilot sub-samples selected for each analysis were determined primarily with the power of the test in mind (Cohen, 1992). For example, the largest possible dataset was selected for the model that included CogScreen-AE variables. Similarly, the incursion management variable was examined in the second analysis where 72 unique scores for pilots aged 40 to 81 were obtained (incursion management scores were only collected in the high-difficulty condition).

Figure 7.1 Order of the main analyses and associated datasets used to test the COMP-A structure. A subset of the sample from analysis two was used to test the COMP-A for subjects aged 60 and older.

¹ Communication with A. Dobbs (creator of the DCAT) suggested that the DCAT might not produce stable results for younger pilots due to their potential for extremely fast response times. The fast responses could be recorded as false alarms by the DCAT software, thereby creating a false positive relationship with age: scores that are theoretically age-normed. A positive correlation between the motor speed and control measure (D1) and age was found, $r(106)=.17$, $p=.07$, $n=108$ suggesting that younger pilots could have been penalized for fast response times. The same correlational analysis was non-significant for pilots aged 40 and older. It was therefore determined to limit the second analysis to pilots aged 40 and older. The exclusion of pilots under the age of 40 also permitted the direct comparison of the present results with the substantial body of work published by the investigators of pilot performance at Stanford University.

7.1 Conception of the COMP-A

The examination of the effects of the DMM on aviation outcomes was undertaken within a comprehensive model of pilot performance. The structure of the model of performance was based on work by Hardy and Parasuraman (1997). This original model was expanded to support a large array of cognition- and system-oriented factors that predicted pilot performance. For ease of reference, the full model of pilot performance was labelled as **COMP-A**: the Cognition-Oriented Model of Performance-Aviation. The name COMP-A reflects the generic nature of the model, such that other complex tasks could be modeled using the same structure (and by modeling an appropriately indexed DMM). For example, a COMP-D model (D for driver) designed to identify at-risk drivers could follow a very similar structure, and include a DMM indexed by driver-related SA, TRP, TM, and PM measures. The COMP-A acknowledges the premise by Hardy and Parasuraman (1997) that both domain-independent and domain-dependent variables be positioned as mediators between individual differences and aviation outcomes. Similarly the DMM, although altered from its original three-factor framework, is influenced by the original DMM as first posited by Herdman and LeFevre (2004). The COMP-A is therefore a modified version of the Hardy and Parasuraman model with the DMM as its central feature. As mentioned, three main hypotheses structured the present work. In addition to testing hypotheses 2 and 3, this chapter also provides empirical support for the specification of latent cognitive variables and the overall structure of the COMP-A.

7.2 Overview of the Two Main Analyses

In both analyses, pilot attributes, domain-independent and domain-dependent cognition, and control variables (mission difficulty and episodic memory) were examined with respect to their association with critical incidents. In the second

analysis, management of the surprise runway incursion was added as an outcome measure. Both analyses also examined two non-causal variables that were hypothesized as plausible predictors (but not effectors) of aviation outcomes. The non-causal predictors were self-rated situation awareness and mental task load (a modified NASA TLX, Hart & Staveland, 1988).

This chapter concludes with three additional statistical analyses. First, a search for instances of homogeneity in the COMP-A was undertaken to determine if causal variables, not present in the COMP-A, might have been omitted from the structural paths. This information was used to explain unexpected relationships found in the model and to point to relevant areas for future research. Second, in an effort to fully interpret the role of the DMM in the COMP-A, explorations of the relationships of SA, TRP, and TM, and PM with other variables in the COMP-A were investigated. The comparison of correlations of SA, TRP, and TM, and PM with antecedents and study outcomes provided greater interpretation of how the DMM functions in the COMP-A. Finally, receiver operating characteristic (ROC) curve analyses permitted a direct comparison between the DMM and other measures currently prescribed to classify pilots as low- or high-risk pilots. The ROC analysis examined the diagnostic utility of the CogScreen-AE measures from the COMP-A and the Logistic Regression Probability Value, a derived CogScreen-AE score (reflecting overall performance), in predicting pilot risk group.

7.3 Domain-Dependent Cognitive Variables in the Model of Pilot Performance

Before discussing the structure of the COMP-A and providing the results of hypothesis testing in the model, a brief discussion of the two domain-dependent cognitive variables in the model is provided below. See Chapter 2 for a full review of

the domain-independent measures of inhibition, visual-spatial attention, working memory, and processing speed.

7.3.1 The DMM

Data modeling, as described in Chapter 6, has thus far generated a formative latent construct representing a plausible mental model for pilots. The DMM is a revised version of the original dynamic mental model first put forth by Herdman and LeFevre (2004). The DMM was modeled as the weighted combination of latent variable scores from SA, TRP, TM, and PM. Like the DMM, SA, TRP, and TM are also latent constructs that required the generation of proxy scores via the development of reliable and valid measurement models. Chapter 6 provided results of the model generating process for SA, TRP, and TM. The premise outlined in Chapters 5 and 6 was that the DMM is a latent construct that serves as a mental model for pilots, and supports basic aviator competencies and safe operation of an aircraft. I posited that a mental model for complex tasks is the mechanism that permits us to perform adequately, as if on “autopilot”. For example, drivers report having driven for substantial amounts of time where they don’t remember being consciously aware of all they did to maintain their goal of staying on the road and getting to their goal location. Presumably, their mind was occupied with other thoughts, yet they remained on task, and on the road. Chapter 5 provided the conceptual arguments for the hypothesis that the DMM construct is comprehensively indexed by SA, TRP, TM, and PM. Thus modeled, the DMM had acceptable reliability and validity. Testing the DMM within a nomological net revealed a significant association with critical incident counts. Furthermore, the DMM was predicted by iRES, a key domain-dependent cognitive factor in pilot performance. The results of the nomological net test demonstrated that the DMM was predicted by

foundational domain-specific cognitive abilities (iRES) and functioned as a predictor of a key aviation outcome variable (critical incidents). Specifically, the higher a pilot's information resolution (iRES), the higher the DMM factor score. In addition, the higher the DMM factor score, the less likely a pilot was to incur a critical event while flying. These results were interpreted such that risk was heightened for pilots who demonstrated poor DMM factor scores.

7.3.2 iRES (“i”= information and “RES” = resolution)

As discussed in detail in Chapter 4, iRES is a measure reflecting the resolution of the information pilots extracted from the environment while flying the two missions. The simulated “world” was populated with information such as *ownership* (heading, airspeed, and altitude), and the *type* and *call-sign* of all other relevant traffic. iRES scores reflected the level of detail pilots could relay regarding this virtual world. To understand iRES, consider the concept of resolution as it pertains to a computer screen. A computer monitor with a million pixels can convey more bits of information than a monitor displaying 500,000 pixels. The amount of information displayed is considered the resolution of the screen. As a group, pilots also displayed varying resolution regarding the detail of information they could convey during the mission. For example, iRES points were obtained by relaying the call-signs of relevant traffic at the airfield. However, empirically, the point system was graded such that more points were awarded in accord with greater levels of correct specificity regarding the call-sign (0 points for no correct details, 1 point for partial correct, and 2 points for fully correct details). In this way, the iRES point system reflected a pilot's resolution of the simulated world. The bits of information relayed by the pilots resulted from

both auditory and visual sources, and both sources were combined to create one iRES score for each mission difficulty condition.

7.3.3 Hypotheses in the COMP-A

The structure of the model was designed to test paths between variables as recommended by Hardy and Parasuraman (1997). Figures 1.1 and 5.1 illustrate the model of pilot performance suggested by Hardy and Parasuraman. Figure 7.2 illustrates the COMP-A model with its 19 hypothesized paths. The “H” assigned to each path represents the hypotheses associated with that path in the structural model and the “+” or “-” associated with the hypothesis reflects the sign of the coefficient associated with that path. A description of each hypothesis is provided in Table 7.1, which includes the expected direction of influence of the predictor on the dependent variable. Variables that are not linked (unless indicated otherwise) are hypothesized to have no causal relationship.

7.4 Description of Exogenous Causal Variables

As shown with Figure 7.2, exogenous variables are not influenced by other variables in the model and are not associated with residual variance estimates (a disturbance term). For example, a directly observed exogenous causal variable in the COMP-A is AGE. A latent exogenous causal variable is EXPERT (pilot expertise). EXPERT is a formative latent variable derived from pilot rating, total hours flown, recent pilot-in-command hours, and simulator hours.

7.4.1 Description of Endogenous Variables

Endogenous variables are influenced by other variables in the model and therefore are associated with error terms (or unexplained variance). The endogenous latent variables

in the COMP-A are the domain-independent latent variables and the DMM (a domain-dependent variable). Single-indicator endogenous variables are iRES, and the outcome measures of critical incidents and incursion management. Other than the outcome measures, the remaining endogenous variables are also predictor variables that are hypothesized to have direct and indirect effects on downstream variables, and might also mediate effects from upstream variables in the COMP-A.

The relationship of age and expertise to domain-independent cognitive variables was examined in both main analyses. Analysis one includes several domain-independent cognitive constructs previously examined with respect to pilot performance. Taylor et al. (2007) predicted that working memory, processing speed, and inhibition were the domain-independent cognitive variables most associated with pilot performance, and the primary mediators of the effects of age on pilot performance. Specifically, older age was thought to negatively impact processing speed and inhibition; slower processing speed, and poorer inhibition and interference control led to working memory declines; and finally, reduced working memory capacity was directly associated with decrements in flight performance. It was proposed that the effects of older age, slower processing speed and poor inhibition indirectly influenced pilot performance via working memory.

In analysis two the more elemental measures of domain-independent cognition collected via the CogScreen-AE were replaced by two domain-independent cognitive variables, yet untested in models of pilot performance. The first was a working memory construct, labelled DCAT in the COMP-A in Figures 7.5 and 7.8. The DCAT was indexed by a span of attentional field measure (a memory and useful field of view task, DCAT subtest 2) and Identification of Driving Situations task (requiring integrated visual and auditory working memory, DCAT subtest 6). The DCAT measure

incorporated age-normed scores and should not be associated with pilot age in the COMP-A. The DCAT includes tasks that are more complex and “realistic” than the computerized working memory tasks from the CogScreen-AE and required pilots to watch short video scenes and make decisions regarding aspects of driving safety in the scenarios. For this reason it was expected that pilot expertise might predict the DCAT. The second measure of domain-independent cognition indexed mental task load via the auditory peripheral detection task. Three manifest indicators reflected the mental task load construct: average response time when the auditory target was detected, individual standard deviation of response times, and the hit rate (% of targets detected). The mental task load measure was the only domain-independent measure to be collected while pilots flew the simulator, and therefore it was hypothesized that it might be predicted by pilot expertise and age. It was also hypothesized that “lower level” cognition, as measured by the mental task load, might predict performance on the more complex DCAT measure.

7.5 Description of Control Variables

The robustness of the COMP-A was examined in light of two control variables: mission difficulty and episodic memory. In the COMP-A, the significance tests and causal directions of the hypothesized structural paths should be stable regardless of the level of workload or a pilot’s retrospective memory. For example, iRES should predict the DMM, regardless of workload condition. Therefore, all measures derived from the simulation protocol were regressed on the level of workload. The “partialling out” of workload and episodic memory effects permits the interpretation that upstream effects of pilot age, expertise and cognition exist regardless of the workload condition. Similarly, the relationships within domain-dependent variables were interpreted as

stable across the low- and high-difficulty conditions. MISSDIFF (mission difficulty) is a single indicator exogenous control variable.

The stability of the COMP-A was tested by controlling for pilots' episodic memory capacity. The ability to maintain accurate representations of past events could have confounded results for measures pertaining to situation awareness. The capacity to relay information about situation awareness was contingent partly on a pilot's retrospective memory mechanisms. Poor episodic memory capacity could interfere with the fidelity of situation awareness measurement. For example, it would not be clear if poor situation awareness scores resulted because a pilot's model was a poor representation of stimuli in the environment, or if a general deficiency in episodic memory impeded the maintenance and relaying of situation awareness information. To control for this confound, an episodic memory control query was applied at the time of each SAGAT testing. The effects from upstream variables on iRES, the DMM, and outcome variables were examined while simultaneously accounting for effects from episodic memory capacity. Likewise, any effects of iRES on the DMM and outcomes, and the DMM on outcomes were interpreted as effects not perturbed by retrospective memory confounds. EPISOD (episodic memory) is a single-indicator exogenous control variable in the COMP-A.

7.6 Description of Additional Predictive but Non-causal Variables

It is possible that some variance in the criterion variables was predicted by non-causal variables. For example, how a pilot rates their own skill, or mental workload is not an "effector" in the COMP-A. The only exceptions to this understanding of non-causal variables were the manifest indicators for TM as a first-order factors in the DMM. However, self-rating for various aspects of pilot competency could potentially explain unique variance in outcome variables. Any factor that assists in identifying pilots at-

risk for a critical incident, for example, should be welcome in a comprehensive model of performance. In the present work, pilot self-rating of situation awareness and mental task load (the NASA TLX) were integral to the flight simulation protocol. These additional non-causal exogenous measures were examined in the second analysis with the sample of 72 older pilots.

7.7 Procedure for the SEM Analysis of the COMP-A

The model in Figure 7.2 labels 19 specific hypotheses that were tested using partial least squares SEM (PLS-SEM). PLS-SEM is a second generation statistical technique that supports the generation and testing of measurement models for latent constructs using factor analysis, within a causal path structure. PLS-SEM was the optimal method for examining direct and indirect effects, and mediating and moderating relationships within a complex structure of latent and directly observed measures (Hair et al., 2012; Kock, 2013). Table 7.1 provides a description of each hypothesis labeled in Figure 7.2. Not labeled in Figure 7.2, but listed in Table 7.1 are four hypotheses pertaining to direct effects of age and expertise on outcome variables, and six hypotheses regarding mediating variables in the model. The stability of the relationships between variables in the model was assessed by simultaneously controlling for the effects of mission difficulty (workload) and pilot episodic memory. In all model tests, the model parameters were estimated using a least squares algorithm (Wold et al., 1984) with the WarpPLS 4.0 software package (Kock, 2013). A “regression” setting was utilized so that latent variable scores in the COMP-A were

determined (i.e. scores were not influenced by other variables that happened to be part of their latent variable block of variables)².

The observed variable raw scores were ranked and standardized, and standard errors were calculated using jackknifed standard error estimation (Kock, 2013). Jackknifing estimation in WarpPLS 4.0 produces stable standard errors for smaller samples and for data with some non-normally distributed variables (Kock, 2013). All measurement models were estimated in an analysis that permitted the “freezing” of latent variable scores. Determined latent variable scores were important when modeling formative indicators in the COMP-A: where DMM and EXPERT reflected formative latent variable scores reflected pilot capacities (traits rather than states). Pilot capacities were hypothesized as stable variables in the COMP-A that would not fluctuate when upstream or downstream variables were added or removed from the structure of the model.

² This aspect of “determined” latent variable scores should not be confused with the separate issue of factor score indeterminacy resulting from common factor analysis methods (Guttman, 1955). Latent variable scores are determined in component-based dimension reduction methods (Velicer & Jackson, 1990), and with PLS methods, as used in the present work.

Table 7.1
Hypothesized COMP-A Path Estimate Hypotheses

Path Estimate Hypotheses	Description
H1–H6	Older age has detrimental effects on performance of domain-independent and domain-dependent cognitive factors.
H7–H9	Slower processing speed, less inhibition, and reduced visual-spatial attention are associated with poorer working memory accuracy.
H10–H11	Slower processing speed, and less inhibition are associated with reduced visual-spatial attention.
H12–H13	Better visual-spatial attention is associated with significantly higher iRES accuracy and better DMM performance.
H14–H15	Higher levels of expertise are associated with significantly higher iRES accuracy and DMM performance
H16	Higher iRES accuracy is associated with significantly better DMM performance.
H17	Better working memory accuracy is associated with better DMM performance.
H18–H19	Better DMM performance is associated with significantly fewer critical incidents and better incursion management.
H20–H21	Higher levels of expertise are associated with fewer critical incidents and better incursion management (not labeled).
H22–H23	Older Age is associated with more critical incidents and but has no direct effect on incursion management (not labeled).
Mediation/Moderation Hypotheses (not labeled)	
A	Total effects of age on outcomes are fully mediated by domain-independent and domain-dependent cognitive factors.
B	Total effects of expertise on outcomes are fully mediated by domain-dependent factors.
C	Total effects of iRES on outcomes are fully mediated by the DMM.
D	Total effects of domain-independent factors on outcomes are mediated by domain-dependent cognitive factors. Corollary: the effects the DMM on the outcomes variables are larger than the combined total effects from the domain-independent variables.
E	Pilot expertise moderates deleterious effects of age on critical incidents. Pilot Age moderates the effect of working memory on incursion management.
Control Variable Hypotheses (not labeled)	
F	Main effects represented by H1–H16 and A, B, C, and D are significant while controlling for mission difficulty and episodic memory.

Figure 7.2. Hypothesized cognition-oriented model of performance-aviation (COMP-A) with CogScreen-AE Domain-Independent Cognitive Measures. Paths from control and non-causal variables to domain-dependent variable and outcomes, and mediating and moderating hypothesis (A-F) are not shown. Variables with a heavy border are latent variables. The measurement models are not shown for the four latent variables. Variables with arrows only going **out** of it are exogenous. Variables with arrows going **into** it are endogenous. The “-” or “+” symbols reflect the hypothesized sign of the path coefficient. Each unit, including the latent variable and its indicators, create a measurement model. Together these units constitute the “outer” models (indicators not shown in this Figure). The directional arrows between the latent and single-indicator variables make up the “inner” path structure, or the structural model. Direct paths between pilot attributes and outcomes are not shown, but were tested. Curved arrows indicate nonlinear relationships between variables were expected.

7.8 COMP-A Analysis One

The first sample used to test the COMP-A was a subset of the full 108 pilots. In this subset ($n = 55$) pilots completed the simulator protocol and both the CogScreen-AE and the DCAT[™] cognitive health screening tools. Earlier, the experimental phase of this work demonstrated that domain-independent cognitive factors such as working memory and attentional field of view had significant relationships with domain-specific cognitive measures and with key study outcomes. Therefore, to maximize the number of cognitive factors available for testing only the sub-sample of pilots who completed the CogScreen-AE, a cognitive assessment battery that produced over 100 raw and derived scores were included in tests of the COMP-A.

For each pilot in the sub-sample of 55 pilots there were two measures for each simulator measure (one from the low-difficulty condition, and one from the high-difficulty condition), providing 110 observations for each measure. Covariates such as age, expertise, and scores from the CogScreen-AE provided 55 observed values. In subsequent path model significance testing, all paths to simulator measures from covariates such as age, expertise, and domain-general cognitive scores required the more stringent $p < .01$ as the cut-off value for significance. This method of significant testing accommodated for any potential inflated relationship that existed between the individual difference covariates and simulator measures due to the small standard errors produced for the covariates. Table 7.2 provides a description of the subsample of $n=55$ pilots utilized in the COMP-A data modeling.

Table 7.2
Description of Subsample of Pilots (n=55) for COMP-A Data Modeling

Pilot Attributes	Mean	Median	Deviation
Age (Years)	48.7	49	15.2
Rating Level	2.6	3	0.9
Recent Pilot-in-Command Hours (past 12 months)	22.4	12	28.9
Total Hours Flown	559.8	240	927.3
Years Licensed	15.2	9	16.3
Simulator Hours	82.5	19	186.6

7.8.1 Outcome statistics for analysis one.

During the simulated flight 32 pilots incurred no critical incidents. Of the remaining 24 pilots 10 incurred 1, 9 incurred 2, 2 incurred 3, and one each incurred 4 and 5 critical incidents. The probability of incurring at least one critical incident was 42%. The range of incursion management scores was 0 to 10. The mode was 10, median was 7, and the arithmetic mean was 5.75.

7.9 Specification of Outer Measurement Models in Analysis One

Before testing the structural paths in the model of pilot performance, measurement (“outer”) models must be constructed and then tested for each latent structure in the “inner” model. Latent variables are constructs that are measured indirectly via their indicators. All the latent variables in the COMP-A, except for the DMM and EXPERT, are reflective latent constructs, meaning the causation flows from the latent construct to the indicator. In contrast, in this work, the DMM and EXPERT are formative latent variables, meaning that the direction of causation starts with the separate elements of the latent variable and flows towards the latent variable. For example, the DMM is “caused” by the linear combination of the weighted sums of its composite parts.

What follows are the results of the measurement model generation, beginning with the major cognitive constructs represented by the CogScreen-AE and the construct representing expertise. Measurement models are tested using factor analysis with methods outlined in WarpPLS 4.0 (Kock, 2013). As is recommended for reflective latent variables, the “Mode A” algorithm was selected and standard errors were produced using the bootstrap algorithm (100 samples, as per Kock, 2013). Sample size necessary for adequate *power of the test* for factor analysis in WarpPLS 4.0 is determined by the larger of either the number of paths running into a variable, or the number of reflective variables, times 10 (Hair, Ringle, & Sarstedt, 2011). Therefore, with a subset sample size of $n=55$ and a model with eight potential manifest indicators the power was considered low, particularly if the effect size is small. Roldán and Sánchez-Franco (2012) have also examined the issue of the power of the test in PLS-SEM and recommend that researchers use the guidelines outlined by Green (1991). According to Roldán and Sánchez-Franco (2012) a power of 0.80 is achieved when the sample size is $n=42$, a maximum of five indicators are utilized, and the effect is large. The risk of low power for the factor analysis meant that one or two indicators could potentially be cast off unnecessarily. Optimizing the power of the test was balanced against the fully specifying reflective measurement models. Because it has also been recommended that only the best indicators, those showing the strongest effects, be retained in measurement models (Hayduk & Littvay, 2012), the higher power was leveraged by selecting the best (in terms of construct and distribution of responses) for the initial measurement model tests. The few instances where more than five indicators were entered into the first measurement model were not considered a threat to the validity of the final measurement models.

“PLS mode A is often referred to as the “reflective” mode. In this mode the outer model weights is calculated through a least squares regression where the latent variable score is the predictor and the indicators are the criteria...The PLS regression ... algorithm iterates until two conditions are met: (a) the weights connecting latent variable scores and indicators are standardized partial regressions coefficients calculated with the indicators as predictors and the latent variables as criteria; and (b) the latent variable scores are exact linear combinations of the indicator scores. In the PLS regression algorithm, the inner model does not influence the outer model. That is, the weights are not influenced by the links connecting latent variables, which are created by the user in Step 4.” (Kock, 2013, p. 22)

7.10 Domain-independent Cognitive Latent Variables

Factor analysis was performed within the WarpPLS 4.0 platform to develop the measurement models for latent constructs that would serve as elements in the COMP-A. The sample of 55 pilots provided unique observations for age, expertise, and the 90 plus raw and derived measures from the CogScreen-AE. Age is clearly a variable measured by the single indicator of years lived, and does not require the generation of a measurement model. Expertise was considered a latent construct because the experience acquired by a pilot was construed as more than their current licensure and rating. Information regarding the length of time pilots had been licensed, their total hours flown, recent pilot-in-command hours, and total hours of simulated flight (all levels/fidelity of flight simulation). Generating a measurement model that was a composite of these indicators introduced a variable in the model that represented a comprehensive and holistic notion of expertise. The generation of the EXPERT variable is discussed later in this section. The large number of measures available from the CogScreen-AE cognitive screening tool permitted the development of four

key domain-general cognitive variables, previously identified in the literature as predictive of pilot performance.

As discussed in Chapter 2, processing speed, inhibition, visual-spatial attention, and working memory have been associated with pilot performance. Both Taylor et al. (2007) and Morrow et al. (2004) found that the deleterious effects of age on pilot performance were partially explained by the concomitant poorer performance by older pilots of domain-general cognitive functions. Specifically it was hypothesized that older age resulted in slower processing speed and poorer cognitive inhibition. The reductions in processing speed and inhibition have a cascade effect of reducing visual-spatial attention (as evidenced by visual tracking ability) and working memory. This chain of cognitive events, thought to explain poorer performance for complex tasks by older adults, has been championed by researcher Salthouse (1992). This sequence of cognitive decline is generally considered an effect of normal aging and affects most individuals to varying degrees with a varying time course.

The measurement models for the domain-independent cognitive constructs were developed from measures produced by the CogScreen-AE. Derived scores produced by the CogScreen-AE, such as “throughput”, “standard deviation”, or “differential” scores, were excluded. Measurement models were developed using a three stage process. First all potential manifest indicators available from the CogScreen-AE that demonstrated a reasonable spread of values were selected as potential indicators for the processing speed (PROSPD), visual-spatial attention (VIS_SPA), and working memory (WM) factors. Selection of the original variables was also based on their descriptions in the CogScreen-AE manual and from personal communication with the developer of CogScreen-AE (Kay, 1995; personal communication, May 2014). As a result 23 variables from the CogScreen-AE output were selected as potential manifest

indicators for the PROSPD, VIS_SPA, and WM latent variables (factors). One measure was recommended to reflect the construct for cognitive inhibition (INHIB). INHIB reflected the perseveration errors from the shifting attention task (see Table 4.5 for details pertaining to CogScreen-AE measures). Because only one measure was selected to represent cognitive inhibition, the generation of a measurement model was not required. The second and third stages for the PROSPD, VIS_SPA, and WM latent variables are described separately below.

Selection of the manifest indicators for PROSPD also began with 16 potential items from the CogScreen-AE. Only those measures that significantly correlated with more than one-third of the other potential PROSPD indicators were retained for the factor analysis. Seven of the 16 processing speed variables correlated significantly with other potential speed variables. The loadings and cross-loadings, as shown in Table 7.3, were acceptable at >0.600 and <0.300 , respectively, for VSCRT, DATDRT, SATACRT, SATINRT, DATSCRT, MTSRT, and PFCRT. The manifest variables for the processing speed latent variable represent correct response times for visual sequencing (alone and as a dual task), monitoring of a moving stimulus, shifting attention, match-to-sample, and letter-number sequencing tasks.

The four indicators selected to represent visual tracking loaded significantly on the VIS_SPA construct (>0.60). Cross-loadings from VIS_SPA indicators to other latent variables were sufficiently low (< 0.40). The DTTAHIT indicator loaded the lowest at 0.657 and also showed the highest cross-loadings with the other domain-independent cognitive variables of WM, PROSPD, and INHIB.

Three indicators were tested in the original measurement model representing the inhibition/interference cognitive variable and tested with the CogScreen-AE shifting attention task: SATDIFAI, SATDIRUL, and SATDIPER. As just listed, the measures

reflect the number of failures to maintain a set, rule shifts successfully completed, and perseverative errors occurring at rule shifts in the “discovery” portion of this test. The present results found that although the measures pertaining to the number of rule shifts completed and the failures to maintain a set strongly loaded on one component (both > 0.900), the measure of perseverative errors did not load significantly (<.400). These results are similar to the principal components analysis conducted by Taylor et al. (2005).

For the WM latent variable, the second stage for selecting the manifest variables involved examining the correlational relationships of manifest variables. First, because the backward digit span accuracy measure was determined a priori to represent WM, only other manifest variables that correlated with the backward digit span measure would be considered for inclusion in the final measurement model for WM. Second, each variable had to be significantly correlated with at least one-quarter of the other variables within the latent variable to be retained for the factor analysis. As discussed in Chapter 4, reflective measurement models, such as those for working memory or processing speed should be represented by indicators that are significantly correlated with one-another. Of the original 16 potential manifest indicators five met the predetermined criteria and were selected for factor analysis. These variables included BDSACC, MATHACC, PFCACC, SDCIRACC, and SATINACC. The loading and cross-loadings of each manifest variable is provided in Table 7.3. Each manifest variable in WM showed significant loadings ($p < .001$) that exceeded the recommended value of 0.60. Cross-loadings from the WM indicators on the PROSPD and INHIB latent variables were acceptably low (all < 0.30). The WM manifest indicators represent accuracy in backward digit span, simple arithmetic operations,

letter-number sequencing, and manipulation of rules in a shifting attention task (e.g., selecting symbols based on colour versus orientation).

Table 7.3

Unrotated Factor Loadings and Oblique-rotated Cross-loadings for Indicators Representing Latent Variables for Working Memory, Processing Speed, Visual-Spatial Attention, and Inhibition.

	PROSPD	WM	INHIB	VIS_SPA	Type (as defined)	SE
VSCRTC	(0.768)	0.058	0.160	0.154	Reflective	0.072
DATDRTC	(0.694)	0.076	0.061	-0.001	Reflective	0.086
SATACRT	(0.760)	0.037	-0.066	-0.141	Reflective	0.067
SATINRT	(0.806)	-0.086	-0.007	-0.021	Reflective	0.061
DATSCRT	(0.733)	0.031	0.085	0.227	Reflective	0.076
MTSRTC	(0.697)	0.187	0.027	-0.151	Reflective	0.077
PFCRTC	(0.599)	-0.349	-0.319	-0.091	Reflective	0.088
BDSACC	-0.074	(0.745)	-0.045	-0.050	Reflective	0.074
MATHACC	-0.184	(0.744)	0.070	0.222	Reflective	0.083
PFCACC	-0.056	(0.760)	0.229	-0.046	Reflective	0.141
SDCIRAC	0.106	(0.612)	-0.181	-0.023	Reflective	0.084
SATINAC	0.256	(0.666)	-0.123	-0.118	Reflective	0.113
SATFAIL	-0.107	-0.172	(0.909)	-0.024	Reflective	0.082
SatRuleR	0.107	0.172	(0.909)	0.024	Reflective	0.040
DTTAABS	0.017	-0.100	-0.170	(0.857)	Reflective	0.061
DTTAHIT	-0.153	-0.523	-0.250	(0.657)	Reflective	0.126
DTTDABS	0.124	0.260	0.163	(0.879)	Reflective	0.066
DTTDHIT	-0.028	0.231	0.193	(0.868)	Reflective	0.066

Note. No. Obs. = 110. Parentheses show indicators in the rows associated with the latent variable at the top of the column. WM = working memory. PROSPD= processing speed. VIS_SPA= visual-spatial attention. INHIB= inhibition. SE= standard error.

The quality of the latent variables produced for WM, PROSPD, INHIB, and VIS_SPA was tested using the quality statistics produced by WarpPLS 4.0 statistical package, as shown in Table 7.4. The average variance extracted (AVE) associated with each construct should be above 0.5, the VIF below 3.3, and Cronbach's alpha and composite reliability above 0.7. The AVE statistic reflects the average strength of the relationship of each indicator with the latent construct. The AVE for PROSPD, WM, INHIB, and VIS_SPA were acceptable at 0.526, 0.500, 0.827, and 0.673. Cronbach's

alpha and the composite reliability coefficients were acceptable, ranging from 0.833 to 0.905, indicating high average degrees of relatedness between indicators within each latent variable (Cronbach, 1951; Sijtsma, 2009). The measurement models of WM, PROSPD, INHIB, and VIS_SPA demonstrate VIFs less than 3.3, indicating that no significant multicollinearity existed in the measurement models (Kock & Lynn, 2012).

Table 7.4
Quality Coefficients for Domain-independent Cognitive Latent Variables

	PROSPD	WM	INHIB	VIS_SPA
R-squared				
Adj. R-squared				
Composite reliab.	0.885	0.833	0.905	0.891
Cronbach's alpha	0.848	0.748	0.790	0.834
Avg. var. extrac.	0.526	0.500	0.827	0.673
Full collin. VIF	1.535	1.573	1.213	1.540

Note. No. of observations= 110. Full coll. VIF= Full collinearity variance inflation factor. R-squared and Adjusted R-Square are not calculated for measurement models.

Correlations between all indicators for the domain-independent cognitive latent variables are shown in Table 7.5. As expected, moderate to high correlations were found between indicators within one measurement model. The backward digit span accuracy, visual sequence reaction time, and the visual tracking indicators tended to show weak to moderate correlations with most other indicators. Discriminant validity for the latent variables is discussed and is shown in Table 7.7.

Table 7.5

Pearson Correlations between all CogScreen Measurement Model Indicators

	VSCRTC	DATDRTC	SATACRT	SATINRT	DATSCRT	MTSRTC	PFRCRTC	BDSACC	MATHACC	PFACC	SOCIRAC	SATINAC	SATFAIL	SaRuler	DTTAABS	DTTAHIT	DTTABS	DTTDHIT
VSCRTC	1.000	0.404	0.527	0.515	0.644	0.412	0.341	-0.367	-0.260	-0.307	-0.147	-0.263	0.187	0.237	0.366	0.204	0.471	0.465
DATDRTC	0.404	1.000	0.304	0.628	0.430	0.427	0.335	-0.217	-0.337	-0.119	-0.103	-0.224	-0.017	0.264	0.234	0.179	0.384	0.362
SATACRT	0.527	0.304	1.000	0.601	0.383	0.563	0.440	-0.288	-0.185	-0.319	-0.098	-0.002	0.132	0.067	0.431	0.208	0.321	0.122
SATINRT	0.515	0.628	0.601	1.000	0.484	0.376	0.423	-0.232	-0.333	-0.450	-0.181	-0.210	0.093	0.242	0.377	0.335	0.443	0.281
DATSCRT	0.644	0.430	0.383	0.484	1.000	0.455	0.291	-0.439	-0.281	-0.282	-0.260	-0.118	0.086	0.209	0.379	0.219	0.514	0.512
MTSRTC	0.412	0.427	0.563	0.376	0.455	1.000	0.322	-0.178	-0.170	-0.111	-0.173	0.042	0.117	0.061	0.185	0.160	0.266	0.246
PFRCRTC	0.341	0.335	0.440	0.423	0.291	0.322	1.000	-0.298	-0.416	-0.457	-0.032	-0.061	-0.002	0.036	0.396	0.329	0.235	0.223
BDSACC	-0.367	-0.217	-0.288	-0.232	-0.439	-0.178	-0.298	1.000	0.438	0.418	0.397	0.349	-0.350	-0.228	-0.326	-0.376	-0.281	-0.270
MATHACC	-0.260	-0.337	-0.185	-0.333	-0.281	-0.170	-0.416	0.438	1.000	0.540	0.231	0.360	-0.323	-0.211	-0.165	-0.304	-0.200	-0.177
PFACC	-0.307	-0.119	-0.319	-0.450	-0.282	-0.111	-0.457	0.418	0.540	1.000	0.309	0.359	-0.226	-0.126	-0.466	-0.505	-0.249	-0.220
SOCIRAC	-0.147	-0.103	-0.098	-0.181	-0.260	-0.173	-0.032	0.397	0.231	0.309	1.000	0.328	-0.364	-0.214	-0.089	-0.183	-0.216	-0.234
SATINAC	-0.263	-0.224	-0.002	-0.210	-0.118	0.042	-0.061	0.349	0.360	0.359	0.328	1.000	-0.312	-0.264	-0.207	-0.349	-0.162	-0.216
SATFAIL	0.187	-0.017	0.132	0.093	0.086	0.117	-0.002	-0.350	-0.323	-0.226	-0.364	-0.312	1.000	0.653	0.043	0.104	0.053	0.067
SaRuler	0.237	0.264	0.067	0.242	0.209	0.061	0.036	-0.228	-0.211	-0.126	-0.214	-0.264	0.653	1.000	-0.044	-0.015	0.149	0.154
DTTAABS	0.366	0.234	0.431	0.377	0.379	0.185	0.396	-0.326	-0.165	-0.466	-0.089	-0.207	0.043	-0.044	1.000	0.643	0.610	0.567
DTTAHIT	0.204	0.179	0.208	0.335	0.219	0.160	0.329	-0.376	-0.304	-0.505	-0.183	-0.349	0.104	-0.015	0.643	1.000	0.316	0.328
DTTABS	0.471	0.384	0.321	0.443	0.514	0.266	0.235	-0.281	-0.200	-0.249	-0.216	-0.162	0.053	0.149	0.610	0.316	1.000	0.874
DTTDHIT	0.465	0.362	0.122	0.281	0.512	0.246	0.223	-0.270	-0.177	-0.220	-0.234	-0.216	0.067	0.154	0.567	0.328	0.874	1.000

Note. No. observations = 110. Correlations above 0.17 are significant at $p < .05$.

Table 7.6

Discriminant Validity for PROSPD, WM, INHIB, and VIS_SPA.

	PROSPD	WM	INHIB	VIS_SPA
PROSPD	(0.725)	-0.449	0.191	0.539
WM	-0.449	(0.707)	-0.403	-0.440
INHIB	0.191	-0.403	(0.909)	0.087
VIS_SPA	0.539	-0.440	0.087	(0.821)

Note. INHIB was not significantly correlated with VIS-SPA. All other correlations, $p < .05$.

7.11 Measurement Models for Self-Rating Scales

Three latent variables representing non-causal, but potentially predictive factors were included in COMP-A analysis one. The measures included self-rated task load and self-rated situation awareness. The three scale items representing task management (part of the DMM) were not included in this factor analysis. An item representing SA for “awareness of the impact of other aircraft on flying the perfect circuit” was posited as single-item non-causal variable (see VanBenthen & Herdman, 2014b). This single self-rated SA item (S4) was not included in the self-rated SA measurement model development. All six indicators for the SELF_WL loaded above 0.60. Six of the possible seven items loaded on SELF_SA. The item pertaining to awareness of the ability to fly the “perfect circuit” loaded less than 0.60 and was removed from the final model. Cross-loadings for all indicators were low, < 0.3200 . The AVEs for SELF_WL and SELF_SA were 0.656, 0.740. SELF_WL, and SELF_SA demonstrated low variance inflation, with all VIFs < 3.3 , indicating that no significant multicollinearity existed in the measurement models (Kock & Lynn, 2012). Latent variable discriminant validity was tested by confirming that the square root of the AVE associated with each latent variable was greater than any correlation between that latent variable and other variables in the model. Model quality indices demonstrated

acceptable reliability with Cronbach's alpha and composition reliability ranging from 0.89 to 0.94 (Table 7.9).

7.12 The Expertise Formative Measurement Model Development

Confounds between pilot age and experience, and the possibility of contrasting effects are difficult to overcome in aviation research. Although older age tends to be associated with poorer performance on cognitive tasks and flight summary scores, progression through the lifespan affords opportunities to generate a rich repertoire of past experience and training. It was the goal of this work to generate a measurement model that represented a comprehensive and holistic notion of expertise. Ericsson and Lehmann (1996) reported that expertise results from years of deliberate practice, and is not merely the additive influence of time engaged in the activity. In light of this explanation of expertise, a measurement model was generated, which included total and recent hours as pilot-in-command, simulated practice time, and the standard measure of licensure and rating. Five possible measures of pilot expertise were available as formative indicators for EXPERT (pilot licensure and rating, recent pilot-in-command hours, total hours flown, total years licensed, and flight simulator hours). The variable reflecting total years licensed was eliminated as a potential indicator for expertise because it showed a strong positive correlation with age and a moderate negative correlation with flight performance (e.g., critical incidents), $r(53) = 0.68$, $p < .001$, and $r(53) = -0.28$, $p < .01$, respectively. As potentially conflated with pilot age, and negatively associated with pilot performance, the number of years licensed did not represent pilot expertise in the same manner as pilot rating or total hours flown. In regards to the specification of the EXPERT measurement model, it was appropriately structured as a *formative* variable because each composite indicator contributed different, but important aspects to the EXPERT construct. Together, the four

composite indicators formed a comprehensive construct representing pilot expertise. The four remaining indicators for expertise that were represented by pilot licensure and rating, recent pilot-in-command hours, total hours flown, and simulator hours loaded significantly on the EXPERT construct. The EXPERT latent variable was tested using formative model quality indices that included indicator weights, *p*-values, variance inflation factors, weight loading signs, VIFs, and effect sizes.

Table 7.7
Quality Indices for Formative Model for Pilot Expertise

	EXPERT	Type (as defined)	SE	P value	VIF	WLS	ES
Rating	(0.414)	Formative	0.040	<0.001	2.391	1	0.370
PIC	(0.287)	Formative	0.074	<0.001	1.193	1	0.178
Hours	(0.375)	Formative	0.056	<0.001	1.984	1	0.305
Sim	(0.261)	Formative	0.092	0.003	1.294	1	0.147

Note. No. obs. = 110. Rating = pilot licensure and rating. PIC = recent pilot-in-command hours. Hours= total hours flown. Sim = total simulator hours. SE= standard error. WLS= weight loading sign. Standard errors estimated using jackknifing (as per Kock, 2013).

As shown in Table 7.7, each element of EXPERT had significant weights ($p<.01$), small VIFs, and substantial effect sizes. The weight loading sign was positive for each formative indicator, suggesting that the direction of causation (from indicator to factor) was supported. Thus the final formative measurement model for EXPERT was defined by pilot licensure and rating, recent pilot-in-command hours, total hours flown, and simulator hours.

7.13 Discriminant Validity of Latent Variables in the COMP-A Analysis One

Evidence of discriminant validity was found by comparing the square-roots of the average variance extracted for each latent variable with all correlations involving that

latent variable (see Table 7.4 and 7.7). Discriminant validity is shown when the square-root of the average variance extracted (AVE) for each latent construct is greater than its correlation with other factors in the model, thus demonstrating that the indicators are well-situated in their present measurement models. Table 7.7 illustrates the acceptable discriminant validity for EXPERT, WM, PROSPD, INHIBITION, and VIS_SPA. Also shown in Table 7.7 are the correlations of each latent variable with the variables representing pilot age, iRES, mission difficulty, and episodic memory, and the DMM. All measurement models show acceptable discriminant validity such that no latent variable has a square-root of the AVE smaller than its correlation with any other variable.

Table 7. 8

Discriminant Validity of Constructs in the COMP-A Analysis One shown by Correlation among Latent Variables with Square-roots of AVEs.

	EXPERT	AGE	iRES	PROSPD	WM	INHIB	VIS_SPA	DMM	MISSDIFF	EPISOD	CRIT_INC	SELF_WL	SELF_SA	SR_S4
EXPERT	(0.735)	0.135	0.393	0.008	0.114	-0.144	-0.065	0.323	0.000	0.145	-0.289	-0.145	0.314	0.375
AGE	0.135	(1.000)	-0.275	0.621	-0.316	0.026	0.636	-0.338	-0.000	-0.119	0.211	0.050	-0.013	0.052
iRES	0.393	-0.275	(1.000)	-0.283	0.215	-0.062	-0.310	0.525	-0.365	0.424	-0.399	-0.358	0.441	0.297
PROSPD	0.008	0.621	-0.283	(0.725)	-0.449	0.191	0.539	-0.299	0.000	-0.024	0.156	0.178	-0.213	-0.198
WM	0.114	-0.316	0.215	-0.449	(0.707)	-0.403	-0.440	0.333	0.000	0.003	-0.108	0.025	0.102	0.145
INHIB	-0.144	0.026	-0.062	0.191	-0.403	(0.909)	0.087	-0.255	0.000	0.000	0.145	-0.102	-0.228	-0.152
VIS_SPA	-0.065	0.636	-0.310	0.539	-0.440	0.087	(0.821)	-0.434	-0.000	-0.095	0.236	0.205	-0.141	-0.107
DMM	0.323	-0.338	0.525	-0.299	0.333	-0.255	-0.434	(0.672)	-0.357	0.270	-0.539	-0.231	0.511	0.348
MISSDIFF	0.000	-0.000	-0.365	0.000	0.000	0.000	-0.000	-0.357	(1.000)	-0.292	0.237	0.419	-0.522	-0.348
EPISOD	0.145	-0.119	0.424	-0.024	0.003	0.000	-0.095	0.270	-0.292	(1.000)	-0.327	-0.214	0.116	0.014
CRIT_INC	-0.289	0.211	-0.399	0.156	-0.108	0.145	0.236	-0.539	0.237	-0.327	(1.000)	0.151	-0.268	-0.189
SELF_WL	-0.145	0.050	-0.358	0.178	0.025	-0.102	0.205	-0.231	0.419	-0.214	0.151	(0.810)	-0.473	-0.297

Note. Square-root of average variance extracted on the diagonals. A latent variable has discriminant validity when the square-root of that latent variable's average variance extracted is larger than its correlation coefficient with any other variable. PROSPD=processing speed. WM= working memory. INHIB= Inhibition/Interference. VIS_SPA= visual-spatial attention. MISSDIFF= mission difficulty. EPISOD = episodic memory. CRIT_INC= critical incidents. Correlations above 0.18 are significant at $p < .05$. Cells with (1.000) values indicate variables with single indicators (e.g., age, iRES).

7.14 Testing the Inner Path Structure of the COMP-A

Once the measurement models for the key elements of the model of pilot performance were determined the structural relationships between each variable, as specified in Figure 7.2, was tested. Figure 7.2 illustrates the hypothesized causal paths in the comprehensive model of pilot performance. Table 7.1 lists the specific hypotheses associated with each predicted path in the model.

Three combinations of pilot groups and outcomes were tested in the COMP-A. Where paths were possible, the three model exemplars tested the hypotheses listed in Table 7.1. The first COMP-A model includes data from 55 pilots who completed the CogScreen-AE cognitive health screening tool. This sample tests the relationship of four key neuropsychological constructs to domain-dependent cognitive variables and the outcome variable of critical incidents. Taylor et al. (2005) suggested that working memory and inhibition could be explained by processing speed mechanisms. This hypothesis was directly tested in the COMP-A. The second and third tests of the COMP-A utilize a larger sample of pilots aged 40 to 81 years who completed the DCAT cognitive health screening tool. Tests two and three determine the predictive utility of the DCAT for critical incidents and incursion management, respectively. All model testing provided opportunities to investigate the DMM as a predictor of aviation outcomes and a mediator of upstream effects on outcomes. The total relative effect sizes (direct and indirect) on the DMM and on aviation outcomes are presented for each predictor in the model. The third test using the COMP-A examined the predictive utility of three workload constructs, and pilot self-ratings of situation awareness.

7.15 Curvilinear relationships between Latent Variables

Once the models were prepared for each variable the “outer” portion of the model was complete. The second step in developing the model involved specifying the inner path structure. This analysis also examined inner paths to determine if they were best equated with a linear or curvilinear relationship. Every inner path in the structural model was tested to determine whether a curvilinear equation best reflected that particular predictor→ criterion variable relationship. Thus, not all *beta* coefficients represented linear relationships. In fact, most relationships were best described as curvilinear. All reflective and formative associations between latent constructs and indicators were linear in nature. As shown in all the COMP-A figures that follow, curved connectors indicate non-linear relationships and straight arrows represent linear relationships. Figure 7.3 illustrates the non-linear relationship between AGE and PROSPD.

Figure 7.3. Curvilinear relationship between AGE and PROSPD. Shown is the best-fitting curve with segment slope coefficients and associated p -values. Plot produced by WarpPLS 4.0 (Kock, 2013).

7.16 Distribution and Normality of Measures in the COMP-A

The distribution of single-indicator and factor scores were examined for both COMP-A analyses. As was discussed in Chapter 6, PLS-SEM algorithms are robust to moderate amounts of non-normality in the data (Goodhue et al., 2012; Kock, 2013). Responses for each measure were transformed into z-scores with a mean of zero and a variance of one (using SPSS 22.0). Each series was also ranked to minimize the effects of outliers. Details regarding the range and means for the raw, unstandardized and unranked data are found in Chapter 3.

7.17 Model Fit and Quality Indices

Model fit and quality indices were produced for each COMP-A analysis. The emphasis on explanations of variance is in accord with the primary function of PLS-SEM, which is to test models that function as predictors of criterion variables of interest. In regards to the model of pilot performance, the interest lies in understanding what variables predict the DMM, and also how well the DMM predicts the two outcome variables of critical incidents and incursion management. Model quality for PLS-SEM models is indexed via coefficients that pertain to the predictive value of upstream variables in the model. The following are indices of predictive utility of a model:

- Average path coefficient (APC), $p < .05$
- Average R-squared (ARS), $p < .05$
- Average adjusted R-squared (AARS), $p < .05$
- Average block VIF (AVIF), acceptable if ≤ 5 , ideally ≤ 3.3
- Average full collinearity VIF, acceptable if ≤ 5 , ideally ≤ 3.3
- Tenenhaus GoF (GoF), small ≥ 0.1 , medium ≥ 0.25 , large ≥ 0.36

Ideally the average path coefficient and the r -squared values are significant, indicating that predictor variables explain significant amounts of variance in criterion variables. The variance inflation factors should be small, demonstrating adequately low levels of multicollinearity. Low levels of multicollinearity suggest that the strength of relationships between predictor and criterion variables has not been inflated due to correlations between other predictors within the measurement models. The Tenenhaus goodness of fit index is an overall indication of how well independent variables are explaining the variance in their associated criterion variables (Kock, 2013).

7.18 Testing Causal Assumptions in the COMP-A

In SEM structures, the arrows between variables represent assumptions regarding causal direction (Pearl, 2012). For example, in the COMP-A, pilot age is assumed to cause changes in domain-independent cognition, and domain-independent cognition effects change in the DMM. Two indices representing the plausibility of each causal assumption in the model were examined for each inner path in the COMP-A: the *r*-squared contributions associated with each path, and the path-correlation signs (where negative values indicate problem paths such as implausible or reversed causal direction). Four model indices also indicated the overall quality of causal specification in the model: Sympton's paradox ratio (SPR), acceptable if ≥ 0.7 , ideally = 1; *r*-squared contribution ratio (RSCR), acceptable if ≥ 0.9 , ideally = 1; statistical suppression ratio (SSR), acceptable if ≥ 0.7 ; and, nonlinear bivariate causality direction ratio (NLBCDR), acceptable if ≥ 0.7 . For example, the NLBCDR tests whether, overall, if reversed links resulted in smaller residual terms associated with the criterion variable. Generally, after testing a bivariate relationship in *both* causal directions, the variable with the smaller residual term is considered the dependent variable, since more of its variance was explained by the predictor than when the link was reversed (Kock, 2013). These causal coefficients are presented with the model fit and quality report for the three COMP-A models presented below.

Misconceptions regarding what SEM analyses can ascertain regarding causal relationships abound (Pearl, 2012). Just as significant correlations do not imply causation between variables, neither do significant *beta* weights, *on their own*, signify sure causal relationships. CB-SEM and PLS-SEM models that specify path arrows to indicate causal relationships do so only under certain assumptions. All paths should be theory-driven and tested for reversed relationships (as the NLBCDR index does).

While path arrows denote the direction of causation, actual claims regarding causal directions should be made cautiously. The presence of endogeneity may undo causal assumptions in a model (Kock, 2013) as would implausible or “theory-poor” directed links. For this reason I have explicitly noted that pilot self-ratings may be predictive of performance, but are not causal in nature. It is not logical that self-ratings would actually cause anything. On the other hand, it is plausible that domain-dependent cognitive performance is caused by general purpose cognitive mechanisms.

7.19 Results of COMP-A Test: Analysis One

Figure 7.4 displays the results of testing the hypothesized structure of the COMP-A, including measurement model and inner path results.

Figure 7.4. Results of measurement model and structural path testing for COMP-A with critical incidents as the outcome variable. No. of Observations = 110. Dashed arrows between variables indicate non-significant direct path coefficients. Curved arrows indicate nonlinear relationships between variables. Outer measurement models for each latent variable are shown as indicated by grey arrows and pale yellow indicator boxes. Directional black arrows between variables indicate the structural (inner) path model. Values shown with each predictor–criterion variable pair are *beta* coefficients that are standardized partial regression coefficients. At the top right of each endogenous latent variable is its total r^2 value (total variance explained by all direct and indirect effects from upstream variables). Near bottom left inset demonstrates effects from control variables on Critical Incidents, iRES, and the DMM (paths not shown on the larger COMP-A model). Bottom variables demonstrate that no additional variance in critical incidents is explained by the self-rating scale constructs.

7.20 Model fit and quality indices for the COMP-A (Analysis One)

The following list is standard output for model fit and quality from the WarpPLS 4.0 software. The quality indices are estimated from coefficients primarily pertaining to explained variance (Kock, 2013).

- Average path coefficient (APC)=0.212, $P=.005$
- Average R-squared (ARS)=0.415, $P<.001$
- Average adjusted R-squared (AARS)=0.392, $P<.001$
- Average block VIF (AVIF)=1.454, acceptable if ≤ 5 , ideally ≤ 3.3
- Average full collinearity VIF (AFVIF)=2.021, acceptable if ≤ 5 , ideally ≤ 3.3
- Tenenhaus GoF (GoF)=0.569, small ≥ 0.1 , medium ≥ 0.25 , large ≥ 0.36
- Simpson's paradox ratio (SPR)=0.933, acceptable if ≥ 0.7 , ideally = 1
- R-squared contribution ratio (RSCR)=0.998, acceptable if ≥ 0.9 , ideally = 1
- Statistical suppression ratio (SSR)=1.000, acceptable if ≥ 0.7
- Nonlinear bivariate causality direction ratio (NLBCDR)=0.967, acceptable if ≥ 0.7

The predictive utility of the COMP-A was demonstrated by acceptable average path coefficients, 0.212, $p<.001$, and average adjusted r -squared, 0.392, $p<.001$. A second index of the explanatory power of a model is the Tenenhaus GoF (Tenenhaus et al., 2005). GoF values below 0.1 indicate that a number of variables in the model are poorly explained by their block of indicators. The COMP-A has a GoF of 0.569, suggesting the model has more than adequate explanatory power. As shown in Figure 7.4, the blocks of predictors for visual-spatial attention, iRES, and the DMM explained between 45% and 54% of the variance associated with their latent variable.

Results also indicated that the direction of causation (e.g., *from* AGE to PROSPD) is justified for all inner paths in the COMP-A, as evidenced by the acceptable NLBCDR value of 0.967. One implausible suggestion was returned in this causation analysis: the reversal of the path between age and Inhibition/Interference.

The latent variable quality coefficient table also contains additional elements that pertain to model quality. Q-squared coefficients (also known as the Stone-Geisser coefficient) that are above zero indicate that the endogenous variable has predictive value in the model (Kock, 2013). Q-squared coefficients range from 0.135 to 0.539, indicating adequate value for all predictor variables.

Table 7.9
Latent Variable Quality Coefficients

	EXPERT	AGE	RES	PROSPD	WM	INHIB	VIS_SPA	DMM	MISSDIFF	EPISOD	CRIT_INC	SELF_WL	SELF_SA	SR_S4
R-squared			0.504	0.404	0.458	0.146	0.448	0.538			0.407			
Adj. R-squared			0.480	0.399	0.438	0.130	0.438	0.506			0.354			
Composite relab.	0.819	1.000	1.000	0.885	0.833	0.905	0.891	0.765	1.000	1.000	1.000	0.919	0.944	1.000
Cronbach's alpha	0.701	1.000	1.000	0.848	0.748	0.790	0.834	0.591	1.000	1.000	1.000	0.892	0.926	1.000
Avg. var. extrac.	0.540	1.000	1.000	0.526	0.500	0.827	0.673	0.452	1.000	1.000	1.000	0.656	0.740	1.000
Full collin. VIF	1.656	2.514	2.049	2.164	1.702	1.407	2.206	2.423	1.871	1.382	1.552	1.615	3.410	2.336
Q-squared			0.499	0.404	0.455	0.135	0.450	0.539			0.415			

Note. R-Squared and Q-Squared coefficients are only calculated for endogenous variables. The average variance extracted for the DMM is not reflective of the quality of a formative latent variable. Single-indicator variables do not have quality coefficients for reliability. Q-squared coefficient reflects the importance of a variable in the model, and should be greater than zero.

7.21 Results of Analysis One: Pilot Attributes

One advantage to modeling multiple variables simultaneously with SEM is the ability to follow the effects of upstream variables through various mediating variables to final downstream criterion variables. In the present work, full mediation of effects refers to the phenomenon where original significant direct effects of A -> B become insignificant (but not necessarily 0) when a mediator (C), A-> C-> B, is added as an intervening variable. Effects of predictors on criterion variables were calculated for

each segment, or groups of segments involving that predictor-criterion pair. For example, the effect of pilot age on working memory included three “routes” that incorporated five different segments. In the COMP-A, the relative total effects (both direct and indirect) of pilot age and expertise were determined for domain-independent and domain-dependent cognitive factors, and for the outcome variable of critical incidents. As shown in Figure 7.5 below, results from the present work suggested that age had strong deleterious effects on processing speed and visual-spatial attention, but not on the measure of cognitive inhibition. Also shown in Figure 7.5, pilot age had a significant total effect on working memory, but this effect was fully mediated by the relationships of age to processing speed and visual-spatial attention, as suggested by Figure 7.4. The total effect size for pilot age on cognitive inhibition was non-significant, $f^2=0.013$, $p=.212$, thus eliminating cognitive inhibition as a mediator of age effects on working memory. The full mediation of the effect of age on working memory was also evidenced by the non-significant direct path from pilot age to working memory, as shown in the COMP-A model in Figure 7.4 (despite significant total effects of pilot age on working memory, $\beta=-0.423$, $p=.01$).

While not predicted by the literature, in the present study pilot expertise had a significant small effect on cognitive inhibition such that higher levels of expertise resulted in better performance on the measures of inhibition. However, cognitive inhibition did not mediate any significant effect of expertise on working memory (indirect effect of expertise--via inhibition--on working memory, $f^2=0.016$). In sum, the transmission of the effects of age through various domain-independent cognitive variables was supported by processing speed and visual-spatial attention. As a criterion variable, almost half the variance in working memory scores was explained by visual-spatial attention, processing speed, and pilot age, $r^2=0.46$. The effects of age

on working memory were fully mediated by processing speed and visual-spatial attention.

Figure 7.5. Relative total effect sizes for AGE, EXPERT, and domain-independent cognitive variables. Effect size (f^2) categorized as per Cohen (1988): n.s. <0.02; 0.02 to 0.14 = small; 0.15 to 0.34 = medium; and >0.34 = large. Solid horizontal bars indicate regions of small, medium, and large effects. The textured area of the bars for AGE and PRO. SPEED indicates the size of the total indirect effects.

The direct effects of pilot age and expertise on iRES and the DMM were tested in the COMP-A. As shown in Figure 7.6, pilot age significantly predicted iRES and the DMM (marginal at $p=.07$). There were significant indirect effects of age on iRES (two mediators including PROSPD and INHIB), $\beta = -0.136$, $p=.007$, and the DMM (11 paths via all the domain-independent variables and iRES), $\beta = -0.217$, $p=.01$. As shown in Figure 7.6, pilot expertise significantly predicted iRES and the DMM. There was a significant indirect effect of expertise on the DMM (three mediators including INHIB, WM, and iRES), $\beta = 0.133$, $p=.002$). The weak total effects of processing speed, inhibition, and working memory on the DMM and critical incidents

suggested that the indirect effects of pilot age on the DMM and critical incidents were not transmitted by processing speed, inhibition, or working memory in any substantial amount.

Age and expertise were also expected to have significant indirect effects on critical incidents. The indirect effects of age on critical incidents was mediated by domain-independent and domain-dependent variables (all paths can be traced in Figure 7.4) and was significant at $\beta = 0.228, p = .006$. Figure 7.6 suggests that the most plausible mediators of the effects of age on critical incidents are the DMM, iRES, and VIS_SPA. Expertise did not significantly moderate the effect of pilot age on critical incidents, $p > .05$.

7.22 Results of Analysis One: Domain-independent Cognitive Measures

Domain-independent cognitive variables were expected to have significant direct effects on iRES and the DMM. It was also possible that domain-independent variables might have significant effects on critical incidents, but that these effects were mediated by iRES and the DMM. As shown in Figure 7.4, VIS_SPA significantly predicted iRES. VIS_SPA had a weak direct effect on the DMM, $\beta = -0.16$ n.s. The overall effect of VIS_SPA on the DMM was also non-significant, $p = .07$. However, the total effects of VIS_SPA on critical incidents was significant (four mediators including iRES, DMM, and WM), $\beta = 0.113, p = .031$. As shown in Figure 7.4, INHIB was significantly associated with the DMM. The indirect path from INHIB to the DMM via WM was not significant. Although the overall effect of INHIB on critical incidents significant, $\beta = 0.096, p = .016$, as shown in the Figure 7.6 the total relative effect size of INHIB for critical incidents was non-significant. The very small relative effects of PROSPD and WM on the DMM and critical incidents is also observed from Figure

7.6. PROSPD and WM had weak but significant relative total effect sizes for the DMM, $f^2 = 0.025$ and 0.03 , but non-significant relative total effect sizes for critical incidents, $f^2 = 0.006$ and 0.004 . PROSPD and WM demonstrated the weakest effects on downstream variables in the COMP-A in comparison to other upstream predictors.

7.23 Results of Analysis One: Domain-dependent Cognitive Measures

The domain-dependent variable of iRES was examined regarding its effect on the DMM and critical incidents. iRES was the third strongest predictor of the DMM (less than pilot age and expertise). Despite a non-significant total effect, $\beta = -0.126$, $p = .088$, iRES demonstrated a non-trivial total effect size of $f^2 = 0.053$ for critical incidents. The DMM mediated the majority of the effects of iRES on critical incidents, with the indirect effect of iRES on critical incidents marginally significant, $\beta = -0.089$, $p = .089$, $f^2 = 0.037$. As shown in Figure 7.6, the strongest predictor in the COMP-A was the DMM. As demonstrated above, the DMM mediated upstream effects from pilot attributes and cognitive variables on critical incidents. The total relative effect size of the DMM for critical incidents was moderate at $f^2 = 0.284$, which explained the greatest share of total explained variance in critical incidents (critical incidents $r^2 = 0.41$).

Prediction of the DMM. Of primary interest in the model of pilot performance are the relationships between upstream variables and the DMM. The r^2 value for the DMM is 0.54 , indicating that more than half the variance in DMM scores was accounted for by the upstream variables in the model. Figure 7.6 demonstrates the relative total effect sizes from each upstream variable hypothesized to have an effect on the DMM (blue bars). The total contributions are calculated as direct and all indirect effects. For example, it may be determined from Figure 7.4 that there are 11

different mediating paths from age to the DMM, $\beta = 0.228$, $p = .009$. The total effect size of age for the DMM was derived from the transmission of age effects along these 11 paths. The combined effect sizes from the domain-independent cognitive variables for DMM scores was less than the effect sizes calculated for pilot age and expertise. The sums of direct and indirect effects of PROSPD, WM, INHIB, and VIS-SPA on the DMM ranged from $f^2 = 0.025$ to 0.097 . In order of strength, COMP-A factors significantly associated with the DMM were pilot attributes, domain-dependent cognition, and finally domain-independent cognition.

Figure 7.6. Relative total effect sizes (direct and indirect) of predictor variables in the COMP-A associated with the DMM (top) and critical incidents (bottom). Effect sizes below 0.02 on the y-axis suggest non-significant effects. Effect sizes from 0.02 to 0.14 are small effects, from 0.15 to 0.34 are moderate effects, and > 0.34 are large effects (f^2 effect sizes as per Cohen, 1988). All effect sizes estimated after controlling for effects of episodic memory and mission difficulty (workload condition) on iRES and DMM.

7.24 Results of Analysis One: Non-causal Measures Predicting Critical Incidents

The COMP-A analysis one included measures of pilot self-rating (SA and Task Load Index). A single item from the SA self-rating scale, “awareness of the impact of other aircraft on the ability to fly the perfect circuit” (SR_S4) was found to be one of only a few measure from an earlier analysis that predicted incursion management (see Van Benthem & Herdman, 2014b), and was therefore permitted to independently predict the aviation outcomes. This single item and the two self-rating scales were considered non-causal predictors of aviation outcome, such that they might predict scores, but did not have causal assumptions associated with the relationships. Figures 7.4 and 7.6 illustrate the nonsignificant relationship between pilot-self rating of task load and of situation awareness.

Table 7.10
Results of Hypothesis Testing for the COMP-A Analysis One

Hypotheses	Description	Decision
H1-H6	Older age has detrimental effects on performance of domain-independent and domain-dependent cognitive factors.	Accept
H7-H9	Slower processing speed, less inhibition, and reduced visual-spatial attention are associated with poorer working memory accuracy.	Accept
H10-H11	Slower processing speed, and less inhibition are associated with reduced visual-spatial attention.	Partial Accept Caveat: Only processing speed was associated with visual-spatial attention.
H12-H13	Better visual-spatial attention is associated with significantly higher iRES accuracy and better DMM performance.	Accept Caveat: Visual-spatial attention effects on the DMM were marginally significant.
H14-H15	Higher levels of expertise are associated with significantly higher iRES accuracy and DMM performance	Accept
H16	Higher iRES accuracy is associated with significantly better DMM performance.	Accept
H17	Better working memory accuracy is associated with better DMM performance.	Accept Caveat: Direction of influence was in expected direction, but weaker than anticipated, $f^2 < 0.04$.
H18-H19	Better DMM performance is associated with significantly fewer critical incidents and better incursion management (incursion hypothesis tested in second analysis only).	Accept Incursion:NA
H20-H21	Higher levels of expertise are associated with fewer critical incidents and better incursion management (not labeled).	Accept
H22-H23	Older Age is associated with more critical incidents and but has no direct effect on incursion management (not labeled).	Accept

Hypotheses	Description	Decision
Mediation Hypotheses (not labeled)		
A	Total effects of age on outcomes are fully mediated by domain-independent and domain-dependent cognitive factors.	Accept Incursion:NA
B	Total effects of expertise on outcomes are fully mediated by domain-dependent factors.	Accept Incursion:NA
C	Total effects of iRES on outcomes are fully mediated by the DMM.	Accept Incursion:NA
D	Total effects of domain-independent factors on outcomes are mediated by domain-dependent cognitive factors. Corollary: the effects the DMM on the outcomes variables are larger than the combined total effects from the domain-independent variables.	Accept Incursion:NA
E	Pilot expertise moderates deleterious effects of age on critical incidents. Pilot age moderates the effect of working memory on incursion management.	Reject
Control Variable Hypotheses (not labeled)		
F	Better episodic memory is significantly associated with better iRES. The high-difficulty condition is significantly associated with lower iRES and DMM scores, and more critical incidents.	Accept Caveat: i) Episodic memory was not significantly associated with the DMM. ii) The effect of mission difficulty on critical incidents was fully mediated by iRES and the DMM.

Note. Hypotheses evaluated based on f^2 relative effect sizes in three models where possible. Grey text indicates testing for COMP-A analysis two only.

7.25 Cognitive Factors Mediating the Effects of Age on the DMM and Critical Incidents: Cognitive Markers for Age

Pilot age has demonstrated persistent effects in models of flight performance despite the inclusion of a broad array of cognitive factors (Taylor et al., 2005). In addition to

the COMP-A, in this analysis simple models were constructed and then compared to examine the attenuation in the effect of pilot age on the DMM and on critical incidents as progressively more complex cognitive mechanisms were added to the models. This method loosely follows Salthouse's (1992) incremental validity analyses used to explore the mediation of age effects on reasoning ability by domain-general cognitive variables. In the Salthouse analysis, variables were added, and remained, in the model successively. The incremental attenuation of r^2 attributed to age on the outcome variable was assigned to the last variable to enter the model. Since cognitive factors are rarely "pure", the order that potential mediating variables are entered into the analysis might influence the mediation capacity of that variable. The mediating effects are likely to change as the order of a variable changes in such an analysis. To overcome this problem, the method of adding a variable individually, noting the reduction of the relative effects of age on the DMM, and then removing the cognitive variable from the model was employed. The base models contained critical incidents regressed on age and the two control variables of expertise and workload. In place of r^2 effects, the present analysis compared the relative total effect size (*f-squared*, as per Cohen, 1988) of age on the DMM and critical incidents in separate models that included processing speed, working memory, inhibition/interference, visual-spatial attention, and iRES (but individually tested with pilot age and the criterion variable). The DMM was also examined as a mediator of age effects on critical incidents. Thirteen models were examined to investigate the attenuation, and thereby mediation, of effects of age on the DMM and critical incidents. With respect to the relationship between age and the DMM, age alone had an effect size of 0.17 (a moderate effect, as per Cohen, 1988). After comparing the 13 models, the largest mediation of these effects was observed for visual-spatial attention, where the effect of age dropped to

0.12. This was followed by iRES, working memory, and processing speed, where 0.03, 0.02, and 0.02 decreases in effect size of age on the DMM were seen. The variable representing cognitive inhibition did not mediate any effects of age on the DMM. In combination, expertise and cognitive factors in this analysis mediated almost 50% of the effects of age on the DMM.

For the outcome variable of critical incidents only the DMM resulted in significant mediation of age effects. Age alone had an effect size of 0.07 for critical incidents. Processing speed, working memory, inhibition/interference, visual-spatial attention, and iRES each resulted in a decrease of $f^2=0.01$ or less, regarding the size of the effect of age on critical incidents. The model with the two control variables of expertise and workload, and age and the DMM resulted in the effect of age decreasing to 0.02 (from the original 0.07). Thus most of the effects of age on critical incidents were mediated by the DMM. These results are not surprising in light the comparison of total effect sizes of all variables on the DMM and critical incidents from the COMP-A, and as shown in Figure 7.6. In sum, for the DMM and critical incidents, a very small effect of age on the criterion variable remained after the inclusion of key cognitive variables in the mediation model.

7.26 Structural Path Testing in the COMP-A: Analysis Two with Domain-Independent Measures from the DCAT and Pilots 40+ Years

Table 7.11 provides a description of the sample of pilots used in the COMP-A analysis two. Only pilot aged 40 years and older were included in this sample. Before reviewing the results of the COMP-A modeling for both outcome variables the results of measurement model development for analysis two is provided.

Table 7.11

COMP-A with Pilots Aged 40 and older using DCAT as Domain-independent Predictors of Pilot Performance (n =72).

	Age	Rating Level	Recent Hours as Pilot-in-Command (12 months)	Total Hours Flown	Years Licensed	Simulator Hours
Mean	54.2	2.5	31.0	849.3	18.4	43.9
Median	52.5	2.0	14.0	332.0	13.5	12.5
Standard Deviation	9.7	0.9	49.0	1460.1	15.2	89.6

Outcome statistics for analysis two. During the simulated flight, 46 pilots incurred no critical incidents. Of the remaining 26 pilots, 11 incurred 1, 10 incurred 2, 3 incurred 3, and one each incurred 4 and 5 critical incidents. The probability of incurring at least one critical incident was 36%. The range of incursion management scores was 0 to 10. The mode was 10, median was 6.5, and the arithmetic mean was 5.78.

7.27 Measurement Models for Domain-independent Cognitive Latent Variables

In the second analysis the domain-independent cognitive variables from the CogScreen-AE were replaced by a latent variable measuring mental workload and by a complex working memory measure indexed by DCAT subtests 2 and 6. The mental task workload measurement model (WL_PDT) was operationalized using two raw (average response time to the auditory tone and hit rate) and one derived measure (intraindividual standard deviation of response time) from the auditory peripheral detection task. Each manifest indicator demonstrated a significant loading with the task workload latent variable, ranging from 0.660 to 0.840. The two indicators for the DCAT model demonstrated acceptable loadings, 0.737. Model quality standards

followed the same protocol as the testing of measurement models for analysis one. Reliability coefficients, Cronbach's alpha and composite reliability were evaluated to examine the interrelationships of the manifest indicators within a latent variable (Cronbach, 1951; Sijtsma, 2009). Reliability coefficients ranged from 0.66 to 0.96. An exception was Cronbach's alpha for the DCAT model, where the reliability coefficient was 0.161 (presumably due to the large standard errors associated with DCAT subtests 2 and 6). The AVE statistic (average variance extracted) reflects the average strength of the relationship of each indicator with the latent construct. The AVE for WL_PDT and DCAT were acceptable at 0.597, 0.544, The measurement models for WL_PDT and DCAT demonstrated low variance inflation, with all VIFs < 3.3, indicating that no significant multicollinearity existed in the measurement models (Kock & Lynn, 2012). Latent variable discriminant validity was tested by confirming that the square root of the AVE associated with each latent variable was greater than any correlation between that latent variable and other variables in the model.

7.28 Measurement Models for Self-rating Scales

The three latent variables representing non-causal, but plausibly predictive factors were also included in COMP-A analysis two. The measures included self-rated mental workload and self-rated situation awareness. A self-rated situation awareness item representing SA for "awareness of the impact of other aircraft on flying the perfect circuit" was posited as single item non-causal variable. This single self-rated SA item (S4) was not included in the full self-rated SA measurement model development. The indicators for the two self-rating scales (SELF_WL and SELF_SA) also loaded significantly on their latent variable. Cross-loadings for all indicators were low, <0.220. The AVEs for SELF_WL and SELF_SA were 0.627, 0.814. SELF_WL, and

SELF_SA demonstrated low variance inflation, with all VIFs < 3.3, indicating that no significant multicollinearity existed in the measurement models (Kock & Lynn, 2012). Latent variable discriminant validity was tested by confirming that the square root of the AVE associated with each latent variable was greater than any correlation between that latent variable and other variables in the model.

Table 7.12

Unrotated Loadings and Oblique-rotated Cross-loadings for Measurement Models in the COMP-A Analysis Two.

	WL_PDT	DCAT	SELF_WL	SELF_SA	Type (as defined)	SE	P value
HRr	(0.840)	-0.095	0.099	-0.006	Reflective	0.058	<0.001
AvRT	(0.806)	-0.076	-0.103	-0.084	Reflective	0.065	<0.001
SD	(0.660)	0.215	0.000	0.110	Reflective	0.100	<0.001
DA2	-0.125	(0.737)	0.000	-0.013	Reflective	0.265	0.003
DA6	0.125	(0.737)	0.000	0.013	Reflective	0.313	0.010
TLXMD	-0.027	0.031	(0.865)	0.015	Reflective	0.069	<0.001
TLXPD	0.260	0.157	(0.628)	0.128	Reflective	0.079	<0.001
TLXTD	0.053	-0.008	(0.835)	-0.015	Reflective	0.061	<0.001
TLXFr	-0.058	-0.090	(0.689)	-0.197	Reflective	0.072	<0.001
TLXEff	-0.014	-0.035	(0.885)	0.027	Reflective	0.073	<0.001
TLXOv	-0.162	-0.031	(0.817)	0.039	Reflective	0.058	<0.001
SAPosOth	0.035	0.048	0.045	(0.863)	Reflective	0.057	<0.001
SAPred	0.039	0.076	0.011	(0.931)	Reflective	0.059	<0.001
SAOtAcAc	-0.055	-0.046	0.052	(0.942)	Reflective	0.055	<0.001
SAOtAcLoc	-0.047	-0.067	0.067	(0.917)	Reflective	0.052	<0.001
SAOv	0.034	-0.008	-0.186	(0.854)	Reflective	0.057	<0.001

Table 7.13

Measurement Model Quality Indices for the COMP-A Analysis Two

	WL_PDT	DCAT	SELF_WL	SELF_SA
R-squared				
Adj. R-squared				
Composite reliab.	0.815	0.704	0.909	0.956
Cronbach's alpha	0.657	0.161	0.877	0.942
Avg. var. extrac.	0.597	0.544	0.627	0.814
Full collin. VIF	1.169	1.157	1.302	1.198

7.29 COMP-A with the DCAT and Mental Task Load (PDT) as Domain-dependent Predictors of Pilot Performance

The COMP-A figures resulting from analysis two are provided below: the first investigated the predictors of critical incidents and the second, the predictors of the surprise runway incursion management. Because of the structural and content similarities, the results of both COMP-A figures are discussed together.

Figure 7.7. COMP-A analysis two: critical incident outcome. Dashed arrows indicate non-significant paths. Curved arrows indicate nonlinear relationships. Single indicator constructs (iRES, AGE, and CRIT_INC) necessarily have loadings of 1.0 and are not labelled. Outer measurement models for each latent variable are shown as indicated by grey arrows and yellow indicator boxes. The measurement models for SA, TRP, and TM are not shown. Black arrows between variables indicate the structural (inner) path model (bold arrows indicate strongest predictor in the criterion block). Values shown with hypothesized paths are standardized β coefficients. Direct path from iRES to CRITICAL INCIDENTS not shown, $\beta = -0.03$. At the top right of each endogenous factor is its total r^2 value (total variance explained by all direct and indirect effects from upstream variables). Near bottom left inset demonstrates effects from control variables: mission difficulty significantly predicted MENTAL TASK LOAD (not shown, $\beta = 0.260$), iRES and the DMM, while episodic memory only predicted iRES. See Appendices F and H for the self-rating scale item legends.

Figure 7.8. COMP-A with the DCAT as predictors and incursion management as the outcome variable. Dashed grey arrows indicate nonsignificant paths. Curved arrows indicate nonlinear relationships between variables. Dashed red arrow indicates significant moderating effect of age on the DCAT to INCURSION path. Direct paths from MENTAL TASK LOAD and iRES to INCURSION were removed in the final model, *betas* = -0.04 and -0.01. Inset near bottom left demonstrates effects from control variables: mission difficulty significantly predicted MENTAL TASK LOAD (not shown, *beta* = 0.260), iRES and the DMM, whereas episodic memory only predicted iRES. See Appendices F and H for the self-rating scale item legends.

7.30 Model Quality Indices for COMP-A Analysis Two: Critical Incidents

- Average path coefficient (APC)=0.185, $P<.001$
- Average R-squared (ARS)=0.316, $P=.651$
- Average adjusted R-squared (AARS)=0.293, $P=1.000$
- Average block VIF (AVIF)=1.275, acceptable if ≤ 5 , ideally ≤ 3.3
- Average full collinearity VIF (AFVIF)=1.632, acceptable if ≤ 5 , ideally ≤ 3.3

- Tenenhaus GoF (GoF)=0.501, small ≥ 0.1 , medium ≥ 0.25 , large ≥ 0.36
- Simpson's paradox ratio (SPR)=0.960, acceptable if ≥ 0.7 , ideally = 1
- R-squared contribution ratio (RSCR)=0.999, acceptable if ≥ 0.9 , ideally = 1
- Statistical suppression ratio (SSR)=1.000, acceptable if ≥ 0.7
- Nonlinear bivariate causality direction ratio (NLBCDR)=0.920, acceptable if ≥ 0.7

7.31 Model Quality Indices for COMP-A Analysis Two: Incursion

- Average path coefficient (APC)=0.190, $P < .001$
- Average R-squared (ARS)=0.298, $P < .001$
- Average adjusted R-squared (AARS)=0.273, $P < .001$
- Average block VIF (AVIF)=1.367, acceptable if ≤ 5 , ideally ≤ 3.3
- Average full collinearity VIF (AFVIF)=1.903, acceptable if ≤ 5 , ideally ≤ 3.3
- Tenenhaus GoF (GoF)=0.482, small ≥ 0.1 , medium ≥ 0.25 , large ≥ 0.36
- Simpson's paradox ratio (SPR)=1.000, acceptable if ≥ 0.7 , ideally = 1
- R-squared contribution ratio (RSCR)=1.000, acceptable if ≥ 0.9 , ideally = 1
- Statistical suppression ratio (SSR)=1.000, acceptable if ≥ 0.7
- Nonlinear bivariate causality direction ratio (NLBCDR)=0.942, acceptable if ≥ 0.7

Model quality indices suggest that both COMP-A models in the second analysis conveyed useful amounts of information regarding the transmission of effects of pilot factors on flight performance. The model for incursion management has slightly lower levels of *r*-squared coefficients and a lower overall goodness of fit, as compared to the model predicting critical incidents. The COMP-A models in analysis two demonstrated plausible directions of causation. In each model there were no plausible suggestions for reversed links, nor were there indications of Simpson's paradox, thus the hypothesized direction of causation was supported in each analysis. In sum both models demonstrated adequate utility as a predictor of pilot performance.

7.32 Results of Analysis Two: Pilot Attributes

In COMP-A analysis two the total effects of pilot age and expertise were determined for domain-independent and domain-dependent cognitive factors, and for the outcome variables of critical incidents (Figure 7.7) and surprise runway incursion management (Figure 7.8). Older age and lower levels of expertise were associated with higher levels of mental workload (WL_PDT). Expertise was also positively associated with DCAT, suggesting that more experienced pilots responded quickly and more accurately to questions pertaining to safety in driving scenarios. Figure 7.9 illustrates the larger relative effect size for age, as compared to expertise, for mental task load. Also shown in Figure 7.9, total relative effects of expertise on DCAT was almost exclusively a direct effect, with a non-significant effect mediated by mental task load.

The direct path from age to iRES was nonsignificant, $p=.09$. The indirect effects of age on iRES were also insignificant (one path via WL_PDT), $\beta = -0.034$, $p=.086$, with the total effect of age on iRES marginal, $\beta = -0.131$, $p=.053$. As shown in Figures 7.7 and 7.8, pilot age significantly predicted the DMM. The indirect path from age to the DMM was significant (4 paths via domain-independent variables and iRES), $\beta = -0.072$, $p=.04$. Pilot expertise significantly predicted iRES and the DMM. The total indirect effects from expertise to the DMM were significant (5 paths via the domain-independent variables and iRES), $\beta = 0.164$, $p=.001$.

Age and expertise were also expected to have significant indirect effects on critical incidents and incursion management. Pilot age had a significant deleterious effect on critical incidents and this effect was fully mediated by domain-independent and domain-dependent variables. The direct path from age to critical incidents was non-significant, $\beta = 0.11$, $p=.19$. The indirect effects of age on critical incidents was mediated along 7 paths that included all upstream variables (paths can be traced

in Figure 7.7) and was marginally significant at $\beta = 0.103$, $p = .055$. Figure 7.7 suggests that the strongest plausible mediators of the effects of age on critical incidents are the DMM, iRES, and DCAT. The negative direct, $\beta = -0.18$, $p < .01$, and total effect size, $f^2 = 0.04$, of pilot age on incursion management were significant. The positive direct, $\beta = 0.19$, $p < .01$ and total effect size, $f^2 = 0.077$, of pilot expertise on incursion management were significant. The indirect effects of age and expertise on incursion management were non-significant, $p > .1$. In neither model did expertise significantly moderate the effects of age, $p > .05$.

Figure 7.9. Relative total effect sizes for AGE, EXPERTISE, and MENTAL TASK LOAD (auditory peripheral detection task) on domain-independent cognitive measures in the COMP-A with 144 observations. Effect sizes below the horizontal bar at 0.02 (y-axis) suggest non-significant total effects. Textured area at the bottom of the EXPERTISE bar indicates mediated effects of EXPERTISE on DCAT. Effect sizes are similar to f -squared coefficients (Cohen, 1988).

7.33 Results of Analysis Two: Domain-independent Cognitive Measures

The direct path from mental task load (PDT) to DCAT was significant, $\beta = -0.25$, $p < .01$. The indirect paths from mental task load to iRES and the DMM were nonsignificant, $p > .05$. The total indirect effect from mental task load to the DMM (via iRES and DCAT) was weak, but significant, $\beta = -0.078$, $p = .021$. As shown in Figures 7.7 and 7.8, DCAT predicted the DMM directly, $\beta = 0.210$, $p < 0.01$. The direct paths from mental task load to critical incidents or incursion management were non-significant and were removed from the final model. The direct path from DCAT to critical incidents and incursion management was non-significant. There was a significant interaction between pilot age and DCAT, such that DCAT scores moderated the negative effect between age and incursion management. Figure 7.12 illustrates the interaction between age and DCAT, where older pilots with low DCAT scores demonstrated poor management of the incursion: whereas younger pilots with low DCAT scores managed the incursion as well as younger pilots with normal DCAT scores.

The total indirect effects of mental task load on critical incidents and incursion management were non-significant. The total indirect effects (one path via the DMM) of DCAT on critical incidents was marginally significant, $\beta = -0.075$, $p = .064$, demonstrating that the DMM fully mediated the significant effects of DCAT on critical incidents. As shown in Figures 7.10 and 7.11, the relative total effect size of DCAT for critical incidents was small but significant, but was non-significant for incursion management.

7.34 Results of Analysis Two: Domain-dependent Cognitive Measures

The domain-dependent variable of iRES was examined regarding its effect on the DMM, critical incidents and incursion management. As shown in Figures 7.10 and 7.11, iRES demonstrated the second largest relative total effects for the DMM, $\beta = 0.22$, $p < 0.01$ (less than pilot expertise but very similar to pilot age effects). Despite a non-significant direct effect on critical incidents, $\beta = -0.026$, $p = .422$, there was a significant indirect effect of iRES on critical incidents that was mediated by the DMM, $\beta = -0.080$, $p = .038$, $f^2 = 0.037$. As shown in Figure 7.10, the strongest predictor in the COMP-A for critical incidents was the DMM. The total relative effect size of the DMM for critical incidents was moderate at $f^2 = 0.195$, which explained the greatest share of total explained variance in critical incidents (critical incidents, $r^2 = 0.40$). The total relative effect sizes for iRES on incursion management was non-significant, and the total relative effect size of the DMM on incursion management was marginal, $f^2 = 0.02$, as illustrated by Figure 7.11.

7.35 Results of Analysis Two: Non-causal Measures Predicting Critical Incidents and Incursion Management

The COMP-A analysis two included measures of pilot self-rating (SA and Task Load Index). The single item from the SA self-rating scale, “awareness of the impact of other aircraft on the ability to fly the perfect circuit” (SR_S4) independently predicted the aviation outcomes. This single item and the two self-rating scales were considered non-causal predictors of aviation outcome, such that they might predict scores, but did not have causal assumptions associated with the relationships. For critical incidents and incursion management, SR_S4 predicted small but significant amounts of variance in the outcome variables, $\beta = 0.210$, $p = .01$. As shown in Figures 7.10 and 7.11, the self-rated situation awareness and task load index constructs did not

demonstrate significant relative total effect sizes for critical incidents. Self-rated SA and WL had significant direct effects on incursion management. Figure 7.11 illustrates the small but significant relative total effect size of self-rated task load index, but not for self-rated SA, for incursion management.

Figure 7.10. Relative total effect sizes for predictors of the DMM (blue) and critical incidents (red) in the COMP-A with 144 observations. WL_PDT = mental workload (peripheral detection task). Effect sizes below 0.02 on the y-axis suggest non-significant effects. Effect sizes from 0.03 to 0.14 are small effects, from 0.15 to 0.34 are moderate effects, and > 0.34 are large effects (f^2 effect sizes as per Cohen, 1988). All effect sizes estimated after controlling for small effects of episodic memory and mission difficulty (workload condition) on iRES and DMM. Due to some multicollinearity the sums of each effect size may be greater than the r^2 value associated with criterion variables. No. of observations = 144.

Figure 7.11. Relative effects sizes for predictors of the DMM (blue) and incursion management (purple) in the COMP-A. Below horizontal bar at 0.02 on the y-axis indicates non-significant effect sizes. Self_WL = self-rated mental workload. Self_SA= self-rated situation awareness. SR_S4= self-rated situation awareness for the impact of other aircraft on the ability to fly the “perfect circuit”. Effect sizes below 0.02 on the y-axis suggest non-significant effects. Effect sizes from 0.03 to 0.14 are small effects, from 0.15 to 0.34 are moderate effects, and > 0.34 are large effects (f^2 effect sizes as per Cohen, 1988). All effect sizes estimated after controlling for small effects of episodic memory and mission difficulty (workload condition) on iRES and DMM. Due to some multicollinearity the sums of each effect size may be greater than the r^2 value associated with criterion variables. No. of observations = 144.

Figure 7.12. Smooth 3-D plot of moderating effect of age on relationship between incursion management scores and DCAT latent variable scores. Older pilots experience less deleterious effects of lower expertise as compared to younger pilots. In regards to incursion management, older pilots appear disproportionately disadvantaged by low DCAT factor scores when compared to younger pilots. Age, incursion and DCAT scores are ranked scores. Plot and legends produced by WarpPLS 4.0 (Kock, 2013).

Table 7.14 below provides an overview of the results of hypothesis testing for the COMP-A (analysis two).

Table 7.14
Results of Hypothesis Testing for the COMP-A Analysis Two

Path Estimate Hypotheses	Description	Decision
H1-H6	Older age has detrimental effects on performance of domain-independent and domain-dependent cognitive factors.	Accept for WL_PDT and DMM Reject for iRES Caveat: DCAT could not be regressed on age
H7-H9	Slower processing speed, less inhibition, and reduced visual-spatial attention are associated with poorer working memory accuracy (DCAT).	Accept WL_PDT predicted DCAT
H10-H11	Slower processing speed, and less inhibition are associated with reduced visual-spatial attention.	NA
H12-H13	Better visual-spatial attention is associated with significantly higher iRES accuracy and better DMM performance.	Accept WL_PDT predicted iRES WL_PDT and DCAT predicted the DMM
H14-H15	Higher levels of expertise are associated with significantly higher iRES accuracy and DMM performance	Accept
H16	Higher iRES accuracy is associated with significantly better DMM performance.	Accept
H17	Better working memory accuracy (DCAT) is associated with better DMM performance.	Accept
H18-H19	Better DMM performance is associated with significantly fewer critical incidents and better incursion management (incursion hypothesis tested in second analysis only).	Accept Caveat: After inclusion of self-rating constructs the relationship of the DMM to incursion management was minimal, $f^2=0.02$.
H20-H21	Higher levels of expertise are associated with fewer critical incidents and better incursion management (not labeled).	Accept
H22-H23	Older Age is associated with more critical incidents and but has no direct effect on incursion management.	Accept: Critical Incidents Reject: Incursion Caveat: Age and Age by DCAT interaction predicted incursion management
Mediation Hypotheses		
A	Total effects of age on outcomes are fully mediated by domain-independent and domain-dependent cognitive factors.	Accept

Path Estimate Hypotheses	Description	Decision
B	Total effects of expertise on outcomes are fully mediated by domain-dependent factors.	Accept Caveat: Some effects of expertise were mediated via domain-independent cognitive factors
C	Total effects of iRES on outcomes are fully mediated by the DMM.	Accept
D	Total effects of domain-independent factors on outcomes are mediated by domain-dependent cognitive factors. Corollary: the effects the DMM on outcomes are larger than combined total effects from domain-independent variables.	Accept for critical incidents Reject for incursion management Corollary: Accept
E	Pilot expertise moderates deleterious effects of age on outcomes. Pilot Age moderates the effect of working memory on incursion management. Control Variable Hypotheses (not labeled)	Reject Accept
F	Better episodic memory is significantly associated with better iRES. The high-difficulty condition is significantly associated with lower iRES and DMM scores, and more critical incidents.	Accept Caveat: i) Episodic memory was not significantly associated with the DMM. ii) The effect of mission difficulty on critical incidents was fully mediated by iRES and the DMM.

Note. Hypotheses testing for COMP-A Analysis Two. Hypotheses evaluated with f^2 relative effect sizes in three models where possible (Cohen, 1988).

7.36 Endogeneity in the COMP-A

SEM models should be tested for the presence of endogeneity (Antonakis, Bendahan, Jacquart, & Lalive, 2012). Endogeneity occurs when there is a spurious relationship between a predictor and criterion variable in the model. Endogeneity suggests that there is an omitted (yet unspecified) reason for the relationship between the predictor and criterion variable. A commonly used hypothetical example of endogeneity is the spurious relationship between ice cream sales and shark attacks at the beach. A model with shark attacks regressed on ice cream sales would seem to indicate that sales

“cause” shark attacks. When a third variable, such as daytime temperature, is added to the model the relationship between ice cream sales and shark attacks disappears. Endogeneity is assumed to exist in a model when the relationship between two variables fades with the specification of the new mediating third variable. Endogeneity is explored by examining the correlations between residual terms for latent constructs in the model. If the residual variances between two causally linked constructs are significantly correlated it is possible that a lurking or confounding variable exists that explains the relationship between the two variables (Antonakis et al., 2012). For example, “...if a latent variable A points at a latent variable B, and the error terms (e)A and (e)B are strongly correlated, then this may be an indication of the existence of a hidden confounder. This hidden confounder may be the real cause behind a significant association between A and B, suggesting a causality problem; namely one in which a link may in fact not be a “true” causal link but rather be due to a third variable, the confounder” (Kock, 2013, p. 65).

Endogeneity was explored for all predictor → criterion relationships in the COMP-A by examining the correlations between residual terms associated with each latent variable. Table 7.15 shows the correlations between residual terms for latent variables in the COMP-A. Of primary concern in the COMP-A (analysis one) would be instances of endogeneity between the DMM, or between critical incident, and their predictors. All correlations between predictor latent variable residual terms with the critical incident residual term were non-significant. No other instance of endogeneity, or confounded relationship between predictor and criterion variables were apparent in the COMP-A.

While not specified in the COMP-A there is a possible omitted cause between the processing speed and inhibition latent variables. The error terms for processing

speed and inhibition were the only two variables to show a significant correlation in the endogeneity analysis, $r(53)=0.233, p=.014$. However, the COMP-A did not specify a causal relationship between processing speed and inhibition. Therefore the domain-independent cognitive variables in the COMP-A appear correctly specified. Future work might examine what other omitted, but causal variable (other than pilot age), might explain both processing speed and inhibition performance.

Table 7.15

Correlations Among Latent Variable Residual Terms with VIFs in the COMP-A (Analysis One)

Correlations among l.v. error terms with VIFs							
	(e)RES	(e)PROSPD	(e)WM	(e)INHIB	(e)VIS_SPA	(e)DMM	(e)CRIT_INC
(e)RES	(1.041)	-0.135	0.115	-0.016	0.044	0.026	0.024
(e)PROSPD	-0.135	(1.085)	0.069	0.233	0.019	-0.011	-0.025
(e)WM	0.115	0.069	(1.077)	0.074	-0.156	-0.024	0.130
(e)INHIB	-0.016	0.233	0.074	(1.064)	0.022	0.023	0.027
(e)VIS_SPA	0.044	0.019	-0.156	0.022	(1.052)	-0.117	0.035
(e)DMM	0.026	-0.011	-0.024	0.023	-0.117	(1.024)	0.064
(e)CRIT_INC	0.024	-0.025	0.130	0.027	0.035	0.064	(1.028)

Notes: Variance inflation factors (VIFs) shown on diagonal. Error terms included (a.k.a. residuals) are for endogenous l.vs.

Note. All p -values for correlations were nonsignificant with the exception of processing speed and inhibition, $r(53)=0.233, p=.014$.

7.37 Control Variables

The inclusion of mission difficulty as a control variable permitted examination of the robustness of the DMM and the utility of the COMP-A in varying workload conditions. Mission difficulty (MISSDIFF) significantly predicted MENTAL TASK LOAD (PDT), $\beta = 0.26, p<.01$ (path not shown in Figures 7.7 and 7.8), the DMM, $\beta = -0.27, p<.01$, and iRES, $\beta = -0.31, p<.01$. The indirect effects of MISSDIFF on critical incidents were fully mediated by MENTAL TASK LOAD, iRES, and the DMM, $\beta = -0.15, p<.024$. Episodic memory controlled for pilot ability to recall details of the mission based on retrospective memory. Control of episodic memory

performance was particularly important for interpretation of situation awareness performance. Episodic memory significantly predicted iRES, $\beta = 0.19$, $p < .01$ and incursion management, $\beta = 0.16$, $p = .05$. Episodic memory did not significantly predict the DMM, nor were any significant effects of episodic memory mediated by iRES to the DMM, $p > .05$.

7.38 Interpreting the Role of the DMM in the COMP-A

The interpretation of formative latent variable relationships has been the topic of much discussion (Andreev, Heart, Maoz, & Pliskin, 2009; Cenfetelli & Bassellier, 2009). While reflective latent structures refer to a single hypothesized construct, and are therefore straightforward to interpret, formative variables reflect a composite of multiple concepts. The DMM was designed to represent a single structure that was also composed of four distinct elements. The DMM was a preferred structure in simple models of aviator performance, as opposed to less parsimonious models where SA, TRP, TM, and PM were posited as individual variables that were predicted by iRES, and functioned as predictors of pilot performance (see Chapter 6). Researchers would nevertheless appreciate understanding how SA, TRP, TM, and PM relate to upstream and downstream variables in the COMP-A. Interpreting the effects of upstream predictors on the composite variables that composed the DMM, and the effect of these composite variables on downstream factors is aided by examining the bivariate correlations of the DMM first-order constructs with all manifest indicators (Table 7.16), and other single-indicator and latent variables in the COMP-A (Table 7.17).

Table 7.16
Bivariate Correlations between Single Indicator and Latent Variables & the DMM

	Rating	PIC	Hours	SIM	AGE	iRES	VSCRTC	DAIDRTC	SATACRT	SATNRT	DATSCRT	MISRTC	PFCRTC	BDSACC	MATHACC	PFCACC	SDCIRAC	SATINAC	
SA	0.079	0.039	-0.135	0.182	-0.223	0.426	-0.091	-0.105	-0.015	-0.164	-0.099	0.017	-0.07	-0.02	0.171	0.156	0.083	0.17	
TRP	0.376	0.231	0.142	0.241	-0.384	0.357	-0.289	-0.187	-0.113	-0.202	-0.21	-0.1	-0.12	0.159	0.067	0.357	0.138	0.12	
TM	0.236	0.154	0.16	0.338	-0.154	0.307	-0.154	-0.156	-0.202	-0.126	-0.347	-0.27	-0.17	0.152	0.118	0.182	0.092	0.057	
PM	0.229	0.1	0.053	0.197	-0.161	0.337	-0.109	-0.071	-0.186	-0.101	-0.156	-0.14	-0.15	0.173	0.195	0.314	0.228	0.166	
	SATFAIL	SATRULE	DTTAABS	DTTAHIT	DTTDABS	DTTDHIT	MISSDIFF	EPISOD	CRIT_INC	INCURS	DA2	DA6	AVRT	SD (RT)	HIT RATE	DMM	TRP	TM	PM
SA	-0.07	-0.086	-0.121	-0.111	-0.168	-0.206	-0.289	0.346	-0.347	0.149	0.167	0.063	-0.12	-0.167	-0.133	-0.601	0.243	0.124	0.32
TRP	-0.08	-0.091	-0.439	-0.233	-0.308	-0.372	-0.301	0.213	-0.356	0.09	0.115	0.133	-0.21	-0.198	-0.348	-0.657	1	0.227	0.31
TM	-0.22	-0.09	-0.37	-0.182	-0.301	-0.408	-0.146	0.02	-0.319	0.095	0.212	-0.03	-0.14	0.001	-0.267	-0.637	0.227	1	0.37
PM	-0.35	-0.2	-0.181	-0.093	-0.152	-0.19	-0.231	0.163	-0.422	0.139	0.204	0.285	-0.32	-0.201	-0.362	-0.78	0.307	0.373	1

Note. Green shaded cells represent significant bivariate correlations at $p < 0.05$ (jackknife resampling method for standard errors). Yellow shaded cells are marginally significant correlations. No. of observations = 110.

The shaded pattern in Table 7.16 indicates that the majority of manifest indicators in the COMP-A showed similar relationships with the first-order latent variables in the DMM. For example, SIM, Age, iRES, PFCACC, DTTDABS, MisDiff, and Critical Incidents were at least marginally correlated with all the DMM components. Rating, DATSCRT, BDSACC, DTTABS, Episodic Memory, SD(RT), and Hit Rate had a marginal or greater association with at least three of the first-order constructs in the DMM. In contrast, SDCIRAC, SATRULE, and DA6 only had a significant relationship with PM, and not the remaining DMM components. The incursion outcome variable had weak bivariate relationships with most of the DMM components, with r ranging from 0.09 to 0.149. The final four columns in the second row demonstrate the significant relationships that all the DMM composite variables had with the DMM and with themselves. The PM latent variable was most consistently significantly associated with all the manifest variables and with components of the DMM.

Table 7.17
Bivariate Correlations between all Possible Single-indicator and Latent Variables and the DMM in the COMP-A.

	EXPERT	AGE	iRES	PROSPD	WM	INHIB	VIS_SPA	DMM	MISSDIFF	EPISOD	CRIT_INC	INCURS	DCAT	WL_PDT
EXPERT	0.735													
AGE	0.135													
iRES	0.393	-0.275												
PROSPD	0.008	0.621	-0.283											
WM	0.114	-0.316	0.215	-0.449										
INHIB	-0.144	0.026	-0.062	0.191	-0.403									
VIS_SPA	-0.065	0.636	-0.31	0.539	-0.44	0.087								
DMM	0.323	-0.338	0.525	-0.299	0.333	-0.255	-0.434							
MISSDIFF			-0.365					-0.357						
EPISOD	0.145	-0.119	0.424	-0.024	0.003		-0.095	0.27	-0.292					
CRIT_INC	-0.289	0.211	-0.399	0.156	-0.108	0.145	0.236	-0.539	0.237	-0.327				
INCURS	0.486	0.109	0.126	0.125	-0.017	-0.092	0.041	0.176		0.079	-0.18			
DCAT	0.19	-0.127	0.151	0.006	0.119	-0.08	-0.095	0.298		0.16	-0.23	0.212		
WL_PDT	-0.204	0.22	-0.378	0.328	-0.3	0.141	0.265	-0.414	0.266	-0.19	0.153	-0.117	-0.229	0.771

Note. Square-root of average variance extracted (latent variables) indicated with blue font on the diagonal. Diagonal grey cells indicate single indicator variable (therefore no average variance extracted). Off-diagonal grey cells indicate no hypothesized (or an implausible) correlation. Green cells indicate significant correlations at $p < .05$. Yellow cells indicate marginal correlations. No. of observations = 110.

As shown in the centre row of Table 7.17, the DMM was significantly associated with all other variables in the DMM (including the two control variables of mission difficulty and episodic memory). The DMM was most strongly associated with iRES and critical incidents, such that higher DMM scores were a result of higher levels of information resolution, and that fewer critical incidents occurred for pilots with the higher DMM scores. This relationship of the DMM to iRES and critical incidents is confirmed by the relative (partialled) total effects shown in Figures 7.10, and 7.11 above. As shown in Table 7.17, there was a weak correlation between the DMM and incursion management in the subsample of 55 pilots.

7.39 Utility of the COMP-A in a Sample of Pilots Aged 60+

A number of sub-analyses were conducted to support the interpretation of the DMM as a domain-dependent cognitive variable with special relevance to the identification of older at-risk pilots. The relevance of cognitive factors in the COMP-A was apparent as the sample of pilots was reduced further to include only pilots aged 60 and older (No. obs. = 42). Figure 7.13 illustrates a simple model where pilot expertise, DCAT, and iRES predicted the DMM. As shown in Figure 7.13, the effects of expertise, DCAT, and iRES on critical incidents were fully mediated by the DMM. Figure 7.14 illustrates the relative total effects of all predictors on critical incidents. Despite the removal of various elements from the original COMP-A structure, the COMP-A for pilots aged 60 and older explained a large amount of variance in critical incidents, $r^2=0.65$. The increase in the r^2 coefficient from 0.40 (in the COMP-A with pilot aged 40 or more years) to 0.65 represented a 50% increase in explained variance.

The model fit and quality indices for the reduced model for pilots aged 60 to 81 are listed below. The average r^2 coefficient for each latent variable is 0.416, indicating that the blocks of predictor variables adequately predicted the variance of each latent

variable. The GoF was also large, 0.583. Similar to all previous COMP-A model quality indices the causal direction assumptions were supported with ideal values of SPR, RSCR and SSR.

- Average path coefficient (APC)=0.240, $P=.004$
- Average R-squared (ARS)=0.416, $P<.001$
- Average adjusted R-squared (AARS)=0.372, $P<.001$
- Average block VIF (AVIF)=1.551, acceptable if ≤ 5 , ideally ≤ 3.3
- Average full collinearity VIF (AFVIF)=1.605, acceptable if ≤ 5 , ideally ≤ 3.3
- Tenenhaus GoF (GoF)=0.583, small ≥ 0.1 , medium ≥ 0.25 , large ≥ 0.36

Figure 7.13. Reduced COMP-A for pilots aged 60 to 81 years. Measurement models for first-order latent variables in the DMM, EXPERTISE and DCAT are not shown. Path from iRES to CRITICAL INCIDENTS is not shown, $\beta = 0.09$, ns.

Figure 7.14. Relative total effects of predictors of critical incidents in the COMP-A model with an older pilot sub-sample from analysis two (Age >59, No. obs. = 42).

Table 7.18

Pearson Correlations for the DMM and Predictor Variables in the Reduced COMP-A Model

Pearson Correlations	DMM	DCAT	iRES	Episodic Memory	Age	Expertise
Critical Incidents	-0.644	-0.541	-0.393	-0.343	0.333	-0.238

Note. Age was a not a variable in the reduced COMP-A model but is shown here to demonstrate the continued relevance of pilot age to critical incidents.

As compared to the zero-order Pearson correlations found in Table 7.17, the relationship of predictors to critical incidents in the COMP-A model for pilots aged 60 and older (shown in Table 7.18) were strong for cognitive variables but weak for pilot expertise. The strongest predictor of critical incidents was the DMM. In a model with

the DMM as the single predictor, the DMM explained 57% of the variance in critical incidents, $\beta = -0.76$.

The zero-order Pearson correlations for predictors of incursion management were also examined in the reduced sample of pilots aged 60 and older and the same pattern was found: the DMM, pilot age, and DCAT relationships to incursion management increased, $r = 0.26$, -0.314 , and 0.305 . Similar to its relationship to critical incidents, the relationship of pilot expertise to incursion management was weaker in the older pilot sample, $r(41) = 0.172$.

7.40 A Threshold for DMM performance as a Function of Age

In order to determine the course of the relationship between pilot age and the DMM in this cross-sectional sample of pilots a “best-fitting” curvilinear plot was produced. Figure 7.15 illustrates the relationship of pilot age to the DMM. The segmented plot representing the best-fitting curve (Kock, 2013) of the multivariate relationship of the DMM to pilot age using the full sample of pilot ages ($N=108$) clearly demonstrates that before the age of 53 the rate of decline in the DMM is minor. After age 53 the slope coefficients for each segment are progressively larger and the reductions in DMM performance associated with older age appear to increase logarithmically, with a six-fold increase in the slope coefficient noted from age 53 to age 81. In contrast, the slope did not appear to change between the ages of 19 to 53.

Figure 7.15. Curvilinear relationship of pilot age to DMM latent variable scores (No. obs = 216). Beyond 53 years the relationship of pilot age to the DMM increases substantially with the relationship appearing logarithmic (note decreasing size for units of change along the x-axis). DMM factor scores were standardized to a mean of zero and SD of 1. Slope coefficients are unstandardized. Plot produced by WarpPLS 4.0 (Kock, 2013).

7.41 Discriminant Utility of the CogScreen-AE Measures and the DMM in Predicting High-Risk Pilots

The CogScreen-AE and the DMM were analyzed using a receiver operating characteristic (ROC) test of discrimination. ROC analyses produce curves that demonstrate the diagnostic utility of a screening tool, or other measure, when classifying subjects into one of two groups. In this test, the sample from the first analysis was used so that CogScreen measures were analyzed. The SPSS 22.0 (IBM, 2014) software package was used for the ROC curve analysis. For each measure analyzed, an ROC curve was produced. An area under the curve coefficient (AUC) is associated with each curve. The AUC represented the overall diagnostic value of that measure in correctly classifying subjects into the low- or high-risk group. Figure 7.16

illustrates the results of the ROC curve analysis for the DMM and the CogScreen-AE derived variable, Logistic Regression Probability Value (LRPV). The LRPV is a score between 0 and 1 that indicates likelihood of brain dysfunction (higher scores mean greater likelihood). The LRPV is known to correlate strongly and positively with pilot age (Kay, 1995).

Each point on a curve represents one particular score from the distribution of scores for that measure. For example, the DMM curve in Figure 7.16 contains a red triangle indicating the location on the curve with the best “trade-off” between sensitivity and specificity. Associated with all scores from the distribution of scores from a measure is a sensitivity and specificity (1-sensitivity, or false positive rate) value. Ideally sensitivity is high, meaning that a high-risk pilot who did incur a critical event is classified thusly. Also ideally, the specificity is high, such that false positives are low and low-risk pilots who did not incur a critical incident are not falsely placed in the high-risk group. The CogScreen-AE measures and the DMM were tested as classifiers of pilot-risk groups. General consensus classifies AUCs below 0.60 as non-useful, 0.61 to 0.70 as poor, 0.71 to 0.90 as good, and above 0.90 as excellent. In reality, it is the implications of falsely identifying a true low-risk pilot as high-risk, and miss-identifying a true high-risk pilot as low-risk that dictate domain-acceptable levels of sensitivity and specificity.

With respect to the DMM, the AUC, indicating overall likelihood of correct classification, was 0.785. Using a cut-off score very close to the mean DMM score (z-score = -0.11, as shown in Figure 7.16) resulted in a sensitivity of 0.71, and specificity of 0.30. This threshold maximized correct classification of pilots who had incurred a critical incident (sensitivity), while minimizing false alarms (specificity). A cut-off score around the total group mean for the DMM would correctly diagnose 70% of the

true high-risk pilots, and would correctly identify 70% of the low-risk pilots as low-risk. The prevalence of a critical incidents was approximately 0.28. This prevalence of critical incidents in this sample of pilots is likely higher than it would be in the general population of pilots. Thus, the positive (or negative) predictive value (PPV or NPV) is biased (Parikh, Parikh, Arun, & Thomas, 2009). Another indicator of the diagnostic utility of a test is the likelihood ratio (+ or -), where LR+ values above 2 indicate tests with potentially clinically relevant predictive power. Using a cut-off score at approximately 0, the DMM LR+ was 2.44.

7.41.1 AUCs for DMM factor scores versus individual SA, TRP, TM, and PM Latent variable scores

The ROC analyses provided an additional opportunity to investigate the individual explanatory power of the four domain-dependent cognitive included in the DMM. An ROC analysis was explored where continuous factor scores from SA, TRP, TM, and PM were plotted in comparison to the null reference line (not illustrated). AUC and significance values are listed in Table 7.19. The four latent variables demonstrated significant AUCs, $p < .01$, but only TM had an AUC coefficient above 0.70. The poorest predictor of risk of critical incidents was the TRP factor scores.

7.41.2 Predictive value of the LRPV and four cognitive factors from the CogScreen-AE

The LRPV and four domain-independent cognitive CogScreen-AE variables in the COMP-A were also examined in the ROC analysis. All results are listed in Table 7.19, but only the ROC curves for the LRPV and DMM are shown in Figure 7.16. The curve for visual-spatial attention was significantly different from the chance reference line, $p = .046$. However, visual-spatial attention had an AUC of 0.378, which indicated that this cognitive variable would be a poor discriminator of pilot risk for critical

incidents. At a sensitivity of 0.80, visual-spatial attention had a corresponding false alarm rate >0.60 . Figure 7.19 demonstrates the high false alarm rate that would accompany the choice of the LRPV as a classifier of critical incident group. To attain a sensitivity of 0.80, there is a corresponding false alarm rate of 0.82, meaning that 82% of the true low-risk pilots would be incorrectly classified as high-risk. The remaining AUCs for the domain-independent cognitive factors of working memory, processing speed, and inhibition were also not significantly different from chance.

Figure 7.16. Receiver operating characteristics (ROC) curve for the DMM (blue curve) and the LRPV (green dashed curve) from the CogScreen-AE predicting low- versus high-risk pilots. Area under the curve coefficient for the DMM = 0.785, $p < .001$. Red triangle indicates DMM factor score with optimal threshold for classifying pilots. Dashed black line is the reference line indicating the null curve, 0.50 (discriminant basically at chance). The LRPV curve is closer to the reference line (null effect) in comparison to the DMM curve. ROC curves for other CogScreen-AE measures (not shown) were similar in appearance to the LRPV. No. obs. = 110.

Table 7.19
AUC Results for the ROC Classifier of Pilot Risk Group

Test Result Variable(s)	Area	Std. Error ^a	Asymptotic Significance ^b	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
DMM	0.785	0.045	0.000	0.696	0.874
SA	0.698	0.053	0.001	0.595	0.801
TRP	0.272	0.051	0.000	0.173	0.371
TM	0.713	0.053	0.001	0.609	0.816
PM	0.636	0.061	0.027	0.516	0.755
LRPV	0.471	0.064	0.642	0.346	0.597
Processing Speed	0.425	0.058	0.222	0.311	0.539
Visual-Spatial	0.378	0.058	0.046	0.264	0.491
Working Memory	0.492	0.062	0.894	0.370	0.614
Inhibition	0.448	0.064	0.395	0.321	0.574

Note. Area under the curve with nonparametric assumption. b. Null hypothesis: true area = 0.50. AUC = area under the curve. ROC = receiver operating characteristic.

7.42 Summary of COMP-A Validation

Three main analyses were undertaken with the purpose of testing hypotheses (2 and 3 below) pertaining to the key role of the DMM in predicting pilot performance:

1. The four domain-dependent cognitive factors of SA, TRP, TM, and PM converge in a multilevel structure to represent the dynamic mental model (DMM) for pilots.

2. The DMM, as a domain-dependent variable, accounts for significantly more variance in the aviation outcomes measures than the variance accounted for by domain-independent cognitive factors.

3. The DMM fully mediates the effects of cognition, age, expertise, and workload factors on aviation outcomes.

The findings from three main analyses examining relationships of the DMM to all other variables in the COMP-A demonstrated the value of a comprehensive domain-dependent cognitive variable for predicting critical incidents. As illustrated in Figures 7.6 to 7.10 and Table 7.17, all variables in the COMP-A were significantly correlated with the DMM (incursion management was marginally correlated at

$r(142) = 0.18$). As hypothesized, the DMM was the strongest predictor of critical incidents when compared to the effects of pilot attributes, domain-independent cognitive variables, iRES, or the control variables of retrospective memory or mission difficulty. In the COMP-A model for incursion management, this outcome was only weakly predicted by the DMM. Rather, the strongest predictors of incursion management were pilot expertise, age, and an age by DCAT interaction. Thus the second major hypothesis was partially accepted, but only for the critical incident outcome.

Table 7.10 and 7.15 itemizes numerous sub-hypotheses pertaining to direct and indirect effects of predictor variables on outcome factors. Of note was the frequent finding that a large share of the significant effects of upstream effects on critical incidents was mediated by the DMM. For example, as shown in Figures 7.6 and 7.10, direct effects of pilot expertise and age were nonsignificant, whereas Table 7.17 demonstrates that the bivariate correlations between these predictors and critical incidents were significant. This mediation of small upstream effects on critical incidents by the DMM was also found for iRES and the domain-independent cognitive variables of visual-spatial attention and DCAT. The full mediation of upstream factors on critical incidents by the DMM was evidence supporting the third major hypothesis in the present work.

Because the prediction of older pilot risk is of particular interest in the aviation literature a simplified COMP-A was modeled using a sub-sample of pilots (aged 60 years and older). In the reduced COMP-A 65% of the variance in critical incidents was predicted by the DMM, DCAT, iRES, expertise, and control variables, with the majority predicted by the DMM alone.

Due to the small direct effect of the DMM on incursion management, only small effects from upstream variables on this outcome measure were mediated by the DMM. For example, paths from pilot age and expertise to critical incidents were significant despite the significant direct paths from these pilot attributes to the DMM. The COMP-A model for incursion management also revealed a significant interaction between pilot age and the DCAT. Older pilots with low DCAT scores demonstrated poor handling of the incursion scenario; whereas younger pilots did not show this trend.

The COMP-A structure was tested using a PLS-SEM platform, where latent variables representing expertise, cognition, and self-rated SA and task workload constructs were examined with respect to their total effects on two aviation outcomes. Thus the COMP-A was an appropriate structure for validating several latent constructs. Expertise, working memory, processing speed, inhibition, and visual-spatial attention were potentially causal constructs that could influence aviation outcomes. Findings showed that measurement models demonstrating acceptable reliability and validity could be appropriately modeled in the PLS-SEM platform. The non-causal constructs from the self-rating scales showed limited value in predicting critical incidents, but demonstrated explanatory value in the COMP-A model for incursion management.

In the tradition of Yesavage et al. (2011), signal detection using receiver operating characteristic analyses investigated the classification properties of cognitive variables in the COMP-A in regards to pilot risk of a critical incident. As the COMP-A results would suggest, only the DMM demonstrated acceptable sensitivity and specificity for stratifying pilots as either having or not having been involved in a critical incident.

In summary, evidence for the second and third hypotheses were found, but not for both outcome variables. While critical incidents was predicted by the DMM, the incursion management measure demonstrated a weaker link with the DMM. As a result, less upstream effects from pilot attributes and domain-independent cognition on incursion management were mediated by the DMM. On the other hand, very large relative effects from the DMM were noted for critical incidents, and, as such, the DMM was a strong mediator of upstream effects on critical incidents. The findings support the notion that predicting relevant outcomes in aviation may rely primarily on domain-dependent measures, or complex domain-independent measures, in accord with the outcome of interest.

Chapter 8

Discussion

Three major assumptions influenced the structure of a cognition-oriented model of pilot performance, the COMP-A.

- First, it was hypothesized that four domain-dependent cognitive factors selected to represent the DMM would converge empirically, and demonstrate good model quality and fit indices. Inherent in this hypothesis was that SA, TRP, TM, and PM individually contributed significantly to the DMM structure.
- Second, the DMM was predicted to act as a central feature in the COMP-A. In this role, the DMM was predicted by upstream factors in the model. In turn, the DMM predicted downstream criterion variables. As an aviation-domain-dependent variable, it was anticipated that the DMM would have a stronger influence on critical incidents and incursion management when compared to the influence of domain-independent cognitive factors on these outcomes. A prerequisite of this hypothesis was that the DMM was a significant predictor of aviation outcomes.
- The third hypothesis predicted that the DMM, as a central variable in the COMP-A, fully mediated the effects of cognition, age, expertise, and workload factors on aviation outcome variables. This final hypothesis was not necessary to prove hypotheses one and two, but it was relevant to the pursuit of a parsimonious and effective cognitive health screening tool for pilots.

Figure 8.1. A generic “COMP” structure with hypothesized inner paths. Only the first-order constructs for the DMM measurement model are shown. See Figure 6.1 for the specification of the full DMM measurement model. Measurement models for other latent constructs are not shown.

8.1 Major Hypothesis One: Dynamic Mental Model Validity

The validity of the DMM was assessed by determining whether the multidimensional model converged with adequate quality and fit. A discussion of the “completeness” and lack of redundancy resulting from the specification of the final DMM also serves to evaluate the validity of the DMM (Petter et al., 2007). A review of DMM completeness was undertaken by comparing descriptions of mental models in general and in the aviation-specific literature with the DMM. A discussion regarding DMM redundancy ensures that the final DMM structure was parsimonious and that relations between DMM composites were not inflated due to issues with multicollinearity.

8.1.1 Convergence of the DMM

The hypothesized DMM structure was based upon the results of several studies from the present work described in Chapter 4, and on theory outlined in Chapter 5. Validating the content and structure of the DMM was undertaken in three main stages. Initial model generation was completed on approximately half the full dataset and then cross-validated on the remaining observations. Cross-validation permitted an examination of whether the final model structure using the first set of data was over-fit to peculiarities of the first dataset, rather than on estimated population parameters. The cross-validation test demonstrated that the final DMM model was robust to new data. These results suggest that the DMM, as indexed by reliable measures of SA, TRP, TM, and PM, could generalize to data collected from a different sample of pilots.

The PLS-SEM and CB-SEM analyses provided empirical support for the DMM specified as a multidimensional reflective-formative construct. The first-order manifest variables loaded significantly on their respective constructs. The PLS-SEM and CB-SEM analyses found that SA was reliably indexed by SA levels 2 and 3 measures. TRP was indexed by deviations from the ideal altitude, and to a lesser extent by deviations from the ideal airspeed. TM was the only measure in the DMM to use self-ratings of performance rather than objective measures collected during flight simulation. Nevertheless, items pertaining to awareness of a) the tasks to be performed, b) how to perform them, and c) the priority for carrying them out were reliable indices of the task management construct, and TM contributed significantly to the DMM. PM was measured by a single indicator, which was a composite score based on remembering to complete radio communication tasks at the correct location in the circuit. Regardless of whether each first-order construct was modeled using PLS composites, or covariance-based common factor algorithms, the individual

measurement models had reliable loadings, and adequate discriminant and convergent validity coefficients.

Comparison of PLS- and CB-SEM validation of the DMM. Despite their unique estimation biases, the CB-SEM and PLS-SEM analyses produced similar patterns of factor loadings for each measurement model. As expected, the CB-SEM software produced first-order loadings somewhat less in magnitude than was produced by PLS-SEM. Also as predicted by Hair et al. (2012), the weights for the second-order model were higher than the estimates produced by PLS-SEM. The CB-SEM analysis included a number of model fit coefficients providing evidence regarding the appropriate structuring of the DMM. In summary, the low inter-correlations between composites, and overall model quality and fit indices demonstrated that the DMM was appropriately specified as a multi-level reflective-formative construct.

8.1.2 Completeness of the DMM Specification

Two principal concerns are vital to the discussion of the results of formative construct modeling (Petter et al., 2007). The first concern pertains to the adequacy of the constructs in fully indexing all relevant aspects of mental model for pilots. Petter et al. suggest that if a critical aspect of a mental model for pilots is deemed missing, then a suitable index of that missing aspect should be found and added to the model. It was the premise of the present work that SA, TRP, TM, and PM comprehensively indexed a mental model for pilots, a model that would support basic aviator competencies. Briefly described, SA referenced situation models of the current and near future position of ownship, other aircraft, and relevant terrain. Automatic perceptual motor responses for safe progression through the “situation” are provided by TRP. Safe maneuvering through an environment includes skills for quickly formulating new

plans of action if required. Knowledge regarding what tasks to perform and how to perform each one with correct prioritization was represented by TM. Last, being aware of environmental or internal cues related to the future execution of acts that were planned in the past was provided by PM.

That SA, TRP, TM, and PM converged empirically does not suggest that their particular manifest indicators (SA2, SA3, T1 etc.) are the *only* possible indices for the DMM. It is possible that combinations of other indices, shown to reliably represent relevant constructs, could also index the DMM. That said, development of a formative model requires confirmation that some relevant aspect of the model is not missing from its structure. The features of mental models, as theorized by the literature and outlined in Table 8.1, can be classified as referring to i) system form and purpose, ii) doing, iii) situation awareness, iv) learning and understanding relations, v) domain-independent cognitive functions, or vi) other general descriptives (e.g., dynamic). To confirm whether the DMM represented key aspects of mental models (as described by the literature) a mapping of the components of the DMM with the features of mental models was undertaken.

The DMM, as indexed by eight manifest indicators, appeared to reflect axiomatic descriptions of mental models. For example, as shown in rows one and two of Table 8.1, Rouse and Morris (1985) suggest mental models refer to the *purpose* and *form* of a system. The manifest indicators of the TM component of the DMM reflect purpose via *awareness of tasks and task priorities*. The SA component indicators address the form of the system via situation models (SA2) that reflect ownership and the spatial awareness of the flight environment.

References in the literature also frequently refer to the mental model as a requisite for “doing”. Besnard (2004) describes mental models as constructed from

procedural knowledge pursuant to attaining goals. Gentner and Stevens (2014) highlight the mental simulation purpose of mental models. The idea of mental simulation was also highlighted by Endsley (1995) and Histon and Hansman (2008) when referring to aviation-related mental models. These concepts of doing appear well-modeled by the DMM structure. TM is a construct that refers to pilot awareness of how, what, and when tasks should be done. SA level 3 and PM support the projection of current states into the future, which indicates that the DMM also reflects the mental simulation aspect of “doing”. The TRP component was also reliant upon knowledge regarding when and how to act upon the system: thus representing “doing”.

Authors who described mental models for pilots also tended to refer to cognitive mechanisms associated with mental models. Decision-making and selective attention (Herdman & LeFevre, 2004; Histon & Hansman, 2008) are two aspects of cognition that appear well covered by the DMM, as they reflect some aspects of all four formative indicators. Problem-solving and long-term memory (Herdman & LeFevre, 2004; Histon & Hansman, 2008) seem less well indexed by the DMM, although some aspects of these cognitive functions should be associated with managing tasks and prospective memory.

Situation awareness and understanding future states of the system were common elements to aviation-specific descriptions of mental models. As shown in Table 8.1, common to mental models in aviation was the idea that long-term memory and attention were essential to the structure and purpose of the mental model. While SA level 1, the perception of information from the environment, was removed from the DMM and recast as iRES, two other aspects of situation awareness remained as part of the DMM: the measures reflecting comprehensive mental representations of

the present and future environment reflected the SA latent variable. The lack of fit for the measures of *heading deviations* and *SA level 1*, and their subsequent removal from the final DMM does not appear to have impacted the DMM's coverage of key aspects of mental models. For example, the manifest indicator representing heading maintenance was removed during model generation, as there was a negative relationship between heading deviations and the two other trajectory measures. The negative relationship with heading deviations might have reflected a trade-off in attention, where greater deviations in heading were incurred to permit greater control on airspeed and altitude, two parameters where excessive deviation is associated with greater risk. Thus the TRP latent variable, which consistently showed relatively strong association with the DMM, still appeared to represent relevant aspects of procedural knowledge, described by the literature as *theoretically* pertinent to mental models. It should be noted that the literature does not contain examples of empirical testing regarding what facets of perceptual-motor coordination, SA, or task management should be present when indexing a mental model for pilots. The weaker aspects of mental model descriptions covered by the DMM were the mental model's reflection of system purpose, a tool for reasoning, workload factors, and perceptual-motor coordination. However, these four aspects of mental models were also less commonly noted in the literature.

Table 8.1

Comparison of the DMM Components to Literature-based Descriptions of Mental Models

Literature-based Descriptions of Mental Models	Primary References	Represented by DMM Components	Coverage by DMM
System-related			
Understanding of system purpose	Cooper, Reimann, and Cronin, (2007), Rouse and Morris (1994), Sheehan and Tessmer (1997)	TM	Weak
Understanding of system form	Cooper, Reimann, and Cronin, (2007), Hardy and Parasuraman (1997) , Rouse and Morris (1994), Sheehan and Tessmer (1997)	SA, TM	Fair
Situation specific	Histon and Hansman (2008) , Herdman and LeFevre (2004) , Sheehan and Tessmer (1997)	SA, TM, TRP, PM	Strong
Doing			
Procedural knowledge/scripts	Endsley (1995) , Sheehan and Tessmer (1997)	TM, SA, TRP	Fair
Mental simulation/Projection to future states	Endsley (1995) , Gentner and Stevens (1983), Histon and Hansman (2008)	SA, TM, PM	Fair
Goal-oriented	Endsley (1995) Histon and Hansman (2008)	TM, PM	Fair
Task specific	Endsley (1995) Histon and Hansman (2008) , Herdman and LeFevre (2004)	TRP, PM	Fair
Implementation process	Histon and Hansman (2008)	TRP, PM	Fair
Instantiated sets of schema (data structures organized in long-term memory)	Endsley, 1995 ; Holland, Holyoak, Nisbett & Thagard (1986); Reason, 1990; Wilson & Rutherford (1989)	TM, PM	Fair
Perceptual-motor coordination	Hardy and Parasuraman (1997) , Herdman and LeFevre (2004)	TRP	Weak
Situation Awareness			
Situation awareness/Internal representation of environment	Hardy and Parasuraman (1997) Histon and Hansman (2008) , Herdman and LeFevre (2004) , Karikawa et al. (2006)	SA, TM, TRP, PM	Strong
Over-simplified and flawed representations of the world	Doyle et al. (2000), Norman (2006)	SA, TM, TRP, PM	Strong

Literature-based Descriptions of Mental Models	Primary References	Represented by DMM Components	Coverage by DMM
Learning and Understanding Relations			
Reflection for what is believed about the world	Rouse and Morris (1994), Johnson-Laird (2004)	SA, TM, TRP, PM	Strong
Mental structure for organizing relevant information about the world	Besnard (2004) Herdman and LeFevre (2004) , Johnson-Laird (2004)	SA, TM, TRP, PM	Strong
Mental structure for discarding irrelevant information about the world	Herdman and LeFevre (2004) , Besnard (2004)	SA, TM, TRP, PM	Strong
Tool for reasoning	Craik (1967); Rasmussen (1979)	SA, TM	Weak
Relations between contents of the world	Vygotsky, Bruner in Lutz and Huit (2004); Rasmussen (1979)	SA, TM, TRP, PM	Strong
Enable and support learning about the world	Vygotsky, Bruner in Lutz and Huit (2004)	SA, TM, TRP, PM	Strong
Cognitive Functions			
Decision-making	Histon and Hansman (2008)	TM, PM	Fair
Problem-solving	Histon and Hansman (2008)	TM	Weak
Selective attention	Histon and Hansman (2008) , Herdman and LeFevre (2004)	SA, TM, TRP, PM	Strong
Long-term memory	Histon and Hansman (2008) , Herdman and LeFevre (2004)	SA, TM, TRP, PM	Strong
Other Characteristics			
Captures variance in human-machine interaction	Histon and Hansman (2008) , Herdman and LeFevre (2004) , Sarter & Woods (1992)	SA, TM, TRP, PM	Strong
High correlation between elements not expected	Herdman and LeFevre (2004)	SA, TM, TRP, PM	Strong
Multidimensional	Herdman and LeFevre (2004)	SA, TM, TRP, PM	Strong
Dynamic	Endsley (1995) , Herdman and LeFevre (2004) , Histon and Hansman (2008)	SA, PM	Fair

Literature-based Descriptions of Mental Models	Primary References	Represented by DMM Components	Coverage by DMM
Develops slowly and with experience	Hardy and Parasuraman (1997), Norman (2006)	SA, TM, TRP, PM	Strong
Reflects individual differences	Hardy and Parasuraman (1997)	SA, TM, TRP, PM	Strong
Responsive to workload	Histon and Hansman (2008), Herdman and LeFevre (2004), Sarter & Woods (1992)	TM	Weak

Note. Coverage scores were labeled as Weak (1), Fair (2-3), or Strong (4) in accord with the number of DMM components that reflect descriptions of the literature-theorized mental model components. Bold-text references refer to aviation-specific descriptions of mental models.

8.1.3 Redundancy in the DMM Specification

A second area for evaluation regarding formative constructs pertains to composite redundancy (Petter et al, 2007). The SA, TRP, TM, and PM constructs were posited as contributing *uniquely* to the DMM. A balance should be struck between avoiding multicollinearity, stemming from indicators with high intercorrelations, and specifying an incomplete array of formative composite indicators, which could leave the formative construct deficient in meaning (Petter et al, 2007).

Factor redundancy is determined after modeling by examining the value of the variance inflation factor (VIF) associated with each composite of the DMM. Variance inflation is indicated by strong correlations between factors and by VIF values above 3 for any factor (Kock & Lynn, 2012). Some overlap between concepts in the DMM was expected; for example, both TM and PM relate to flight tasks. The TM measurement model indexes pilot subjective ratings of their awareness of tasks and the PM variable represents actual performance on a prospective memory task.

However, results found that all VIFs were low, and correlations between DMM components were low or non-significant.

The evaluation of the DMM as a predictor and mediating variable in the COMP-A is discussed subsequently and speaks further to the DMM's completeness and utility in predicting aviation outcome measures.

8.1.4 Summary of the DMM Validation

The DMM appears to reflect core aspects of mental models as described by the literature. A reflective-formative multidimensional model that was specified with eight manifest indicators addressed construct completeness while avoiding composite redundancy and multicollinearity issues. Although the addition of unique observed indicators, perhaps pertaining to knowledge of system purpose or form, might have added additional facets to the DMM, over-specification can lead to poorly constructed latent variables. Focussing on fewer but higher quality indicators is preferred (Hayduk & Littvay, 2012). The careful selection of manifest indicators and a model generation process, carried out with strict attention to reliability and validity, resulted in a formative measurement model with strong model quality and fit coefficients.

The predictive role of the DMM in a nomological net also provided evidence of the construct validity of the DMM. Before posited as the central feature of the COMP-A the DMM was examined in a small nomological net to test for construct and criterion validity. The DMM demonstrated significant relationships with a relevant antecedent, iRES, and with the criterion variable, critical incidents. The results of testing alternate models, by removing the first-order constructs from the DMM (one by one) to allow them to individually predict critical incidents, suggested that the four-construct structured DMM delivered a parsimonious construct, useful for predicting

an important aviation outcome. Confidence in the results of the more complex COMP-A, where the DMM acted as a direct predictor of aviation outcomes and a mediator of upstream effects, appeared warranted after testing and validation procedures.

8.2 Major Hypothesis Two: Modeling the DMM in the COMP-A

As an aviation-domain-dependent variable, the DMM was hypothesized to have a stronger influence on critical incidents and incursion management when compared to domain-independent cognitive factors. A prerequisite of this hypothesis was that the DMM was also a significant predictor of aviation outcomes.

8.2.1 Predictors of the DMM

Effects of pilot expertise on the DMM. As expected, higher levels of expertise were associated with better performance for iRES and the DMM. Pilot expertise was the second largest predictor of the DMM (after pilot age). As shown in Figures 7.4, 7.7, and 7.8, a clear path of effects from expertise, to iRES, and to the DMM was observed.

Effects of pilot age on the DMM. The relative total effects of age associated with the DMM were moderate. The COMP-A revealed that about half of the total effects of age on the DMM were mediated by a combination of three domain-independent variables and iRES. The significant effect of pilot age on GA flight performance has consistently been shown in the large-scale GA pilot and aging program of research centred at Stanford University and the Palo Alto Department of Veterans Affairs (cf. earlier work by Taylor et al. [1994] and including recent work by Kennedy, Q. et al. [2013]). The deleterious effects of age have been found for various aspects of flight performance, such as course deviations, cockpit monitoring and communication tasks, landing decisions, and approach parameters. Other studies have

also found that older age was correlated with poorer performance for detection of cockpit malfunctions or loss of separation with other traffic (Coffey et al., 2007) and with flight path deviations (Causse et al., 2011).

Domain-independent cognitive variables as mediators of expertise and age effects on iRES and the DMM. Expertise was significantly associated with the more task-relevant domain-independent measures (WL_PDT, INHIB, and DCAT), thus these measures mediated some of the effects of expertise on iRES and the DMM. The factor representing pilot expertise was a formative variable that included pilot rating, total and recent hours flown, and simulator hours. This combination of experience was selected as each variable appeared to represent a unique and relevant facet of pilot expertise. That expertise can influence the more aviation-related domain-general cognitive mechanisms was demonstrated by Nunes and Kramer (2009) in a study investigating the effects of age on air traffic controllers. Significant relationships were found between expertise (comparing controllers and age-matched non-controllers) and inhibitory control and a span of attentional field test. These results, in conjunction with the current findings, demonstrated that expertise could be associated with more cognitive resources for secondary tasks, better inhibitory control, and better visual-spatial attention: all domain-independent measures. Thus, higher levels of expertise appear to augment lower level aspects of cognition when indexed by tasks with domain-dependent features (e.g. tracking stimuli, decision-making etc.).

Age effects on iRES and the DMM were also mediated by domain-independent variables. The largest mediation of the effects of age on the DMM was observed for visual-spatial attention. Visual-spatial attention has been studied in pilots by Tsang and Shaner (1998) where older age was associated with reduced performance on two visual tracking tasks for both pilots and non-pilots. Spatial

abilities were also highlighted as mediators of pilot performance by Morrow et al. (2003), where older pilots demonstrated poorer performance than younger pilots on read-back accuracy for ATC tasks. With respect to the remaining domain-independent cognitive factors and iRES, comparatively small mediation of the effect of age on the DMM was found for iRES, working memory, and processing speed. The variable representing cognitive inhibition did not mediate any of the effects of age on the DMM.

It has frequently been speculated that cognitive processing speed may be the primary mediator of age effects on complex behaviours (Kennedy, Q. et al., 2013; Salthouse, 1996; Taylor et al., 2000). In contrast, the present results indicated that although processing speed was correlated with age and study outcome measures, it was visual-spatial attention that transmitted the majority of indirect effects of age on flight performance. In addition, although the DMM was predicted either directly or indirectly by domain-independent variables, the least of these were processing speed, working memory, and inhibition. These results vary from those of Taylor et al. (2000) who found that a composite variable composed of CogScreen-AE speed and working memory measures (comparable to the present working memory and processing speed latent variables) was the primary predictor of summary flight scores (that also included a task-related performance component similar to the DMM). The restricted age of the Taylor et al. sample might be the reason for the differences in the results found in the present work. In the Taylor et al. sample of pilots aged 50 years and older, 33% of the variance in the flight summary scores was explained by the speed and working memory composite. A further 12% of the variance in the flight summary scores was explained by four other composite variables representing “attribute identification” (comparable to the cognitive inhibition latent variable), motor

performance (from the Path-Finder task), tracking (comparable to the visual-spatial attention latent variable), and visual associative memory (three symbol-digit coding measures). Similar to the present findings, Taylor et al. (2000) observed that pilot age accounted for a significant amount of variance in flight summary scores beyond that accounted for by the four cognitive composites. The age-restricted analysis (ages 60 to 81) from the current work demonstrated results in line with those of Taylor et al.. In the subsample of pilots aged 60 and older DCAT, comparable to the speed and working memory factors from Taylor et al., showed a strong association with the DMM, $\beta = 0.57, p < .01$. Despite their differences, both the findings from Taylor et al. (2000) and the present work demonstrate the importance of cognitive factors in predicting pilot performance for a sample of older pilots.

8.2.2 Predictors of Aviation Outcomes in the COMP-A

Effects of pilot expertise on outcomes. The mediation of expertise effects on critical incidents, and incursion management by INHIB and DCAT was not significant. The lack of significant mediation of expertise effects on aviation outcomes by domain-independent variables resulted from the small effects of these variables on outcome variables. The DMM fully mediated the effect of expertise on critical incidents, whereas iRES only partially mediated the effect of expertise on the DMM.

As was shown earlier by Van Benthem and Herdman (2014b), in comparison to other variables, expertise (pilot rating) was the strongest predictor of incursion management. That pilot expertise was associated with better performance was not unexpected as previous work has shown that a variety of indices of expertise were associated with better flight performance (Causse et al., 2011; Morrow et al., 2003; Taylor et al., 2005; 2007; Yesavage et al., 2011). There is interesting precedent in the

literature where more experienced pilots made better overall decisions, yet responded slower to tasks than novice pilots (Stokes & Downs, 1998). The incursion management scenario was designed so that, if participants detected the incursion at an early point in the approach, the avoidance maneuver could be undertaken with reasonable speed and adequate time to proceed to a controlled go-around, *and* complete a radio call informing circuit traffic of their intention. The timing facet of the incursion management score reflected a safe and reflective switch to new procedure, and did not reflect observed motor speed, *per se*. Therefore it appears that more experienced pilots noted the runway incursion earlier on approach, and could therefore proceed to a new course of action with adequate time and complete the associated radio call.

Effects of age on outcomes. The DMM, and iRES to a lesser extent, fully mediated the significant effects of age on critical incidents. The finding that older age was associated with a greater number of critical incidents was expected, as flight simulation studies have shown poorer decision-making for older pilots (Kennedy, Q. et al., 2010). Database investigations of actual aviation incidents, although not conclusive due to issues with flight time exposure for older pilots, have typically pointed to older pilots being involved in relatively more accidents, odds ratio = 2.2 for pilots over aged 54 compared to the 25 to 34 year-old reference group (Li et al., 2005).

Some effects of age on incursion management were mediated by the DMM, whereas a large portion of the age effect on incursion management could be attributed to the interaction between age and the DCAT. In accord with this interaction, older pilots with low DCAT scores were significantly more likely to have lower incursion management scores. In contrast, younger pilots' performance handling the surprise

incursion was not as strongly modulated by their DCAT scores. Due to their non-significant relationships with the outcome variables, processing speed, working memory, and inhibition did not mediate any relevant effects from pilot age or expertise on critical incidents or incursion management. Visual-spatial attention was the only domain-independent cognitive factor to have a significant relative total effect size for critical incidents.

Interaction effects of expertise and age. While interaction effects were tested at each opportunity, expertise did not significantly moderate the effect of age on outcome variables in the present work. The lack of substantial interaction between pilot age and expertise is also supported by other analyses of pilot age and flight performance. Taylor et al. (2005) found that older pilots performed worse on the communication task and that higher levels of expertise did not protect older pilots from this association. In the later Taylor et al. report (2007), higher-rated pilots also showed fewer declines over time, but expertise offered some protection against age-related declines in the communication task performance. Reading charts and interpreting ATC messages were performed better by younger, rather than older pilots: an age by expertise interaction was not significant (Morrow et al., 2003). However, some of the effects of age on performance may have been attenuated for older pilots with more flying experience. It was also speculated that memory and spatial abilities may have mediated performance on the ATC tasks (Morrow et al., 2003). Yesavage et al. (2011) also reported that higher rating was associated with better summary flight scores for all pilot age groups, but expertise did not significantly mitigate the effects of age. Kennedy, Q. et al. (2010) demonstrated that expertise provided an advantage for older pilots on the simulated flight control task of banking performance in a holding pattern. For other flight performance parameters measured in the Kennedy, Q.

et al. (2010) study (e.g., approach, landing decisions) expertise did not attenuate the negative association between age and performance. When recent pilot-in-command hours replaced pilot rating as the proxy for pilot expertise, the interaction between age and expertise for the banking skill measure was no longer significant (Kennedy, Q. et al., 2010). In sum, results from the present research are similar to previous work, where the effects of older age were not seen for key aspects of pilot performance.

Confounding factors have proven difficult to control in analyses of age, expertise, and flight performance (Hardy & Parasuraman, 1997). Unless major confounds between age and expertise are accounted for it is difficult to make claims about the unique effects of either factor upon pilot performance. In studies of simulated pilot performance, pilot rating has been the most common index of expertise (Hardy & Parasuraman, 1997), although total hours flown (Causse et al, 2011) have also been incorporated as a control variable, where age associations with study variables conditioned on the reported hours flown. The present work attempted to control for potential confounds between age and expertise by defining expertise as a formative latent construct that was a weighted sum of important facets of pilot expertise. The composite expertise factor, formed by rating, hours flown, recent pilot-in-command hours, and total simulator hours, demonstrated that although pilot rating was the strongest contributor to the construct, the three remaining concepts also provided significant aspects to the EXPERT latent variable.

Explanation for the minimal role of processing speed in the COMP-A.

Domain-independent variables mediated very small effects of age on the DMM. This was found despite the large direct effect of age on processing speed. The minimal transmission of age effects on study outcomes by domain-independent measures is contrary to Taylor et al. (2005) who suggested that processing speed and cognitive

inhibition were possible mediators of the effects of age on study outcomes. In the present SEM results, the weak total effects of processing speed and inhibition on the DMM and critical incidents suggested that the indirect effects of pilot age were not transmitted by processing speed or inhibition in any substantial amount. Although strong effects of age on processing speed are common in the literature, the transmission of age effects on aviation outcomes are likely not mediated by processing speed due to the almost zero relative total effects of processing speed on the DMM and critical incidents. This result is not surprising, for the reasons discussed in Chapter 2. On this view, processing speed is not a true causal factor in relation to complex cognitive behaviors. It is plausible that age effects associated with complex cognitive performance are mediated by concomitant age-associated changes in brain morphology, and are not caused by a general slowing in processing speed (Rabbitt et al., 2007). Multivariate analyses have shown that processing speed might explain variance in some aspects of cognition, such as working memory, but it does not explain other measures of intelligence (Rabbitt et al., 2007). In accord with Rabbitt et al. (2007, p. 368), presumed links between processing speed and performance (such as aviation performance, Taylor et al., 2005) may represent an unwarranted association between “functionally unrelated processes of change in the aging brain”.

A biological explanation for nonsignificant processing speed effects in the COMP-A. While the relative total effects of processing speed for the DMM and study outcomes were nonsignificant, there remained the direct correlations between processing speed and the DMM and critical incidents to explain. A possible explanation for the weak association between processing speed and the DMM, or critical incidents, may be related to the various changes to white and grey matter in the brain that accompany older age. Both processing speed reductions and poorer DMM

performance may be a result of reduced efficiency of brain networks. As postulated in Chapter 5, the architecture of the DMM assumed coordinated function of the multiple brain networks that mediate SA, TRP, TM, and PM (allowing for some overlap of areas). It is known that the elegance with which brain networks operate to perform complex functions deteriorates with older age. Some descriptions of these changes to functional networks include *PASA* (Davis, Dennis, Daselaar, Fleck, & Cabeza, 2008) and *HAROLD* (Cabeza, 2002), where changes to networks reflect more posterior-anterior shifts in activation, and reduced hemispheric asymmetry of activated brain regions in older adults. Thus longer response times typically noted in processing speed tasks are potential artifacts of the increased time needed for neural processing in the expanded networks in older subjects.

Kennedy, K. et al. (2015) investigated how brain function was affected by older age. In this study, brain activation during easy versus hard language tasks under varying task load conditions was studied. For the easy task, less restricted activation of the default node network as a function of older age was observed, whereas younger subjects showed the expected deactivation of the default node network (i.e. older subjects did not show this anticipated pattern of deactivation to the same degree). For the older subjects, expanded, and thus potentially slower networks, appear to serve a compensatory role for managing the task: the less efficient networks found in the older subjects appeared to mediate equivalent accuracy for behavioural measures when compared to the scores from the younger group. However, behavioural response times increased significantly and linearly with older age (Kennedy, K. et al., 2015). Furthermore, when additional recruitment of cognitive control areas was expected, as a function of task load, older subjects had reduced activation in brain regions that did show recruitment for the younger subjects (Kennedy, K. et al., 2015). Finally, and

relevant to the potentially long range networks plausibly required of the DMM, studies of the default node network found that long range paths within the network were more vulnerable in older subjects (Tomasi & Volkow, 2011). It could be speculated that just as inadequate communication between long range circuitry in the default node network occurred for older subjects, so too might other long range networks that mediate the DMM be degraded in older subjects. Thus, the expansion of neural circuits, either compensatory or otherwise, may be the cause of longer response times for older participants. The results of Kennedy, K. et al. (2015) and Tomasi and Volkow (2011) suggest that deterioration of highly skilled tasks, such as flying an aircraft, may result from age-related changes to communication within networks of brain regions. Thus older adults experience changes in network patterns, which may mediate both the reduction in DMM performance and the increased cognitive processing times.

Control variables in the COMP-A. Mission difficulty acted as a control variable in the COMP-A to ensure that the causal relationships specified in the model were robust in both low- and high-difficulty conditions. In both main analyses mission difficulty predicted the DMM and iRES. The total effects of mission difficulty on critical incidents were fully mediated by iRES and the DMM. This result answered a lingering question resulting from the experimental phase of this work, which queried whether workload effects on critical incidents were mediated by domain-dependent cognitive factors. In both analyses (see Figures 7.4 and 7.7) similar effects of mission difficulty on critical incidents were observed, and these effects were mediated (also similarly) by iRES and the DMM. The influence of workload states on pilot performance was expected, and follows findings from Svensson et al. (1999) who found that a variety of workload measures, both subjective and objective were

correlated with pilot performance. In a CB-SEM model of simulated fighter pilot performance high-difficulty mission conditions predicted an index of SA. The SA measure was a significant predictor of performance. However, it was not clear whether all the effects of mission difficulty on pilot performance were mediated by SA (Svensson et al., 1999). In the present sample of pilots, mission difficulty effects were fully mediated by domain-dependent variables. Further clarification of the effects of workload on pilot performance was illustrated by the finding in the present work that mission difficulty simultaneously significantly predicted iRES and the DMM; thus suggesting that workload effects on the DMM were only partially mediated by iRES. The inclusion of mission difficulty in the COMP-A also had a practical use in that it confirmed the experimental protocol where two distinct workload conditions were designed to emerge from the differences in the circuit features at the Yorkton, SK and Kaslo, BC aerodromes.

The second control variable was the episodic memory measure. Episodic memory was used in the SEM model as a control variable, to increase the *isolability* of SA and iRES from other non-SA factors (such as long-term or episodic memory). The episodic memory measure resulted from one query that took place at each of the two SAGAT probes occasions. This was a complex query, as auditory and spatial information had to have been stored and maintained for several minutes, without instructions from the experimenter to do so. This retrospective memory measure did not predict the DMM, yet it did correlate significantly with SA and TRP individually. Episodic memory also predicted iRES, critical incidents, and incursion management. The effects of episodic memory appeared to be partially mediated by iRES. The relationship of episodic memory to aviation outcomes in this study was not expected, and perhaps represents a novel finding. Until this point, there has been no direct link

made between retrospective memory and simulated flight performance for older GA pilots.

8.2.3 Diagnostic value of the DMM

Decades of work on driver safety among older drivers has shown that various cognitive measures correlate with driving performance. Nevertheless, measures that correlate moderately with performance have not been shown to consistently discriminate between low- and high-risk older drivers (Crizzle, Classen, Bedard, Lanford, & Winter, 2012; Dobbs & Shergill, 2013). Understandably, efforts are recently being made to examine measures for their diagnostic potential to classify older adults as low- or high-risk for a driving incident before recommending these measures be used as decision criteria for driver risk (Bouillon, Mazer, & Gelinas, 2006; Crizzle et al., 2012, Dobbs & Shergill, 2013). As frequently referenced in the present work, numerous studies in the aviation domain have also searched for domain-independent measures that predict pilot performance. For example, a study of airline pilots examined pilot attributes, domain-independent, and domain-dependent variables on simulated B727 performance (Hyland, Kay, & Deimler, 1994). Older age (range 41 to 71 years) was significantly associated with poorer performance on both the domain-independent and domain-dependent measures. Older age was also associated with lower subjective ratings by raters for simulator measures, but was not associated with objective flight performance in the B727 simulator in this sample of high-expertise professional pilots. One set of domain-independent measures, from CogScreen-AE, was also correlated with the subjective ratings of pilot performance. The conclusion was drawn that the accuracy (predominantly working memory) and processing speed

CogScreen-AE measures could discriminate between low- and high-performing pilots (Hyland et al., 1994).

While processing speed and working memory have generally been shown to be correlated with some facets of pilot performance, the extent of these measures to accurately diagnose at-risk pilots is not known. Some work has been initiated in this regard. Yesavage et al. (2011) found that a slower than median-split rate of decline (over three years), for a sample of older pilots, could be predicted by initial performance in the simulator, and by ROC-determined cut-scores from a processing speed and an executive function measure (both derived from CogScreen-AE). The median-split rate of decline was based on 23 simulator measures, forming four basic categories of performance. The “decision-tree” produced for that analysis resulted in 55 out of 92 pilots correctly identified as having a *slower* than median rate of decline and 60 out of 92 pilots correctly identified as having a faster than median rate of decline (an average of 63% accuracy). A follow-up analysis in that study found that processing speed was only associated with the landing approach measures of performance. A reasonable conclusion from the Yesavage et al. (2011) study is that processing speed and executive functions provide a means for correctly classifying pilots (mean age= 58 years), regarding their rate of performance change over time, at a level 13% over the chance rate (50%). Issues with their findings are the relatively low diagnostic indices of sensitivity and specificity (approximately 0.60), the low likelihood ratio ($LR+ = 1.68$ [calculated for this work using an online diagnostics tool]), and a lack of a meaningful criterion variable. It is not known whether a rate of decline that is steeper than the average rate translates into increased pilot risk.

The efforts to discover diagnostically meaningful measures should be applauded, as it is critical for the equitable assessment of older pilots. However, the search for a

diagnostic tool has so far relied primarily on processing speed and working memory measures, which, though likely to have negative correlations with age, may not reflect pilot risk for poor outcomes. In Chapter 9 recommendations regarding the development of domain-dependent diagnostic tools are discussed.

8.2.4 Mental task load in the COMP-A

When the DMM was conceptualized in Chapter 5, pilot mental task load was removed as an indicator of the DMM and repositioned as the domain-independent cognitive variable (the PDT) in the COMP-A. In this role, mental task load was defined as an index of cognitive reserve, such that performance on the PDT reflected the mental capacity available for domain-independent functions (e.g., detecting and responding to task irrelevant auditory stimuli). This PDT index of mental task load was significantly correlated with most variables in the COMP-A, but was most strongly associated with the DMM. Despite the significant correlation of the PDT with the DMM, it explained relatively small amounts of variance in the DMM, in comparison to other predictors. In sum, the objective measure of mental task load correlated as expected with age, expertise and a host of domain-independent and domain-dependent cognitive variables, but was a poor predictor of critical incidents and incursion management. This finding suggests that in comparison to other cognitive constructs and to pilot attributes mental task load is of less importance than expected in comprehensive models of performance.

8.2.5 Summary of predictors of the DMM and aviation outcomes

Structural equation modeling of the COMP-A with the DMM as the principle variable demonstrated the value of domain-dependent cognitive factors in predicting aviation

outcomes. Domain-independent factors, such as processing speed and working memory, predicted some variance in the performance of pilots. Yet, the utility of these domain-independent variables was overshadowed by the predictive value demonstrated by the domain-dependent factors of iRES and the DMM. Visual-spatial attention was similar to iRES in its role as a predictor of the DMM in both COMP-A analyses. However iRES was a better predictor of critical incidents. Likewise, in the second analysis, iRES and DCAT predicted the DMM almost equally, but again, iRES was a better predictor of critical incidents. Regarding incursion management, a different pattern emerged, where the DCAT explained more variance than iRES, albeit a very small amount.

The difference in the DMM's ability to predict critical incidents versus incursion management demonstrates where the DMM's application to pilot safety was most useful. The differential utility of cognitive variables in predicting aviation outcomes was discussed in Taylor et al. (2000), where authors suggested that to fully appreciate predictors of pilot performance multiple criterion variables should be evaluated. Also echoing Hardy and Parasuraman (1997), Taylor et al. recommended that predictors of pilot performance should include both domain-general and domain-specific cognitive variables.

In sum, Figures 7.6 and 7.10 clearly portray the DMM as the principle variable in the COMP-A when critical incidents was the criterion variable. The relative total effects of the DMM for critical incidents was moderate to high; despite the inclusion of pilot age, a comprehensive latent variable representing expertise, four domain-independent cognitive variables, and iRES.

The DMM was also predicted either directly or indirectly by all upstream variables. Pilot age, expertise, iRES, visual-spatial attention, DCAT, and mental task

load demonstrated the largest effects on the DMM. The ROC diagnostic analyses indicated that the DMM factor scores were better predictors of pilot risk than any of the DMM's individual components (SA, TRP, TM, and PM).

8.3 Major Hypothesis Three: The DMM as a Mediator of Effects in the COMP-A

The DMM was predicted by pilot attributes (e.g., age, expertise) and by domain-independent measures, especially visual-spatial attention. iRES, a domain-dependent index of auditory and visual information processed and retained for short periods during flight, was also closely associated with the DMM. As hypothesized, the DMM fully mediated the effects of cognition, age, expertise, and workload factors on aviation outcome variables. As shown in Figures 7.4 and 7.7, the DMM mediated upstream effects for critical incidents. The DMM also fully mediated the effects of mission difficulty on critical incidents. Only the retrospective memory control variable, which did not significantly predict the DMM, had a significant direct effect on critical incidents in the COMP-A.

Despite its strong effects in two of the COMP-A models, the DMM did not appear to be a universal predictor of aviation outcomes. The third hypothesis was not confirmed for the incursion management outcome. In the present work, prediction of incursion management proved to be a more difficult modeling endeavour. In the COMP-A model the DMM did not fully mediate the effects from expertise or pilot age on incursion management. The relative total effects of the DMM for incursion management appeared to be tempered primarily by pilot expertise, and by the explanatory power of the interaction between pilot age and the DCAT latent variable.

Thus, the DMM, conceived as a mental model necessary for basic aviator competency, was an excellent predictor of critical incidents, but was not as useful in

predicting a less commonly studied aviation outcome: handling a surprise runway incursion. Below, the differential effects of the DMM on outcomes in the COMP-A is interpreted in light of the “working mental model” for ATC by Histon and Hansman (2008) where libraries of mental models were emphasized.

8.3.1 The “Working Mental Model” for ATC

Histon and Hansman (2008) describe “working mental models” for air traffic controllers (ATC) that are central to ATC cognitive processes. Working mental models are dynamic, task, and situation specific, and play a role in predicting future states of the system. According to Histon and Hansman (2008, p.71), mental models support the “generation and maintenance of situation awareness as well as various decision-making and implementation processes.” The working mental model is one component in the larger ATC cognitive system (Histon & Hansman, 2008, p. 72). In accord with Histon and Hansman, as different situations arise during flight, pilots will select from a “library” of mental models. Each mental model is populated by skills and knowledge designed to handle the current situation. For example, in ATC environments a basic mental model will support the monitoring of current and future status of numerous aircraft. Off-nominal events, or abrupt changes in task workload may necessitate the transition to alternative mental models stored in long-term memory. Structures, more complex than the COMP-A, that operationalize numerous mental models would be more likely to predict off-nominal events, such as surprise runway incursions. It would seem that dealing with a sudden and unexpected runway incursion upon final approach might require access to a specialized procedure: a procedure that was not an element of the base model DMM. This suggestion is supported by the incursion management COMP-A where the DMM directly influenced the outcome to a lesser

degree and mediated small-sized effects from upstream variables, but significant effects from the retrospective memory variable explained unique variance in incursion management scores. Thus, the DMM, as a mental model for basic pilot competencies, weakly predicted the incursion management variable, but retrospective memory did predict incursion management. The association of retrospective memory with incursion management reflects pilot ability to quickly retrieve a new mental model, one that supports the handling of a surprise event upon approach.

8.4 Summary of the DMM and COMP-A Validation

The COMP-A was based on a cognitive systems engineering design where cognitive factors were central to modeling the human-machine system. The constituent variables selected for the COMP-A followed recommendations by Hardy and Parasuraman (1997) that pilot attributes, system factors, domain-independent and domain-dependent cognitive factors predict relevant aviation outcomes. A key factor in the COMP-A was a variable representing a mental model for pilots. Again, following the recommendations of Hardy and Parasuraman (1997) and Herdman and LeFevre (2004) a mental model for pilots was posited centrally in a cognition-oriented model of flight performance. Using this structure, three main hypotheses were validated.

First, the DMM as a separate entity in the COMP-A was found to be a conceptually and empirically valid model. The DMM appeared to reflect key facets of mental models, as described by the literature, including specific literature pertaining to the role of mental models in aviation-related human-machine systems. The DMM, constructed as a reflective-formative multidimensional model showed good model fit and acceptable quality coefficients. To ensure validity of the DMM specification, PLS-SEM and CB-SEM model generating and testing methodologies were used. The

DMM composites (SA, TRP, TM, and PM) were constructed from well-fitting measurement models that individually showed relevant contributions to the DMM, and significant relationships to upstream and downstream variables in the COMP-A.

Second, the COMP-A appeared well-specified in regards to the relationships between the DMM and upstream predictors, and downstream criterion variables. The direction of causal assumptions (e.g., iRES predicted the DMM, and the DMM did not predict iRES) were also verified with four causality indices. The inclusion of control variables adjusted for possible confounds in interpreting the relationship between the DMM and other variables. For example, retrospective memory did not jeopardize the interpretation of a causal relationship between iRES and DMM. The inclusion of control variables also ensured that the relationships of the DMM to other variables were robust to two levels of mission difficulty. In addition, no concerns regarding endogeneity were found for the COMP-A, meaning that no glaring omissions of causal variables were noted. The COMP-A appeared to be a valid environment for testing the second hypothesis that the DMM was a significant predictor of outcome variables and was also significantly predicted by upstream variables.

Regarding the COMP-A outcome variables, the DMM was the primary predictor of critical incidents. In contrast, the DMM explained a small amount of variance in the runway incursion outcome variable. Thus, the DMM may not be the primary predictor for all types of aviation outcomes. The management of a sudden off-nominal event was less well-predicted by the DMM than it was by pilot expertise and, for older pilots, working memory function. This differentiation in the prediction of pilot outcomes by cognitive factors was noted by Causse et al. (2011) where lateral deviations from the ideal trajectory were significantly correlated with reasoning, age, and working memory. In contrast, a second outcome variable, a decision-making task

(go versus no-go) related to the calculation of a crosswind factor, was associated with expertise and working memory and “set-shifting” (a task particularly reliant upon executive functions) (Causse et al, 2011). In the present work, set-shifting capacity was indexed by INHIB, a measure of inhibition and interference. As shown in Table 7.17, incursion management scores, which reflected an ability to switch from the impending touch and go to a go-around procedure, was not predicted by INHIB, but it was predicted by the DCAT latent variable. Despite the difference between this work, and the work of Causse et al. (2011) regarding the interference measure, both studies found that working memory indices were predictive of a task requiring a shift from a planned to a new procedure. Appropriate responses by pilots were reliant upon quick mental calculations of information detected from the flight environment (e.g., detection of a runway incursion on approach).

Third, in the COMP-A models that predicted critical incidents the DMM fully mediated significant effects from upstream variables. Less effects from upstream variables were mediated via the DMM in the model predicting incursion management. When the DMM did not mediate effects from antecedents in the COMP-A, this was associated with the very small effects of these antecedents for the outcome measures.

The COMP-A was also an appropriate environment for examining the relationships between age and domain-independent cognitive factors. The present work showed that age had detrimental associations with all the domain-independent variables. The COMP-A demonstrated that the effect of age on working memory was fully mediated by processing speed, inhibition, and visual-spatial attention.

Expertise was found to positively influence some of the more complex domain-independent variables, as well as the domain-dependent factors of iRES and the DMM. Expertise was not found to interact with pilot age in a way that would protect

older but more experienced pilots from the age-related reductions that were observed for number of flight performance variables. Thus expertise was beneficial for older and younger pilots alike. The finding that expertise was positively associated with domain-independent cognitive mechanisms suggested that expertise had “privileged access” for influencing some cognitive variables usually considered impervious to expertise effects. The complex measures representing inhibitory control (INHIB: CogScreen-AE), span of attentional field, and situational judgement (Subtest 2 and 6, DCAT), and mental task load (WL_PDT) were significantly associated with the comprehensive expertise, particularly in the reduced COMP-A model of pilots aged 60 years and older. The total effect sizes for these domain-independent variables on critical incidents were minimal. The DCAT latent variable and its interaction with pilot age were instrumental in predicting incursion management. Thus, although expertise effects on the DMM were partially mediated by the relationship between expertise and complex domain-independent measures, there was little evidence that expertise could influence risk of critical incidents via these domain-independent cognitive variables. In regards to incursion management, the DCAT did appear to mediate some of the effects of pilot expertise.

The COMP-A also included pilot self-ratings of SA and workload. Because of the importance of pilot self-evaluation (particularly for guiding general aviation pilot competency checks) it was anticipated that these non-causal but potentially predictive variables might be relevant in the COMP-A. Overall, self-rated SA and workload predicted little to no variance in critical incident rates. The exception was the self-rated SA item pertaining to pilots’ awareness of other aircrafts’ impact on their ability to fly the “perfect circuit”. Some utility was also found for the pilot self-rating of mental task load in the model predicting incursion management.

Finally, a reduced COMP-A model was produced to examine key predictors of critical incidents for pilots aged 60 to 81 years. In this older pilot analysis the relationships between age and the DMM with critical incidents was stronger than was found for the sample that included pilots 40 to 60 years. It appeared that the robust association between age and critical incidents in older pilots may have been mediated by the relationship of DMM to critical incidents, a relationship that became particularly strong after the age of 54 years. The role of cognitive variables in the COMP-A increased in a sample of pilots aged 60 and older, whereas the predictive utility of pilot expertise for critical incidents and incursion management was reduced. While this subsample was relatively small it demonstrated the value of a model of pilot performance focused on domain-relevant cognitive variables. Two-thirds of the variance in critical incidents was explained by a simple model that included expertise, the DMM, DCAT, iRES, and two control variables that contributed minimally. When considering just older pilots, a very large portion of variance in critical incidents could be explained by three latent cognitive variables and the expertise latent variable. The relationships revealed in the COMP-A model of pilot performance, and the findings from the ROC curve analysis demonstrated the value of the DMM. The minor predictive utility of domain-independent cognitive variables has implications regarding the identification of older pilots who may be at risk.

Chapter 9

Implications and Recommendations

Around dinnertime on June 3, 2010 a 78-year-old pilot and his 69-year-old passenger left their fishing lodge and flew over pristine Lac Berté in northern Quebec. After just a short time in the clear skies, the seaplane and its two occupants crashed violently into the mirror-like surface of the lake. Despite the pilot's history of diabetes, hypertension, and heart disease, the Transportation Safety Board of Canada found no sign of sudden incapacitation of the highly experienced pilot; nor was there any sign of engine or structural failure in the well-maintained aircraft. Fatal aviation accidents rarely leave so few clues, and the cause of this crash remains undetermined to this day. Although just a few experts were involved in solving the mystery at Lac Berté, preventing such tragedies should be the concern of all stakeholders in aviation safety. The knowledge gained from the present research addresses our ability to identify at-risk pilots before the occurrence of a critical event.

Based on a Transportation Safety Board aviation accident report (2010 - A10Q0087).

The following chapter discusses state of the art methods for identifying at-risk pilots and positions the outcomes of the present research in this context. After evaluating risk-classification techniques, a broad outline of steps that could be taken to apply the findings from the present work to general aviation is provided.

A common method for identifying cognitive risk factors has been to investigate correlations between executive cognitive functions and flight performance factors that are associated with poor outcomes. For example, Causse et al. (2011) found that better reasoning skills were correlated with accurate decisions to abort a landing where the crosswind exceeded the safety threshold for that aircraft. Unfortunately the reasoning measure was not investigated regarding its sensitivity and specificity for classifying

pilots as low- or high-risk for a poor landing outcome. On the other hand, Yesavage et al. (2011) reported a 60% diagnostic success rate for classifying a general decline in pilot performance (year 0 to year 3) in reference to the sample's median rate of decline using processing speed and executive function measures. A link between these flight summary scores and the probability of a poor outcome (e.g., a critical incident) could not be established from the Yesavage et al. analysis.

Researchers in the aviation domain who have attempted to predict outcomes of pilot training programs have also noted that domain-general cognitive assessment tools such as CogScreen-AE (Kay, 1995), MAB-II (Hill & Jackson, 1984), and MicroCog (Powell, Kaplan, Whitla, Catlin, & Funkenstein, 1993) were better at predicting class rank and academic grades than they were of pass/fail of the program (King et al., 2012). Even less is known regarding the ability of these tests to accurately identify actual risk for older pilots. It is possible that a combination of variables, as suggested by the COMP-A, will be necessary for accurately predicting cognitively at-risk pilots. *As illustrated by the DMM in the COMP-A, the present research suggests that the classification of pilots as low- or high-risk for a critical incident will be better when using domain-dependent cognition rather than domain-independent cognitive mechanisms.*

9.1 Current Protocols for Assessing Risk for Older GA Pilots

As shown in Figure 9.1, the current FAA guidelines prescribe that pilots with suspected cognitive health issues be tested with a large array of standardized neurocognitive tests (FAA, 2015). The list of required neurocognitive test battery for pilots with possible neurocognitive impairment demonstrates a strong reliance on domain-general cognition as markers for pilot risk. Standardized test results are sent to

the FAA where, presumably, a preponderance of evidence from the full list of tests indicate to reviewers a pilot's risk should they be permitted to fly.

- What is required in the "core test battery?"** The core test battery listed below provides a standardized basis for the FAA's review of cases, and must include:
- CogScreen-Aeromedical Edition (CogScreen-AE).
 - The complete Wechsler Adult Intelligence Scales (Processing Speed and Working Memory Indexes must be scored).
 - Trail Making Test, Parts A and B (Reitan Trails A & B should be used since aviation norms are available for the original Reitan Trails A & B, but not for similar tests [e.g., Color Trails; Trails from Kaplan-Delis Executive Function, etc.])
 - Executive function tests to include:
 - (1) Category Test or Wisconsin Card Sorting Test, and
 - (2) Stroop Color-Word Test
 - Paced Auditory Serial Addition Test (PASAT).
 - A continuous performance test (i.e., Test of Variables of Attention [TOVA], or Conners' Continuous Performance Test [CPT-II], or Integrated Visual and Auditory Continuous Performance Test [IVA+]), or Gordon Diagnostic System [GDS].
 - Test of verbal memory (WMS-IV subtests, Rey Auditory Verbal Learning Test, or California Verbal Learning Test-II),
 - Test of visual memory (WMS-IV subtests, Brief Visuospatial Memory Test-Revised, or Rey Complex Figure Test),
 - Tests of Language including Boston Naming Test and Verbal Fluency (COWAT and a semantic fluency task),
 - Psychomotor testing including Finger Tapping and Grooved Pegboard or Purdue Pegboard.
 - Personality testing, to include the Minnesota Multiphasic Personality Inventory (MMPI-2). (The MMPI-2-RF is not an approved substitute. All scales, subscales, content, and supplementary scales must be scored and provided. **Computer scoring is required.** Abbreviated administrations are not acceptable.)

Figure 9.1. Recommended neurocognitive battery for any suspected neurological impairment. Impairments include "Head trauma, stroke, encephalitis, multiple sclerosis, other suspected acquired or developmental conditions, and medications used for treatment, may produce cognitive deficits that would make an airman unsafe to perform pilot duties." The 2015 Guide for Aviation Medical Examiners available from the Federal Aviation Administration (2015, p.219). Retrieved from http://www.faa.gov/about/office_org/headquarters_offices/avs/offices/aam/ame/guide/, Page 219. Core tests for any suspected neuropsychological impairment.

Although the CogScreen-AE has shown some association with "line-check" performance of airline pilots (Hoffmann, Hoffmann, & Kay, 1998), its ability to identify GA pilots who are at risk for a critical incident due to neurocognitive impairment or cognitive decline has not been validated. Furthermore, a strong reliance on measures that emphasize cognitive speed may result in a bias against older pilots. For example, the CogScreen-AE Professional Manual indicates that pilots who obtain an LRPV score higher than 0.60 have a high likelihood of cognitive dysfunction (Kay, 1995). Notably however, in the present work, all but one pilot over the age of 53 who

completed the CogScreen-AE had an LRPV score above 0.60 (see Figure 9.2). Thus the use of the LRPV to identify at-risk pilots older than 53 years might result in over-identifying a large portion of GA pilots. Clearly, not all pilots above the age of 53 years in the present study incurred a critical incident or scored below 6 points in the incursion scenario (a score of 6 or lower indicating high-risk as identified by Van Benthem & Herdman, 2014b). Figure 9.3 illustrates the pattern of results in the present study regarding pilot age and the two outcome variables.

Figure 9.2. Relationship of CogScreen-AE LRPV. Note that only one pilot over the age of 53 in the present sample achieved an LRPV below the threshold of 0.60. The higher the LRPV is above the 0.60 threshold, the greater the likelihood of brain dysfunction (Kay, 1995).

The strong association of the LRPV with pilot age and a lack of baseline norms for GA pilots beyond age 50 years (norms associated with meaningful outcomes) suggest that results of CogScreen-AE testing should be used cautiously when predicting pilot risk for the older general aviation population.

Figure 9.3. Relationship of DMM factor scores to critical incidents and incursion management as a function of pilot age. Note that the grey-filled circles indicate pilots aged 44 years and above. Not all older pilots incurred critical incidents (any value above 1 on the y-axis in this plot), nor did they all score below 6 for incursion management, No. observations = 110.

As shown in Figure 9.1, the “trail making test” (Reiten, 1958) and the Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005) are also part of the FAA recommended test battery for neurocognitive assessment of pilots (FAA, 2015). In a related domain, however, the utility of these two tests in predicting driver outcomes is weak (Crizzle, et al., 2012; Rapoport et al., 2013). In a 2012 report on aging pilots and risk management to the International Congress of Aviation and Space Medicine, the Australian Civil Aviation Authority recommended the MoCA (and not the CogScreen-AE) as the neurocognitive test for use in the evaluation of older pilots with suspected cognitive decline (Fitzgerald, Cheng, Clem, Drane, & Hegde, 2012). However, the report does not make it clear why the MoCA was selected over other tests, or how effective the MoCA has been in identifying at-risk pilots.

9.1.1 ROC Curve Analyses Comparing CogScreen-AE Factors and the DMM as Classifiers of Critical Incident Risk

The COMP-A provides the opportunity to test the intra-domain relationships between cognitive variables such as processing speed, working memory, cognitive inhibition, and visual-spatial attention. These domain-general cognitive variables are the predominant cognitive health factors used to determine pilot fitness to fly in the CogScreen-AE (Kay, 1995). However, an ROC analysis undertaken with the present sample of pilots revealed the underwhelming relationship of domain-independent cognition to aviation outcomes. As shown in Table 7.19, and Figure 7.18, the ROC curve demonstrated that the DMM, a domain-dependent cognitive measure, showed more sensitivity and specificity in predicting a critical incident when compared to domain-independent cognitive measures from the CogScreen-AE.

The CogScreen-AE derives indices of performance called the Taylor Factor Scores. The five Taylor Factor Scores were also subjected to additional ROC analyses to determine if the specialized algorithms used by the CogScreen-AE might yield better results regarding the stratification of pilots as low- or high-risk for a critical incident. Only the Taylor Speed/Working Memory Factor approached significance, $AUC = 0.62$, $p = .06$, (compared with $AUC = 0.79$, $p < .001$ for the DMM) with the AUC confidence interval including the null reference line. As shown in Figure 9.4, the best cut-score trade-off for the Taylor Speed/Working Memory measure resulted in a sensitivity of 0.70 and a false alarm rate of 0.43. The Taylor Speed/Working Memory combination of subtests demonstrated a clear improvement over the use of latent variables derived from the CogScreen-AE (see Table 7.19), yet was unsuccessful in classifying pilots as low- or high-risk without falsely identifying a large number of low-risk pilots as high-risk.

Figure 9.4. Receiver operating characteristics (ROC) curve for the DMM (blue curve) and Taylor Speed and Working Memory (WM) factor (green curve) predicting critical incidents vs. no critical incident pilots. Orange triangles indicate optimal DMM and Taylor Speed and WM scores for classifying pilots as at-risk for a critical incident. Black line is reference line indicating the null 0.50 discriminant curve (basically chance). No. obs. = 110.

The finding that domain-dependent measures may be necessary to predict aviation outcomes is a challenge for stakeholders in aviation safety. The present results inform current protocols by suggesting that methods of assessing pilots that rely heavily on domain-independent measures may not be appropriate for identifying pilots who might be at risk, particularly older GA pilots. On the other hand, an assessment that generates measures similar to those used to model the DMM, namely SA, TRP, TM, and PM, could more optimally identify at-risk pilots, when compared to current screening tools.

9.2 Learning from the Domain of Older Driver Assessment

The domain of *driver research* is similar to aviation with respect to the complex array of cognitive functions that support operator competencies. Assessment of the competencies of older drivers shows that even though a domain-independent cognitive measure might correlate with driver performance this does not immediately entail the diagnostic utility of this measure regarding driver risk (Dobbs & Shergill, 2013). One cognitive screening test of older driver competence that has shown good diagnostic utility is the DCAT (Dobbs, Heller, & Schopflocher, 1998; Korner-Bitensky & Sofer, 2009). A computerized test of cognition, the DCAT is able to adequately classify drivers with cognitive impairment who made “discriminating” errors (e.g., unsafe left-hand turns) and those who made regular types of errors (e.g., rolling stop etc.) (Dobbs et al., 1998). An important feature of the DCAT is its inclusion of subtests that are highly-relevant to actual driving behaviour. Three of these are subtests 2, 3, and 6. These subtests reflect span of attentional field, complex judgement (mimicking the ability to make accurate spatial judgements for left-hand turns), and the situational judgement test that includes 10 videos of actual driving scenes. For example, with subtest 6 drivers watch videotaped traffic scenes and then make timed judgements regarding some aspect of safety in the scenes-. Importantly, the video scene is viewed from the exact perspective of the driver (Dobbs et al., 1998). In the present work, subtests 2 and 6 were combined to create one DCAT factor score. The DCAT latent variable was significantly associated with critical incidents and was also particularly useful in explaining the association between age and incursion management. In summary, the findings from the present work, the lack of acceptable domain-independent classifiers of older pilot risk, and the results of the search for older driver

determinants of risk conflict with current protocols for neurocognitive assessment of the aging aviator.

9.3 Alternatives to Domain-Independent Cognitive Predictors of Pilot Performance

9.3.1 Pilot Age

Age is used as a classification factor in GA regulations, with private pilots aged 40 years and older requiring more frequent medical certification (every two years) than younger pilots (every five years). Additionally, for Canadian commercial aviation there must be at least one pilot under 60 in the cockpit at all times, and pilots over 65 cannot hold the captain position. Despite the often deleterious influences of age found for various components of flight performance, the full mediation of age effects on critical incidents by the DMM suggested that age, by itself, is not the determining factor in pilot safety through the lifespan. Other factors, associated with advancing age, may reveal why older pilots tended to demonstrate poorer flight performance when compared to younger pilots. As described in Chapter 7 there was also the potential for age-related endogeneity in the COMP-A. It was possible that a third, unobserved, variable was missing: a variable that could explain the relationship between processing speed and cognitive inhibition. Although endogeneity between these domain-independent measures did not exist because no causal path was theorized to exist between them, there is still the possibility that some other predictor could explain the general trend for all the domain-independent measures to correlate with each other (see Table 7.17). Biomarkers for age are age-related physical changes that can occur naturally, or via disease processes. Some have suggested that age, as a predictive construct, could be eliminated in models of behaviour if a full array of biomarkers for age can be identified.

Biomarkers of Age in Aviation Research. Advances in understanding why older pilot age is associated with declines in performance have been made by examining potential biomarkers for age. Ideally, if a marker for age could be found that predicts performance more strongly than age alone does, then age could eventually be removed from the COMP-A (and replaced with one or more better biomarkers of age). The following studies are valuable for their ability to explain the relationship between pilot age and performance. For example, hippocampal volume was associated with better performance, but only for pilots with advanced rating (Adamson et al., 2010b). In other work, Apolipoprotein E epsilon 4 expression (APOE E), a marker associated with Alzheimer's Disease, was present in otherwise healthy pilots (Adamson et al., 2010a). This study did not test simulated pilot performance, but a relationship was found between the APOE E allele expression and performance on a visual-paired associate test. This visual-paired associate test was the same CogScreen-AE subtest utilized as an observed indicator in the working memory latent variable in the present study (SDCIRAC). While Adamson et al. (2010a) did not find differences in hippocampal volume between APOE E carriers, those pilots with the allele did have significantly lower performance on the visual-paired associate test. The relatively small total effect sizes of the CogScreen-AE measures for aviation outcomes temper the association of the APOE E protein expression and pilot performance. In a separate analysis, Adamson et al. (2010b) found that whole brain size, although negatively associated with age, was not predictive of simulated pilot performance, and did not explain the deleterious effect of age on simulator performance. The main effects of carrying the APOE E allele were not found for a summary flight performance score (Taylor et al., 2011). However, an unusual interaction was found such that only the carriers of the APOE E in the highest-rating group of pilots showed

no longitudinal decline in flight summary scores (from baseline to year 2). This interaction was not observed for any other pilot rating group. In summary, the examination of a very global measure, such as brain volume, and discrete measures such as hippocampal volume and APOE E expression have not revealed main effects, nor have they accounted for the effects of age associated with simulated flight performance. These findings suggest that other strategies, such as examining the relationship of the integrity of *networks of brain areas* that mediate complex cognitive behaviours (e.g., methods similar to Kennedy, K. et al. [2015] for lexical decision-making) to flight performance might yield more productive insights regarding biological markers associated with age and pilot performance.

9.3.2 The DMM as an Alternative to Age and Domain-Independent Cognitive Factors

The DMM was structured as four domain-dependent cognitive mechanisms that converge to form a mental model for pilots that supports basic aviator competency. The DMM fully mediated the effects of age and of pilot expertise on critical incidents. This means that theoretically, older pilots with higher DMM factor scores may have a reduced risk for critical incidents than their peers with low DMM performance (see Figure 9.3). Nevertheless, the strong association with age and the DMM highlights the relative importance of domain-independent cognition, in comparison to pilot attributes or domain-general cognitive variables, when predicting pilot risk. The present findings regarding DMM performance and age take on additional significance when compared with the distribution of pilot age in the primary association of pilots in Canada. As shown in Figure 9.5, the average age of pilot members of the Canadian Owners and Pilots Association (COPA) coincides with the sharp reduction in DMM performance of pilots after the age of 53 years from the present work (results of this study are

overlaid on the distribution of pilot age from the 2012 COPA Survey). The drift towards lower DMM performance for younger pilots is slow. In contrast, the DMM rate of decline increased six-fold from age 53 to 81 years in the present sample. Figure 9.4 suggests that the DMM might function as a marker for age, but would be more predictive in models of pilot performance that limit the age range to pilots aged 50 and older.

In summary, there are a number of plausible explanations for the effects of age on pilot performance. In light of the present work, a leading hypothesis might be that poor construction of the DMM, perhaps mediated by age-induced changes to brain morphology, structure and function, is the root cause of performance decrements seen in older pilots.

Figure 9.5. Plot of DMM scores from the present work superimposed on histogram of COPA pilot ages. Left y-axis= percent of COPA membership survey respondents per age group. Right y-axis=DMM factor scores (ranked and standardized). Black curve represents estimated average DMM factor scores per age group in the present work. Rate of decline of DMM factor scores by pilot age was fairly consistent from age 19 to 53 years. In this sample, beyond 53 years of age the rate of decline in DMM performance increases 5-fold until age 81. Adapted from Christopher, K. (2012). Report on the COPA 2012 Membership Survey. Age Distribution. p. 43. DMM factor score right-hand y-axis, vertical bar at age 52, and DMM score curve added.

9.4 Recommendations for Identifying Older At-Risk Pilots

Findings from the present work suggest that a screening tool designed to identify older at-risk pilots should incorporate measures that index domain-dependent variables such as SA, TRP, TM, and PM. Results from the COMP-A modeling are summarized in the form of recommendations regarding the assessment of the aging aviator.

The following recommendations are based on the results of testing the COMP-A, which included 216 observations derived from 108 pilots. In summary, a cognitive health screening tool should incorporate information regarding pilot attributes, domain-dependent cognition, and complex visual-spatial and working memory

measures (like the DCAT). Self-rating scales that index task management and awareness of the impact of other aircraft on the ability to fly the perfect circuit might also be useful for diagnosing pilots as high-risk for critical incidents or poor outcomes associated with off-nominal events. Table 9.1 lists the suggested elements of a cognitive health screening tool for aviation.

Table 9.1
Suggested Elements of a Cognitive Health Screening Tool for Aviation

Assessment Variables	Details
Expertise	Parameters include license and rating, total hours and recent pilot-in-command hours, and total simulator hours.
Age	Weighting schemes regarding age may be tried in order to maximize their predictive utility.
Complex Working Memory and Span of Attentional Field	Tasks that index a span of attentional field and a domain-relevant situational judgement task. This can be integrated seamlessly into the testing environment.
The DMM	Circuit tasks to elicit SA, TRP, TM, and PM.
Self-Rating Items	Pilot self-assess mental task load, task management, impact of other aircraft on the ability to fly the perfect circuit.
Platform	Preferably realistic controls and environmental stimuli (e.g., sound).

An overview of a program of research and development for the purposes of developing a cognitive health screening tool for the aging aviator is provided below:

1. Develop a cognitive health screening tool based on the content as described in Table 9.1. Developing an immersive testing platform may require a specialized team of simulation engineers and software designers and experts in cognition and aviation. The testing platform should permit the collection of data that reflect SA, TRP, TM,

and PM measures. Weights or cut-scores from a variety of factors from the COMP-A should maximize accurate classification of pilots as low- or high-risk. Findings from the present dataset can be used as starting values.

2. Pilot test the screening tool with approximately 40 pilots against full simulator performance to confirm the high correlation between the new DMM scores and the present sample DMM scores.

3. Once validated (with the pilot test) conduct a full validation of the screening tool with a large sample (200 pilots of varying expertise). Full validation activities include:

- a. Comparison of screening tool results with results of actual check-ride scores
- b. Development of cut-scores for diagnosing low- and high-risk via ROC analyses

4. Cross-validate cut-score algorithms with refined screening tool with second large sample of pilots (ideally another 200 pilots).

5. Final cognitive health screening tool will identify pilot risk with respect to critical incident involvement and the outcome of an off-nominal event.

6. Risk should ideally be categorized as low, moderate, or high. Therefore, in the absence of any other evidence to the contrary,

- a. low-risk pilots should be considered as having adequate cognitive skills for continued flying.
- b. high-risk pilots should be considered as not having the adequate cognitive skills for continued flying.
- c. moderate-risk pilots will require additional testing of a practical and/or neurocognitive nature before a decision regarding cognitive competency for continued flying can be made.

9.5 Final Note

Modeling a broad array of predictors in the COMP-A using a CSE framework demonstrated the value of domain-dependent cognitive factors in predicting aviation outcomes. Domain-independent factors, such as processing speed and working memory, predicted minimal amounts of variance in critical incident counts. For the majority of these general purpose cognitive measures, their usefulness in predicting aviation outcomes was overshadowed by the predictive value demonstrated by the DMM, a domain-dependent factor. These findings suggest that addressing splinter skills (through specialized online games and activities, for example) may not result in better outcomes for pilots. Although short-term improvements in working memory or speed of processing may be attained, further research is required to determine whether augmentation of isolated mechanisms of working memory or processing speed will transfer to reduced pilot risk. The current findings also suggest that mitigating risk for older pilots will require other strategies, such as engaging in pilot continuing education, maintaining regular flying opportunities, and limiting the complexity of the flight environment (e.g. familiar terrain).

In sum, the DMM represents a complex cognitive mechanism that supports basic aviator competencies. Two other domain-dependent cognitive variables were also significant predictors of aviation outcomes: namely iRES and episodic memory. Data modeling in the COMP-A and the findings of the experimental phase of the present work demonstrated a clear differentiation between the diagnostic utility of domain-independent and domain-dependent cognitive factors regarding pilot risk. These results inform current methods for identifying at-risk pilots and suggest that future work investigate alternatives to conventional batteries of domain-independent cognitive function.

References

- Adamson, M. M., Bayley, P. J., Scanlon, B. K., Farrell, M. E., Hernandez, B., Weiner, M. W., Yesavage, J., & Taylor, J. L. (2012). Pilot expertise and hippocampal size: associations with longitudinal flight simulator performance. *Aviation, Space, and Environmental Medicine*, 83(9), 850-857.
- Adamson, M. M., Landy, K. M., Duong, S., Fox-Bosetti, S., Ashford, J. W., Murphy, G. M., Weiner, M. W., & Taylor, J. L. (2010a). Apolipoprotein E ε4 influences on episodic recall and brain structures in aging pilots. *Neurobiology of Aging*, 31(6), 1059-1063.
- Adamson, M. M., Samarina, V., Xu Xiangyan, Huynh, V., Kennedy, Q., Weiner, M., Yesavage, J., & Taylor, J. L. (2010b). The impact of brain size on pilot performance varies with aviation training and years of education. *Journal of the International Neuropsychological Society*, 16(3), 412-423.
doi:10.1017/S1355617710000111
- Air Safety Institute. (2012). *22nd Joseph T. Nall Report: General aviation accidents in 2010*. Bruce Landsberg. Retrieved from <http://www.aopa.org//media/Files/AOPA/Home/News/All%20News/2012/December/Ambushed%20by%20Ice/11nall.pdf>
- Andreev, P., Heart, T., Maoz, H., & Pliskin, N. (2009). Validating formative partial least squares (PLS) models: methodological review and empirical illustration. Retrieved from <http://aisel.aisnet.org/icis2009/193/>
- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2012). Causality and endogeneity: Problems and solutions. *The Oxford Handbook of Leadership and Organizations*. Oxford University Press, Oxford.
- Ba, Y., & Zhang, W. (2011). A Review of Driver Mental Workload in Driver-Vehicle-Environment System. In *Internationalization, Design and Global Development* (pp. pp. 125-134). Orlando, FL.
- Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology*, 8(4), 485.
- Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist*, 54(7), 462.

- Bazargan, M., & Guzhva, V. S. (2011). Impact of gender, age and experience of pilots on general aviation accidents. *Accident Analysis and Prevention*, 43(3), 962-970. doi:10.1016/j.aap.2010.11.023
- Becker, J.-M., Klein, K., & Wetzels, M. (2012). Hierarchical Latent Variable Models in PLS-SEM: Guidelines for Using Reflective-Formative Type Models. *Long Range Planning*, 45(5-6), 359-394. doi:10.1016/j.lrp.2012.10.001
- Bellenkes, A. H., Wickens, C. D., & Kramer, A. F. (1997). Visual scanning and pilot expertise: the role of attentional flexibility and mental model development. *Aviation, Space, and Environmental Medicine*. Retrieved from <http://doi.apa.org/?uid=1997-06674-001>
- Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *The Journal of General Psychology*, 39(1), 15-22.
- Besnard, D., Greathead, D., & Baxter, G. (2004). When mental models go wrong: co-occurrences in dynamic, critical systems. *International Journal of Human-Computer Studies*, 60(1), 117-128.
- Bielak, A. A., Hultsch, D. F., Strauss, E., MacDonald, S. W., & Hunter, M. A. (2010). Intraindividual variability is related to cognitive change in older adults: evidence for within-person coupling. *Psychology and Aging*, 25(3), 575.
- Boehm-Davis, D. A., Casali, J. G., Kleiner, B. M., Lancaster, J. A., Saleem, J. J., & Wochinger, K. (2007). Pilot performance, strategy, and workload while executing approaches at steep angles and with lower landing minima. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(5), 759-772.
- Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychological Methods*, 16(3), 265.
- Bolstad, C. A. (2001). Situation Awareness: Does It Change With Age? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 45, pp. 272-276). Retrieved from <http://drivingassessment.uiowa.edu/drivingmetrics/Bolstad%20HFES%202001%20.pdf>
- Bolstad, C. A., & Hess, T. M. (2000). Situation awareness and aging. *Situation Awareness Analysis and Measurement*, 277-302.

- Bolstad, C. A., Endsley, M. R., Costello, A. M., & Howell, C. D. (2010). Evaluation of Computer-Based Situation Awareness Training for General Aviation Pilots. *The International Journal of Aviation Psychology*, 20(3), 269-294.
doi:10.1080/10508414.2010.487013
- Bouillon, L., Mazer, B., & Gelinas, I. (2006). Validity of the Cognitive Behavioral Driver's Inventory in Predicting Driving Outcome. *American Journal of Occupational Therapy*, 60(4), 420-427.
- Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends in Cognitive Sciences*, 14(6), 277-290.
- Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology*, 47(3), 191.
- Brunyé, T. T., & Taylor, H. A. (2008). Working memory in developing and applying mental models from spatial descriptions. *Journal of Memory and Language*, 58(3), 701-729. doi:10.1016/j.jml.2007.08.003
- Cabeza, R. (2002). Hemispheric asymmetry reduction in older adults: the HAROLD model. *Psychology and Aging*, 17(1), 85.
- Cain, B. (2007). *A review of the mental workload literature*. Defence Research and Development (Canada). Retrieved from <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA474193>
- Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data*. Cambridge university press.
- Cass, E. M. (2010). *Can situation awareness be predicted? Investigating relationships between CogScreen-AE and pilot situation awareness*. Carleton University. Retrieved from <https://curve.carleton.ca/system/files/theses/30326.pdf>
- Causse, M., Dehais, F., & Pastor, J. (2011). Executive functions and pilot characteristics predict flight simulator performance in general aviation pilots. *The International Journal of Aviation Psychology*, 21(3), 217-234.
- Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of formative measurement in information systems research. *Management Information Systems Quarterly*, 33(4), 7.

- Christopher, K. (2012). *Report on the COPA 2012 membership survey*. Ottawa, ON.
Retrieved from <http://www.copanational.org/files/2012Survey4Website.pdf>
- Coffey, E., Herdman, C. M., Brown, M., & Wade, J. (2007). Age-related changes in detecting unexpected air traffic and instrument malfunctions. In *Proceedings of the 14th International Symposium on Aviation Psychology*. Dayton, Ohio.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Edition* (Third edition). Mahwah, N.J: Routledge.
- Converse, S. (1993). Shared mental models in expert team decision making. *Individual and Group Decision Making: Current Issues*, 221.
- Cooper, A., Reimann, R., & Cronin, D. (2007). *About face 3: the essentials of interaction design*. John Wiley & Sons.
- Craik, K. J. W. (1943). *The Nature of explanation*. Cambridge, Eng: U.P.
- Crizzle, A. M., Classen, S., Bedard, M., Lanford, D., & Winter, S. (2012). MMSE as a predictor of on-road driving performance in community dwelling older drivers. *Accident Analysis & Prevention*, 49, 287-292.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.
- Davis, S. W., Dennis, N. A., Daselaar, S. M., Fleck, M. S., & Cabeza, R. (2008). Que PASA? The posterior-anterior shift in aging. *Cerebral Cortex*, 18(5), 1201-1209.
- Dehais, F., Causse, M., Vachon, F., Régis, N., Menant, E., & Tremblay, S. (2014). Failure to Detect Critical Auditory Alerts in the Cockpit Evidence for Inattentional Deafness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 56(4), 631-644.
- Delprado, J., Kinsella, G., Ong, B., & Pike, K. (2013). Naturalistic measures of prospective memory in amnesic mild cognitive impairment. *Psychology and Aging*, 28(2), 322.
- Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research*, 61(12), 1203-1218.

- Diamantopoulos, A., & Temme, D. (2013). MIMIC models, formative indicators and the joys of research. *AMS Review*, 3(3), 160–170.
- Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: an alternative to scale development. *Journal of Marketing Research*, 38(2), 269–277.
- Dismukes, R. K. (2008). Prospective memory in aviation and everyday settings. *Prospective Memory: Cognitive, Neuroscience, Developmental, and Applied Perspectives*, 411-428.
- Dismukes, R. K., & Berman, B. (2010). Checklists and monitoring in the cockpit: Why crucial defenses sometimes fail. *National Aeronautics and Space Administration Technical Memorandum NASA/TM-2010-216396*.
- Dismukes, R., & Nowinski, J. (2007). Prospective memory, concurrent task management, and pilot error. In A. Kramer, D. Wiegmann, & A. Kirlik (Eds.) *Attention: From Theory to Practice*. (pp. 225-236). New York: Oxford.
- Dobbs, A. R., Heller, R. B., & Schopflocher, D. (1998). A comparative approach to identify unsafe older drivers. *Accident Analysis & Prevention*, 30(3), 363-370.
- Dobbs, B. M., & Shergill, S. S. (2013). How effective is the Trail Making Test (Parts A and B) in identifying cognitively impaired drivers? *Age and Ageing*, 42(5), 577-581. doi:10.1093/ageing/aft073
- Dodhia, R. M., & Dismukes, R. K. (2009). Interruptions create prospective memory tasks. *Applied Cognitive Psychology*, 23(1), 73-89.
- Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2. doi:10.1186/2193-1801-2-222
- Doyle, J. K., & Ford, D. N. (1998). Mental models concepts for system dynamics research. *System Dynamics Review*, 14(1), 3-29.
- Doyle, J. K., Ford, D., Radzicki, M., & Trees, W. S. (2001). *Mental Models of Dynamic Systems*. (UNESCO Encyclopedia of Life Support Systems.). UK: EOLSS Publications.
- DriveABLE Testing Ltd. (1997). *Evaluations for at-risk experienced drivers*. Edmonton, Alberta.: DriveABLE Testing Ltd.
- Eckert, M. A. (2011). Slowing Down: Age-Related Neurobiological Predictors of Processing Speed. *Frontiers in Neuroscience*, 5. doi:10.3389/fnins.2011.00025

- Einstein, G. O., & McDaniel, M. A. (2005). Prospective Memory: Multiple Retrieval Processes. *Current Directions in Psychological Science*, 14(6), 286-290.
- Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 32, pp. 97-101). SAGE Publications. Retrieved from <http://pro.sagepub.com/content/32/2/97.short>
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 32-64.
- Endsley, M. R. (1999). Situation awareness and human error: Designing to support human performance. In *Proceedings of the High Consequence Systems Surety Conference*. Lawrence Erlbaum Associates. Retrieved from <http://209.238.175.8/Papers/pdf/Sandia99-safety.pdf>
- Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. *Situation Awareness Analysis and Measurement*, 10.
- Endsley, M. R., & Bolstad, C. A. (1994). Individual Differences in Pilot Situation Awareness. *The International Journal of Aviation Psychology*, 4(3), 241-264. doi:10.1207/s15327108ijap0403_3
- Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, 47(1), 273-305.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191.
- Federal Aviation Administration. (2010). *Annual Runway Safety Report*. Retrieved from http://www.faa.gov/airports/runway_safety/news/publications/media/Annual_Runway_Safety_Report_2010.pdf
- Federal Aviation Administration. (2015). Guide for Aviation Medical Examiners. U.S. Department of Transportation. Retrieved from https://www.faa.gov/about/office_org/headquarters_offices/avs/offices/aam/ame/guide/

- Fitzgerald, D., Cheng, I., Clem, P., Drane, M., & Hegde, A. (2012). Ageing Pilots: Aviation Risk Management. Retrieved from http://www.casa.gov.au/wcmswr/_assets/main/avmed/download/icasml_ageing.pdf
- Flight Safety Foundation. (2005). Memory lapses usually do not signal serious medical problems. *Human Factors & Aviation Medicine*, 52(4), June 20, 2013.
- Fornell, C., & Bookstein, F. L. (1982). Two Structural Equation Models: LISREL and PLS Applied to Consumer Exit-Voice Theory. *Journal of Marketing Research*, 19(4), 440-452. doi:10.2307/3151718
- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 39-50.
- General Accounting Office. (2012). *General aviation: Additional FAA efforts could help identify and mitigate safety risks*. Washington: United States Government. Retrieved from <http://www.gao.gov/assets/650/649219.pdf>
- Gentner, D., & Stevens, A. L. (2014). *Mental Models*. Psychology Press.
- Gilbert, S. J., Hadjipavlou, N., & Raoelison, M. (2013). Automaticity and Control in Prospective Memory: A Computational Model. *PLoS ONE*, 8(3), e59852. doi:10.1371/journal.pone.0059852
- Goh, J., & Wiegmann, D. (2002). Human factors analysis of accidents involving visual flight rules flight into adverse weather. *Aviation, Space, and Environmental Medicine*, 73(8), 817-822.
- Goodhue, D. L., Lewis, W., & Thompson, R. (2012). Does PLS have advantages for small sample size or non-normal data? *Mis Quarterly*, 36(3), 891-1001.
- Götz, O., Liehr-Gobbers, K., & Krafft, M. (2010). Evaluation of structural equation models using the partial least squares (PLS) approach. In *Handbook of partial least squares* (pp. 691-711). Springer.
- Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, 26(3), 499-510.
- Gugerty, L. J., & Tirre, W. C. (2000). Individual differences in situation awareness. *Situation Awareness Analysis and Measurement*, 249-276.
- Guttman, L. (1955). The Determinacy of Factor Score Matrices with Implications for Five Other Basic Problems of Common-Factor Theory¹. *British Journal of Statistical Psychology*, 8(2), 65–81. doi:10.1111/j.2044-8317.1955.tb00321.x

- Gutzwiller, R. S., & Clegg, B. A. (2013). The Role of Working Memory in Levels of Situation Awareness. *Journal of Cognitive Engineering and Decision Making*, 7(2), 141-154. doi:10.1177/1555343412451749
- Guynn, M. J. (2003). A two-process model of strategic monitoring in event-based prospective memory: Activation/retrieval mode and checking. *International Journal of Psychology*, 38(4), 245.
- Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2013). *A primer on partial least squares structural equation modeling (PLS-SEM)*. SAGE Publications, Incorporated.
- Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, 19(2), 139–152.
- Hair, J. F., Sarstedt, M., Ringle, C. M., & Mena, J. A. (2012). An assessment of the use of partial least squares structural equation modeling in marketing research. *Journal of the Academy of Marketing Science*, 40(3), 414–433.
- Hardy, D. J., & Parasuraman, R. (1997). Cognition and flight performance in older pilots. *Journal of Experimental Psychology: Applied*, 3(4), 313.
- Hardy, D. J., Satz, P., D'Elia, L. F., & Uchiyama, C. L. (2007). Age-related group and individual differences in aircraft pilot cognition. *The International Journal of Aviation Psychology*, 17(1), 77-90.
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload*. Amsterdam: North Holland Press.
- Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning and Motivation*, 22, 193-225.
- Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, 12(1), 159.
- Herdman, C., & LeFevre, J. (2004). *Tactical Aviation Mission System Simulation Situational Awareness Project*. (No. DRDC-CR-2007-002). Defence Research and Development Canada. Carleton University, Ottawa (Ontario). DTIC Document.
- Hill, T. D., & Jackson, D. N. (1984). Brief Form of the Multidimensional Aptitude Battery (MAB). Retrieved from <http://eric.ed.gov/?id=ED250383>

- Histon, J. M., & Hansman, R. J. (2008). Mitigating Complexity in Air Traffic Control: The Role of Structure-Based Abstractions. Retrieved from <http://18.7.29.232/handle/1721.1/42006>
- Hoffmann, C. C., Hoffmann, K. P., & Kay, G. G. (1998). The role that cognitive ability plays in CRM. In *NATO Symposium, Human Factors and Medicine, Panel on Collaborative Crew Performance in Complex Operational Systems, Edinburgh, Scotland*. Retrieved from [http://ftp.rta.nato.int/publiC/PubFullText/RTO/MP/RTO-MP-004/\\$MP-004-37.pdf](http://ftp.rta.nato.int/publiC/PubFullText/RTO/MP/RTO-MP-004/$MP-004-37.pdf)
- Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. MIT Press.
- Hollnagel, E., & Woods, D. D. (1983). Cognitive systems engineering: New wine in new bottles. *International Journal of Man-Machine Studies*, 18(6), 583-600.
- Hollnagel, E., & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC Press.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Articles*, 2.
- Hou, M., Kobierski, R. D., & Brown, M. (2007). Intelligent adaptive interfaces for the control of multiple UAVs. *Journal of Cognitive Engineering and Decision Making*, 1(3), 327-362.
- Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology*, 13(4), 373-386.
- Hyland, D., Kay, E., & Deimler, J. (1994). *Age 60 rule research, Part IV: Experimental evaluation of pilot performance*. (No. FAA report no. FAA-AM-94-23). Washington, DC: Department of Transportation, Federal Aviation Administration; 1994. Available from: National Technical Information Service, Springfield, VA 22161.
- IBM (2014). IBM SPSS statistics version 21. *Boston, Mass: International Business Machines Corp.*

- Jahn, G., Oehme, A., Krems, J. F., & Gelau, C. (2005). Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(3), 255-275.
- Jannoo, Z., Yap, B., Auchoybur, N., & Lazim, M. (2014). The Effect of Nonnormality on CB-SEM and PLS-SEM Path Estimates. *International Journal of Mathematical, Computational Science and Engineering*, Vol:8. Retrieved from <http://www.waset.org/publications/9997382>
- Jarmasz, J., Herdman, C. M., & Johannsdottir, K. R. (2005). Object-Based Attention and Cognitive Tunneling. *Journal of Experimental Psychology: Applied*, 11(1), 3.
- Johnson, W., Logie, R. H., & Brockmole, J. R. (2010). Working memory tasks differ in factor structure across age cohorts: Implications for dedifferentiation. *Intelligence*, 38(5), 513-528. doi:10.1016/j.intell.2010.06.005
- Johnson-Laird, P. (2004). The history of mental models. *Psychology of Reasoning: Theoretical and Historical Perspectives*, 179.
- Jonassen, D., & Tessmer, M. (1996). An outcomes-based taxonomy for the design, evaluation, and research on instructional systems. *Training Research Journal*, 2(1996), 97.
- Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. *Aviation, Space, and Environmental Medicine*.
- Jones, N., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental models: an interdisciplinary synthesis of theory and methods. Retrieved from <http://ro.uow.edu.au/smartpapers/81/>
- Jöreskog, K. G. (1967). A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Bulletin Series*, 1967(2), 183-202.
- Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica*, 86(2), 199 - 225. doi:10.1016/0001-6918(94)90003-5
- Karikawa, D., Takahashi, M., Ishibashi, A., Wakabayashi, T., & Kitamura, M. (2006). Human-Machine System Simulation for Supporting the Design and Evaluation of Reliable Aircraft Cockpit Interface. In *SICE-ICASE, 2006. International Joint Conference* (pp. 55-60). doi:10.1109/SICE.2006.315259
- Kay, G. (1995). *CogScreen Aeromedical Edition Professional Manual* (Professional Manual). Florida: Psychological Assessment Resources, Inc.

- Kennedy, K. M., Rodrigue, K. M., Bischof, G. N., Hebrank, A. C., Reuter-Lorenz, P. A., & Park, D. C. (2015). Age trajectories of functional activation under conditions of low and high processing demands: An adult lifespan fMRI study of the aging brain. *NeuroImage*, *104*, 21 - 34. doi:10.1016/j.neuroimage.2014.09.056
- Kennedy, Q., Taylor, J., Heraldez, D., Noda, A., Lazzeroni, L., & Yesavage, J. (2013). Intraindividual variability in basic reaction time predicts middle-aged and older pilots' flight simulator performance. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *68*(4), 487-494.
doi:10.1093/geronb/gbs090
- Kennedy, Q., Taylor, J. L., Reade, G., & Yesavage, J. A. (2010). Age and Expertise Effects in Aviation Decision Making and Flight Control in a Flight Simulator. *Aviation Space and Environmental Medicine*, *81*(5), 489-497.
doi:10.3357/ASEM.2684.2010
- Kieras, D. E., & Bovair, S. (1984). The Role of a Mental Model in Learning to Operate a Device*. *Cognitive Science*, *8*(3), 255-273.
- Kihlstrom, J. F. (2010). The psychological unconscious. In *Handbook of personality. Theory and research*. Oliver P. John, Richard W. Robins, Lawrence A. Pervin (Eds.). Guilford Press.
- King, R. E., Retzlaff, P., Barto, E., Ree, M. J., Teachout, M. S., & Carretta, T. R. (2012). *Pilot Cognitive Functioning and Training Outcomes* (No. AFRL-SA-WP-TR-2012-0004). School of Aerospace Medicine Wright Patterson, AFB, OH.
- Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*. New York, NY, USA: Guilford Press. Retrieved from
<http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10267361>
- Kock, N. (2013). WarpPLS 4.0 user manual. *Laredo, TX: ScriptWarp Systems*.
- Kock, N. (2014). Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of E-Collaboration (IJeC)*, *10*(1), 1–13.
- Kock, N., & Lynn, G. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, *13*(7).

- Korner-Bitensky, N., & Sofer, S. (2009). The DriveABLE Competence Screen as a predictor of on-road driving in a clinical sample. *Australian Occupational Therapy Journal*, 56(3), 200-205.
- Lassiter, D. L., Morrow, D. G., Hinson, G. E., Miller, M., & Hambrick, D. Z. (1996). Expertise and Age Effects on Pilot Mental Workload in a Simulated Aviation Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 40(3), 133-137. doi:10.1177/154193129604000306
- Lee, N., Cadogan, J. W., & Chamberlain, L. (2014). Material and efficient cause interpretations of the formative model: resolving misunderstandings and clarifying conceptual language. *AMS Review*, 4(1-2), 32-43.
- Li, G., Baker, S. P., Qiang, Y., Grabowski, J. G., & McCarthy, M. L. (2005). Driving-while-intoxicated history as a risk marker for general aviation pilots. *Accident Analysis and Prevention*, 37(1), 179-184. doi:10.1016/j.aap.2004.04.005
- Logie, R., Maylor, E., Della Sala, S., & Smith, G. (2004). Working memory in event- and time-based prospective memory tasks: Effects of secondary demand and age. *European Journal of Cognitive Psychology*, 16(3), 441-456.
- Lutz, S., & Huitt, W. (2004). Connecting cognitive development and constructivism: Implications from theory for instruction and assessment. *Constructivism in the Human Sciences*, 9(1), 67-90.
- Maguire, E. A., Woollett, K., & Spiers, H. J. (2006). London taxi drivers and bus drivers: A structural MRI and neuropsychological analysis. *Hippocampus*, 16(12), 1091-1101. doi:10.1002/hipo.20233
- McCulloch, P., & Nelder, J. A. (1989). Generalized linear models. *Statistics in the 21st Century*, CRC Press LLC and American Statistical Association, Boca Raton, Alexandria, 387-396.
- McDaniel, M. A., & Einstein, G. O. (2000). Strategic and automatic processes in prospective memory retrieval: A multiprocess framework. *Applied Cognitive Psychology*, 14(7), S127-S144.
- McVay, J. C., & Kane, M. J. (2012). Why does working memory capacity predict variation in reading comprehension? On the influence of mind wandering and executive attention. *Journal of Experimental Psychology: General*, 141(2), 302.
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81.

- Mohler, S. R. (1998). Smoking tobacco may add to risk of dementia. *Human Factors & Aviation Medicine*, 45(5), June 20, 2013.
- Morrow, D. G., Menard, W. E., Stine-Morrow, E. A. L., Teller, T., & Bryant, D. (2001). The Influence of Expertise and Task Factors on Age Differences in Pilot Communication. *Psychology and Aging*, 16(1), 31-46. doi:10.1037/0882-7974.16.1.31
- Morrow, D. G., Ridolfo, H. E., Menard, W. E., Sanborn, A., Stine-Morrow, E. A. L., Magnor, C., Herman, L., Teller, T., & Bryant, D. (2003). Environmental Support Promotes Expertise-Based Mitigation of Age Differences on Pilot Communication Tasks. *Psychology and Aging*, 18(2), 268-284. doi:10.1037/0882-7974.18.2.268
- Morrow, D., Leirer, V., Altiteri, P., & Fitzsimmons, C. (1994). When expertise reduces age differences in performance. *Psychology and Aging*, 9(1), 134.
- Morrow, D., Yesavage, J., Leirer, V., & Tinklenberg, J. (1993). Influence of aging and practice on piloting tasks. *Experimental Aging Research*, 19(1), 53-70. doi:10.1080/03610739308253923
- Muthén, L. K., & Muthén, B. O. (1998). MPlus. *Statistical Analyses with Latent Variables. User's Guide*, 3. Retrieved from http://www.statmodel.com/virg_nov_course.shtml
- Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., ... Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4), 695-699.
- National Transportation Safety Board. (2012). *Annual aviation statistics for 2011 released: No fatalities on U.S. airlines or commuters, general aviation accidents increased*. NTSB. Retrieved from <http://www.nts.gov/news/2012/120427.html>
- Neisser, U. (1976). Cognition and reality. Principles and implication of cognitive psychology. *San Francisco: WH Freeman Nd Company*.
- Norman, D. A. (1983). Some observations on mental models. In D. Gentner and A.L. Stevens eds. *Mental Models*. Lawrence Erlbaum Associates Inc. pp7-14.
- Nowinski, J. L., & Dismukes, K. (2005). Effects of ongoing task context and target typicality on prospective memory performance: The importance of associative cueing. *Memory*, 13(6), 649–657.

- Nowinski, J. L., Holbrook, J. B., & Dismukes, R. K. (2003). Human memory and cockpit operations: An ASRS study. In *Proceedings of the 12th International Symposium on Aviation Psychology* (pp. 888-893).
- Nunes, A., & Kramer, A. F. (2009). Experience-Based Mitigation of Age-Related Performance Declines: Evidence From Air Traffic Control. *Journal of Experimental Psychology. Applied*, 15(1), 12–24. doi:10.1037/a0014947
- O'Hare, D. (2006). Cognitive functions and performance shaping factors in aviation accidents and incidents. *The International Journal of Aviation Psychology*, 16(2), 145-156.
- Oman, C. M., Kendra, A. J., Hayashi, M., Stearns, M. J., & Burki-Cohen, J. (2001). Vertical navigation displays: Pilot performance and workload during simulated constant angle of descent GPS approaches. *The International Journal of Aviation Psychology*, 11(1), 15-31.
- Oster, C. V., Strong, J. S., & Zorn, C. K. (1992). *Why airplanes crash: Aviation safety in a changing world*.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs. *Journal of Cognitive Engineering and Decision Making*, 2(2), 140-160. doi:10.1518/155534308X284417
- Parikh, R., Parikh, S., Arun, E., & Thomas, R. (2009). Likelihood ratios: Clinical application in day-to-day practice. *Indian Journal of Ophthalmology*, 57(3), 217-221. doi:10.4103/0301-4738.49397
- Patten, C. J., Kircher, A., Östlund, J., & Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident Analysis & Prevention*, 36(3), 341-350.
- Pearl, J. (2012). *The causal foundations of structural equation modeling*. DTIC Document. Retrieved from <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA557445>
- Petter, S., Straub, D., & Rai, A. (2007). Specifying formative constructs in information systems research. *Mis Quarterly*, 31(4), 623-656.

- Powell, D. H., Kaplan, E. F., Whitla, D., Catlin, R., and Funkenstein, H. H., (1993). MicroCog: Assessment of Cognitive Functioning (Version 2.1) Manual. Psychological Corporation, San Antonio, Texas.
- Rabardel, P. (1995). *Les hommes et les technologies*. Paris: Armand Colin. Retrieved from <http://ergoserv.univ-paris8.fr/site/groupe/modele/articles/public/art372105503765426783.pdf>
- Rabbitt, P., Scott, M., Lunn, M., Thacker, N., Lowe, C., Pendleton, N., ... Jackson, A. (2007). White matter lesions account for all age-related declines in speed but not in intelligence. *Neuropsychology*, 21(3), 363.
- Ramsey, F., & Schafer, D. (2012). *The statistical sleuth: a course in methods of data analysis*. Cengage Learning.
- Rapoport, M. J., Naglie, G., Weegar, K., Myers, A., Cameron, D., Crizzle, A., ... others. (2013). The relationship between cognitive performance, perceptions of driving comfort and abilities, and self-reported driving restrictions among healthy older drivers. *Accident Analysis & Prevention*, 61, 288-295.
- Rasmussen, J. (1979). *On the structure of knowledge: a morphology of mental models in a man-machine system context*. National Laboratory. Retrieved from http://www.risoe.dk/rispubl/reports_INIS/RISOM2192.pdf
- Rebok, G. W., Qiang, Y., Baker, S. P., & Li, G. (2009). Pilot age and error in air-taxi crashes. *Aviation, Space, and Environmental Medicine*, 80(7), 647.
- Reid, G. B., Shingledecker, C. A., & Eggemeier, F. T. (1981). Application of conjoint measurement to workload scale development. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 25, pp. 522-526). Sage Publications. Retrieved from <http://pro.sagepub.com/content/25/1/522.short>
- Reiten, R. M. (1958). Trail Making Test: Manual for Administration, Scoring and Interpretation. Indianapolis, IN: Indiana University Medical Center.
- Rendell, P. G., & Craik, F. I. M. (2000). Virtual week and actual week: Age-related differences in prospective memory. *Applied Cognitive Psychology*, 14(SpecIssue), S43-S62. doi:10.1002/acp.770
- Rendell, P. G., & Thomson, D. M. (1999). Aging and prospective memory: Differences between naturalistic and laboratory tasks. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, 54B(4), P256-P269.

- Ringle, C. M., Götz, O., Wetzels, M., & Wilson, B. (2009). On the use of formative measurement specifications in structural equation modeling: A Monte Carlo simulation study to compare covariance-based and partial least squares model estimation methodologies. *METEOR Research Memoranda (RM/09/014)*. Retrieved from http://mpa.ub.uni-muenchen.de/15390/1/MPRA_paper_15390.pdf
- Roldán, J. L., & Sánchez-Franco, M. J. (2012). Variance-based structural equation modeling: guidelines for using partial least squares in information systems research. *Research Methodologies, Innovations and Philosophies in Software Systems Engineering and Information Systems*, 193–221.
- Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use*. DTIC Document. Retrieved from <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA227864>
- Rose, N. S., Rendell, P. G., McDaniel, M. A., Aberle, I., & Kliegel, M. (2010). Age and individual differences in prospective memory during a “Virtual Week”: the roles of working memory, vigilance, task regularity, and cue focality. *Psychology and Aging*, 25(3), 595.
- Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., ... Eickhoff, S. B. (2012). Modelling neural correlates of working memory: a coordinate-based meta-analysis. *Neuroimage*, 60(1), 830-846.
- Rouse, W. B., & Morris, N. M. (1985). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3), 349.
- Salthouse, T. A., Mitchell, D. R., Skovronek, E., & Babcock, R. L. (1989). Effects of adult age and working memory on reasoning and spatial abilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3), 507.
- Salthouse, T. A. (1991). *Theoretical perspectives on cognitive aging*. Psychology Press.
- Salthouse, T. A. (1992). Influence of processing speed on adult age differences in working memory. *Acta Psychologica*, 79(2), 155-170.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103(3), 403.

- Salthouse, T. A. (2000). Aging and measures of processing speed. *Biological Psychology*, 54(1), 35-54.
- Salthouse, T. A. (2006). Mental exercise and mental aging evaluating the validity of the “use it or lose it” hypothesis. *Perspectives on Psychological Science*, 1(1), 68-87.
- Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, 30(4), 507–514.
- Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society*, 16(05), 754-760.
doi:10.1017/S1355617710000706
- Salthouse, T. A. (2011). Neuroanatomical substrates of age-related cognitive decline. *Psychological Bulletin*, 137(5), 753.
- Sanjram, P. K., & Khan, A. (2011). Attention, polychronicity, and expertise in prospective memory performance: Programmers’ vulnerability to habit intrusion error in multitasking. *International Journal of Human-Computer Studies*, 69(6), 428-439. doi:10.1016/j.ijhcs.2011.01.006
- Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology*, 1(1), 45-57.
- Sarter, N. B., & Woods, D. D. (1992). Mode error in supervisory control of automated systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 36, pp. 26-29). SAGE Publications. Retrieved from <http://pro.sagepub.com/content/36/1/26.short>
- Sarter, N. B., & Woods, D. D. (1994). Pilot interaction with cockpit automation II: An experimental study of pilots’ model and awareness of the flight management system. *The International Journal of Aviation Psychology*, 4(1), 1-28.
- Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots’ monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(3), 347-357.
- Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23-74.

- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, 84(1), 1.
- Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A Practical Guide to Calculating Cohen's f^2 , a Measure of Local Effect Size, from PROC MIXED. *Frontiers in Psychology*, 3. doi:10.3389/fpsyg.2012.00111
- Shao, B., Guindani, M., & Boyd, D. (2014, May 14). Causes of fatal accidents for instrument-rated commercial pilots. Aerospace Medical Association Annual Scientific Meeting, San Diego, CA. Retrieved from https://s3.amazonaws.com/gdprod/orgs/223/gatherings/1080/related_links/10739/attachments/SEf-zwmj21ib_3zM_dwLSQ/377.pdf?1395231932
- Sheehan, J., & Tessmer, M. (1997). A Construct Validation of the Mental Models Learning Outcome Using Exploratory Factor Analysis. Retrieved from <http://eric.ed.gov/?id=ED409870>
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2), 127.
- Shiffrin, R. M., & Schneider, W. (1984). Automatic and controlled processing revisited.
- Sijtsma, K. (2009). On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. *Psychometrika*, 74(1), 107-120. doi:10.1007/s11336-008-9101-0
- Smith, K., & Hancock, P. A. (1995). Situation Awareness Is Adaptive, Externally Directed Consciousness. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37(1), 137-137.
- Snow, M. P., & Reising, J. M. (2000). Comparison of two situation awareness metrics: SAGAT and SA-SWORD. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, pp. 49–52). SAGE Publications.
- Stokes, A. F., & Downs, J. (1998). CRT text VS. VHF voice: Effects of Communication Modality in Single-Pilot Cockpits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 42(4), 414–418. doi:10.1177/154193129804200403

- Sulistyawati, K., Wickens, C. D., & Chui, Y. P. (2011). Prediction in situation awareness: Confidence bias and underlying cognitive abilities. *The International Journal of Aviation Psychology*, 21(2), 153-174.
- Svensson, E., Angelborg-Thanderz, M., & Wilson, G. F. (1999). *Models of pilot performance for systems and mission evaluation-psychological and psychophysiological aspects*. DTIC Document.
- Svensson, E., & Wilson, G. (2002). Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. *The International Journal of Aviation Psychology*, 12(1), 95-110.
- Taylor, J. L., Kennedy, Q., Noda, A., & Yesavage, J. A. (2007). Pilot age and expertise predict flight simulator performance - A 3-year longitudinal study. *Neurology*, 68(9), 648-654.
- Taylor, J. L., O'Hara, R., Mumenthaler, M. S., Rosen, A. C., & Yesavage, J. A. (2005). Cognitive ability, expertise, and age differences in following air-traffic control instructions. *Psychology and Aging*, 20(1), 117-133. doi:10.1037/0882-7974.20.1.117
- Taylor, J. L., O'Hara, R., Mumenthaler, M. S., & Yesavage, J. A. (2000). Relationship of CogScreen-AE to flight simulator performance and pilot age. *Aviation Space and Environmental Medicine*, 71(4), 373-380.
- Taylor, J. L., Kennedy, Q., Adamson, M. M., Lazzeroni, L. C., Noda, A., Murphy, G. M., & Yesavage, J. A. (2011). Influences of APOE epsilon 4 and Expertise on Performance of Older Pilots. *Psychology and Aging*, 26(2), 480-487. doi:10.1037/a0021697
- Taylor, J. L., Yesavage, J. A., Morrow, D. G., Dolhert, N., Brooks, J. O., & Poon, L. W. (1994). The Effects of Information Load and Speech Rate on Younger and Older Aircraft Pilots' Ability to Execute Simulated Air-Traffic Controller Instructions. *Journal of Gerontology*, 49(5), P191-P200. doi:10.1093/geronj/49.5.P191
- Tenenhaus, M., Vinzi, V. E., Chatelin, Y.-M., & Lauro, C. (2005). PLS path modeling. *Computational Statistics & Data Analysis*, 48(1), 159-205.
- Tognoli, E., & Kelso, J. S. (2014). The metastable brain. *Neuron*, 81(1), 35-48.
- Tomasi, D., & Volkow, N. D. (2011). Aging and functional brain networks. *Molecular Psychiatry*, 17(5), 549-558.

- Transport Canada. (2012). Part IV - Personnel Licensing and Training. Retrieved from <http://www.tc.gc.ca/eng/civilaviation/regserv/cars/part4-standard-standard%20428-3870.htm#schedule3>
- Transportation Safety Board of Canada. (2006). *Safety Issues Investigation Report SII 405-01*. Retrieved from http://www.tsb.gc.ca/eng/rapports-reports/aviation/etudes-studies/siia0501/siia0501_sec2.asp
- Transportation Safety Board of Canada. (2013). Statistical Summary, Aviation Occurrences 2011. Retrieved from <http://www.tsb.gc.ca/eng/stats/aviation/2013/ssea-ssao-2013.asp>
- Travers, B. G., Bigler, E. D., Tromp, D. P., Adluru, N., Froehlich, A. L., Ennis, C., ... others. (2014). Longitudinal processing speed impairments in males with autism and the effects of white matter microstructure. *Neuropsychologia*, 53, 137-145.
- Tsang, P. S., & Shaner, T. L. (1998). Age, Attention, Expertise, and Time-Sharing Performance. *Psychology and Aging*, 13(2), 323-347. doi:10.1037/0882-7974.13.2.323
- Uttl, B. (2008). Transparent meta-analysis of prospective memory and aging. *PLoS One*, 3(2), e1568.
- Uttl, B. (2011). Transparent meta-analysis: does aging spare prospective memory with focal vs. non-focal cues? *PloS One*, 6(2), e16618.
- Van Benthem, K., & Herdman, C.M. (2013). Individual pilot factors predict simulated runway incursion outcomes. Proceedings of the *17th International Symposium on Aviation Psychology*, May 5–9, 2013 Dayton, Ohio.
- Van Benthem, K., & Herdman, C. M. (2014a). Congruence of self-rating with pilot performance: Validation of a task load index for human-machine systems. Presentation for the *85th Aerospace Medical Association Annual Scientific Meeting*, May 2014, San Diego, CA.
- Van Benthem, K., & Herdman, C. M. (2014b). Individual pilot factors predict simulated runway incursion outcomes. In Michael Vidulich, Pamela Tsang, & John Flach (Eds.), *Advances in Aviation Psychology, Volume 1*, p. 197–214. Ashgate Books.

- Van Benthem, K., Herdman, C. M., Brown, M., & Barr, A. (2011a). The Relationship of age, experience and cognitive health to private pilot situation awareness performance. *Proceedings of the 16th International Symposium on Aviation Psychology*, May 2–5, 2011 Dayton, Ohio.
- Van Benthem, K., Herdman, C. M., & Garrett, S. (2011b). Flight instructors and pilot examiners perceptions of older and younger pilot competency and safety concerns. *Proceedings of the 16th International Symposium on Aviation Psychology*, May 2–5, 2011, Dayton, Ohio.
- Van Benthem, K., Herdman, C. M., & LeFevre, J. (2011c). Prospective memory tasks in aviation: effects of age and working memory. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds). *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2335–2340). Austin, Texas: Cognitive Science Society.
<http://mindmodeling.org/cogsci2011/papers/0556/index.html>
- Van Benthem, K. D., Herdman, C. M., Tolton, R. G., & LeFevre, J. A. (2015). Prospective Memory Failures in Aviation: Effects of Cue Salience, Workload, and Individual Differences. *Aerospace Medicine and Human Performance*, 86(4), 366-373.
- Van Benthem, K., Tolton, R.G., & Herdman, C. M. (2013). Coherence between pilot self-rating and objective measures of situation awareness performance. Presentation for *the 84th Aerospace Medical Association Annual Scientific Meeting, May 2013*, Chicago, Illinois.
- Van Benthem, K., Herdman, C. M., LeFevre, J., & Tolton, R.G. (2012a). Prospective memory for general aviation pilots in low and high workloads. Presentation for *the 83rd Aerospace Medical Association Annual Scientific Meeting, May 15, 2012*, Atlanta, Georgia.
- Van Benthem, K., Herdman, C. M., Tolton, R.G., & Barr, A. (2012b). Working memory, age, and experience influences on simulated critical incidents for general aviation pilots. Poster for the *22nd Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*, June 7, 2012, Kingston, Ontario.
- Van Rossum, G. (1995). *Python tutorial. Technical Report CS-R9526*. Amsterdam.
- Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, 25(1), 1-28.

- Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, 122(3), 231.
- Vidulich, M. A., & Hughes, E. R. (1991). Testing a subjective metric of situation awareness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (35), 1307-1311. Sage Publications.
- Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness. *Handbook of Human Factors and Ergonomics*, 243-273.
- Von Neumann, J. (1958). *The Computer and the Brain*. New Haven/London: Yale University Press. Retrieved from <http://www.leydesdorff.net/vonneumann/>
- West, R., & Bowry, R. (2005). Effects of aging and working memory demands on prospective memory. *Psychophysiology*, 42(6), 698-712.
- Wickens, C. D. (2008). Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3), 397-403.
doi:10.1518/001872008X288420
- Wickens, C. D., & Long, J. (1995). Object versus space-based models of visual attention: Implications for the design of head-up displays. *Journal of Experimental Psychology: Applied*, 1(3), 179.
- Wiegmann, D. A., Goh, J., & O'Hare, D. (2002). The role of situation assessment and flight experience in pilots' decisions to continue visual flight rules flight into adverse weather. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 44(2), 189-197.
- Wiggins, M. W. (2011). Vigilance decrement during a simulated general aviation flight. *Applied Cognitive Psychology*, 25(2), 229-235.
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594.
- Wilson, G. F. (2002). An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology*, 12(1), 3 - 18. doi:10.1207/S15327108IJAP1201_2
- Wilson, J. R., & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 31(6), 617-634.

- Wold, S., Ruhe, A., Wold, H., & Dunn, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735-743.
- Woods, D. D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering*. CRC Press.
- Yesavage, J. A., Jo, B., Adamson, M. M., Kennedy, Q., Noda, A., Hernandez, B., Taylor, J. L. (2011). Initial Cognitive Performance Predicts Longitudinal Aviator Performance. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 66B(4), 444-453. doi:10.1093/geronb/gbr031
- Yesavage, J. A., Taylor, J. L., Mumenthaler, M. S., Noda, A., & O'Hara, R. (1999). Relationship of age and simulated flight performance. *Journal of the American Geriatrics Society*.
- Zhang, K., & Wickens, C. D. (1987). A Study of the Mental Model of a Complex Dynamic System: The Effect of Display Aiding and Contextual System Training. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 31(1), 102-106. doi:10.1177/154193128703100123

Appendix A

A: Psychology Ethics Review Board Approval

Clearance Certificate

x

Carleton University
Psychology Research Ethics Board

Certificate of Ethics Clearance

Principal Investigator

Department

Study Number

Chris Herdman

Psychology

10-034

Institution(s) where research will be conducted:

Carleton University

Co-Investigators and other researchers:

Researcher	Study Role	Position
Chris Herdman	Faculty Sponsor	Faculty

Study Title:

The Usefulness of a Cognitive Assessment Battery in Evaluating General Aviation Performance

Approval Date:

08/21/2012

Validity Term:

1 Year

Approval Type:

Final

Submitted Date	Study Component	Approval Date
09/12/2011	Addendum	09/13/2011
08/26/2010	Report	09/01/2010
08/19/2011	Report	08/30/2011
08/17/2012	Report	08/21/2012

Comments:

Certification

The protocol describing the above-named project has been reviewed by Carleton University Psychology Research Ethics Board and the research procedures were found to be acceptable on ethical grounds for research involving human participants.

Chair, Ethics Committee for Psychology Research

This Certificate of Clearance is valid for the above term provided there is no change in the research procedures.

Close

Print

Appendix B

B: Participant Informed Consent Form

Study: The Usefulness of a Cognitive Assessment Battery in Predicting General Aviation Performance

Faculty Sponsor: Dr. Chris Herdman, Department of Psychology, Carleton University, tel. 613-520-2600 x.8122

The purpose of this informed consent form is to ensure that you understand both the purpose of the study and the nature of your participation. The informed consent must provide you with enough information so that you have the opportunity to determine whether you wish to participate in the study. This study has received clearance by the Carleton University Psychology Research Ethics Board (Ethics Approval: 10-034). Please ask the researcher to clarify any concerns that you may have after reading this form.

Research Personnel: In addition to the Faculty Sponsor named above, the following people are involved in this research and may be contacted at any time should you require further information about this study.

Name	Title	Department	Email	Phone
Kathy Van Benthem	Ph.D. Student	Institute of Cognitive Science	kathy_vanbenthem@carleton.ca	613-355-2515
Anne Barr	Researcher	n/a	abarr@connect.carleton.ca	520-2600 x.2496
Rani Tolton	Researcher	n/a	ra.tolton@sympatico.ca	520-2600 x.2496

Other Contacts: Should you have any ethical concerns regarding this study, please contact Dr. Avi Parush. For any other concerns about this study, please contact Dr. Jo-Anne Lefevre.

Name	Position	Phone
Dr. Avi Parush	Chair, Psychology Research Ethics Board	520-2600 x.6026
Dr. Jo-Anne Lefevre	Director, Institute of Cognitive Science	520-2600 x.2693

Purpose: The purpose of this study is to assess the effects of mental workload on situation awareness and flying performance using a Cessna aircraft simulator.

Task: Your first task will be to complete a brief questionnaire about your experience as a general aviation pilot and a 20-minute cognitive assessment using touch screen technology. You will then be instructed in the requirements for flying a "perfect circuit." Your next task will be to fly six circuits (including touch-and-go landings) using a medium-fidelity Cessna simulator. Other aircraft will be gradually introduced into this circuit pattern as you progress through the experiment. Notification of these additional aircraft will be made via radio calls. Each circuit will require approximately six minutes to complete. While you are flying the circuits, the simulation will "freeze" twice, at which point you will be required to remember the details of your circuit and the other aircraft in the circuit.

Locale, Duration, and Compensation: Testing will take place in VSIM 1214 at Carleton University and will take approximately 2.5 hours. You will not be compensated monetarily or otherwise for participating in this experiment. This research is not affiliated with any flying club, flight school or aviation program. There is no compensation to you by the university, flight club, flight school or aviation program for your participation.

Potential Risks/Discomfort: There are no potential psychological risks associated with participation in this experiment. Please note that your performance on the task in this experiment does not provide an indication of your skills as a pilot. However, if you feel anxious and/or uncomfortable about your performance in this experiment or feel effects of the simulator, please bring your concerns to the researcher's attention.

Anonymity/Confidentiality: All data collected in this experiment will be kept strictly confidential through the assignment of a coded number and securely stored on a local computer for a maximum of ten years. Similarly, this Informed Consent form will be kept for a maximum of ten years before being destroyed. The information provided will be used for research purposes only. You will not be identified by name in any reports produced from this study. Further, the information is made available only to the researchers associated with this experiment.

Right to Withdraw/Omit: You have the right to withdraw from this experiment at any time without penalty. Your participation in this experiment is completely voluntary.

I have read the above description of the study examining the usefulness of a cognitive assessment battery in predicting general aviation flight performance. By signing below, this indicates that I agree to participate in the study, and this in no way constitutes a waiver of my rights.

Name: _____
Signature: _____

Date: _____
Witness: _____

Appendix C

C: Participant Recruitment Poster

Full-Scale Cessna Simulator Study

The ACE Lab at Carleton University is looking for participants for a study to explore cognition and general aviation.

If you are: a pilot or a pilot in training
medically licensed to fly
at least 18 years of age and
interested in flying a full-scale Cessna 172 simulator

Please contact one of us to learn more and/or to arrange a time to participate in the study.

The study session is approximately four hours and takes place at Carleton University.

Participation is completely voluntary. You will not be compensated monetarily or otherwise for participating in this study but parking costs will be paid by the study.

Thank you,
Kathleen Van Benthem
kvbenthe@connect.carleton.ca
PhD Program

Rani Tolton
rgtolton@gmail.com
Masters Program

Appendix D

D: Sample Data Capture from Flight Simulation Protocol

Time	Longitude	Altitude (ft)	Airspeed (knots)
03/12/2010 16:40	116.926	2357.775	0.0152
03/12/2010 16:40	116.926	2357.778	0.0366
03/12/2010 16:40	116.926	2357.782	0.0611
03/12/2010 16:41	116.926	2357.782	0.3858

Appendix E

E: Pilot Debriefing Form

The Usefulness of a Cognitive Assessment Battery in Predicting General Aviation Flight Performance

Thank you for your participation in this study. The purpose of this study is to determine the usefulness of a cognitive assessment battery in predicting general aviation performance. In this experiment, general aviation performance, specifically situation awareness and decision-making was assessed in three ways. First, it was measured by recall of the location and details of your own and other aircraft in your flying environment. Second, it was measured in terms of flying the simulator in the “perfect circuit.” Thirdly, information was collected with respect to responses to the unexpected runway incursion.

Being able to predict general aviation performance using a standardized and objective assessment tool (such as the computerized test used in this experiment) is critically important. Knowing the capabilities and limitations of the human mind in this context is important for the safety and well-being of the general aviation community. If you are interested in learning more about using cognitive testing to predict general aviation flight performance, then please see the following:

This study has received clearance by the Carleton University Psychology Research Ethics Board (Ethics Approval: 10-034). Should you have any ethical concerns regarding this study, please contact Dr. Shelley Brown (Chair, Psychology Research Ethics Board, 613-520-2600 ext. 1505). Should you have any other concerns about this study, please contact Dr. Jo-Anne LeFevre, (Director, Institute of Cognitive Science, 613-520-2600 ext. 2693) or any of the following individuals:

<u>Name</u>	<u>Title</u>	<u>Department</u>	<u>Study Role</u>	<u>Contact Info.</u>
Kathy Van Benthem	Ph.D. Student	Cognitive Science	Principal Researcher	613-355-2515
Anne Barr	Researcher	n/a		520-2600 x.2496
Rani Tolton	Researcher	n/a		
Dr. Chris Herdman	Professor	Psychology	Faculty Advisor	520-2600 x.8122

Appendix F

F: Pilot Demographic and Expertise Form

Gender: M / F

Is your vision 20/20? yes / no

If no,

Do you wear corrective eyewear? yes / no

If yes,

What kind of eyewear? glasses / contact lenses

Do you have a current medical certification? yes/no Category

Flexibility Limitations _____ (i.e. monocular vision)

Month/Year of Birth month _____ year __ 19 _____

Current License: _____

Current Rating _____

Number of years with a valid pilot's license:

Recreational:

Private:

Commercial:

Airline

Military

Approximate number of flight hours:

Aircraft Type	Flight Hours

How many hours have you been Pilot in Command (PIC) in the last year?

--

When was the last time you flew a circuit?

--

Do you have any experience with Yorkton Airport in Saskatchewan?

Yes / No

Do you have any experience with Kaslo Airport in British Columbia?

Yes / No

If yes, please provide details (e.g., instructed there, stopped for refueling).

Are you comfortable flying a Cessna 172?

Yes / No

What percentage of your flight time do you spend in an uncontrolled aerodrome?

_____%

Have you used flight simulation software previously?

Yes/No

If yes,

For approximately how many hours?

Briefly describe the simulator system you have used most

Appendix G

G: NASA TLX Mental Task Load Self-Rating Scale

Place an X on each scale at the point that represents your workload for each listed factor. If necessary, refer to the Workload Scale Descriptions for definitions of each factor.

Factor	Description	Item Code
Mental demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching)? Was the task easy or demanding, simple or complex, exacting or forgiving?	TLX_1
Physical demand	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?	TLX_2
Temporal demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely (i.e., low temporal demand) or rapid and frantic (i.e., high temporal demand)?	TLX_3
Frustration level	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?	TLX_4
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?	TLX_5
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?	NA
Overall workload	What was your overall (average) workload for this part of the flight when all factors are taken into consideration?	TLX_6

Mental demand	LOW	--- --- --- --- --- --- --- --- --- ---	HIGH
Physical demand	LOW	--- --- --- --- --- --- --- --- --- ---	HIGH
Temporal demand	LOW	--- --- --- --- --- --- --- --- --- ---	HIGH
Frustration level	LOW	--- --- --- --- --- --- --- --- --- ---	HIGH
Effort	LOW	--- --- --- --- --- --- --- --- --- ---	HIGH
Performance	GOOD	--- --- --- --- --- --- --- --- --- ---	POOR
Overall workload	LOW	--- --- --- --- --- --- --- --- --- ---	HIGH

Appendix H

H: SAGAT Probe Answer Form

Participant #

Aircraft Locator #1

1. Where you are now at the time of the pause **Mark "ME Now"** + Current Heading Speed Altitude
2. Location of other aircraft at pause (location in circuit) **Mark "Other"** + call sign and type of other(s)
3. Where you were when other aircraft first entered/announced himself in the circuit - **Mark "Me -0"**
4. Where you will be in the circuit in about 2 minutes (location in circuit)- **Mark "Me-2 min"**
5. Where the other aircraft will be in the circuit in 3 minutes (location in circuit)- **Mark "Other-3 min"**

Now- complete this circuit with a touch and go and go around for one more full circuit and then a full stop

Appendix I

I: Situation Awareness Self-Rating Scale

Please rate your situation awareness for the following items by circling a value from 1 to 7, where 1 represents “very poor” situation awareness and 7 represents “very good” situation awareness. Circle N/A if awareness of an item did not apply to you during this part of the experiment.

	Very Poor						Very Good		<u>Item Code</u>
Awareness of the navigation environment (within the scenario circuits)	1	2	3	4	5	6	7	N/A	SR_S1
Awareness of the tasks to be performed	1	2	3	4	5	6	7	N/A	T_1
Awareness of time to complete specific tasks	1	2	3	4	5	6	7	N/A	T_2
Awareness of time to complete the “perfect” circuit	1	2	3	4	5	6	7	N/A	SR_S4
Awareness of task priorities	1	2	3	4	5	6	7	N/A	T_3
Awareness of your aircraft’s position relative to other aircraft in the circuit	1	2	3	4	5	6	7	N/A	SR_S5
Awareness of impact of other aircraft on ability to fly the “perfect circuit”	1	2	3	4	5	6	7	N/A	SR_S6
Ability to predict future circuit events	1	2	3	4	5	6	7	N/A	SR_S7
Awareness of other aircraft activities	1	2	3	4	5	6	7	N/A	SR_S8
Awareness of other aircraft location(s)	1	2	3	4	5	6	7	N/A	SR_S9
Your overall level of situation awareness	1	2	3	4	5	6	7	N/A	SR_S10

Appendix J

J: Surprise Runway Incursion Scoring System

Score Component	Range	Interpretation
Radio Call	0-1	0= no/late radio call made 1= call made with adequate time
Time to Detect	0-3	0= no detection of runway incursion 1=detected too late 2=detected just in time 3=detected adequately
Time to Evasive Action	0-3	0= no evasive action 1=evasive action too late 2=evasive action just in time 3=evasive action adequately
Quality of Evasion Action	0-3	0=minimal separation or contact 1= poor separation/evasion 2=adequate separation/evasion 3=high quality separation/evasion

Appendix K

K: CogScreen-AE Subtest Abbreviations

BDSMPAN	Backward Digit Span Maximum Span
MATHACC	Math Accuracy
VSCACC	Visual Sequence comparison Accuracy
SDCCOR	Symbol Digit Coding Correct Responses
SDCACC	Symbol Digit Coding Accuracy
SDCIRACC	Symbol Digit Coding Immediate Recall
MTSACC	Matching to Sample Accuracy
MANACC	Manikin Accuracy
DATSCACC	Divided Attention Sequence Comparison Accuracy
ASCACC	Auditory Sequence Comparison Accuracy
SDCDRACC	Symbol Digit Coding Delayed Recall Accuracy
SATADACC	Shifting Attention Test Arrow Direction Accuracy
SATACACC	Shifting Attention Test Arrow Colour Accuracy
SATINACC	Shifting Attention Test Instruction Accuracy
SATDIACC	Shifting Attention Test Discovery Accuracy
DTTPAAACC	Dual Tracking Task Previous Number Alone Accuracy
DTTPDACC	Dual Tracking Task Previous Number Dual Accuracy
PFNACC	Pathfinder Number Accuracy
PFLACC	Pathfinder Letter Accuracy
PFCACC	Pathfinder Comparison Accuracy

MATHRTC	Math Reaction Speed
VSCRTC	Visual Sequence Comparison Reaction Speed
SDCRTC	Symbol Digit Coding Response Reaction Speed
MTSRRTC	Matching to Sample Reaction Speed
MANRTC	Manic Reaction Speed
DATIRTC	Divided Attention Visual Monitoring alone Reaction Speed
DATSCRTC	Divided Attention Sequence Comparison Reaction Speed
DATDRTC	Divided Attention Dual Reaction Speed
ASCRTC	Auditory Reaction Speed
SATADRTC	Shifting Attention Test Arrow Direction Reaction Speed
SATACRTC	Shifting Attention Test Arrow Colour Reaction Speed
SATINRTC	Shifting Attention Instruction Reaction Speed
SATDIRTC	Shifting Attention Test Discovery Reaction Speed
DTTPARTC	Dual Task Previous Number Alone Reaction Speed
DTTPDRTC	Dual Task Previous Number Dual Reaction Speed
PFNRTC	Pathfinder Number Reaction Speed
PFNLRTC	Pathfinder Letter Reaction Speed
PFCRTC	Pathfinder Combined Reaction Speed
PFNTOT	Pathfinder Number Total Time to Complete PFNACC
PFLTOT	Pathfinder Letter Total Time to Complete PFLACC
PFCOT	Pathfinder Combined Total Time to Complete PFCACC
DTTDHIT	Dual Task Boundary Hits
DTTDABS	Dual Task Tracking Alone
DTTAABS	Dual Task Tracking Dual
DTTAHIT	Dual Task Boundary Hits Single Task

Appendix K

Appendix L

L: Abbreviated Output from MPlus (v.7.11)

```
TITLE:           Multi-dimensional DMM Model !stand pre-processed ranked warp pls
DATA:           FILE IS DMM_Nov10noHdg.csv; ! n= 216 observations
VARIABLE:       NAMES = alt spd s2 s3 t1 t2 t3 PM;

Analysis: TYPE=general;
            Estimator = ml;
MODEL:
SA by s2@1 s3* ; ! create the reflective lv "situation awareness"
TRP by alt@1 spd* ; ! create the reflective lv "task relevant performance"
TM by t1* t2@1 t3* ; ! create the reflec lv " task management"
DMM by ; !create the formative lv "dynamic mental model"
DMM ON PM@1 SA@1 TM@1 TRP@1 ; ! specify the formative lv "dmm" equal weights
DMM@0; ! set DMM residual variance to 0- this creates the fully indexed DMM
[DMM@0]; ! set DMM mean at 0

PM WITH SA; ! allow PM to covary with SA - check to be sure r < 0.40 approx.
TRP with SA; ! allow TRP to covary with SA
TRP with PM; ! allow TRP to covary with PM

THE MODEL ESTIMATION TERMINATED NORMALLY
MODEL FIT INFORMATION
Number of Free Parameters 29
Loglikelihood

H0 Value -2216.911
H1 Value -2208.914

Chi-Square Test of Model Fit

Value 15.994
Degrees of Freedom 15
P-Value 0.3825

RMSEA (Root Mean Square Error Of Approximation)

Estimate 0.018
90 Percent C.I. 0.000 0.068
Probability RMSEA <=0 .05 0.814

CFI/TLI

CFI 0.998
TLI 0.996

Chi-Square Test of Model Fit for the Baseline Model

Value 478.006
Degrees of Freedom 28
P-Value 0.0000
```

SRMR (Standardized Root Mean Square Residual)

Value 0.030

MODEL RESULTS

Two-Tailed

Estimate S.E. Est./S.E. P-Value

SA BY

S2 1.000 0.000 999.000 999.000

S3 1.072 0.239 4.481 0.000

TRP BY

ALT 1.000 0.000 999.000 999.000

SPD 0.584 0.191 3.061 0.002

TM BY

T1 1.208 0.112 10.799 0.000

T2 1.000 0.000 999.000 999.000

T3 0.901 0.092 9.783 0.000

DMM ON

SA 1.000 0.000 999.000 999.000

TM 1.000 0.000 999.000 999.000

TRP 1.000 0.000 999.000 999.000

DMM ON

PM 1.000 0.000 999.000 999.000

PM WITH

SA 0.265 0.071 3.739 0.000

TM 0.170 0.059 2.894 0.004

TRP WITH

SA -0.138 0.061 -2.274 0.023

PM -0.326 0.072 -4.523 0.000

TM WITH

SA 0.061 0.051 1.206 0.228

TRP -0.098 0.056 -1.745 0.081

Means

PM 0.000 0.068 0.005 0.996

Variances

PM 0.995 0.096 10.392 0.000

SA 0.561 0.148 3.800 0.000

TRP 0.781 0.260 3.007 0.003

TM 0.573 0.096 5.996 0.000

Residual Variances

ALT 0.214 0.243 0.880 0.379

SPD 0.729 0.108 6.723 0.000

S2 0.435 0.127 3.428 0.001

S3 0.351 0.142 2.478 0.013

T1 0.160 0.061 2.611 0.009

T2 0.422 0.057 7.399 0.000

T3 0.531 0.061 8.733 0.000

STDYX Standardization

Two-Tailed

Estimate S.E. Est./S.E. P-Value

SA BY

S2 0.750 0.085 8.777 0.000

S3 0.805 0.089 9.016 0.000

TRP BY

ALT 0.886 0.138 6.411 0.000

SPD 0.517 0.094 5.477 0.000

TM BY

T1 0.916 0.034 26.749 0.000

T2 0.759 0.039 19.341 0.000

T3 0.683 0.044 15.669 0.000

DMM ON

SA 0.449 0.050 8.924 0.000

TM 0.454 0.045 9.999 0.000

TRP 0.530 0.076 6.981 0.000

DMM ON

PM 0.599 0.050 11.928 0.000

PM WITH

SA 0.355 0.070 5.035 0.000

TM 0.225 0.070 3.194 0.001

TRP WITH

SA -0.209 0.099 -2.102 0.036

PM -0.370 0.079 -4.690 0.000

TM WITH

SA 0.108 0.085 1.270 0.204

TRP -0.146 0.084 -1.749 0.080

Means

PM 0.000 0.068 0.005 0.996

R-SQUARE

Observed Two-Tailed

Variable Estimate S.E. Est./S.E. P-Value

ALT 0.785 0.245 3.205 0.001

SPD 0.268 0.098 2.739 0.006

S2 0.563 0.128 4.389 0.000

S3 0.647 0.144 4.508 0.000

T1 0.839 0.063 13.375 0.000

T2 0.576 0.060 9.670 0.000

T3 0.467 0.060 7.835 0.000

Copyright (c) 1998-2013 Muthen & Muthen

Appendix M

M: Comparison of Correlations of PLS-SEM and CB-SEM DMM Component Scores with Pilot Attributes and Simulator Performance

		DMM_FS_Mplus	DMM_FS_Warp	Age	CriticalIncidents	Incursion	RatingLevel	HrsPIC	TotalHours	SimExperience
Spearman's rho	Correlation Coefficient	1.000	.610**	-.090	-.213**	.214*	.138*	.288**	-.058	.164
	Sig. (2-tailed)		.000	.190	.002	.026	.043	.002	.553	.089
	N	216	216	216	216	108	216	108	108	108
DMM_FS_Warp	Correlation Coefficient	.610**	1.000	-.215**	-.343**	.168	.296**	.263**	.157	.369**
	Sig. (2-tailed)	.000		.001	.000	.082	.000	.006	.106	.000
	N	216	216	216	216	108	216	108	108	108
Age	Correlation Coefficient	-.090	-.215**	1.000	.238**	.117	.035	-.123	.455**	-.243*
	Sig. (2-tailed)	.190	.001		.000	.227	.610	.204	.000	.011
	N	216	216	216	216	108	216	108	108	108
CriticalIncidents	Correlation Coefficient	-.213**	-.343**	.238**	1.000	-.169	-.334**	-.254**	-.155	-.398**
	Sig. (2-tailed)	.002	.000	.000		.080	.000	.008	.109	.000
	N	216	216	216	216	108	216	108	108	108
Incursion	Correlation Coefficient	.214*	.168	.117	-.169	1.000	.388**	.221*	.228*	.201*
	Sig. (2-tailed)	.026	.082	.227	.080		.000	.021	.018	.037
	N	108	108	108	108	108	108	108	108	108
RatingLevel	Correlation Coefficient	.138*	.296**	.035	-.334**	.388**	1.000	.493**	.693**	.269**
	Sig. (2-tailed)	.043	.000	.610	.000	.000		.000	.000	.005
	N	216	216	216	216	108	216	108	108	108
HrsPIC	Correlation Coefficient	.288**	.263**	-.123	-.254**	.221*	.493**	1.000	.366**	.207*
	Sig. (2-tailed)	.002	.006	.204	.008	.021	.000		.000	.032
	N	108	108	108	108	108	108	108	108	108
TotalHours	Correlation Coefficient	-.058	.157	.455**	-.155	.228*	.693**	.366**	1.000	.042
	Sig. (2-tailed)	.553	.106	.000	.109	.018	.000	.000		.668
	N	108	108	108	108	108	108	108	108	108
SimExperience	Correlation Coefficient	.164	.369**	-.243*	-.398**	.201*	.269**	.207*	.042	1.000
	Sig. (2-tailed)	.089	.000	.011	.000	.037	.005	.032	.668	
	N	108	108	108	108	108	108	108	108	108