Game-Based Assessment of Cognitive Status in Clinical and Non-Clinical Samples

by

Tiffany Tong

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

© Copyright by Tiffany Tong 2018
Game-Based Assessment of Cognitive Status in Clinical and Non-Clinical Samples

Tiffany Tong

Doctor of Philosophy

Department of Mechanical and Industrial Engineering
University of Toronto
2018

Abstract

In this dissertation I validated a serious game for cognitive assessment that was originally developed in my Masters thesis (Tong, 2014). I used a whack-a-mole game that requires response inhibition, by adding a distractor character that should not be hit. I evaluated the game as a potential screening tool for the clinically relevant condition of delirium, and I examined how game performance was correlated with the Mini-Mental State Examination (MMSE), a standard assessment of cognitive ability in the elderly. Patients 65 years and older presenting at the Sunnybrook emergency department participated in three clinical studies and response time (RT) on the game was found to be significantly correlated with MMSE scores, and sufficiently predictive of whether the person had delirium as determined by the Confusion Assessment Method. Comparing game performance at 8-hour intervals, I found that game RT showed good test-retest reliability in the second clinical study. In the third clinical study, I found that false alarm rate provided a useful supplement to median correct RT in predicting clinical measures. Following the clinical studies, I evaluated the game with healthy adults in a pilot study to determine how the duration of character appearance affected whether the task was speeded or
not. I found that a character appearance duration of around 800 msec marked the transition from a speeded to an unspeeded task. I then carried out a final experiment with healthy adults where the two appearance durations used were 800 msec (borderline speeded) and 1200 msec (unspeeded). The experiment assessed game performance for eight combinations of character appearance duration and two other dichotomous factors, and assessed performance on four tasks that measure different aspects of inhibition ability (Stroop task, Flanker task, anti-saccade task, Go/No-Go discrimination task). False alarms increased significantly in the condition where targets were least frequent and distractors were visually similar to the target. Game performance was strongly correlated with performance on the Go/No-Go discrimination task but not with the other three inhibition tasks used. The results of this thesis point the way towards the development of serious games that can supplement current clinical assessments.
Acknowledgments

Dr. Mark Chignell, words cannot begin to describe my gratitude to you for taking a chance on me. Without your wisdom, guidance, and mentorship, this thesis would not have been possible.

I am grateful to the members of my thesis committee, Dr. Birsen Donmez and Dr. Kelly Lyons, and my external examiner, Dr. Bruce Morton, who provided guidance and feedback on my thesis. I would also like to thank my clinical advisor, Dr. Jacques Lee, for providing mentorship and the opportunity to carry out research in a clinical setting. I also extend my appreciation to Dr. Jacques Lee’s research team at Sunnybrook Research Institute.

To the members of the Interactive Media Lab: thank you for the great memories and friendship.

To my family and friends: thank you for your endless patience, encouragement, and support of my academic pursuit.

The serious game evaluated in this dissertation was jointly created between the Interactive Media Lab at the University of Toronto, under the direction of Dr. Mark Chignell, and our collaborator, Dr. Jacques Lee and his research team at Sunnybrook Research Institute. The game was originally developed and evaluated for its usability in my Masters thesis (Tong, 2014). This game underwent continual iterative design that was driven by invaluable feedback on the feasibility and usability of the game in a clinical context from Dr. Lee's research team.
# Table of Contents

Acknowledgments........................................................................................................ iv

Table of Contents.......................................................................................................... v

List of Tables .................................................................................................................. xii

List of Figures ................................................................................................................ xviii

List of Appendices ............................................................................................................ xxi

List of Abbreviations ....................................................................................................... xxii

List of Units ..................................................................................................................... xxiv

Chapter 1......................................................................................................................... 1

1 Introduction .................................................................................................................. 1

1.1 Problem Statement ................................................................................................. 1

1.2 Research Questions and Overview of Studies ....................................................... 1

1.3 Road Map ................................................................................................................ 3

Chapter 2......................................................................................................................... 5

2 Background ................................................................................................................ 5

2.1 Cognitive Assessments ......................................................................................... 5

2.1.1 MMSE ............................................................................................................... 5

2.1.2 MoCA .............................................................................................................. 7

2.1.3 CAM ............................................................................................................... 7

2.1.4 Discussion (Advantages and Disadvantages) ................................................ 8

2.2 Cognitive Psychology Tasks and Executive Functioning ..................................... 9

2.2.1 Inhibition Ability and Tasks ........................................................................... 9

2.2.2 Inhibition across the Lifespan ....................................................................... 15

2.2.3 Discussion (Advantages and Disadvantages) ............................................... 16

2.3 Computer Cognitive Screening ........................................................................... 17
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1 Overview</td>
<td>17</td>
</tr>
<tr>
<td>2.4 Discussion (Advantages and Disadvantages)</td>
<td>18</td>
</tr>
<tr>
<td>2.5 Serious Games for Cognitive Screening</td>
<td>18</td>
</tr>
<tr>
<td>2.5.1 Introduction</td>
<td>18</td>
</tr>
<tr>
<td>2.5.2 Serious Games for Cognitive Screening</td>
<td>19</td>
</tr>
<tr>
<td>2.5.3 Discussion (Advantages and Disadvantages)</td>
<td>24</td>
</tr>
<tr>
<td>2.6 Summary</td>
<td>25</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>27</td>
</tr>
<tr>
<td>3 Methodology</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Description of the game</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Design and Development Process</td>
<td>27</td>
</tr>
<tr>
<td>3.2.1 Design</td>
<td>27</td>
</tr>
<tr>
<td>3.2.2 Development</td>
<td>28</td>
</tr>
<tr>
<td>3.3 Measurements</td>
<td>28</td>
</tr>
<tr>
<td>3.4 User Requirements</td>
<td>28</td>
</tr>
<tr>
<td>3.5 Game Architecture and Design</td>
<td>29</td>
</tr>
<tr>
<td>3.5.1 Level Design</td>
<td>30</td>
</tr>
<tr>
<td>3.5.2 Character Type</td>
<td>30</td>
</tr>
<tr>
<td>3.5.3 Character Appearance</td>
<td>31</td>
</tr>
<tr>
<td>3.5.4 Character Size</td>
<td>31</td>
</tr>
<tr>
<td>3.5.5 Game Length</td>
<td>31</td>
</tr>
<tr>
<td>3.5.6 Data Storage</td>
<td>31</td>
</tr>
<tr>
<td>3.6 Game Hardware</td>
<td>32</td>
</tr>
<tr>
<td>3.7 Clinical Study Methodology</td>
<td>32</td>
</tr>
<tr>
<td>3.8 Experimental Study Methodology</td>
<td>33</td>
</tr>
<tr>
<td>3.9 Game Variants</td>
<td>33</td>
</tr>
</tbody>
</table>
Chapter 4: Assessing Cognitive Status in Elderly Adults with Incident Delirium

4.1 Concurrent Validation Study

4.1.1 Motivation

4.1.2 Research Question

4.1.3 Methodology

4.1.4 Requirements Analysis

4.1.5 Statistical Analysis

4.1.6 Results

4.1.7 Detection of Abnormal State Using Serious Game Performance

4.1.8 Discussion

4.1.9 Conclusion

4.2 Reliability Study

4.2.1 Research Question

4.2.2 Methodology

4.2.3 Results

4.2.4 Discussion

4.2.5 Conclusion (for the reliability study)

4.3 Conclusions (Across the First Two Clinical Studies)

Chapter 5: Re-Assessing the Concurrent Validity of the Serious Game in Elderly Adults

5.1 Research Questions

5.2 Methodology

5.2.1 Statistical Analysis

5.3 Results

5.3.1 Description of Sample
| 6.6.3 | Relationship Between Game Performance and Cognitive Ability | 97 |
| 6.6.4 | Predicting Cognitive Performance Using Game Performance | 97 |
| 6.6.5 | Design Recommendations | 98 |
| 6.6.6 | Limitations | 98 |
| 6.6.7 | Further Research | 99 |
| 6.7 | Conclusion | 99 |

Chapter 7

| 7 | Conclusion | 100 |
| 7.1 | General Discussion of the Results and their Implications | 100 |
| 7.2 | Contributions | 102 |
| 7.3 | Limitations | 103 |
| 7.4 | Future Work | 104 |
| 7.5 | Final Words | 104 |

References

Appendices

| 8 | Appendices | 142 |
| 8.1 | Appendix A. Literature Review on Computerized Cognitive Assessments | 142 |
| 8.1.1 | Translated Computerized Assessment Tools | 142 |
| 8.1.2 | Computerized Assessment Tools | 142 |
| 8.1.3 | Discussion (Advantages and Disadvantages) | 151 |
| 8.1.4 | Literature Review on Other EF (Updating and Shifting) | 152 |
| 8.2 | Appendix B. Literature Review Supporting Materials | 157 |
| 8.3 | Appendix C. Supporting Materials for the Concurrent Validation Study in Assessing Cognitive Status in Elderly Adults with Incident Delirium | 158 |
| 8.3.1 | Supporting Methodology Materials | 158 |
| 8.3.2 | Supporting Results Materials | 160 |
8.4 Appendix D. Supporting Materials for the Test-Retest Reliability Study in Assessing Cognitive Status in Elderly Adults with Incident Delirium .................................................. 163

8.5 Appendix E. Supporting Materials for Re-Assessing the Concurrent Validity of the Serious Game in Elderly Adults .................................................................................. 168

8.6 Appendix F. Supporting Materials for the Construct Validation of a Serious Game for Assessment of Executive Functioning ................................................................. 171

8.6.1 Experimental Materials .................................................................................. 171

8.6.2 Inhibition Task Implementation .................................................................... 174

8.6.3 Supporting Methodology Materials ............................................................... 175

8.6.4 Supporting Results Materials ....................................................................... 176

8.7 Appendix G. Literature Review on Response Time as a Measure of Executive Function .......................................................... 178

8.7.1 Executive Function ....................................................................................... 178

8.7.2 Brain Injury ................................................................................................... 179

8.7.3 Dementia ....................................................................................................... 180

8.7.4 Mild Cognitive Impairment ........................................................................... 181

8.7.5 Using RT IIV as a Predictor of Cognitive Ability and Performance .......... 182

8.7.6 Measures of RT IIV and Methodological Issues ........................................... 182

8.8 Appendix H. Assessing Cognitive Status in Elderly Adults with Prevalent Delirium Study ........................................................................................................ 185

8.8.1 Assessing Cognitive Status in Elderly Adults with Prevalent Delirium ...... 185

8.9 Appendix I. Supporting Materials for Assessing Cognitive Status in Elderly Adults with Prevalent Delirium Study ........................................................................ 194

8.9.1 Game variants used in the validation study .................................................. 195

8.10 Appendix J. Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting Study .................................................................................. 197

8.10.1 Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting ................................................................................................. 197

8.11 Methodology .................................................................................................. 198

8.11.1 Recruitment ................................................................................................ 198
8.11.2 Cognitive Task Performance While Walking.........................................................203

8.12 Appendix K. Supporting Materials for Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting Study.................................................212

8.12.1 Experimental Materials......................................................................................212

8.12.2 Supporting Materials.........................................................................................215

8.12.3 Supporting Methodology Materials....................................................................218

8.12.4 Supporting Results Materials.............................................................................218

8.13 Appendix L. Delirium Index.................................................................................223
List of Tables

Table 3-1. Summary of cognitive assessments administered in the three clinical studies. .......... 33

Table 3-2. Summary of game measures collected based on version of game evaluated in the three clinical studies.................................................................................................................................................................................. 34

Table 4-1. Summary of completion rates for standard cognitive assessment scores. ............... 40

Table 4-2. Correlations comparing game performance to the standard cognitive assessments. * \( p<0.05 \), ** \( p<0.01 \), *** \( p<0.001 \). MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.............................................................................................................................................................................................................. 42

Table 4-3. Partial correlations that control for CRT reaction time on game performance and standard cognitive assessments and control for game median correct RT on standard cognitive assessments. * \( p<0.05 \), ** \( p<0.01 \), *** \( p<0.001 \). MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time. ................................................................................................................................................................. 43

Table 4-4. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (≥24) or presence (<24) of dementia as assessed by the MMSE. MMSE = Mini-Mental State Examination, and CRT = Choice Reaction Time................................................................. 44

Table 4-5. Mann-Whitney U test results comparing game performance based on the absence (≥23) or presence (<23) of cognitive impairment as assessed by the MoCA. MoCA = Montreal Cognitive Assessment, and CRT = Choice Reaction Time. ........................................................................................................................................... 46

Table 4-6. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (CAM negative) or presence (CAM positive) of delirium as assessed by the CAM. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, and CRT = Choice Reaction Time. .................................................................................................................................................. 47
Table 5-1. Correlations comparing game performance to the standard cognitive assessments. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, and DI = Delirium Index. ................................. 67

Table 5-2. Mann-Whitney $U$ test results comparing cognitive assessment performance based on the absence ($\geq 24$) or presence ($<24$) of dementia as assessed by the MMSE. MMSE = Mini-Mental State Examination. ............................................................................................................. 69

Table 6-1. Summary of hypothesis testing. ......................................................................................................................... 88

Table 6-2. Two-tailed correlations ($n=30$) between cognitive measures (Pearson’s $r$ appear on the top row and Spearman’s rho are on the bottom row in italics, in each cell). Significant values greater than $\pm 0.5$ are bolded. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. AS = anti-saccade task, F = Flanker task, GNG = Go/No-Go discrimination task, and S = Stroop task. ........................................... 90

Table 6-3. Two-tailed correlations ($n=30$) between game performance measures and response inhibition assessed by the combined metric on block 1 of the Go/No-Go discrimination task (Equation 6-3). Significant values greater than $+/−0.5$ are bolded. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. ................................................................................................................................. 93

Table 6-4. Two-tailed correlations between the game performance metric (Equation 3-2) for each game block and response inhibition assessed by the combined metric on block 1 of the Go/No-Go discrimination task (Equation 6-3). Significant values greater than $\pm 0.5$ are bolded. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. ................................................................................................................................. 94

Table 8-1. Comparative summary of standard cognitive assessments. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method. ................................................................................................................................. 157

Table 8-2. Comparison between traditional paper-and-pencil cognitive assessments and the use of serious games for cognitive screening. ................................................................................................................................. 157

Table 8-3. Detailed information on the different game variants played by patients. ........................................................................... 158

Table 8-4. Summary of study sample demographics and cognitive assessment scores. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion
Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.

Table 8-5. Correlations comparing game performance to the standard cognitive assessments. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.

Table 8-6. Partial correlations that control for CRT reaction time on game performance and standard cognitive assessments. CRT = Choice Reaction Time, MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, and DVT = Digit Vigilance Test.

Table 8-7. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (≥24) or presence (<24) of dementia as assessed by the MMSE. MMSE = Mini-Mental State Examination.

Table 8-8. Mann-Whitney U test results comparing game performance based on the absence (≥23) or presence (<23) of cognitive impairment as assessed by the MoCA. MoCA = Montreal Cognitive Assessment.

Table 8-9. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (CAM negative) or presence (CAM positive) of delirium as assessed by the CAM. CAM = Confusion Assessment Method.

Table 8-10. Demographics of the sample study.

Table 8-11. Relationships between sessions on serious game median correct RT, was determined using two-tailed Pearson’s r correlations.

Table 8-12. Relationships between sessions on serious game median target offset for correct trials, was determined using two-tailed Pearson’s r correlations.

Table 8-13. Relationships between sessions on serious game median correct RT, was determined using two-tailed Spearman’s r correlations.
Table 8-14. Relationships between sessions on serious game median target offset for correct trials, was determined using two-tailed Spearman’s $r$ correlations. ......................................................... 165

Table 8-15. Detailed information on the different game variants played by patients.............. 168

Table 8-16. Summary of study sample demographics and cognitive assessment scores. MMSE = Mini-Mental State Examination, CAM = Confusion Assessment Method, and DI = Delirium Index. .......................................................................................................................... 169

Table 8-17. Spearman’s $\rho$ correlations comparing game performance to the standard cognitive assessments. The correlation coefficient is listed on the first line, followed by the $p$-value and $n$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MMSE = Mini-Mental State Examination, CAM = Confusion Assessment Method, and DI = Delirium Index. .......................................................................................................................... 169

Table 8-18. Summary of cognitive task and measure..................................................................... 175

Table 8-19. Summary of exploratory factor analysis results for the game median correct RT data $(n=30)$........................................................................................................................................ 176

Table 8-20. Summary of exploratory factor analysis results for the game false alarm rate data $(n=30)$........................................................................................................................................ 176

Table 8-21. Two-tailed Pearson’s $r$ correlation between game performance and cognitive measures. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MMSE = Mini-Mental State Examination, CAM = Confusion Assessment Method, and DI = Delirium Index................................................................. 190

Table 8-22. Summary of study sample demographics and cognitive assessment scores. MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, and DI = Delirium Index. .......................................................................................................................... 194

Table 8-23. Modified technology questionnaire questions and responses. ............................. 195

Table 8-24. Game settings for the different versions of the game. The target size for all levels was held constant at 175 px. ........................................................................................................... 195

Table 8-25. Two-tailed Pearson’s $r$ correlation between game performance and cognitive measures. *Correlations involving the CAM were calculated using point-biserial correlations. *
Table 8-26. Two-tailed Spearman’s rho correlation between game performance and cognitive measures. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, and DI = Delirium Index.

Table 8-27. Two-tailed Pearson’s $r$ correlation analysis between the median correct RT game-based cognitive assessment played by while sitting with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

Table 8-28. Two-tailed Pearson’s $r$ correlation analysis between the median correct RT on game-based cognitive assessment while walking with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

Table 8-29. Two-tailed Pearson’s $r$ correlation analysis between performance measures on the cognitive task and DT cost of playing the serious game using stride variability. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

Table 8-30. Two-tailed Pearson’s $r$ correlation analysis between performance measures on the cognitive task and DT cost of playing the serious game using stride time. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

Table 8-31. Two-tailed Pearson’s $r$ correlation analysis between the median target offset game-based cognitive assessment played by while sitting with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

Table 8-32. Two-tailed Spearman’s rho correlation analysis between the median correct RT game-based cognitive assessment played by while sitting with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.
Table 8-33. Two-tailed Spearman’s \( \rho \) correlation analysis between the median target offset game-based cognitive assessment played by while sitting with cognitive measures. \( n = 20 \) for the correlations with all cognitive tasks except for the WCST, where \( n = 19 \). \( * p < 0.05, ** p < 0.01, *** p < 0.001 \). WCST = Wisconsin-card sorting task. ................................................................. 219

Table 8-34. Two-tailed Pearson’s \( r \) correlation analysis between the median correct target offset on game-based cognitive assessment while walking with cognitive measures. \( n = 20 \) for the correlations with all cognitive tasks except for the WCST, where \( n = 19 \). \( * p < 0.05, ** p < 0.01, *** p < 0.001 \). WCST = Wisconsin-card sorting task. ................................................................. 219

Table 8-35. Two-tailed Spearman’s \( \rho \) correlation analysis (\( n = 20 \)) between the median correct RT on game-based cognitive assessment while walking with cognitive measures. \( * p < 0.05, ** p < 0.01, *** p < 0.001 \). WCST = Wisconsin-card sorting task. ................................................................. 220

Table 8-36. Two-tailed Spearman’s \( \rho \) correlation analysis (\( n = 20 \)) between the median correct target offset on game-based cognitive assessment while walking with cognitive measures. \( * p < 0.05, ** p < 0.01, *** p < 0.001 \). WCST = Wisconsin-card sorting task. ................................................................. 220

Table 8-37. Two-tailed Spearman’s \( \rho \) correlation analysis between performance measures on the cognitive task and dual-task (DT) cost of playing the serious game using stride variability. \( * p < 0.05, ** p < 0.01, *** p < 0.001 \)................................................................. 221

Table 8-38. Two-tailed Spearman’s \( \rho \) correlation analysis (\( n = 18 \)) between performance measures on the cognitive task and dual-task (DT) cost of playing the serious game using stride time. \( * p < 0.05, ** p < 0.01, *** p < 0.001 \)................................................................. 221
List of Figures

Figure 1-1. Overview of the four studies (three clinical and one experimental) carried out in this thesis (note that “clinical reliability study” is referred to elsewhere in this thesis as the “second clinical study” and the ”clinical validation study 2” is elsewhere referred to as the “third clinical study”. The “experimental study” is elsewhere referred to as the “fourth study”. .......................... 3

Figure 3-1. Screenshot of the whack-a-mole game. .......................................................................................................................... 27

Figure 3-2. Process diagram of the serious game. .......................................................................................................................... 30

Figure 4-1. Histogram of DI scores for CAM negative and CAM positive patients. DI = Delirium Index. CAM = Confusion Assessment Method. .......................................................................................................................... 42

Figure 4-2. Scatterplot illustrating the differences on game median correct RT based on MMSE score (≥24 = normal cognitive function or possible dementia; <24 = signs of dementia). MMSE = Mini-Mental State Examination. .......................................................................................................................... 45

Figure 4-3. Scatterplot illustrating the differences on game median correct RT based on MoCA score (≥23 = normal cognitive function or possible MCI; <23 = signs of MCI). MoCA = Montreal Cognitive Assessment, MCI = mild cognitive impairment .......................................................................................................................... 46

Figure 4-4. Dot plot illustrating the differences on game median correct RT based on CAM groups (CAM negative=delirium absent; CAM positive=delirium present). CAM = Confusion Assessment Method. .......................................................................................................................... 48

Figure 4-5. Overview of the sample population used in the validation and reliability study. In the reliability study, of the 127 enrolled patients, only 114 patients completed the study. (Right) The diagram on the right illustrates the overlap in 56 patients between the validation and reliability study. (Left) The diagram on the left illustrates the number of patients that completed each follow up session. .......................................................................................................................... 56

Figure 4-6. Bar chart illustrating the mean of median correct RT across each follow up session. Error bars represent 95% CI. .......................................................................................................................... 58
Figure 4-7. Histogram illustrating the median correct RT difference between successive sessions. Each bin represents 100 msec. ................................................. 59

Figure 5-1. Scatterplot illustrating the differences on the performance metric based on MMSE score (≥24 = normal cognitive function or possible dementia; <24 = signs of dementia). MMSE = Mini-Mental State Examination. ............................................................. 68

Figure 6-1. The impact of character appearance duration to mean signal detection sensitivity (d’) value. ........................................................................................................ 73

Figure 6-2. Interaction plot comparing character appearance duration to mean of median correct RT. The different lines represent target frequency. ............................................................. 74

Figure 6-3. Interaction plot comparing character speed to mean probability of a miss. The different lines represent target frequency. ........................................................................ 75

Figure 6-4. Interaction plot for median correct RT based on character appearance duration, grouped by target frequency. .................................................................................. 84

Figure 6-5. Interaction boxplot for miss rates based on character appearance duration, grouped by target frequency. .................................................................................. 85

Figure 6-6. Interaction boxplot for combined performance metric based on character appearance duration, grouped by target frequency. ................................................................ 85

Figure 6-7. Interaction boxplot for median correct RT based on distractor type, grouped by target frequency........................................................................................................ 87

Figure 6-8. Interaction boxplot for false alarm rates based on distractor type, grouped by target frequency ........................................................................................................ 87

Figure 8-1. Barplot illustrating the number of assessments completed at each session. ........ 163

Figure 8-2. Scatterplots depicting the serious game median correct RT between: (A) initial enrolment compared to follow up session 1 (r=0.927, p<0.001), (B) follow up session 1 compared with 2 (r=0.833, p=0.001), and (C) follow up session 2 compared with 3 (r=0.547, p=0.0656). ........................................................................................................ 164
Figure 8-3. Scatterplots depicting the serious game median target offset for correct trials between: (A) initial enrolment compared to follow up session 1 ($r=0.822, p<0.001$), (B) follow up session 1 compared with 2 ($r=0.818, p=0.002$), and (C) follow up session 2 compared with 3 ($r=0.492, p=0.105$).

Figure 8-4. Bar chart illustrating the mean of median correct target offset (px) between initial enrolment compared to follow up session 3. Error bars are 95% CI.

Figure 8-5. Scatterplot illustrating the differences on game correct RT performance based on MMSE score ($\geq 24 =$ normal cognitive function or possible dementia; $< 24 =$ signs of dementia). MMSE = Mini-Mental State Examination.

Figure 8-6. Scatterplot illustrating the differences on game false alarm rate performance based on MMSE score ($\geq 24 =$ normal cognitive function or possible dementia; $< 24 =$ signs of dementia). MMSE = Mini-Mental State Examination.

Figure 8-7. Visual of game board setup with a grid size of 3x3 and character size of 175 px.

Figure 8-8. Diagram illustrating the relationship between different measures of inhibition.

Figure 8-9. Boxplot of median correct RT per game level.

Figure 8-10. Boxplot of median correct RT per game level while walking.

Figure 8-11. Illustration of how the smartphone was affixed to the individual during the gait analysis portion of the study using a smartphone holder.
List of Appendices

8.1 Appendix A. Literature Review on Computerized Cognitive Assessments. .................. 142

8.2 Appendix B. Literature Review Supporting Materials. .................................................. 157

8.3 Appendix C. Supporting Materials for the Concurrent Validation Study in Assessing Cognitive Status in Elderly Adults with Incident Delirium...................................................... 158

8.4 Appendix D. Supporting Materials for the Test-Retest Reliability Study in Assessing Cognitive Status in Elderly Adults with Incident Delirium...................................................... 163

8.5 Appendix E. Supporting Materials for Re-Assessing the Concurrent Validity of the Serious Game in Elderly Adults ........................................................................................................ 168

8.6 Appendix F. Supporting Materials for the Construct Validation of a Serious Game for Assessment of Executive Functioning.............................................................. 171

8.7 Appendix G. Literature Review on Response Time as a Measure of Executive Function. 178

8.8 Appendix H. Assessing Cognitive Status in Elderly Adults with Prevalent Delirium Study. 185

8.9 Appendix I. Supporting Materials for Assessing Cognitive Status in Elderly Adults with Prevalent Delirium Study........................................................................................................ 194

8.10 Appendix J. Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting Study.............................................................................................................. 197

8.12 Appendix K. Supporting Materials for Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting Study........................................................................ 212

8.13 Appendix L. Delirium Index.......................................................................................... 223
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACC</td>
<td>Anterior cingulate cortex</td>
</tr>
<tr>
<td>AD</td>
<td>Alzheimer’s Disease</td>
</tr>
<tr>
<td>ANAM</td>
<td>Automated Neuropsychological Assessment Metrics</td>
</tr>
<tr>
<td>ADAS-Cog</td>
<td>Alzheimer’s Disease Assessment Scale-cognition sub-scale</td>
</tr>
<tr>
<td>ANT</td>
<td>Attention network task</td>
</tr>
<tr>
<td>BADS</td>
<td>Behavioural Assessment of the Dysexecutive Syndrome</td>
</tr>
<tr>
<td>CAM</td>
<td>Confusion Assessment Method</td>
</tr>
<tr>
<td>CANS-MCI</td>
<td>Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment</td>
</tr>
<tr>
<td>CANTAB</td>
<td>Cambridge Neuropsychological Test Automated Battery</td>
</tr>
<tr>
<td>CIHI</td>
<td>Canadian Institute for Health Information</td>
</tr>
<tr>
<td>CNTB</td>
<td>Computerized Neuropsychological Test Battery</td>
</tr>
<tr>
<td>COGDRAS</td>
<td>Cognitive Drug Research Computerized Assessment System</td>
</tr>
<tr>
<td>COGDRAS-D</td>
<td>Cognitive Drug Research Computerized Assessment System-Dementia</td>
</tr>
<tr>
<td>CogScreen-AE</td>
<td>CogScreen – Aeromedical Edition</td>
</tr>
<tr>
<td>CRT</td>
<td>Choice Reaction Time</td>
</tr>
<tr>
<td>CSI</td>
<td>Cognitive Stability Index</td>
</tr>
<tr>
<td>CST</td>
<td>Cognitive Screening Test</td>
</tr>
<tr>
<td>CV</td>
<td>Coefficient of variation</td>
</tr>
<tr>
<td>DI</td>
<td>Delirium Index</td>
</tr>
<tr>
<td>DLB</td>
<td>Dementia with Lewy Bodies</td>
</tr>
<tr>
<td>DT</td>
<td>Dual-task</td>
</tr>
<tr>
<td>DTW</td>
<td>Dual-task while walking</td>
</tr>
<tr>
<td>DVT</td>
<td>Digit Vigilance Test</td>
</tr>
<tr>
<td>ED</td>
<td>Emergency department</td>
</tr>
<tr>
<td>EF</td>
<td>Executive function</td>
</tr>
<tr>
<td>ERP</td>
<td>Event related potential</td>
</tr>
<tr>
<td>FAB</td>
<td>Frontal assessment battery</td>
</tr>
<tr>
<td>FLB</td>
<td>Frontal lobe dementia</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional magnetic resonance imaging</td>
</tr>
<tr>
<td>fNIR</td>
<td>Functional near-infrared</td>
</tr>
<tr>
<td>HDS-R</td>
<td>Revised Hasegawa Dementia Scale</td>
</tr>
<tr>
<td>IIV</td>
<td>Intra-individual variability</td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile range</td>
</tr>
<tr>
<td>ISD</td>
<td>Intra-individual standard deviation</td>
</tr>
<tr>
<td>IVar</td>
<td>Intra-individual variance</td>
</tr>
<tr>
<td>LTC</td>
<td>Long-term care</td>
</tr>
<tr>
<td>MCI</td>
<td>Mild cognitive impairment</td>
</tr>
<tr>
<td>MMSE</td>
<td>Mini-Mental State Examination</td>
</tr>
<tr>
<td>MoCA</td>
<td>Montreal Cognitive Assessment</td>
</tr>
<tr>
<td>OT</td>
<td>Occupational therapist</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal components analysis</td>
</tr>
<tr>
<td>PET</td>
<td>Positron emission transmission</td>
</tr>
<tr>
<td>RA</td>
<td>Research assistant</td>
</tr>
<tr>
<td>RASS</td>
<td>Richmond Agitation-Sedation Scale</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>----------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>RAVLT</td>
<td>Rey auditory verbal learning test</td>
</tr>
<tr>
<td>RT</td>
<td>Response time</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SRT</td>
<td>Simple reaction time</td>
</tr>
<tr>
<td>SUS</td>
<td>System Usability Scale</td>
</tr>
<tr>
<td>TENI</td>
<td>Test de Evaluación Neuropsicológica Infantil</td>
</tr>
<tr>
<td>TMT</td>
<td>Trail making test</td>
</tr>
<tr>
<td>VR</td>
<td>Virtual reality</td>
</tr>
<tr>
<td>WAIS</td>
<td>Wechsler Adult Intelligence Scale</td>
</tr>
<tr>
<td>WCST</td>
<td>Wisconsin card sorting task</td>
</tr>
</tbody>
</table>
## List of Units

<table>
<thead>
<tr>
<th>Unit</th>
<th>Equivalent Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>msec</td>
<td>millisecond</td>
</tr>
<tr>
<td>px</td>
<td>pixel</td>
</tr>
</tbody>
</table>
Chapter 1

1 Introduction

1.1 Problem Statement

Cognitive ability is a requirement for performing tasks, whether they are relatively simple such as walking, or bathing, all the way up to complex tasks that require problem solving. In gerontological contexts, assessing cognitive status is important in distinguishing between normal decline due to aging and abnormal deterioration due to diseases or brain injury. Existing measures of cognitive status used in healthcare have been developed especially for use in clinical contexts. Clinical measures are traditionally paper-and-pencil based tests that can be costly to administer (copyright fees and personnel costs), potentially time consuming, and require administration by trained personnel.

An alternative approach to paper-based cognitive assessment is to use computerized versions that can be self-administered with no, or minimal, assistance. The translation of existing assessments into digital formats is an example of this approach. Another method (which is the focus of this dissertation) is to extract cognitive features of existing screening tools and design serious games to assess cognitive status. Serious games are games whose primary purpose is not entertainment, but in the context of cognitive screening, they can be more motivating and enjoyable. The goal of the research reported in this dissertation was to demonstrate the use of a game-based approach to cognitive assessment that could potentially be used both for healthy adults across the life span and for older adults in clinical contexts.

1.2 Research Questions and Overview of Studies

The main research questions addressed in this dissertation were framed with respect to the whack-a-mole game that I developed (as an instance of other games of its type that may be developed in the future) and are as follows:

1. How does performance on the serious game played by older adults in a clinical setting compare to existing methods of cognitive assessment in a clinical setting (i.e., the Mini-Mental State Exam, the Montreal Cognitive Assessment, the Confusion Assessment Method, and the Delirium Index)?
2. How reliable are game performance measures when measured repeatedly on the same
participants in a clinical setting?

3. Is game performance correlated with general inhibitory executive function, or is it
associated with specific subtypes of inhibition, and if so, which one(s)?

The first two research questions were addressed through three studies carried out in a clinical
setting with older adults. The remaining research question was addressed using a fourth study,
with healthy adults in a laboratory setting.

An overview of the four studies (three clinical and one experimental) and the different samples
are shown in Figure 1-1. Two clinical studies were reported in Chapter 4 with an overlap in
sample population. In the first clinical study evaluating the validation of the game, there were
147 participants enrolled but only 141 that completed the study, of which 56 patients overlapped
(shaded area in Figure 1-1) with the second clinical study (which also collected data from 71
non-overlapping participants) assessing the reliability of the game. Adding in the 132
participants in the third clinical study and the 30 participants in the fourth study, this thesis
covers data on 418 enrolled participants, of which 361 participants completed the study and
played serious game. In addition, data from six participants in a pilot study informed the
selection of character appearance durations that were used in the fourth study.

<table>
<thead>
<tr>
<th>Clinical Validation Study 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Chapter 4)</td>
</tr>
<tr>
<td><em>n=147 enrolled</em></td>
</tr>
<tr>
<td><em>n=141</em></td>
</tr>
<tr>
<td>n=85 * (non-overlap)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical Reliability Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Chapter 4)</td>
</tr>
<tr>
<td><em>n=127 enrolled</em></td>
</tr>
<tr>
<td><em>n=114</em></td>
</tr>
<tr>
<td>n=56 * (overlap)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Clinical Validation Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Chapter 5)</td>
</tr>
<tr>
<td><em>n=170 enrolled</em></td>
</tr>
<tr>
<td><em>n=132</em></td>
</tr>
<tr>
<td>n=71 * (non-overlap)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experimental Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Chapter 6)</td>
</tr>
<tr>
<td><em>n=30 enrolled</em></td>
</tr>
</tbody>
</table>

* Number of participants that completed the study.

a Number of participants that completed validation study 1.
b Number of participants that completed the reliability study.
Figure 1-1. Overview of the four studies (three clinical and one experimental) carried out in this thesis (note that “clinical reliability study” is referred to elsewhere in this thesis as the “second clinical study” and the ”clinical validation study 2” is elsewhere referred to as the “third clinical study”. The “experimental study” is elsewhere referred to as the “fourth study”.

1.3 Road Map

This thesis begins with a literature review covering how cognition can be measured in both clinical and non-clinical contexts. Next, a review of existing computer-based assessments is presented followed by description of a representative set of existing serious games for cognitive assessment. Chapter 3 focuses on the methodology of the serious game, how it was designed and developed.

Evaluation of the Serious Game in a Clinical Context

Chapter 4 describes two clinical studies, with the first being an evaluation of the concurrent validity of the serious game in a clinical context (hospital emergency department) used by elderly, emergency patients to screen for incident delirium, which is delirium that occurs over time. Performance on the game was compared to standard clinical assessments. In the second clinical study, test-retest reliability of game performance was assessed, also in the emergency department setting.

Chapter 5 describes the evaluation of the serious game in a third clinical study involving elderly emergency patients. In this case, false alarm rate is added as an additional measure of game performance.

Evaluation of the Serious Game Construct Validity

Chapter 6 describes the construct validation of the serious game through an experimental study carried out with healthy participants. The effect of different parameters of the serious game in terms of impact on game difficulty was evaluated and game performance was compared to performance on tasks representing different subtypes of inhibition ability.

Conclusion:
The thesis concludes in Chapter 7 with a discussion of the implications and contributions of the research. Chapter 7 also discusses the limitations of this research and suggests topics for future research.
Chapter 2

2 Background

2.1 Cognitive Assessments

Assessing cognitive status is an important indicator of functional ability, independence, and mortality (Inouye, Bogardus, Baker, Leo-Summers, & Cooney, 2000). The under-diagnosis of cognitive impairments can lead to increased economic burden, hospitalization, and even death (Inouye et al., 2000). In geriatric healthcare there are standard cognitive assessment tools that screen for cognitive impairment, such as the Mini-Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975), Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005), and Confusion Assessment Method (CAM) (Inouye et al., 1990).

2.1.1 MMSE

The MMSE is a paper-and-pencil based questionnaire graded out of 30-points that evaluates the following cognitive domains: orientation, registration, attention and calculation, recall, and language (Folstein et al., 1975). On the 30-point scale, scores ranging between 0 and 17 suggest severe cognitive impairment; scores between 18 and 23 suggest moderate or mild cognitive impairment (MCI), and scores between 24 and 30 suggest no cognitive impairment. Scores on the MMSE are often used to rate dementia severity from mild, moderate, to severe (Tombaugh & McIntyre, 1992). Moreover, administration of the MMSE requires a trained administrator to ask the patient a set of questions and one question requires the patient to draw on the test sheet. The assessment can be completed within 5 to 10 minutes (Tombaugh & McIntyre, 1992). It is copyrighted and (at the time of this writing) the MMSE can be purchased for $81.00 USD for a package of 50 test booklets (PAR, 2018). Assuming one test booklet is administered per patient, the cost for each administration is $1.62 USD per patient (i.e., $81.00 USD/50 test booklets = $1.62 USD).

2.1.1.1 Reliability of the MMSE

The test-retest reliability of the MMSE has been evaluated through the administration of the test on two successive days (Folstein et al., 1975) with a correlation of $r=0.887$ being observed between the two sets of scores. Other researchers have also evaluated the test-retest reliability of
the MMSE in both short (i.e., twice on the same day with the same and different MMSE evaluators) and long-term intervals (i.e., two months to two years between successive administrations). In studies with a test-retest interval of two months or less, the reliability coefficient ranged from 0.80 to 0.95 in “cognitively intact” adults (Tombaugh & McIntyre, 1992). In intervals greater than two months, the reliability coefficients ranged from 0.79 to 0.92. In longer test-retest intervals such as one year, the reliability coefficient ranged from 0.227 to 0.50 (Mitrushina & Satz, 1991; Olin & Zelinski, 1991). In an interval of two years, the reliability coefficient was 0.38 (Mitrushina & Satz, 1991).

2.1.1.2 Validity of the MMSE

The validity of the MMSE has been examined by evaluating the test’s sensitivity, specificity, and concurrent validity. Sensitivity refers to ability to correctly identify individuals having a condition (true positive amongst all those with the disease). In comparison, specificity refers to ability to correctly identify individuals not having a condition (true negative amongst all those who are disease free). The validity of the MMSE was originally established by comparing MMSE performance between three samples: (1) controls, (2) adults with dementia or depression, and (3) adults with repeated hospital admissions with dementia or depression (Folstein et al., 1975). Patients diagnosed with dementia or depression was preselected for the study. The original study by (Folstein et al., 1975) does not describe the methodology or diagnostic test originally used to diagnosis these patients. Mean scores on the MMSE were highest in the control group (mean = 27.6), followed by adults with depression in sample B (mean MMSE score = 25.1), and by adults with depression and cognitive impairment (mean MMSE score = 19.0), and finally adults with dementia scored the lowest (mean MMSE score = 9.7).

Concurrent validity refers to the extent to which performance on one test corresponds to another test that has been previously established to measure the same construct. Folstein et al. (1975) originally examined the concurrent validity of the MMSE by comparing performance on the MMSE with the Verbal and Performance components of the Weschler Adult Intelligence Scale (WAIS). The correlation between MMSE scores and the Verbal component was $r=0.6776$, and correlation between MMSE scores and the Performance component was $r=0.660$, which demonstrated the concurrent validity of the MMSE with an existing measure of cognitive status.
2.1.2 MoCA

The MoCA is another example of a paper-and-pencil based cognitive assessment scored out of 30-points, that takes 10-minutes to complete (Nasreddine et al., 2005). There are eight sections on the MoCA, which includes tasks on: visuospatial/executive function, naming, memory, attention, language, abstraction, delayed recall, and orientation. It was designed for physicians to screen for MCI, a transitional state between normal cognitive functioning and dementia. Similar to the MMSE, a trained test administrator asks a series of questions to a patient.

2.1.2.1 Reliability of the MoCA

The test-retest reliability of the MoCA was evaluated with 26 adults with a strong correlation ($r=0.92$, $p<0.001$) being found between the two administrations of the test, which were on average 35 days apart (Nasreddine et al., 2005).

2.1.2.2 Validity of the MoCA

The validity of the MoCA was evaluated by comparing performance on this assessment with scores on the MMSE. Nasreddine et al. (2005) compared three groups of adults: (1) controls, (2) adults with MCI, and (3) adults with Alzheimer’s Disease (AD). Patients were diagnosed with MCI by a trained neurologist or geriatrician and evaluated with a standard mental status battery (Nasreddine et al., 2005). The control group had the highest number of scores within the normal range for both the MMSE and MoCA. Adults in the AD group scored within the abnormal group for both the MMSE and MoCA. In contrast, 73% of adults with MCI scored within the normal MMSE range but were abnormal on the MoCA. Nasreddine et al. (2005) demonstrated agreement between MoCA and MMSE scores for 274 adults, $r=0.87$, $p<0.001$.

2.1.3 CAM

The CAM is a paper-and-pencil based binary assessment of the presence or absence of delirium (Inouye et al., 1990). Delirium is an acute confusional state that is a complication of other illness or acute medical care. Delirium is very common in the hospital setting, and can increase mortality rates, the length of hospital stay, and threatens an individual’s ability to live independently. Delirium is characterized by acute onset and fluctuating attention, and cognitive status (e.g., memory, orientation, and language). In contrast to the CAM which is dichotomous (where CAM positive indicates the likely presence of delirium), the Delirium Index (DI)
Appendix L) (McCusker, Cole, Bellavance, & Primeau, 1998) provides a rating of the severity of symptoms. The CAM was designed using the Diagnostic and Statistical Manual of Mental Disorders Third Edition, Revised criteria for delirium and is divided into four categories: acute onset, inattention, disorganized thinking and altered level of consciousness. It is designed for use by trained administrators, and takes approximately 5 to 10 minutes to complete, which is similar to the completion times for the MMSE and MoCA (Inouye et al., 1990). The CAM is based on observations on a patient by a trained administrator, and does not require any response from the patient.

2.1.3.1 Reliability of the CAM

The inter-rater reliability of the CAM has been assessed by comparing individual ratings by two raters who assessed a patient at the same time while one of them carried out an interview with the patient. An inter-rater reliability for the presence or absence of the delirium was 100% (κ=1.0) (Inouye et al., 1990). However, the test-retest reliability of the CAM was not evaluated (Inouye et al., 1990), as delirium is characterized by a fluctuating course.

2.1.3.2 Validity of the CAM

Face validity refers to the extent that a test appears to effectively measure the intended construct. Content validity refers to the degree to which a test measures all aspects of a given construct. The face and content validity of the CAM was evaluated by a panel of experts that reviewed whether the CAM addresses characteristics of delirium (Inouye et al., 1990). The concurrent validation of the CAM was originally assessed by comparing the outcome on the CAM with a psychiatrist and geriatrician’s assessment of patients at two separate sites. Both the psychiatrist and geriatrician assessment were blinded to each other’s ratings. The sensitivity of the CAM ranges from 94 to 100% and the specificity ranges from 90 to 95% (Inouye et al., 1990).

2.1.4 Discussion (Advantages and Disadvantages)

Current paper-and-pencil based cognitive screening methods are only minimally interactive, creating little in the way of engagement or entertainment, which may decrease a patient’s motivation to repeat a necessary assessment. They are typically initiated by a healthcare professional rather than sought out by individuals, and they are generally not designed for self-administration or for use by non-clinicians. Some tools such as the CAM are subjective
assessments, which may result in administrator bias (Inouye et al., 1990). It may not be feasible for a test administrator to repeatedly assess individuals for changes in their cognitive status over time. The lack of frequent assessment may result in under-diagnosis of a condition such as delirium, where cognitive status can fluctuate over the course of a day, making it difficult to detect early stages of delirium and initiate preventative measures (Inouye et al., 1990).

Other concerns with existing screening methods include the lack of alternate versions, which can lead to practice/learning effects in tests such as the MMSE. Assessments such as the MMSE and MoCA contain a drawing component, which requires motor skills and movement ability that some patients may not have.

2.2 Cognitive Psychology Tasks and Executive Functioning

This section of the literature review focuses on the concept of executive functions (EFs) as a measure of cognitive status, and covers a description of EFs and how they can be measured.

EFs are responsible for the control and regulation of complex cognitive tasks (Miyake et al., 2000a). According to Miyake et al. (2000a), EFs are chiefly composed of three broad functions: inhibition, shifting, and updating, with a revised model by Miyake et al. (2000a), identifying inhibition as a general EF, with shifting and updating as more specific functions. EFs are important in carrying out a range of activities from problem solving and planning to physical activities such as walking and driving. EFs have been shown to decrease with age (Fisk & Sharp, 2004) and loss of EF ability can lead to the inability to perform activities of daily living (Martyr & Clare, 2012), typically leading to loss of independence or institutional care.

In this thesis, I focus on inhibition tasks since inhibition ability is a major component of executive functioning and it is the component of executive functioning that is targeted by the serious game used in this thesis. Please refer to Appendix A for a discussion of other types of EF (shifting and updating).

2.2.1 Inhibition Ability and Tasks

Inhibition is an EF involved in inhibiting perceptual, cognitive or motor processing and different cognitive tasks tend to require inhibition of varying subgroups of those processes. Inhibitory EF is measured by a variety of tasks such as the Stroop, Go/No-Go discrimination, Flanker, and anti-
saccade tasks. There is considerable evidence that there are different subtypes of inhibition (Dempster, 1992; Diamond, 2013; Friedman & Miyake, 2004; Harnishfeger, 1995; Nigg, 2000). (Diamond, 2013) suggests that inhibition ability can be divided into response inhibition (i.e., self-control, which is the ability to suppress responses to temptations, and avoid acting impulsively), and interference control (selective attention) ability, which is the ability to focus attention to targets while suppressing attention to other distractions/stimuli. Interference control is often referred to as attention control or attention inhibition (Diamond, 2013). Research by Friedman and Miyake (2004) suggests that inhibition and interference-control are inhibition-related processes that can be divided into three categories: (1) pre-potent response inhibition (i.e., the ability to actively suppress information and pre-potent responses), (2) resistance to distractor interference (i.e., the ability to resolve interference and ignore irrelevant information), and (3) resistance to proactive interference (i.e., the ability to focus or select attention to relevant stimuli). In Friedman and Miyake's (2004) concept of inhibition, pre-potent inhibition can be measured using the anti-saccade task (Hallett, 1978), stop-signal task (Logan, 1994), and Stroop task (Stroop, 1935). Their second category of resistance to distractor interference can be measured using the Flanker task (Eriksen & Eriksen, 1974), word naming task (Kane, Hasher, Stoltzfus, Zacks, & Connelly, 1994), and shape matching task (DeSchepper & Treisman, 1996). Their third and final category of inhibition, resistance to pre-potent interference, can be measured using the Brown-Peterson variant (Kane & Engle, 2000), AB-AC-AD task (Rosen & Engle, 1998), and the cued recall task (Tolan & Tehan, 1999).

Work by Nigg (2000) suggests that inhibition can be divided into four processes: (1) interference control (i.e., the ability to suppress responses due to stimulus competition), (2) cognitive inhibition (i.e., the ability to suppress information from working memory), (3) behavioural inhibition (i.e., the ability to suppress pre-potent responses), and (4) oculomotor inhibition (i.e., the suppression of reflexive eye movements/saccades). Nigg's (2000) four category taxonomy is related to Harnishfeger's (1995) classification of inhibition into three dimensions. The first dimension of Harnishfeger's (1995) work describes a type of inhibition that is intentional (i.e., suppressing thoughts and actions) or unintentional (i.e., inhibition that occurs before conscious awareness). The second dimension describes a type of inhibition that is either behavioural or cognitive. For example, behavioural inhibition is reflected in motor responses whereas cognitive inhibition is involved in mental processes. The third and final dimension of Harnishfeger's

The Stroop task (Stroop, 1935) is used as a measure of inhibition ability, where a participant is presented with a word that may be written in a compatible, or incompatible ink/font colour. For instance, the word “red” written in green ink is an example of an incompatible stimulus. The goal is either to identify the semantic meaning of the word (its “reading”) or to identify the ink colour of the word depending on the trial. Congruent trials are trials where the word and font colour match, (i.e., the word “blue” is written in blue ink) and incongruent trials are trials where the word and ink colour do not match (i.e., the word “red” is written in green ink). Response conflict often occurs when a participant incorrectly reads the word instead of the ink colour. As reading is a relatively automated task for an adult native speaker, reading the ink colour in an incongruent trial is difficult to suppress and as a result correct responses on incongruent trials tend to be slower. The “Stroop interference effect” is operationalized as the difference in response time (RT) between incongruent and neutral trials (trials with non-colour words such as “the”, and “when”). A common interpretation for this response conflict is that it is due to participants’ automatically reading the word rather than the ink colour (MacLeod & MacDonald, 2000). In contrast, Stroop facilitation is the RT difference between congruent and neutral trials.
Performance on the Stroop task is related to interference/inhibition control (Diamond, 2013), and attention control inhibition (i.e., the ability to ignore unrelated stimuli from disrupting task performance) (Tiego, Testa, Bellgrove, Pantelis, & Whittle, 2018).

Previous research on the Stroop task has demonstrated that bilinguals have better performance on the task (Bialystok, Craik, & Luk, 2008; Mohamed Zied et al., 2004) compared to monolinguals in children (Bialystok & Martin, 2004; Martin-Rhee & Bialystok, 2008), young adults (Bialystok et al., 2008; Costa, Hernández, & Sebastián-Gallés, 2008), and older adults (Bialystok & Martin, 2004; Mohamed Zied et al., 2004). Bilinguals demonstrate a smaller Stroop interference effect and have more efficient interference suppression. Further research on the effect of bilingualism on the Stroop task has been investigated using behavioural measures such as RT and accuracy and electrophysiological measures such as event-related brain potentials (ERPs) (Kousaie & Phillips, 2012). Kousaie and Phillips (2012, 2017) investigated the bilingual effect on the Stroop, Simon and Flanker tasks for adults. All three of these tasks were expected to require response inhibition and interference suppression (Kousaie & Phillips, 2017). On the Stroop task, monolinguals demonstrated a larger N2 amplitude (Kousaie & Phillips, 2012), whereas bilinguals had a larger P3 amplitude (Kousaie & Phillips, 2017). Both N2 and P3 are event-related potentials that can be measured using electroencephalography. The N2 is a negative-waveform whose activity is thought to be associated with conflict monitoring and activity in the anterior cingulate cortex (ACC) (Mathalon, Whitfield, & Ford, 2003). The P3 is a positive-waveform that is thought to be related to working memory (Donchin, 1981) and implicated in resource allocation (van Veen & Carter, 2002). Both the behavioural and electrophysiological performance on the Stroop task demonstrated that bilinguals (versus monolinguals) detect conflict earlier and need to allocate fewer resources to deal with it.

The beneficial effects of bilingualism on the Stroop task have been challenged. For example, research carried out by Kousaie and Phillips (2017) did not demonstrate an advantage of bilingualism on task performance (for neither RT nor accuracy across all trial types; neutral, incongruent, congruent). This negative result may have occurred because of the large number of trials used in the study by Kousaie and Phillips (2017). Research by Hilchey and Klein (2011) has shown that the advantage of bilinguals (versus monolinguals) in the Stroop task decreases with practice.
Other research on the Stroop task using RT has demonstrated that there is greater anterior cingulate cortex (ACC) activation on incongruent trials compared to congruent trials (Pardo, Pardo, Janer, & Raichle, 1990). The ACC appears to be involved in the selection (Petersen, Fox, Posner, Mintun, & Raichle, 1988), preparation and execution of motor responses in decision-making processes (Zysset, Müller, Lohmann, & von Cramon, 2001). In the Stroop task, Botvinick, Cohen, and Carter (2004) have suggested that there is a conflict monitoring unit that receives input from a response unit, which becomes active during response competition between word naming and colour identification. This formulation predicts that preceding trials influence activation of ACC. Increased ACC activation is then expected to elicit more focus on the color-naming task, thereby reducing the Stroop interference on the following trial.

The Go/No-Go discrimination task is another task used to measure inhibition, where a participant is presented with either a go or no-go stimulus. The goal is to respond only to go stimuli and inhibit responses to no-go stimuli. There is often a higher frequency of go trials than no-go trials, which creates a response bias towards go trials. Incorrect responses to no-go trials may also be related to a participant’s impulsivity (Perry & Carroll, 2008). Performance on the Go/No-Go discrimination task is related to response inhibition, which is a form of inhibition that requires an individual to inhibit a prepotent response (Diamond, 2013; Verbruggen & Logan, 2008). The Go/No-Go discrimination and Stroop tasks activate different areas of the brain. Using functional magnetic resonance imaging (fMRI), it has been shown that in the Stroop task, color naming elicits more activity in the left dorsolateral prefrontal cortex, versus more activity in the ACC on incongruent trials (MacDonald, Cohen, & Stenger, 2000). Studies involving both the Stroop and Go/No-Go discrimination task, have demonstrated that these two tasks measure different subtypes of inhibition: response selection/interference control and response inhibition, respectively (Bender, Filmer, Garner, Naughtin, & Dux, 2016; Dimoska et al., 2011; Hung, Gaillard, Yarmak, & Arsalidou, 2018; Noreen & MacLeod, 2015). Again using fMRI, the Go/No-Go discrimination task has been linked to the activation in the right lateral prefrontal cortex in no-go trials (Simmonds, Pekar, & Mostofsky, 2008). Neurophysiological research on the Go/No-Go discrimination task has shown that performance on this task activates the ACC. Increased activity in the ACC is associated with responding to low-frequency stimuli, and processing of no-go stimuli is associated with a larger N2 ERP (Braver, Barch, Gray, Molfese, & Snyder, 2001). It has been suggested that processes associated with N2 are required to suppress
The anti-saccade task is another measure of inhibition ability, where a participant is presented with a fixation point, and is shown a stimulus on one side (i.e., left) of the fixation target. The objective is to make a saccade in the opposite side (i.e., right) of the stimulus (Hallett, 1978). Failure to make an anti-saccade is associated with the inability to override the reflex to make a saccade towards the target (Bialystok, Craik, & Ryan, 2006). The anti-saccade task is sensitive to lesions in the frontal lobe (Miyake et al., 2000b) as would be expected if it requires EF. Additionally, neurons in the frontal eye fields and superior colliculus are activated during the anti-saccade task, which has been described by Munoz and Everling (2004) as a competition between two processes: a process that is started when a stimulus is present that initiates the automatic pre-potent response, and a second process which is initiated by the inversion of the stimulus to initiate a voluntary anti-saccade. The process that initiates the automatic pro-saccade must be inhibited in order to allow the slower, and voluntary, anti-saccade response to reach the threshold required for responding. Errors on the anti-saccade task have been attributed to lack of inhibition in saccade neurons in the frontal eye fields and superior colliculus, leading to an inability to divert the eyes away from the stimulus as required by the anti-saccade task (Munoz & Everling, 2004). Performance on the anti-saccade has been proposed to measure response inhibition (Ettinger et al., 2018; Tiego et al., 2018).

The Flanker task is another measure of inhibition ability (Eriksen & Eriksen, 1974), where a stimulus is flanked on both sides by distractor stimuli. The objective is to identify the central stimulus. This task requires selective attention to focus on the central stimulus while ignoring the surrounding stimuli. In this task, participants are prone to direct their attention and vision (i.e., a pro-saccade) toward the distractor stimuli presented. In incompatible trials, participants are required to look in the opposite direction (i.e., make an anti-saccade). This anti-saccade results in slower RT and higher error rates, which suggests that participants still process the flanking stimuli (Kousaie & Phillips, 2012). Previous literature has demonstrated that incongruent trials activate the ACC more than congruent trials do, which is consistent with much other research that indicates that the ACC is needed to monitor response conflict (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Bunge, Hazeltine, Scanlon, Rosen, & Gabrieli, 2002; Casey et al., 2000; Durston et al., 2003; Hazeltine, Poldrack, & Gabrieli, 2000; van Veen, Cohen, Botvinick,
Activation of the ACC also occurs in non-motor tasks (e.g., when error feedback is shown in the Wisconsin Card Sorting task (WCST) (Monchi, Petrides, Petre, Worsley, & Dagher, 2001)). Previous research has found that the conflict effects found in the Stroop and Flanker tasks, respectively, do not correlate with each other (Fan, Flombaum, McCandliss, Thomas, & Posner, 2003), suggesting that the two tasks are assessing different kinds of inhibition. However, other studies suggest that both performance on the Stroop and Flanker tasks measure the ability to ignore distracting information and response interference (Rey-Mermet & Gade, 2017; Stahl et al., 2013).

2.2.2 Inhibition across the Lifespan

EF changes across the lifespan can be mapped to an inverted U-shape curve (Dempster, 1992), where EF improves from childhood to adolescence, and then declines with age (Mayr, Spieler, & Kliegl, 2001; McDowd & Shaw, 2000). The age-related decline in EF is postulated to be associated with changes in the prefrontal cortex (Salthouse, Atkinson, & Berish, 2003). In the remainder of this section I consider how aging affects inhibition ability.

Studies investigating inhibition across the lifespan include work by Comalli, Wapner and Werner (1962), and Cepeda, Kramer and Gonzalez de Sather (2001). Comalli et al. (1962) demonstrated a larger Stroop interference effect in children with a steady decrease in adolescents followed by a decline in older adults. Previous research has shown that older adults exhibit longer RT and more errors compared to younger adults on incongruent trials (Cohn, Dustman, & Bradford, 1984; Houx, Jolles, & Vreeling, 1993; Kieley & Hartley, 1997; Spieler, Balota, & Faust, 1996). Within the older adult cohort, institutionalized older adults demonstrate longer RT on incongruent trials compared to older adults living in a community (Comalli, P. E., Krus, D. M., & Wapner, 1965).

Livesey & Morgan (1991) carried out research examining the age differences in inhibition ability on the Go/No-Go discrimination task comparing children in two age groups, aged 4 and aged 5. They found that explicitly instructing children to inhibit their response to no-go stimuli, resulted in improvement in performance. Moreover, using fMRI, Casey et al. (1997) investigated age differences comparing the performance on the Go/No-Go discrimination task with young and old adults. They observed activation in the prefrontal cortex with significantly higher volume activation in young adults compared to older adults. Significant differences were also observed in the mean false alarm rate, whereby children erroneously hit no-go targets more than adults.
Altogether, these studies suggest that the Go/No-Go discrimination task highlights the lack of inhibition control in children, which develops over time, suggesting that performance on this task is age-related and can be used to assess inhibition ability over the lifespan.

Research results relating to age-related performance on the Flanker task have been mixed. For example, research using functional near-infrared spectroscopy (fNIR) has demonstrated no difference in performance between younger and older adults (Kawai, Kubo-Kawai, Kubo, Terazawa, & Masataka, 2012). However, Nieuwenhuis et al. (2002), demonstrated using the Flanker task that younger adults have higher error rates and faster RT than older adults. Other ageing research on the Flanker task has been studied using a similar task called the attention network task (ANT), which combines the Posner spatial cueing and Flanker task. In the ANT, younger adults have demonstrated decreased alerting attention (e.g., maintaining an alert state) compared to older adults (Jennings, Dagenbach, Engle, & Funke, 2007). In contrast, no age-related differences in other attention components of the ANT such as orienting have been observed between younger and older adults (Fernandez-Duque & Black, 2006; Jennings et al., 2007). In addition, age-related performance on executive attention in the ANT were found by Fernandez-Duque and Black (2006) but were not found by Jennings et al. (2007).

Age-related inhibition performance has also been studied using the anti-saccade task. Work by Butler, Zacks and Henderson (1999), and Olincy, Ross, Youngd and Freedman (1997) have demonstrated that older adults tend to make more errors on the anti-saccade task compared to younger adults. Research by Olincy et al. (1997) observed a linear trend in the latency rate of initiating an anti-saccade, which increased with age. Based on these findings, they suggested that the time to process and inhibit a response in the incorrect target direction is longer in older adults and declines with age, whereas eye movement accuracy is not age-dependent. The anti-saccade task is another task that supports the concept of age-related decline in inhibition ability.

2.2.3 Discussion (Advantages and Disadvantages)

Tasks designed to measure EF have been studied across the lifespan using both behavioural measures (e.g., RT and accuracy) and neuroimaging techniques such as fMRI, and fNIR, to help identify the cognitive processes responsible for EF activity. The methodologies and implementation of these tasks are often well documented, which may provide a good basis to create a serious game using the psychological properties of each task.
Although there are many benefits to using existing neuropsychological tasks to measure EF, there also exist disadvantages that should be considered. For instance, these tasks are often administered on a computer with assistance of a trained administrator to help with setup and instructions. This requirement may limit their use in clinical settings. There is a scarcity of studies that compare performance on EF related tasks with performance on clinical cognitive assessment such as the MMSE, MoCA, and CAM. There also exist different implementation of EF tasks, which are often not validated, making it difficult to compare results obtained between different studies that use different implementations of tasks.

2.3 Computer Cognitive Screening

This section will present a discussion of computerized tools for cognitive screening. Individual computerized cognitive screening tools are reviewed in Appendix A. This section begins with a brief overview of the different types of digital screening tools. Next, it presents the advantages and disadvantages of using computerized tools for cognitive screening.

2.3.1 Overview

The two primary approaches to address challenges associated with paper-and-pencil based assessments are (1) digitizing existing cognitive screening tools, and (2) creating new screening tools. In brief, computerized cognitive assessments have been designed to screen for general cognitive status (e.g., Automated Neuropsychological Assessment Metrics, CogScreen – Aeromedical Edition, Cambridge Neuropsychological Test Battery, CNS Vital Signs and NIH Toolbox, CogTest Online, CogState, and Cognitive Stability Index). Many of these computerized assessments aim to evaluate a core set of cognitive domains such as working memory, attention, visuospatial memory, cognitive flexibility, and language. Some of these assessments have also been used with adults having cognitive impairments despite them not being designed for use by this population. In addition, there are also computerized cognitive assessments designed to screen for specific conditions such as dementia (e.g., Alzheimer’s Disease Assessment Scale-Cognitive Subscale), and MCI (e.g., Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment, MCI Screen). A literature review on individual computerized cognitive assessments is presented in Appendix A.
2.4 Discussion (Advantages and Disadvantages)

The computerization of cognitive assessments offers many advantages over paper-and-pencil analogs. For example, computerized assessments can record ambient information such as the date and time, and precise information such as RT and accuracy (Wild, Howieson, Webbe, Seelye, & Kaye, 2008). They can also offer potential resource and cost savings in terms of reducing the need to be administered in conventional clinical settings with trained test administrators. In addition, the digital format of computerized assessments typically makes it easier to create randomized or alternate versions of a test (Gualtieri & Johnson, 2006).

Unlike paper-and-pencil tests, data generated from computerized assessments do not need to be transcribed into a digital format. This may facilitate easier data sharing between patients and healthcare professionals. Computerized assessments also have the potential to be self-administered or administered with minimal assistance from a care provider such as a family member, which enables patients to complete assessments in their own homes, outside of conventional clinical settings such as hospitals, which may be inaccessible or difficult to access. The ability to administer computerized assessments that require little to no assistance on readily-available technologies (such as tablets and computers) enables care providers to ask patients to complete self-administered assessments at any time. This ease of access may also encourage patients to repeat the assessments over time, which provides more information about the patient’s cognitive status.

Although there are many advantages to using computerized assessments, it is important to consider their disadvantages. The validity and reliability of digitized tests should be considered as well as their usability. Simply computerizing assessments may make tasks easier to administer but they may still be uninteresting for the patient. Some computerized assessments (e.g., NIH Toolbox) require a subscription fee, which may also deter widespread use.

2.5 Serious Games for Cognitive Screening

2.5.1 Introduction

An alternative to computerized assessments for cognitive screening is the use of serious games, which are games whose primary purpose is not entertainment (Charsky, 2010). This section will provide an overview of existing cognitive screening games, as well as (non-serious) computer
games that have been used for cognitive screening although not necessarily developed for this particular purpose.

2.5.2 Serious Games for Cognitive Screening

Some serious games for cognitive screening have been evaluated for their validity and test-retest reliability such as Space Code and Space Matrix (e.g., McPherson & Burns (2007), McPherson & Burns (2008)). In contrast, games for cognitive screening that have not been formally validated include the Child Neuropsychological Assessment Test (Tenorio Delgado, Arango Uribe, Aparicio Alonso, & Rosas Díaz, 2016), and BrainBaseline (Lee et al., 2012).

Space Code evaluates processing speed, while Space Matrix evaluates working memory (McPherson & Burns, 2007). The validity of both games were assessed by comparing game performance to a series of psychological tasks that measure working memory (e.g., picture swaps task, dot matrix task, and Raven’s advanced progressive matrices short form) and processing speed (e.g., digit-symbol task, visual matching task, and a decision speed task). Space Code exhibited the strongest interclass correlation with processing speed tasks: visual matching task (interclass correlation $= 0.48 – 0.57$), digit-symbol task (interclass correlation $= 0.40$), and visual matching task (interclass correlation $= 0.55$). Space Matrix had the strongest correlation with the working memory tasks: dot matrix task (interclass correlation $= 0.66$) and with the Raven’s short form task (interclass correlation $= 0.44$). In addition, the predictive validity of Space Code and Space Matrix were correlated against student school grades. Performance on Space Matrix was significantly correlated with Preparatory School grades, and but not with High School grades. In contrast, Space Code was not significantly correlated with any school grades (McPherson & Burns, 2007). Overall, Space Code and Space Matrix demonstrated a certain amount of concurrent validity against corresponding psychological tasks, while performance on Space Matrix showed predictive validity with school grades, and performance on Space Code showed validity against psychological tasks.

Research by Brown et al. (2014) used a crowdsourcing approach where 20,800 participants were able to complete four mobile games on working memory, attentional blink, selective stop-signal attention, and decision-making. In their study, participants performed at least one of four tasks to completion. The effect sizes on the crowd-sourced mobile-game versions of the working memory, attentional blink, and decision-making tasks were comparable to laboratory results. The
stop-signal task results were not comparable, as smaller effect sizes were observed in the crowd-sourced sample. They suggest that decreased performance and smaller effect sizes may have been due to noise in the user’s environment that would otherwise be controlled for in a laboratory setting. Overall, this research by Brown et al. (2014) demonstrates the ability to use crowd-sourcing as a method for administering cognitive tasks with comparable results (on some tasks) to those obtained in a laboratory setting.

Computerized cognitive assessments have also been designed for children, such as the Child Neuropsychological Assessment Test (Test de Evaluación Neuropsicológica Infantil, TENI, its Spanish acronym) (Tenorio Delgado et al., 2016). The TENI suite is administered on a touchscreen device and includes eight games that cover 10 subtests including attention (focused and sustained), rapid naming, hand-eye coordination, visuo-constructional skills, visual episodic memory, working memory, behavioural inhibition, seriating and theory of mind. A trained administrator provided instructions verbally to the child and the TENI was evaluated with 524 children. The reliability of the eight games ranged from 0.7 to 0.9 (the time between the test and re-test assessments was not indicated). The validity of the game was assessed by comparing correlations between the subtests and tests of cognitive function. Significant correlations were observed with tests for attention and visuospatial skills, as well as tests of EF. Research using TENI demonstrated that a game-like battery of cognitive assessments could be used with children between the ages of 3 to 9 years, if a trained administrator was present.

Similar to the crowdsourcing approach taken by Brown et al. (2014), Lee et al. (2012) released an iPad application called BrainBaseline (available between December 2010 to April 2011). BrainBaseline consisted of 13 subtests including: visual short-term memory, spatial working memory, n-back, simple RT, Go/No-Go RT, digit symbol substitution, Posner cueing, Flanker, Stroop, attentional blink, visual search, task switch, and trail-making task (TMT). They divided their sample into six age groups based on decades (e.g., 20s, 30s). Lee et al. (2012) carried out a principal components analysis (PCA), which resulted in two components, the first component was related to processing speed, and the second component corresponded to memory. Their findings suggested that task performance increased with age, with participants in their 20s having the best performance, with performance already declining in the 30s group to the over 60 group.
Overall, there are a mix of serious games for cognitive screening that have been evaluated with healthy children and adults. Some of these games have been evaluated for concurrent validity against measures of EF (e.g., McPherson and Burns (2007), and McPherson and Burns (2008)), whereas others have not (e.g., Tenorio Delgado et al. (2016), Brown et al. (2014), and Lee et al. (2012)).

2.5.2.1 Serious Games for Older Adults and Cognitively Impaired Adults

This section will describe serious games that have been designed for cognitive screening. Serious games designed for use by adults with cognitive impairments such as MCI and AD include the ‘Kitchen and Cooking’ game by Manera et al. (2015), and games designed using virtual reality (VR) (e.g., Tarnanas, Laskaris, et al. (2015), and Werner, Rabinowitz, Klinger, Korczyn and Josman (2009)). Card games have also been used a platform to measure cognitive status. For example, Kurata et al. (2012) created a card game called “Ryokansan” to screen for early stage dementia while the classic card game FreeCell have been modified and used to measure cognitive status in older adults by Jimison, Pavel, McKanna, and Pavel (2004).

A feasibility study by Manera et al. (2015) was carried out with a sample of 21 older adults (age 60 to 90) comprising nine adults with MCI, and 12 adults with AD. They compared performance on a tablet-based serious game, “Kitchen and Cooking”, with clinical and neuropsychological assessments. Their serious game focused on the use of EF, planning abilities, attention, object recognition, and praxis. Performance on the game was compared to the MMSE, instrumental activities of daily living, activities of daily living, TMT part A, Victoria Stroop Test, and Visual Association Test (Lindeboom, 2002). Performance on the game was significantly correlated with the MMSE ($r=-0.68, p=0.001$), TMT part A ($r=-0.59, p=0.006$), and Victoria Stroop Test (Word/Dot index, $r=-0.70, p=0.001$, and Interference/Dot index, $r=-0.55, p=0.013$). Manera et al. (2015) also noted that there was variability in the time spent playing the serious game outside of the meetings with clinicians, which they suggest highlights the importance of adapting to patients’ interests and level of impairment in designing serious games for training.

Another serious game in the pilot stage, called Smart Aging, was evaluated by (Zucchella et al., 2014) with healthy individuals ($n=50$). This 3D game used a computer and touchscreen monitor to simulate a virtual loft apartment, and was designed to identify MCI through a series of tasks that simulate daily activities. The gaming platform was designed to evaluate the following
cognitive domains: EF, attention, memory, and visuospatial orientation. Their preliminary results indicated that the task difficulty needed to be lowered, and that some older adults experienced difficulty with the technology. This research highlights the importance of piloting technology with target users to understand their needs and abilities.

In addition to common game platforms such as computers and touchscreen devices that have been explored in previous research for cognitive assessment, VR has also been used by healthy and cognitively impaired older adults. Research by Tarnanas et al. (2015) used a serious game to assess virtual action planning with VR. They evaluated their game with 25 patients with MCI and 25 age-matched controls. Their goal was to identify decline in spatial navigation, prospective memory, and EF. Participants were administered the Cognitive Difficulties Scale, TMT, WAIS, MMSE, Stroop Color-Word Test, WCST, Rey auditory verbal learning test (RAVLT), Geriatric Depression Scale, and an auditory oddball paradigm task and were asked to play the VR game. In the game, participants were provided with instructions on how to locate five artifacts in the VR museum. There were significant correlations between presence of MCI and performance on the RAVLT, and between total RT on the game and performance on the Cognitive Difficulties Scale in MCI participants. Further game results indicated that healthy older adults had fewer errors; better recall and shorter total recall time on the game compared to participants with MCI.

Another example of a VR game used by older adults for cognitive assessment is the virtual action planning supermarket game studied by Werner et al. (2009). This game was designed to diagnose MCI in older adults. Werner et al. (2009) evaluated the feasibility and ecological validity of the tool, by comparing performance on the VR game between 30 older adults with MCI, and 30 healthy older adults who were matched for sex and age. Participants were administered the MMSE and Behavioural Assessment of Dysexecutive Syndrome (BADS), and were asked to complete a short form questionnaire about their experience with the game. The BADs is a cognitive assessment designed to screen for changes in EF that may result in disturbances in skills and performance required in daily activities. Findings from the study revealed that performance on the BADs was correlated with four out of eight of the VR game measures for participants in the MCI group (with correlations ranging between $r = -0.63$ and $-0.58$). Results from the short form questionnaire indicated that both participants with MCI and no MCI found the VR game enjoyable. One potential drawback of this approach is that it requires special VR equipment.
Jimison, Pavel, McKanna, and Pavel (2004) adapted the game FreeCell to monitor the cognitive status of older adults. In their version of the game, hints were provided after 30 seconds of inactivity, which were based on the next suggested move provided by an automated solver. They compared how an adult played the game to “standard” performance based on an automated solver. Participants were also administered a cognitive battery that consisted of tests of verbal memory, language, attention, working memory and processing speed. Jimison, Pavel, McKanna and Pavel (2004) calculated a performance metric called game efficiency, which is the slope of performance based on the number of moves and game difficulty. Using this metric, they were able to use a threshold of 0.4 to discriminate between participants with MCI (lower efficiency) and no MCI (higher efficiency). The cut-off used to distinguish between participants with and without MCI was not indicated. The authors propose that monitoring cognitive status through game play is an unobtrusive method that can identify long-term trends in cognitive performance, and assist in early detection of cognitive decline.

Research by Kurata et al. (2012) created a serious game administered on a touchscreen device to diagnose early stage dementia. Their platform, “Ryokansan”, consisted of four games that required the user to flip over cards, find errors, re-arrange images, and identify a target. Participants were administered the MMSE, Revised Hasegawa Dementia Scale (HDS-R) and frontal assessment battery (FAB). Their results revealed that the MMSE score for AD patients was correlated with all games: card flipping ($r=0.45$, $p<0.01$), finding errors ($r=0.41$, $p<0.01$), picture arrangement ($r=0.59$, $p<0.01$), and accuracy rate for target identification ($r=0.76$, $p<0.01$). Overall, the researchers suggest that performance on their serious game can be useful for assessing early signs of dementia.

As shown in the above review of serious games for cognitively impaired adults, researchers have assessed concurrent validity of serious games by comparing game performance to measures of EF (e.g., Stroop task) and clinical assessments such as the MMSE (e.g., Manera et al. (2015), and Kurata et al. (2012)). Some games have not yet been validated but have been assessed in terms of their feasibility and usability, which are important prerequisites to the use of games that employ new technology such as VR (e.g., Werner et al. (2009)), but to also make them usable (and preferably enjoyable) for the target population.
2.5.3 Discussion (Advantages and Disadvantages)

Serious games for cognitive screening have the potential to be engaging, which can promote repeated use and prolonged usage. Collecting repeated data over time can provide a longitudinal view of a person’s cognitive status, which is difficult and costly to achieve in a clinical setting. Most clinical staff are not trained in administration of psychometric tools such as the MMSE, and their clinical care responsibilities may preclude frequent cognitive assessments.

Game design and gamification techniques can be used to make serious games more enjoyable and promote their use. Another key feature is that serious games can be randomized, making them play differently each time, so that required responses are not predictable. This avoids the learning and practice effects that occur when a person uses the same cognitive assessment (e.g., a paper-and-pencil test) more than once. Serious games can also be expanded for social interaction, although changes to make a game more motivating or interesting would have to be assessed in terms of any possible impact they may have on the validity of the game as a cognitive assessment tool.

Game-based cognitive screening tools can also leverage technology such as computers, smart devices, and gaming consoles. This variety of platforms enables new interaction methods for users to assess their cognition and can incorporate physical operations, which requires cognitive control of the motor system. Portable platforms such as laptops and smart devices enable game-based cognitive screening to be carried out in a non-clinical setting such as a user's home, which is particularly beneficial when users do not live in close proximity to medical clinics or hospitals. Moreover, the use of technology enables automatic data recording of ambient data such as location and time as well as numerical measures such as RT. Once game performance data, and associated location and time data is collected, it can be published on a secure online site and automatically shared with care providers.

Serious games can be designed for self-administration that is carried out independently, or with minimal assistance. In the case of people living independently, cognitive screening based on serious games (versus standard assessments) can be a more enjoyable, and ongoing activity, where the users/patients are in control of how often they are screened.
Despite the many benefits of serious games, it is important to consider potential disadvantages. For example, some individuals may find games too childish or they may not wish to play through a serious game to assess their cognitive status or may be uncomfortable using technology. For some, cognitive screening may only have face validity, or be acceptable, if a trained person in a clinical context carries it out. There may also be barriers associated with technology use such as usability issues, which would prevent the game-based assessment from being played. It is particularly important that game performance (and resulting interpretations of cognitive status) is not underestimated due to usability or accessibility issues. It also may not be possible to provide some users with the devices required to run a game-based cognitive screening system. Serious games for cognitive assessment will likely work best when they are designed to be interesting and usable by a well-defined subgroup of people (e.g., people over the age of 65 who are in reasonably good physical shape but with some degree of cognitive impairment) and when they are implemented on technologies that are broadly available to the target subgroup, not only in clinical settings, but preferably in home settings as well.

A key concern with serious games is the tendency for them to have insufficient data concerning their validity and reliability as a cognitive screening tool. Often inappropriate methods are used to demonstrate validity, along with relatively small or unrepresentative samples. For instance, researchers may have participants with, and without a dementia diagnosis play a serious game. Claims of game validity based on a statistically significant difference in performance between the two groups is problematic, however, since one might just as easily find a statistically significant difference on measures such as height and weight since there may be a tendency for people with dementia to be more physically frail. Thus, serious games should be validated in terms of the quality of predictions that they make at the individual, not group, level. Distinguishing between groups of people, with and without dementia is a relatively easy task that could potentially be done efficiently with many measures that would not require the development of a serious game.

2.6 Summary

In clinical settings, cognitive status is often measured using standard tests such as the MMSE, MoCA and/or CAM, all of which require trained test administrators. Computerized assessments and game-based tools also exist as methods for cognitive screening but are not widely used in
clinical contexts as of this writing. Current computerized cognitive screening and game-based assessments generally have insufficient validation and reliability assessment.

While many serious games and computerized assessments have been developed for cognitive screening, none of them have become standards of care in the way that the MMSE has become a standard for clinical assessment of cognitive status. In my view, there are two main requirements that need to be addressed in creating a serious game that can be successfully used in clinical assessment. The first issue is to develop a game that fits the clinical context and can be used successfully within that context. In practice, this means iteratively designing the game with clinicians, repeatedly testing the game in a clinical context, and refining it in that context until it becomes usable in that setting. This requires a multi-year commitment to evaluating and refining the game in the clinical context. The second requirement is to validate the game in the clinical context not just once but multiple times, so that reliable measures of effect size and of predictive performance, such as specificity and sensitivity, can be demonstrated across multiple studies.

Motivated by the trends and observations noted in the preceding literature review, the research reported below provides an example of how to validate a serious game for cognitive screening both in a clinical context and with a healthy sample. The approach that I use combines the advantages of computerized assessments (i.e., ability to record RT and accuracy with precise measurement, and providing alternate versions of assessments) while also addressing current shortfalls in serious games research (i.e., insufficient validation and reliability, and insufficiently large and clinically relevant samples). In the following two chapters (Chapters 4 and 5), I report on clinical studies that evaluated the whack-a-mole game both as a cognitive assessment tool and as a delirium screening tool. In addition to these clinical evaluations, I also sought to improve the psychometric foundations of the serious game, by demonstrating the subtype of inhibition ability that it is most closely related to (discussed in Chapter 6). My goal in this later work was to avoid an approach that has frequently been used in past research, where cognitive properties assessed by a serious game for cognitive assessment is identified through design intention rather than through construct validation in a formal experiment.
Chapter 3

3 Methodology

3.1 Description of the game

I developed a serious game to assess cognitive status in elderly adults with a focus on detecting small changes in cognition for conditions such as delirium. The serious game mimics features of the classic psychological Go/No-Go discrimination task (Yechiam et al., 2006), a measure of inhibition ability. As implemented, the game is similar to the game whack-a-mole (Figure 3-1).

![Figure 3-1. Screenshot of the whack-a-mole game.](image)

3.2 Design and Development Process

3.2.1 Design

The original version of the serious game was designed during my Master's of Applied Science thesis described in Tong (2014), and evaluated in a usability study with healthy adults (Tong & Chignell, 2014). Since that implementation, the game has gone through an iterative design process (with evaluation and re-design at each iteration) in a clinical setting. Each game design iteration was informed through user requirements gathering by engaging with stakeholders including an emergency medicine physician, a clinical psychologist, and clinical research
assistants (RAs) at Sunnybrook Research Institute located in Toronto, Ontario, Canada. These stakeholders have experience working with elderly adults with cognitive impairments.

The concept of the game was motivated by an interest in detecting the risk of delirium in an ED using an interactive method that could easily be repeated. Based on previous literature, performance on the CAM has been compared to RT performance on a choice reaction time (CRT) task (David Peter Lowery, Wesnes, Brewster, & Ballard, 2008). Through a joint collaboration between the University of Toronto and Sunnybrook Research Institute, I developed a serious game based on a CRT task, and Go/No-Go discrimination task. Compared to existing serious games for cognitive screening discussed in the literature review (Chapter 2 – Section 2.5), this game was designed and continually reiterated upon using a user-centered approach.

3.2.2 Development

The game was developed in Java using the Android software development kit, and is playable on Android-based smart devices such as phones and tablets.

3.3 Measurements

The initial serious game measured RT and target offset. RT was measured as the time between the appearance of a character on the game board and the user's touch. Target offset was measured as the pixel (px) deviation between the center of the character to the user's touch point. In later versions of the game, error rates were also captured with false alarm and miss rates being calculated across blocks of trials. Over the course of this dissertation a number of different versions of the game were developed based on user requirements.

3.4 User Requirements

I conducted regular meetings to assess user requirements and sought feedback from a range of healthcare professionals that have experience working with older adults with cognitive impairments, including emergency physicians, clinical psychologists, RAs, nurses, and occupational therapists (OTs). Meetings were held with emergency physicians clinical psychologists, and RAs generally once every two months. Regular, weekly meetings were held with RAs. Feedback from older adults on the game play was recorded by RAs and OTs. Family members and care providers of the older adults also provided feedback on the experience to RAs.
I then iteratively incorporated feedback into the design of the game. Both RAs and OTs provided feasibility (i.e., duration of the task, technology to use) and usability notes (i.e., text size, character design) on their experience administrating and assisting older adults play the game.

### 3.5 Game Architecture and Design

- Once signed into the application, the user is shown a menu of game levels. The purpose of signing into the application provides a method to identify each participant/user. This identifier is then linked to the participant’s game play data.
- Each game level varies based on the size of the game board, number of characters that appear on the game board at once, the types of characters, size of the characters, and the length of the game. A description of these variables are in the following section on level design.
- After selecting a game level to play, a set of instructions for that game level is shown to the user. Once the user has finished reading these instructions, they dismiss this dialog to proceed to the game.
- The game begins with a 3000 msec count down timer. On the game board, the user's score, and time remaining in the game are displayed. The score increments by 1 point for each target character hit, and decrements by 1 point for each non-target character hit. The game ends once the timer reaches 0 msec.
- At the end of the game, the user is given the option to replay the current level or to return to the menu.
- To exit the application, the user can select the "sign off" icon on the top of the menu.
- A process diagram visualizing this process can be seen in Figure 3-2.
3.5.1 Level Design

Each level of the serious game can vary based on the size of the game board, number of characters that appear on the game board at once, the types of characters, size of the characters, and the length of the game. The ability to vary these parameters enables the game to be adjusted to increase the perceived difficulty of the task, and based on the needs of the user.

3.5.2 Character Type

Characters can either be targets (e.g., go-stimuli), which are meant to be hit, or non-targets (e.g., no-go stimuli), which a user is meant to inhibit their response to. In this serious game, targets could either be moles or raccoons, and non-targets could were butterflies and moles with hats. The selection of these characters were a result of user requirements meetings. Specific feedback included selection of colours that were easier for older adults to perceive, as well as accounting for potential red-green colour blindness, which occurs in 5-8% of the males.
3.5.3 Character Appearance

The number of characters that can appear simultaneously on the game board varied from one to three. This function was added as a means to increase the difficulty level of the game, specifically with respect to executive function. Easier levels would have only one character on the game board at a time compared to multiple characters on harder levels, further testing executive function.

3.5.4 Character Size

The size of each character, measured in pixels, varied between 175 px, 200 px and 250 px.

3.5.5 Game Length

The length of each game varied, measured in seconds. The duration that a character appeared for on the game board also varied. For example, in the game versions used in the three clinical studies described in Chapters 4 and 5, the character appearance duration ranged from 2000 msec to 2005 msec. In the game variants used in Chapter 5, the maximum game length varied from 2 minutes to 2.5 minutes. The impact of character appearance duration was investigated in further detail in the experimental study described in Chapter 6. The inter stimulus interval was set to 1000 msec.

3.5.6 Data Storage

Data was stored and recorded in a structured query language database inside of the Android application. Information on the date and time a stimulus appeared, and when the stimulus was hit was recorded. The location, in terms of the x and y-coordinate, of where the character appeared on tablet screen was logged. The x and y-coordinate of where the user touched the tablet and the x and y-coordinate where the user released their finger from the tablet was recorded. Other data (permitting on the device) such as the hit pressure and size was measured. The distance between the center of a stimulus to the user’s hit point, measured in pixels, was calculated as the user’s target offset for each hit. The time difference between the appearance of a stimulus and when a user hit the target, measured in seconds or milliseconds, was calculated. Note, the tablet contained no personal health information to allay concerns over privacy breach. This was a fundamentally important user requirement.
3.6 Game Hardware

The serious game was played on two possible 10.1-inch tablets manufactured by Samsung (Samsung, 2006). The tablets models were the Samsung Galaxy Tab 4, Samsung Galaxy Tab A tablets. The Samsung Galaxy Tab 4 tablets measured 243.4mm (width) x 176.4 mm (height) x 7.9mm (depth) (Samsung, 2016). The Samsung Galaxy Tab A tablets measured 100 mm (height) x 61.1 mm (width) x 32 mm (depth) (Samsung, 2018).

3.7 Clinical Study Methodology

In the three clinical studies carried out and described in Chapters 4, and 5, the following general methodology was performed. All participants were recruited from Sunnybrook Health Sciences Centre ED located in Toronto, Ontario, Canada. Participants were first administered a battery of cognitive assessments by a trained clinical RA (Table 3-1). Measures of global cognitive status included the MMSE and MoCA. Measures of delirium and delirium severity were measured using the CAM and DI, and behavior was measured using the Richmond Agitation-Sedation Scale (RASS) (Sessler et al., 2002). Attentional performance was measured using the digit vigilance test (DVT) (Kelland Manchester, Hampshire, Lewis, & Kelland, 1996; Lewis & Rennick, 1979), and CRT task. A CRT task was administered in the clinical validation study described in Chapter 4 to investigate the relationship between RT from a CRT task and measures of cognitive status. Previous research by Lowery, Wesnes, Brewster and Ballard (2008), compared RT performance on a CRT task with the CAM in patients at risk of developing delirium. They observed that patients with delirium (as assessed using the CAM) had longer RT, increased variability of RT, and lower accuracy on the CRT compared to non-delirious patients.

The serious game was played on a 10-inch Samsung Galaxy Tab 4 10.1 tablet (Samsung, 2016). Participants received instructions on how to play the game and interact with the tablet. There was no limit on the number of attempts to play the game. Participants were invited to provide open feedback at the end of the study. At the end of each session, the RA informally interviewed the participant on his/her experience with the game. In addition, RAs provided their own feedback and comments on their experience with the game and their observations of the interaction between each participant and the game. The game play took 8 minutes to complete, and total participant time for all psychometric scales was less than 1 hour. The RAs recorded the date of the ED visit, whether study participation was refused, and the cognitive assessment scores.
Usage notes were also recorded and later used to infer usability problems as well as evidence for enjoyment and engagement.

**Table 3-1. Summary of cognitive assessments administered in the three clinical studies.**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Study</th>
<th>Assessments Administered</th>
</tr>
</thead>
</table>
| 4       | Clinical Validation Study #1 | • Mini-Mental State Examination (MMSE)  
• Montreal Cognitive Assessment (MoCA)  
• Confusion Assessment Method (CAM)  
• Delirium Index (DI)  
• Richmond Agitation-Sedation Scale (RASS)  
• Digit Vigilance Test (DVT)  
• Choice Reaction Time (CRT) |
| 4       | Clinical Reliability Study   | • Mini-Mental State Examination (MMSE)  
• Confusion Assessment Method (CAM)  
• Delirium Index (DI) |
| 5       | Clinical Validation Study #2 | • Mini-Mental State Examination (MMSE)  
• Confusion Assessment Method (CAM)  
• Delirium Index (DI)  
• Richmond Agitation-Sedation Scale (RASS) |

### 3.8 Experimental Study Methodology

The methodology and study design for the experimental study carried out with healthy participants is described in detail in Chapter 6 – Section 6.4.

### 3.9 Game Variants

In the version of the game evaluated in the first clinical validation study and clinical reliability study described in Chapter 4, game performance was measured using median correct RT. The software was designed to collect signal detection measures such as hits (i.e., correctly responding to a go-stimulus such as a mole or raccoon) and false alarms (e.g., incorrectly hitting a no-go stimulus such as a mole wearing a hit or a butterfly) (Stanislaw & Todorov, 1999). Through an iterative, user-centered design process, the game was revised based on user feedback from clinicians, clinical RAs, and patients to increase the usability of the game. As a result of this process, the game was continually reiterated upon during the data collection process in the first clinical validation study and clinical reliability study. This resulted in 14 game versions, which varied based on character size, and game board size (Appendix C – Table 8-3).

Following the two clinical studies carried out in Chapter 4, the game was redesigned to record data on signal detection values such as misses and correct rejections (Stanislaw & Todorov, 1999). A miss occurs when a user does not respond to a go stimulus (e.g., mole, raccoon), and a
correct rejection occurs when a user correctly inhibits their response to a no-go stimulus (e.g., mole wearing a hat, and butterfly). The collection of these two additional measures was then used to calculate false alarm rate, and a combined performance metric described below. This revised version of the game was evaluated in the second clinical validation study and experimental study described in Chapters 5 and 6, respectively. In the second clinical validation study, there were three variants of the game (Appendix E – Table 8-15). A summary of the different game performance measures per study is found in Table 3-2.

Table 3-2. Summary of game measures collected based on version of game evaluated in the three clinical studies.

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Study</th>
<th>Game Performance Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Clinical Validation Study #1</td>
<td>• Median correct RT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Signal detection values: hits and false alarms</td>
</tr>
<tr>
<td>4</td>
<td>Clinical Reliability Study</td>
<td>• Median correct RT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Signal detection values: hits and false alarms</td>
</tr>
<tr>
<td>5</td>
<td>Clinical Validation Study #2</td>
<td>• Median correct RT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Signal detection values: hits, false alarms, misses and correct rejections</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• False alarm rate (Equation 3-1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Combined performance metric (Equation 3-2)</td>
</tr>
<tr>
<td>6</td>
<td>Experimental Study</td>
<td>• Median correct RT</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Signal detection values: hits, false alarms, misses and correct rejections</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• False alarm rate (Equation 3-1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Combined performance metric (Equation 3-2)</td>
</tr>
</tbody>
</table>

A combined performance metric for the game was created to combine the median correct RT and false alarm rate (Equation 3-1) for each patient using a z-score transformation (Equation 3-2). In my previous research, I used a combined performance metric that combines both RT and accuracy into a single measure, described by Chignell et al. (2015). This measure transforms RT and accuracy using z-scores, which results in a distribution-free metric. This measure can cope with a speed-accuracy trade-off, where longer RT is expected to lead to higher accuracy, and faster RT is expected to lead to lower accuracy. In a trade-off condition, a linear relationship is expected between speed RT and errors, resulting in a positive slope and a correlation approach 1 for the resulting line (or a slope approaching -1 if accuracy was used instead of error as the measure).

$$false\ alarm\ rate = \frac{total\ false\ alarms}{total\ false\ alarms + total\ correct\ rejections}$$
Equation 3-1. False alarm rate calculation.

\[ \text{combined performance metric} = -z(\text{median correct RT}) - z(\text{false alarms rate}) \]

Equation 3-2. Combined performance metric.
Chapter 4

4 Assessing Cognitive Status in Elderly Adults with Incident Delirium

This chapter describes the concurrent validation and evaluation of the test-retest reliability of the serious game in a sample of elderly adults selected from a hospital emergency department (ED). The first part of the chapter describes the validation of the game, and the second part of the chapter describes the evaluation of the test-retest reliability of the serious game. The chapter begins with the motivation for the study, followed by the study design, results, and a discussion of the findings.

4.1 Concurrent Validation Study

4.1.1 Motivation

Delirium is a condition that is characterized by acute onset and fluctuating attention and cognitive status (Inouye et al., 1990). The changing nature of delirium makes it difficult to recognize and diagnose in busy clinical environment such as a hospital ED. In a hospital ED, the prevalence of delirium has been estimated to range between 10 to 17% (Elie et al., 2000; Hustey & Meldon, 2002; Hustey, Meldon, & Palmer, 2000; Kakuma et al., 2003; Lewis, Miller, Morley, Nork, & Lasater, 1995). Delirium is also an independent risk factor for mortality - patients with delirium have twice the risk of death (Cole & McCusker, 2002). In terms of the economic burden of this condition, delirium is estimated to cost $6.9 billion in the US (Francis & Kapoor, 1992; Inouye, 2006) and care costs are doubled for patients with delirium. Loss of independence and admission to a nursing home are also doubled after an episode of delirium. In addition, patients with delirium can remember their experiences and have reported post-traumatic stress disorder. The problem of delirium will only to continue to grow and become present with our rapidly ageing population. This highlights the importance of identifying signs of delirium and implementing appropriate interventions.

The ED is a high-risk environment for developing incident delirium, which is delirium that develops while a patient is waiting to be admitted to the hospital. In contrast, prevalent delirium is delirium that is present in patients on admission to the hospital (Anderson, 2005). Émond et al.
(2017) recently demonstrated that an additional 7-11% of patients develop incident delirium in the ED. Using data from 2016-2017 on over 11.2 million visits to the ED, the Canadian Institute for Health Information (CIHI) reported that ED visits have increased 11% from the previous year. In this sample, 90% of patients were admitted and reported leaving the ED in 7.8 hours. The remaining 10% of patients waited longer than 7.8 hours in the ED. For adults over the age of 65 years, CIHI indicated that patients waited even longer, and 90% of older adults completed their ED visit within 36.3 hours (Canadian Institute for Health Information, 2017). Based on the longer wait times in the ED experienced by older adults and the risk of incident delirium, the recognition of delirium is an imperative as unrecognized delirium can triple a patient’s mortality risk (Kakuma et al., 2003). The mortality risk is six-times greater in patients with unrecognized delirium that are discharged home compared to non-delirious patients. The recognition of delirium in the ED is crucial to providing the appropriate medical care and assistance to patients (Inouye & Charpentier, 1996). The ability of a screening tool such as a serious game to identify early signs of delirium could alert ED staff to initiate the proper prevention strategies.

In the ED, delirium is only recognized in approximately in only 17-25% of cases. One of the clinical features of delirium is that it fluctuates over time, which may contribute to the low recognition rate in the ED. Delirious patients may have a transient confusion initially and thus serial assessments are required. This can be particularly challenging in the ED environment, as patient may have many transitions in care during the initial. In clinical contexts, the CAM is used as the screening standard for delirium. It requires a trained administrator, and is a subjective dichotomous measure of delirium. In terms of feasibility and hospital ED resources, it may not be possible to administer the CAM to each and every patient that enters the ED, let alone repeat the CAM multiple times with the same patient to detect fluctuating changes in their cognition during their wait in the ED, which frequently last over 12 hours (Canadian Institute for Health Information, 2017). The potential to use a serious game as a quick screening tool that automatically detects early signs of delirium has the potential to reduce the mortality risk associated with unrecognized delirium, and loss of independence in older adults (Inouye et al., 1999). The goal is not to replace the CAM but instead to complement the CAM. The proposal is to have patients play the game while they wait in the ED, which can detect changes in their performance, which could subsequently alert ED staff attention to risk of delirium, or approaching delirium, in the patient. Thus, the motivation to validate a serious game in a clinical
setting with elderly adults with incident delirium is to provide healthcare professionals with a non-subjective measure of cognitive status that is not labour-intensive to use. This method requires minimal assistance to administer, can easily be repeated, and provides an engaging and fun method of cognitive screening. A validated serious game-based cognitive tool can provide ED staff with invaluable information regarding a patient’s cognitive status over the course of their stay in the ED while waiting to be admitted to the hospital or discharged. The serious game automatically records ambient data such as the time, and game scores, which can be used to help ED health professionals track changes in a patient’s game performance, which may be related to changes in their cognitive health that may be suggestive of a fluctuating course, a hallmark of delirium. Decline in game performance would then automatically trigger intervention techniques aimed at preventing delirium. This serious game can also offer an enjoyable medium for patients to pass the time while being cognitively active as they are waiting to be admitted to the hospital.

4.1.2 Research Question

The research question that this section will address is as follows:

1. How does performance on the serious game played by older adults in a clinical setting agree (in terms of correlation measures) with existing methods of cognitive assessment in a clinical setting (i.e., MMSE, MoCA, CAM, and DI)?
2. Can performance on the serious game be used as a screening tool for the onset of delirium?

The first research question was explored by examining the concurrent validity of the serious game using a correlation analysis (using both Pearson’s $r$ and Spearman’s $\rho$) between standard measures of cognitive status (e.g., MMSE, MoCA, CAM, and DI) with game performance (e.g., median correct RT). Game median correct RT for each participant was calculated by looking only at their correct trials and calculating the median value across all trials (regardless of game level).

The second research question was assessed by exploring the relationship between game performance and cognitive status. Specifically, we assessing if there was a cut-off value that could distinguish between those who were at risk for delirium and those who were not.
4.1.3 Methodology

Working with Dr. Jacques Lee and his team of clinical RAs in the ED, I conducted a prospective observational clinical study with participants recruited from the Sunnybrook Health Sciences Centre ED located in Toronto, Ontario, Canada under a research protocol approved by the Sunnybrook Health Sciences Centre (protocol #070-2013). A research ethics cover sheet along with the protocol from Sunnybrook Health Sciences Centre (protocol #070-2013) was also approved by the research ethics board at the University of Toronto (protocol #28953). Participants who were 70 years or older and who were present in the Sunnybrook Health Sciences Centre ED for a minimum of four hours were recruited for the study. The inclusion criteria were chosen to increase the range of participants at risk of developing delirium. Exclusion criteria included participants who were (1) critically ill (defined by the Canadian Triage Acuity Scale score of 1), (2) in acute pain (measured using the Numeric Rating Scale with a score greater than or equal to two out of 10), (3) receiving psychoactive medications, (4) judged to have a psychiatric primary presenting complaint, (5) previously enrolled, (6) blind, or (7) unable to speak English, follow commands, or communicate verbally. These exclusion criteria were chosen for practical and ethical reasons.

Clinical RAs administered standard cognitive assessments including the MMSE, CAM, DI, RASS, DVT, and a CRT task. Each participant was then asked to play the serious game and provide feedback.

4.1.4 Requirements Analysis

I conducted requirements analysis with a physician and a team of RAs at the participating institution. The members on this team included physicians, and RAs who worked in emergency services with elderly adults. Insight on the needs of patients and caregivers were provided by these informants, based on their previous experience, and on their experience with the tablet technology in this study.

4.1.5 Statistical Analysis

The cognitive data and serious game results were non-normally distributed based on visual inspection of the data. Tests for normality, including the Kolmogorov-Smirnov and Shapiro-Wilk tests, were not used due to the large sample size in this study because they are known to
result in oversensitivity to relatively small departures from normality (Field, 2013). Transformations of the data were not performed because some of the measures, such as the CAM and DI, are binary/categorical and cannot follow a normal distribution. My interest was in correlations as a measure of the effect size of the underlying relationship between game performance and the cognitive assessments, but I used nonparametric correlation measures for some of the comparisons that involved categorical or narrow ordinal scales. Correlations between the dichotomous CAM and the other measures were assessed using point-biserial correlations (Field, 2013). Correlations involving the DI and RASS (and not involving the CAM) were assessed using Spearman’s rho as the DI and RASS use a small number of ordered categories. The remaining comparisons used Pearson’s r correlations.

The serious game software was designed to collect data whenever a patient touched the tablet during specific times and within specific areas of the screen. Game performance was measured based on a participant’s median correct RT. The target offset on correct trials were also calculated, which can be found in Appendix C, but are not discussed in the results below.

4.1.6 Results

4.1.6.1 Description of Sample

147 participants (80 males and 67 females) were recruited between the ages of 70 and 94 years (mean age = 80.61 years, standard deviation = 6.08). One participant was excluded for not completing any of the cognitive assessments and five people did not play the serious game (of whom two were CAM-positive), leaving 141 participants who completed the study (age range = 70-94 years, mean age = 80.64 years, standard deviation = 6.09).

Some participants declined to complete some of the cognitive assessments entirely or declined to answer certain questions. Completion of an assessment was at the discretion of the participant. The completion rate of each test is shown in Table 4-1. All participants completed the CAM, DI, and RASS. The serious game had a combined completion rate of 96.6% (141/146), whereas the completion rates for the other assessments were lower with DVT having the worst completion rate at 36.3% (37/102).

<table>
<thead>
<tr>
<th>Cognitive assessment</th>
<th>Completion rate, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4-1. Summary of completion rates for standard cognitive assessment scores.
Mini-Mental State Examination (MMSE) 145/146 (99.3)
Montreal Cognitive Assessment (MoCA) 108/146 (73.9)
Confusion Assessment Method (CAM) 146/146 (100.0)
Delirium Index (DI) 146/146 (100.0)
Richmond Agitation-Sedation Scale (RASS) 146/146 (100.0)
Digit Vigilance Test (DVT) a 37/102 (36.3)
Choice Reaction Task (CRT) a 82/99 (83.0)
Serious Game 141/146 (96.6)

a This assessment was introduced later in the study.

Because the DVT and CRT assessments were initiated partway through the study, the denominators in calculating completion rates for those measures (102 and 99, respectively) were lower than for the other tests (which were initiated at the start of the study).

There were a number of people in the sample with low MMSE and MoCA scores (down to 9 and 8, respectively). There were 129 participants who were negative for the CAM and 12 participants who were positive (a positive result on the CAM suggests that the participant has delirium). Moreover, the DI scores ranged from 0 to 10 (the score indicates the severity of delirium), RASS scores ranged from –2 to 1 (a score >0 suggests that the patient is agitated and a score <0 suggests that the patient is sedated), DVT scores ranged from 81 to 103, and CRT accuracy ranged from 34% to 95%. A distribution of the DI scores is shown against the CAM scores in Figure 4-1, which illustrates that CAM positive patients generally had higher DI scores compared to CAM negative patients who had lower DI scores. The combined median RT on the CRT was 1200 msec (IQR 400). The overall median correct RT on the serious game was 900 msec (IQR 300). A summary of the scores on the cognitive assessments can be found in Appendix C.
4.1.6.2 Comparison Between Serious Game Performance and Standard Cognitive Assessments

A correlation analysis revealed significant relationships between game median correct RT and scores on the six cognitive assessments: MMSE, MoCA, CAM, DI, RASS, and CRT (Table 4-2). Note that information about which types of correlation were used for each comparison is shown in the footnotes to Table 4-2.

As a follow-up to the correlation analyses in Table 4-2, I carried out the same analysis using Spearman’s rho correlations instead of Pearson’s r correlations. All significant correlations between the cognitive assessments and game correct RT were also observed to be significant using Spearman’s rho.

Table 4-2. Correlations comparing game performance to the standard cognitive assessments. * p<0.05, ** p<0.01, *** p<0.001. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.
<table>
<thead>
<tr>
<th>Game Median Correct RT</th>
<th>MMSE</th>
<th>MoCA</th>
<th>CAM</th>
<th>DI</th>
<th>RASS</th>
<th>DVT</th>
<th>CRT RT</th>
<th>CRT Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Median Correct RT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMSE</td>
<td>1</td>
<td>-0.617***</td>
<td>-0.357***</td>
<td>0.421***</td>
<td>-0.336***</td>
<td>0.303***</td>
<td>-0.0938</td>
<td>0.588***</td>
</tr>
<tr>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p=0.587</td>
<td>p&lt;0.001</td>
<td>p=0.0193</td>
<td></td>
</tr>
<tr>
<td>MoCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.630***</td>
<td>-0.693***</td>
<td>-0.689***</td>
<td>0.339***</td>
<td>0.200***</td>
<td>-0.503***</td>
<td>0.307**</td>
<td></td>
</tr>
<tr>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p=0.24</td>
<td>p&lt;0.001</td>
<td>p=0.005</td>
<td></td>
</tr>
<tr>
<td>CAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.515***</td>
<td>-0.644***</td>
<td>-0.434***</td>
<td>-0.237*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p&lt;br&gt;0.001</td>
<td>p=0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.418***</td>
<td>-0.037</td>
<td>0.272**</td>
<td>-0.160</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p&lt;0.001</td>
<td>p&lt;0.001</td>
<td>p=0.02</td>
<td>p&lt;0.001</td>
<td>p=0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RASS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>-0.124</td>
<td>0.129</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p=0.17</td>
<td>p=0.16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DVT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.045</td>
<td>-0.237</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p=0.80</td>
<td>p=0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRT RT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRT Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Correlations involving the CAM were calculated using point-biserial correlations. Correlations involving the DI and RASS (and not involving the CAM) were assessed using Spearman’s rho. All other correlations were calculated using Pearson’s r.

bCannot be computed because at least one of the variables is constant. There were no CAM positive patients that completed the DVT.

In order to examine the separate contributions of speed of processing and EF on cognitive assessment scores, I looked at the partial correlations of serious game and CRT performance (controlling for each other) with the clinical assessments (Table 4-3). The partial correlations with game median correct RT (controlling for CRT) remained significant for the MMSE, CAM, and RASS, but not for the MoCA, and DI. On the other hand, the partial correlations involving CRT, but controlling for serious game performance median correct RT were non-significant (Table 4-3), suggests that the relationship between CRT and the clinical measures can be fully accounted for by shared variance with game performance, whereas a significant part of the correlation of game performance with the clinical assessments is unique variance that is separate from the variance that is shared with the CRT.

**Table 4-3. Partial correlations that control for CRT reaction time on game performance and standard cognitive assessments and control for game median correct RT on standard**
cognitive assessments. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.

| Assessment | Control for CRT RT | | Control for Median Game Correct RT | | |
| --- | --- | --- | --- | --- | --- | --- |
| | Game Median Correct RT | | | CRT RT | | |
| | $\rho$ | $p$ | $\rho$ | $p$ | $\rho$ | $p$ |
| MMSE | -0.245 * | 0.031 | -0.126 | 0.273 | 0.157 | 0.169 |
| MoCA | -0.079 | 0.524 | -0.180 | 0.142 | 0.103 | 0.403 |
| CAM | 0.288 * | 0.010 | 0.011 | 0.920 | -0.027 | 0.814 |
| DI | 0.190 | 0.095 | 0.171 | 0.136 | -0.136 | 0.235 |
| RASS | -0.317 ** | 0.004 | 0.126 | 0.267 | 0.036 | 0.755 |
| DVT | -0.143 | 0.436 | 0.085 | 0.643 | -0.288 | 0.109 |

4.1.7 Detection of Abnormal State Using Serious Game Performance

A Mann-Whitney $U$ test (Table 4-4) was performed to investigate the difference between cognitive ability and serious game performance when the MMSE score was $\geq 24$ (normal cognitive function or possible MCI) versus when that score was below 24 (signs of dementia) (Folstein et al., 1975; O’Connor et al., 1989). The MMSE was chosen as the grouping criterion because it was a standard in screening for dementia at the time this research was carried out. The test results suggest that there was a significant difference on the CRT in terms of RT between participants with dementia (MMSE <24) and no dementia (MMSE $\geq$24) (O’Connor et al., 1989). In addition, there was a significant difference between MMSE groups in terms of game median correct RT ($U=377.0, z=-4.5; p<0.001$). For Table 4-4, the corresponding scatterplot (Figure 4-2) is shown. Figure 4-2 shows the distribution of game median RT versus MMSE (“dementia” scores are indicated by circles) where a tendency for lower MMSE scores are associated with longer game median correct RTs can be seen.

Table 4-4. Mann-Whitney $U$ test results comparing cognitive assessment performance based on the absence ($\geq$24) or presence (<24) of dementia as assessed by the MMSE.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MMSE &lt;24</th>
<th>MMSE $\geq$24</th>
<th>$U^a$</th>
<th>$p$</th>
<th>$z$</th>
<th>$r$</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Mean (SE)</td>
<td>n</td>
<td>Mean (SE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game Median Correct RT</td>
<td>18</td>
<td>1600 (200)</td>
<td>122</td>
<td>900 (200)</td>
<td>377.0</td>
<td>&lt;0.001</td>
<td>-4.5</td>
</tr>
<tr>
<td>CRT Accuracy</td>
<td>8</td>
<td>0.7 (0)</td>
<td>73</td>
<td>0.8 (0)</td>
<td>181.0</td>
<td>&lt;0.001</td>
<td>-1.7</td>
</tr>
<tr>
<td>CRT RT</td>
<td>8</td>
<td>2200 (300)</td>
<td>73</td>
<td>1300 (0)</td>
<td>104.0</td>
<td>&lt;0.001</td>
<td>-2.9</td>
</tr>
</tbody>
</table>

$a$ Table has been reordered based on the $U$ statistic value according to estimated $P$ value.
RT measures are reported in milliseconds, CRT accuracy reflects proportion of correct responses.

Figure 4-2. Scatterplot illustrating the differences on game median correct RT based on MMSE score (≥24 = normal cognitive function or possible dementia; <24 = signs of dementia). MMSE = Mini-Mental State Examination.

Similar to the analysis reported in Table 4-4, a Mann-Whitney U test (Table 4-5) was performed to investigate the difference between cognitive ability and serious game performance when the MoCA score was 23 and above (normal cognitive function) versus below 23 (MCI) (Luis, Keegan, & Mullan, 2009). The MoCA was chosen as the criterion in this comparison because it is a de facto standard in screening for MCI versus normality. There was a significant difference ($U=307.0$, $z=–3.2$; $p=0.03$) on the CRT RT between participants with cognitive impairment (MoCA <23) and no impairment (MoCA ≥23). There was also a significant difference between MoCA groups for game median correct RT ($U=890.0$, $z=–2.6$; $p=0.001$). For Table 4-5, the corresponding bivariate relationship scatterplot is illustrated in the scatterplot in Figure 4-3. This figure illustrates a tendency for lower MoCA scores to be associated with longer median correct RTs, although that relationship appeared to be weaker for the MoCA than it was for the MMSE.
Table 4-5. Mann-Whitney U test results comparing game performance based on the absence (≥ 23) or presence (< 23) of cognitive impairment as assessed by the MoCA. MoCA = Montreal Cognitive Assessment, and CRT = Choice Reaction Time.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MoCA, &lt;23</th>
<th>MoCA ≥ 23</th>
<th>U^a</th>
<th>p</th>
<th>z</th>
<th>r</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean (SE)</td>
<td>n</td>
<td>Mean (SE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game Median Correct RT</td>
<td>38</td>
<td>1100 (70)</td>
<td>67</td>
<td>900 (20)</td>
<td>890.0</td>
<td>0.01</td>
<td>-2.554</td>
</tr>
<tr>
<td>CRT Accuracy b</td>
<td>26</td>
<td>0.8 (0.02)</td>
<td>44</td>
<td>0.9 (0.02)</td>
<td>439.5</td>
<td>0.11</td>
<td>-1.6</td>
</tr>
<tr>
<td>CRT RT</td>
<td>38</td>
<td>1000 (70)</td>
<td>67</td>
<td>900 (20)</td>
<td>307.0</td>
<td>0.03</td>
<td>-3.2</td>
</tr>
</tbody>
</table>

^a Table has been reordered based on the U statistic value according to significance.

^b RT measures are reported in milliseconds, CRT accuracy reflects proportion of correct responses.

Figure 4-3. Scatterplot illustrating the differences on game median correct RT based on MoCA score (≥ 23 = normal cognitive function or possible MCI; < 23 = signs of MCI). MoCA = Montreal Cognitive Assessment, MCI = mild cognitive impairment.

Another Mann-Whitney U test (Table 4-6) was performed to investigate the difference between cognitive ability and serious game performance when delirium was present (CAM positive) versus absent (CAM negative). The CAM was chosen as the grouping factor as it is the gold standard in screening for delirium. The test indicated a significant difference on the MMSE,
MoCA, RASS, and DI between participants with delirium (CAM positive) and no delirium (CAM negative). In addition, there was a significant difference between CAM groups in terms of median correct RT on the serious game ($U=207.0$, $z=-4.2$, $p<0.001$). For Table 4-6, a scatterplot of the relationship between CAM groups and median game correct RT is shown in Figure 4-4. These between-group differences in game median correct RT and MMSE are consistent with findings by Lowery (2010), where CAM-negative participants demonstrated faster RT (and faster correct RT) and higher MMSE scores compared to CAM-positive participants.

**Table 4-6. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (CAM negative) or presence (CAM positive) of delirium as assessed by the CAM. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, and CRT = Choice Reaction Time.**

<table>
<thead>
<tr>
<th>Assessment</th>
<th>CAM Negative</th>
<th>CAM Positive</th>
<th>$U^a$</th>
<th>$p$</th>
<th>$z$</th>
<th>$r$</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>RASS</td>
<td>142 -0.03 (0.02)</td>
<td>14 -0.8 (0.2)</td>
<td>288.0</td>
<td>&lt;0.001</td>
<td>-7.8</td>
<td>0.62</td>
<td>0.0–0.0</td>
</tr>
<tr>
<td>Game Median Correct RT</td>
<td>129 900 (40)</td>
<td>12 1700 (200)</td>
<td>207.0</td>
<td>&lt;0.001</td>
<td>-4.2</td>
<td>0.4</td>
<td>0.2–1.0</td>
</tr>
<tr>
<td>MoCA</td>
<td>101 23.8 (0.4)</td>
<td>7 14.3 (2.0)</td>
<td>60.5</td>
<td>&lt;0.001</td>
<td>-3.7</td>
<td>0.36</td>
<td>21.0–26.0</td>
</tr>
<tr>
<td>MMSE</td>
<td>131 27.6 (0.2)</td>
<td>14 18.4 (1.3)</td>
<td>38.0</td>
<td>&lt;0.001</td>
<td>-5.9</td>
<td>0.49</td>
<td>26.0–29.0</td>
</tr>
<tr>
<td>DI</td>
<td>131 0.6 (0.1)</td>
<td>14 6.9 (0.5)</td>
<td>24.5</td>
<td>&lt;0.001</td>
<td>-6.6</td>
<td>0.55</td>
<td>0.0–1.0</td>
</tr>
<tr>
<td>CRT RT</td>
<td>78 1300 (60)</td>
<td>4 2600 (50)</td>
<td>45.0</td>
<td>0.02</td>
<td>-2.4</td>
<td>0.26</td>
<td>1.0–1.4</td>
</tr>
<tr>
<td>CRT Accuracy</td>
<td>78 0.8 (0.01)</td>
<td>4 0.7 (0.1)</td>
<td>91.5</td>
<td>0.17</td>
<td>-1.4</td>
<td>0.15</td>
<td>0.8–0.9</td>
</tr>
</tbody>
</table>

$^{a}$Table has been reordered based on the $U$ statistic value according to significance. No Mann-Whitney U test analysis was carried out for the DVT because there were no CAM-positive participants who completed the DVT. Additional assessments are included for comparison.

$^{b}$RT measures are reported in milliseconds, CRT accuracy reflects proportion of correct responses.
Figure 4-4. Dot plot illustrating the differences on game median correct RT based on CAM groups (CAM negative=delirium absent; CAM positive=delirium present). CAM = Confusion Assessment Method.

To verify that our results were not dependent on a specific analytic methods, I replicated all the Mann-Whitney U tests in Table 4-4 and Table 4-5 with their parametric equivalent, in this case the independent samples t-test. The pattern of significant and non-significant effects was identical for both tests.

4.1.7.1.1 Predicting Delirium Status Using Serious Game Performance

In the preceding section, I examined the relationship between game performance and current standards for clinical assessment with respect to MCI, delirium, and dementia. In this section, I examined the question of how well the serious game performance predicts CAM status (delirium).

I examined different possible game correct RT cut-off values for distinguishing between people who should be screened for possible delirium (using the CAM) and those who should not. Setting a relatively long median correct RT for the decision threshold (≥1880 msec) resulted in good specificity (126 of 129, 97.67% CAM-negative patients were correctly identified) but relatively poor sensitivity (only five of 12, 41.67% CAM-positive patients were correctly identified).
On the other hand, using a more stringent median correct RT cut-off of 1130 msec, there was both good sensitivity (10 of 12, 83.33% CAM-positive patients were correctly identified) and good specificity (110 of 129, 85.27% CAM-negative patients were correctly identified).

4.1.7.2 Usability Issues and Evidence of Enjoyment and Engagement

The following brief notes recorded by the RAs during participant use of the serious game are indicative examples of enjoyment and engagement that were observed: “Loved the game, she was playing games on her iPhone before I approached her,” “Enjoyed the game, he would play on his own,” “Too easy but don’t make it too challenging, like the game,” and “Really loved the tablet, wanted to keep playing even after testing was over.” However, usability problems were also observed. Some participants placed their palm on the tablet while trying to interact with the serious game. This confused the software because it was unclear which hit points were intentional versus accidental. Some participants claimed that the game was too easy and suggested that I include more difficult levels to make it more interesting. Participants also expressed an interest in playing games such as crossword puzzles. Anecdotally, the RAs who supervised the data collection at the hospital reported that this game was easier to administer and more fun to complete compared to standard cognitive assessments such as the MoCA, and DVT.

4.1.7.3 Ergonomic Issues

While interacting with the tablets, the participants assumed numerous positions, such as being seated, lying down, standing, or walking around. Each of these positions had different ergonomic requirements and some brief recommendations are provided in the Discussion section. Some participants were also frail and required the assistance of the RA to hold the tablet for them.

4.1.8 Discussion

Performance on the serious game in terms of median correct RT was significantly correlated with MMSE, MoCA, CAM, DI, and CRT scores for elderly ED participants and differences were in the expected direction (slower game median correct RT for people with possible MCI and dementia). The correlations suggest a relationship between longer median correct RT on the game and lower cognitive assessment scores. In these cases, longer median correct RT may have been due to other reasons such as the presence of physical impairment, pain, and fatigue. These correlations demonstrate the potential value of serious games in clinical assessment of cognitive
status. The correlations between the standard cognitive tests observed in this study are similar to results seen in other research. For example, correlations of $r=0.43$ and $r=0.60$ between MMSE and MoCA scores for healthy controls and patients with MCI, respectively, have been found (Trzepacz, Hochstetler, Wang, Walker, & Saykin, 2015). In this study, I observed a correlation of $r=0.63$ ($p<0.001$) between the MMSE and MoCA scores. Overall, the correlation of the serious game with existing methods of clinical cognitive assessment appears to be almost as strong as the correlations of the clinical assessment methods with themselves.

In the partial correlation analysis, I observed that the serious game correlates with the MMSE, CAM, and DI, but that part of that correlation is attributable to speed of processing (CRT speed). Thus, serious game performance in this case involved both speed of processing and EF components. Both components are likely involved in the correlation of the serious game with the MMSE. However, only the speed of processing component appears to be involved in the correlation with the MoCA. Crucially, the partial correlations of serious game performance (controlling for CRT RT) were higher than the corresponding partial correlations for CRT (controlling for serious game performance) indicating that the serious game is an overall better predictor of cognitive status than simple processing speed as measured by the CRT task.

One of the goals of this study was to develop a method for predicting the presence of delirium using this serious game. In this study, I found that a median correct RT cut-off of 1130 msec implied relatively good sensitivity and specificity in the clinical decision. However, 10 of the 129 (7.8%) participants were above the median cut-off and only five of these were CAM-positive. Thus, in a clinical setting, the question remains of how to deal with people who are identified as CAM-positive using this RT cut-off value. One approach would be to give those people the CAM and then treat the CAM-positive patients using standard protocols for patients with delirium. The value of the serious game in this case is that it would allow (based on screening with the serious game) a high rate of delirium detection using the CAM in a small percentage of patients (assuming that the current results generalize to other contexts). While screening of 10% of patients would have worked well based on retrospective analysis of the current data, in practice of 15-20% may provide a reasonable compromise between efficiency (in terms of how many CAM assessments have to be administered, and completeness, ensuring that as many cases of delirium as possible are identified). Ideally, a suitably adapted serious game would also detect risk of delirium onset so that prevention strategies could be used on targeted
patients before they developed delirium, but that prospect was beyond the scope of the research reported in this dissertation.

During this study, many ergonomic issues were observed that could arise during the administration of the serious game. For instance, there were a variety of positions and methods used to interact with the tablet-based serious game. For participants who are sitting down, I recommend a tablet case that has a hand holder or kickstand to allow them to interact with the tablet in multiple ways. In contrast, for participants lying down on a bed, it may be difficult for them to hold the tablet to play the serious game; thus, a stand affixed to a table or intravenous pole that holds up the tablet would be appropriate. Furthermore, the ergonomic solutions that are adopted should meet hospital standards on hygiene and sanitization for technology. For patients with hand injuries or visual disabilities, the serious game may not be a usable option. One strategy for screening out such people in future studies would be to remove those patients who scored lower than 20 on the Manual Function Test on both limbs (Miyamoto, Kondo, Suzukamo, Michimata, & Izumi, 2009).

User-centered design and ergonomic interventions were both key in ensuring that the serious game was usable with a challenging user group (elderly patients) and in the fairly unique and demanding context of a hospital ED. The touch interface was modified so that it was more forgiving of the kinds of gestures made by elderly users when interacting with the game and the game play was modified so that users with a wide range of ability could play the game. Ergonomic issues that were dealt with in this research included the form factor of the device and the selection and use of accessories to facilitate interactions with the device in different contexts.

Based on this study, I present the following recommendations for enhancing tablet-based user interaction between elderly adults and touch-based technologies:

- Accept multiple gestures, including taps and swipes, as input to maximize interaction.
- Provide a stylus for users who have difficulties interacting with the tablet with their fingers.
- For time-sensitive tasks, the time limit should be increased to allow older or more frail users a chance to interact with the software.
- Tablet screen protectors should be installed to provide more friction between a user’s hand and the screen.
• A variety of ergonomic stands and mounts should be available to accommodate various interaction positions.

• Serious games for cognitive assessment should incorporate validated psychological task components (e.g., EFs) and should be easily playable for independent use.

• Assess the validity of the game across different subgroups of patients. Consider the possibility of using multiple versions of a game, or multiple games, to accommodate the different characteristics and needs of different types of patient.

4.1.8.1 Limitations

The usability and validation results obtained apply to elderly adults in an emergency setting. Further research would be needed to generalize these results to different types of patients and clinical settings. The design of this study was cross-sectional, so each participant/patient was only studied during one ED visit and played the game only once. Future research may assess the reliability of the game when played repeatedly by the same patient in the ED. One other limitation is that only one game was examined in this research (the whack-a-mole game that I developed). Other serious games should also be explored to determine which games work best with different types of patients.

This work is an initial validation study of the serious game for cognitive screening, where the game was only administered once. One of the goals of this research is frequent cognitive screening, which can potentially lead to learning effects on the game. In the study reported later in this chapter I assess the reliability of the game-based screening tool and whether or not learning or fatigue effects are present when the game is played multiple times.

Since this study was carried out in a clinical setting, the supervising physician and his team had a large say in how the game should be designed for use by elderly patients. Due to earlier research, results showing that slower correct RT was related to the presence of delirium I focused on game median correct RT. In the design of this game, and I did not implement the ability to measure signal detection measures. The iterative design nature of this study also meant that there were a number of different versions of the serious game that were used although in all versions the fundamental requirement of the game as a Go/No-Go discrimination task was preserved. In total, there were 14 variants of the game (Table 8-3), thus, not all patients played the same variants of the game. The changes to the game versions were relatively minor and should not affect the
nature of the task, for example the character size increased from 200 px to 250 px or the game board size changed from 2x2 to 3x3. All participants had the opportunity to play a game level with distractors.

4.1.9 Conclusion

This validation study provides several lines of evidence to suggest that serious games are a promising methodology for cognitive screening in clinical settings, even in the high-acuity time-pressured ED environment. Importantly I demonstrated the feasibility of implementing a serious game for cognitive screening in a clinical environment. To the best of my knowledge, this is the first time that a serious game for cognitive assessment has been tested in an ED and with a full battery of standard cognitive assessment methods for comparison. Based on these results, ergonomically appropriate serious games have the potential to revolutionize cognitive assessment of the elderly in clinical settings, allowing assessments to be more frequent, more affordable, and more enjoyable.

This research also provides a case study in the development of an interactive serious game for cognitive screening that may be used independently and repeatedly, thus promoting patient-centered health and safety. I have demonstrated in this study that elderly adults older than age 70 years can successfully play the serious game in an ED and that median correct RT performance on the game can be used as an initial screen for cognitive status.

These findings do not yet demonstrate that the serious game evaluated here is ready to be used to screen for delirium in the ED. Only 12 CAM-positive patients were observed in the study and game median correct RT was predictive of CAM status. However, due to the known underreporting of delirium in the ED, an efficient and usable method of screening for delirium is clearly needed. In this study, a game median correct RT cut-off of 1130 msec produced a sensitivity of 83% and a specificity of 86% when used retrospectively as a screen for CAM-positive status. Although further research is needed, it seems possible that a suitably revised and validated game might be able to identify approximately 80% to 90% of CAM-positive cases while requiring the screening of no more than approximately 15-20% of cases.

Outside the ED, the use of the serious game for ongoing patient-administered assessment would ideally involve patients who remain actively engaged with their support network (e.g., family
and care providers) and with healthcare professionals. For instance, if patients perform poorly on the serious game or notice a decline in their performance, they could discuss these results with their care providers, which might lead to interventions such as changes to medication or lifestyle that could slow observed declines.

### 4.2 Reliability Study

In the previous section, the concurrent validity of the serious game was discussed. This section will examine the test-retest reliability (e.g., Anastasi, (1988)) of the game, as it is an important criterion in a cognitive test for detecting risk of delirium onset, where multiple administrations of the test to the same participant may be required. This study (with a sample that includes data collected from 56 patients in the validation study reported in the previous section) extends the concurrent validation study carried out in the previous section by focusing on the test-retest reliability of the game-based assessment with older adults in the ED that were screened for incident delirium. This is a further investigation of the serious game’s psychometric properties, with the motivating application being screening for delirium in an ED.

#### 4.2.1 Research Question

The research question that this section will address is as follows:

1. How reliable is game performance when measured repeatedly on the same participants in a clinical setting?

This research question was examined by looking at the test-retest reliability of the game by comparing the correlations between game performance and measures of cognitive status across follow up sessions.

#### 4.2.2 Methodology

**4.2.2.1 Patient Selection**

The research ethics protocol submitted to Sunnybrook Health Sciences Centre (protocol #070-2013) was amended during the validation study discussed in the previous section to invite patients to participate in the reliability study. As the amendment was approved, and the protocol changed, part way through the validation study, the final 56 patients out of the 147 patients in the validation study were invited to participate in the additional sessions required for the reliability
study (if they stayed more than eight hours in the ED). However, not all of these 56 patients in the overlapping sample contributed reliability data due to reasons such as refusing to participate in a follow up session or being discharged. The enrolment for the validation study ended with a sample size of 147 patients. The recruitment efforts then focused on recruiting patients for the reliability study, which eventually included 71 patients that were not in the validation study.

Again, not all the additional 71 patients enrolled actually participated in the reliability study due to similar reasons (e.g., refusal to participate and or being discharged before follow up). Thus, out of the 56 patients recruited from the validation study, plus the 71 additional patients that were recruited for the reliability study, 114 patients actually participated in the reliability study. Figure 4-5 illustrates the overlapping and non-overlapping participants in the validation and reliability study, as well as the number of patients who actually completed the study. A breakdown of the number of patients that completed each follow up session is in Figure 4-5.

The inclusion criteria for the reliability study were the same as for the validation study described in the methodology above. The same exclusion criteria from the validation study were also used, with the addition of the following new criteria. Patients were screened with the CAM and removed from the study if they were found to be CAM positive in the initial session, or in any of the follow up sessions. Since patients with delirium will typically have fluctuating cognitive status, they were not considered in assessing test-retest reliability.
4.2.2.2 Test-Retest Reliability

The test-retest reliability of the serious game was assessed by conducting follow-ups at regular time intervals a minimum of eight hours apart, on CAM negative patients only. Patients varied in the number of follow-ups they participated in, depending on their total time in the ED. During each follow up, a RA administered the MMSE, DI, CAM, and DVT, and asked the patient to play the serious game. For each patient, there were a maximum of five follow-up sessions, in addition to the initial enrolment into the study where the first set of assessments and initial game play occurred. Reliability of game scores between pairs of sessions was tested using both Pearson’s $r$ and Spearman’s $\rho$ correlations.

4.2.2.3 Statistical Analysis

The normality of MMSE scores and of the serious game median correct RT data was assessed. The target offset was calculated for correct trials only and details about target offset results can be found in Appendix D. Data normality was visually inspected using histograms, P-P and Q-Q plots. Due to the large sample size, tests such as the Kolmogorov-Smirnov and Shapiro-Wilk were not carried out as they are overly sensitive with large sample sizes (Field, 2013). MMSE scores were treated as interval data. Median correct RTs were used to summarize the RT data in order to reduce the impact of positive skew and outliers on analyses with the RT data. In addition, non-parametric tests were also used as an alternative interpretation of the data that do not make normality assumptions (Spearman’s $\rho$, and the Wilcoxon signed-ranks test).

4.2.3 Results

4.2.3.1 Study Sample

114 patients (61 females, and 53 males) participated in the reliability study, between the ages of 70 and 104 years (standard deviation = 7). The average length of stay in the ED was 16.3 hours (standard deviation = 9.0) (Table 8-10).
Of the 114 participants who played the serious game in the initial enrollment session, 47, 23, and 16 patients (who tested as CAM negative) played the game in follow up sessions 1, 2, and 3, respectively. At initial enrolment, MMSE scores ranged from 12 to 30, and MoCA scores ranged from 10 to 30.

The completion rate of the MMSE, DI, CAM, and DVT decreased from the initial enrolment to subsequent follow up sessions (Figure 8-1). The number of participants that completed each assessment between adjacent sessions was very low, thus the test-retest reliability was not calculated.

4.2.3.2 Test-Retest Reliability

The test-retest reliability of serious game performance was investigated by calculating two-tailed Pearson’s $r$ and Spearman’s $\rho$ correlations between pairs of sessions. A table of correlations (for the median correct RT) is shown in Appendix D. The correlations with median correct RT between the initial enrolment and follow up session 1, and between follow up sessions 1 and 2 were significant ($p<0.05$) with $r$-values ranging between 0.834 and 0.886 (Table 8-11). Scatterplots for the three adjacent pairs of sessions are shown in Figure 8-2 (median correct RT).

The test-retest reliability of the serious game median correct RT was recalculated using Spearman’s $\rho$ correlations (Table 8-13) at each follow up. The $\rho$-values were significant ($p<0.05$) between the initial enrolment and follow up session 1, and between follow up sessions 1 and 2 for both median correct RT ($\rho$-values ranging between 0.804 and 0.818).

4.2.3.3 Practice and Fatigue Effects

I carried out inferential tests to assess the statistical significance of possible practice/learning effects. Three paired $t$-tests (two-tailed) were carried out to determine if there was a difference in game median correct RT between (1) initial enrolment and follow up 1, (2) follow up sessions 1 and 2, and (3) follow up session 2 and 3. Bar charts corresponding to these comparisons are shown in Figure 4-6, which shows a decreasing trend in median correct RT across the sessions.
There was no significant difference between the initial enrolment and first follow up session or between the first and second follow up sessions or between the second and third follow up sessions with median correct RT.

As a follow up, Wilcoxon signed-rank tests were carried out between adjacent sessions. There was no significant difference between the initial enrolment and first follow up session or between the second and third follow up sessions. However, there was a significant difference between follow up sessions 1 and 2: ($z=-2.329$, $p=0.020$).
Figure 4-7. Histogram illustrating the median correct RT difference between successive sessions. Each bin represents 100 msec.

Figure 4-7 visualizes potential practice and fatigue effects across successive follow up sessions based on differences (within participants) in median correct RT between adjacent sessions. The histograms were scaled to be on the same x-axis with the same time bin sizes and were lined up vertically to facilitate visual comparison. I assumed that patients with a difference in median correct RT that was greater than zero in the subsequent session were in poorer condition or were experiencing fatigue. In contrast, for patients with a difference less than zero there was likely a practice effect, as they were speeding up in the subsequent session. Since the distribution of differences in median correct RT between adjacent sessions (within individuals) is reasonably well balanced around the no difference (0 msec) point there is little evidence of a genuine learning effect in game median correct RT performance. Instead, the reduction in game median correct RT in later sessions (as indicated by significant Wilcoxon signed-rank tests) is likely due, in most part, to slower participants dropping out of the study either because they were treated more quickly or because they were less willing or able to participate in the later sessions.
4.2.4 Discussion

The present findings demonstrate the test-retest reliability of the game-based screening tool with an elderly emergency population. I observed a significant relationship between two adjacent pairs of administrations of the serious game (initial enrolment with follow session 1, and follow up sessions 1 to 2) with $r$-values generally ranging between 0.83 and 0.89 across the various pairs of median correct RT (Table 8-11). Similar Spearman’s rho-values ranges were observed for game median correct RT (Table 8-13). One possible explanation why there was no significant relationship between the second and third follow up session may have been due to the small number of participants who stayed in the ED long enough to participate in the third follow up session ($n=12$).

There were a wide range of MMSE (12–30) and MoCA (10–30) scores at initial enrolment. However, patients with possible dementia (MMSE <24) (O’Connor et al., 1989) and MCI (MoCA <23) (Luis et al., 2009) were still able to play the serious game. This suggests that the game can be played by people with cognitive impairments and may potentially be able to distinguish between different levels of impairment.

The game median correct RT performance over sessions, within patients, was relatively consistent with the histograms of within patient median correct RT differences between adjacent sessions (Figure 4-7) balanced around the zero difference value. Thus, there is no evidence for the presence of a practice/learning effect, with changes in the means of median correct RT in later sessions, were likely due to sicker or lower functioning patients either being discharged or refusing to participate in further testing.

It appears that the sample of patients who remained in the hospital over an extended period of time, remained CAM negative, and who were able to play the game, were comparatively fit. They had better game performance than the patients who dropped out of the study after the earlier sessions. Within subjects, there was no tendency for either speeding up or slowing down across sessions. While it is possible that practice/learning effects (speeding up) precisely matched the fatigue effects (slowing down) between sessions, a more parsimonious explanation would be that there were no significant practice or fatigue effects in this case, which would be a beneficial property of the game if it can be further verified in future research.
The test-retest reliability correlations obtained with game median correct RT in this study were comparable with the test-retest correlations obtained by Tombaugh (2005) in his study of the test-retest reliability of the MMSE. However, in the present study, follow up sessions were separated by eight hours, whereas in Tombaugh's (2005) study the different time periods were separated by one or more years. In addition, the MoCA has been shown to have a high test-retest reliability of 0.92 over a period of around a month (Nasreddine et al., 2005), which is significantly longer than the 8 hours separation between follow-ups in the present study. In cases where short test-retest time intervals were used, test-retest reliability of the MMSE was reported to be much higher (between 0.8 and 0.95). However, previous published values may tend to overestimate the value of the test-retest reliability of the MMSE. First, the MMSE studies generally selected people who were not assessed to be cognitively impaired (e.g., the participants would have had MMSE scores relatively close to the top end of the MMSE scale). This would have resulted in a compressed scale, making it more likely that scores within individuals would tend not to change (in contrast to game median correct RT, where the median correct RTs may vary over many milliseconds). Second, since the MMSE items are identical and since the focus is on unimpaired individuals, scores will tend to stay the same or possibly improve due to practice effects, and the ability to remember items on the test. In contrast to the MMSE, a patient cannot remember an exact sequence of events in the game used in my research, since where and when targets appear was varying, and randomized. In this study, I found little evidence of practice effects in game median correct RT performance, when differences between administrations of the game, conducted at varying times, were assessed within individuals (Figure 4-7).

4.2.4.1 Limitations and Future Work

This study focused on patients admitted to one hospital ED. It is possible that somewhat different findings might have been obtained in a different ED. In future studies, research should explore other target patient populations of interest so as to improve the generalizability of the results and expand the settings where this can be useful. While no strong learning (practice) effects were observed in game performance between sessions, it is likely that there was an initial learning effect when patients first started using the game. The initial learning effect was not assessed in this study. Instead, patients were allowed to do some initial practice with the game, with feedback and encouragement provided by the RAs, before the patient played the game to make the assessment. In contrast to the apparent lack of practice effects for the serious game median
correct RT, practice effects have been found in “standard” cognitive assessments such as the MoCA and MMSE (Tombaugh & McIntyre, 1992).

While significant, the test-retest reliabilities observed here for the game are lower than the corresponding reliabilities reported for the MMSE (for short time periods of up to a few months). However, the game provides a much wider range of scores, since RT is measured in milliseconds and it is not possible to memorize the answers to questions as it may be for tests such as the MMSE.

The inter-rater reliability for the CAM was not assessed in this study. However, the CAM has been shown to have a high inter-rater reliability (Bhat and Rockwood, 2007). Moreover, there was a high rate of loss-to-follow-up due to carrying out assessments every eight-hours. Future studies should consider using shorter latencies between sessions (e.g., every one-hour) in order to improve the chances of detecting delirium.

4.2.5 Conclusion (for the reliability study)

This study has demonstrated that the game-based screening tool is a reliable tool (in terms of short-term test-retest reliability) for measuring cognitive status, and that it can be used independently by patients in emergency care after a few minutes of training (at most). The game is also usable, and can be self-administered by patients. While the game appears to have lower test-retest reliability than the MMSE, it does not seem to have practice effects if people are given a short amount of initial training, and it provides a wider range of scores and no opportunities for memorizing answers. While the serious game has yet to be fully validated for routine assessment of cognitive decline, this research opens the way to further explore self-administration by patients of a cognitive screening tool that is able to track their progress over time. In busy environments such as the ED, this type of serious game for self-administered cognitive assessment could, in the future, assist both healthcare providers and patients by providing critical information on a patient’s cognitive status over time. A suitably revised version of the game used in this dissertation research should be a useful supplement to tests such as the MoCA, and MMSE in situations where it may be difficult or impractical to use those existing assessments.
4.3 Conclusions (Across the First Two Clinical Studies)

In this chapter, I have addressed the first research question of how the serious game played by older adults in a clinical setting compares to existing methods of cognitive assessment in a clinical setting. The concurrent validity of the tool was shown using a correlation analysis. Performance on the serious game correlated significantly with the MoCA \((r= -0.357, p<0.001)\) and MMSE \((r= -0.617, p<0.001)\), and correlated (point-biserial correlation) with the CAM \((r= 0.421, p<0.001)\) and with other cognitive assessments. The second research question of how reliable game performance measures are when measured repeatedly on the same participants in a clinical setting was examined in the second clinical study. The test-retest reliability of the tool was demonstrated using a correlation analysis between adjacent follow up sessions \((r=0.8)\).
Chapter 5

5 Re-Assessing the Concurrent Validity of the Serious Game in Elderly Adults

This chapter describes an additional clinical evaluation of the concurrent validity of the serious game in a sample of elderly adults selected from a hospital ED. The game was revised to record new performance measures, including false alarm rates. The motivation for doing this was to see if false alarm rates are also related to performance on standard cognitive assessments and to see if they can usefully supplement median correct game RT in assessing cognitive status, and provide a stronger relation with the DI assessment. In addition, the study reported in this chapter attempted to replicate my finding in the first clinical study that the serious game RT was significantly correlated with MMSE scores.

5.1 Research Questions

The research questions that this chapter will address are as follows:

1. Does performance on the serious game played by older adults in a clinical setting agree with (in terms of correlation measures) existing methods of cognitive assessment in a clinical setting (i.e., MMSE, CAM, and the DI)? (Seeking to replicate the findings of the previous clinical study.)
2. Does game false alarm rate by itself predict clinical assessments?
3. Does game false alarm rate when combined with median correct game RT predict the clinical assessments?
4. Does game false alarm rate when combined with median correct game RT improve prediction of the clinical assessments compared to median correct game RT by itself?

These research questions were addressed by examining the concurrent validity of the serious game using a correlation analysis (using both Pearson’s $r$ and Spearman’s $\rho$) between
standard measures of cognitive status (e.g., MMSE, CAM, DI) with measures of game performance (median correct RT, false alarm rate, combined performance metric).

5.2 Methodology

The research protocol was approved by Sunnybrook Health Sciences Centre (protocol #035-2017). Adults who were 65 years or older and who were present in the ED for a minimum of four hours were recruited for the study. Exclusion criteria included patients who (1) live in a full care nursing home, (2) have a critical illness rendering them unable to communicate or provide consent, (3) have visual impairment that makes them unable to use the tablet, and (4) have other communication difficulties preventing use of the tablet.

Clinical RAs administered standard cognitive assessments including the MMSE, CAM, and DI. Each participant was then asked to play the serious game and provide feedback.

5.2.1 Statistical Analysis

The cognitive data and serious game results were non-normally distributed based on visual inspection of the data. Tests for normality, including the Kolmogorov-Smirnov and Shapiro-Wilk tests, were not used due to the large sample size in this study because they are known to result in oversensitivity to relatively small departures from normality (Field, 2013). Transformations of the data were not performed because some of the measures, such as the CAM and DI, are binary/categorical and cannot follow a normal distribution. My interest was in correlations as a measure of the effect size of the underlying relationship between game performance and the cognitive assessments, but I used nonparametric correlation measures for some of the comparisons that involved categorical or narrow ordinal scales. Correlations between the dichotomous CAM and the other measures were assessed using point-biserial correlations (Field, 2013). Correlations involving the DI (and not involving the CAM) were assessed using Spearman’s rho because the DI uses a small number of ordered categories. The remaining comparisons used Pearson’s r correlations.

The serious game software was redesigned to collect data on signal detection values such as hits, false alarms, correct rejections, and misses (Stanislaw & Todorov, 1999). Game performance was measured based on a participant’s median correct RT, false alarm rates (Equation 3-1), and a combined performance metric (Equation 3-2).
In total, there were three variants of the game (Table 8-15), thus, not all patients played the same variants of the game. The changes to the game versions were relatively minor and were unlikely to have affected the psychological characteristics of the task. For example, the game board size changed from 2x2 to 3x3. All participants had the opportunity to play a game level with distractors. The target offset on correct trials was also calculated, and analyses relating to this measure can be found in Appendix E.

5.3 Results

5.3.1 Description of Sample

In this study, 170 patients were enrolled, (82 males and 88 females) between the ages of 65 and 100 years (mean age = 77.77 years, standard deviation = 8.46). Of the 170 patients who were enrolled, 132 patients played the serious game.

The MMSE scores ranged between 12 and 30, with a mean score of 26.5, and 79 participants did not complete the MMSE. There were 152 enrolled participants who were negative for the CAM and 2 participants who were positive (a positive result on the CAM suggests that the participant has delirium), and 16 participants without a CAM score. Only one CAM-positive patient played the game. Moreover, the DI scores ranged from 0 to 12 (the score indicates the severity of delirium). There were 30 participants without a DI score. The overall median correct RT on the serious game was 1100 msec (IQR 400). The overall mean false alarm rate was 0.09 (standard deviation = 0.1). A summary of the scores on the cognitive assessments can be found in Appendix E.

5.3.2 Comparison Between Serious Game Performance and Standard Cognitive Assessments

Correlation analysis replicated a significant relationships between game median correct RT and scores on two of the cognitive assessments: MMSE, and DI (Table 5-1). The game false alarm rate also had a statistically significant relationship with the MMSE, and DI. Analyses with the CAM were not carried out since there was only one CAM-positive patient who played the game. Note that information about which types of correlation were used for each comparison is shown in the footnotes of Table 5-1.
As a follow-up to the Pearson’s $r$ correlation analysis in Table 5-1, I carried out the same analysis using Spearman’s $rho$ correlations (Table 8-17). All significant Pearson’s $r$ correlations between the cognitive assessments and game median correct RT, game false alarm rate, were also observed to be significant using Spearman’s $rho$.

To compare the strength of each correlation between each game performance measure with performance on the MMSE, Fisher’s $z$-transformations were carried out. The correlation between the median game correct RT and MMSE ($r=-0.508$) was not significantly lower than the correlation between the game combined performance metric with MMSE ($r=0.605$), $z=-0.796$, $p=0.213$. The correlation between the game false alarm rate and MMSE ($r=-0.572$) was not significantly lower than the correlation between the game combined performance metric with MMSE ($r=0.605$), $z=-0.26$, $p=0.397$.

Table 5-1. Correlations comparing game performance to the standard cognitive assessments. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, and DI = Delirium Index.

<table>
<thead>
<tr>
<th></th>
<th>Game Median Correct RT</th>
<th>Game False Alarm Rate</th>
<th>Game Combined Performance Metric</th>
<th>DI $^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Median Correct RT</td>
<td>0.0443</td>
<td>-0.655 ***</td>
<td>-0.508 ***</td>
<td>0.367 **</td>
</tr>
<tr>
<td></td>
<td>$p=0.679$</td>
<td>$p&lt;0.001$</td>
<td>$p&lt;0.001$</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td></td>
<td>$n=90$</td>
<td>$n=131$</td>
<td>$n=83$</td>
<td>$n=118$</td>
</tr>
<tr>
<td>Game False Alarm Rate</td>
<td>-0.800 ***</td>
<td>-0.472 **</td>
<td>0.328 ***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.002$</td>
<td>$p&lt;0.001$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n=90$</td>
<td>$n=56$</td>
<td>$n=80$</td>
<td></td>
</tr>
<tr>
<td>Game Combined Performance Metric</td>
<td>0.605 **</td>
<td>-0.506 **</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.002$</td>
<td>$p=0.0030$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n=56$</td>
<td>$n=118$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMSE</td>
<td>1</td>
<td>-0.659 **</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.0016$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n=85$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DI</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

67
Correlations involving the DI were assessed using Spearman’s rho. All other correlations were calculated using Pearson’s r.

### 5.3.3 Detection of Abnormal State Using Serious Game Performance

A Mann-Whitney U test (Table 5-2) was performed to investigate the difference between cognitive ability and serious game performance when the MMSE score was 24 and above (normal cognitive function or possible dementia) versus when that score was below 24 (signs of dementia) (Folstein et al., 1975; O’Connor et al., 1989). The MMSE was chosen as the grouping criterion because it was a standard in screening for dementia at the time this research was carried out. There was a significant difference between MMSE groups in terms of game correct RT ($U=219.0$, $z=–3.9; p<0.001$), for false alarm rates ($U=123.0$, $z=3.2; p=0.001$), and for the combined performance metric ($U=168.0$, $z=–4.4; p=0.001$). Figure 5-1 shows the distribution of combined performance metrics versus MMSE scores, where lower MMSE scores tend to be associated with lower combined performance values.

![Figure 5-1](image.png)

Figure 5-1. Scatterplot illustrating the differences on the performance metric based on MMSE score ($\geq24 = \text{normal cognitive function or possible dementia}; <24 = \text{signs of dementia}$). MMSE = Mini-Mental State Examination.
Table 5-2. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (≥24) or presence (<24) of dementia as assessed by the MMSE.

MMSE = Mini-Mental State Examination.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MMSE &lt;24</th>
<th>MMSE ≥24</th>
<th>U</th>
<th>p</th>
<th>z</th>
<th>r</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean (SE)</td>
<td>n</td>
<td>Mean (SE)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game Median Correct RT</td>
<td>17</td>
<td>1.3 (0.06)</td>
<td>74</td>
<td>1.0 (0.03)</td>
<td>219.0</td>
<td>&lt;0.001</td>
<td>–3.9</td>
</tr>
<tr>
<td>Game Performance Metric</td>
<td>17</td>
<td>-1.2 (0.4)</td>
<td>74</td>
<td>0.4 (0.1)</td>
<td>168.0</td>
<td>&lt;0.001</td>
<td>–4.4</td>
</tr>
<tr>
<td>Game False Alarm Rate</td>
<td>17</td>
<td>0.2 (0.04)</td>
<td>74</td>
<td>0.07 (0.01)</td>
<td>123.0</td>
<td>0.001</td>
<td>–3.2</td>
</tr>
</tbody>
</table>

5.4 Discussion

Performance on the serious game, in terms of median correct RT, and false alarm rates, and the game performance metric (which combined z-scores for median correct RT and false alarm rates) were all significantly correlated with MMSE, and DI scores. Differences were in the expected direction (slower game correct RT for people with possible dementia, and higher false alarm rates for people with possible dementia). The MMSE and DI were strongly correlated with each other ($r=–0.659$), with the combined game performance metric also being significantly correlated with DI, although not as strongly ($rho=0.506$). Overall, the correlation of the serious game with existing methods of clinical cognitive assessment appears to be almost as strong as the correlations of the clinical assessment methods with themselves.

The strong correlations between the game false alarm rates and the MMSE and DI demonstrate that this aspect of game performance is also predictive and, along with median correct RT, should be explored in future research in assessing cognitive status. Individuals with higher false alarm rates tended to have lower cognitive performance, and the strongest relationship between game performance and cognitive status was for the game performance metric that combined both median correct RT and false alarm rate. Further research exploring the game performance measures in a larger sample can be carried out to determine which game measures are most predictive of cognitive status.

5.5 Conclusions

This work demonstrates the concurrent validity of the game with an additional ED sample (non-overlapping with the studies reported in Chapter 4) of 132 elderly patients who played the game. The present study also showed the effect of an additional game measure, false alarm rate, plus the effect of the game performance metric that was derived using this new measure along with
correct game RT. This research provides a case study of how median correct RT and false alarm rates on a serious game can be combined to create a better estimate of cognitive status.
Chapter 6

6 Construct Validation of a Serious Game for Assessment of Executive Functioning

This chapter reports on a study of the construct validity of the serious game, using a healthy adult sample. The purpose is to investigate how varying the different game parameters in the game are related to measures of general EF and inhibition. In order to distinguish between different forms of inhibition, I used four different inhibition tasks (Stroop, Go/No-Go discrimination, Flanker, anti-saccade) to provide a more detailed assessment of which type of inhibition is measured by game performance. These tasks were described in the literature review chapter (section 2.2.1).

6.1 Motivation

The cognitive screening tools administered in the two clinical validation studies described in Chapters 5 and 6, such as the MMSE and MoCA, both contain components designed to assess EF using tasks such as copying two pentagons on the MMSE, and a shortened TMT part B on the MoCA. As demonstrated in the two clinical validation studies, lower cognitive performance was associated with decreased performance on the serious game. Previous literature has demonstrated that lower EF is associated with lower cognitive performance as measured using MMSE, MoCA, and CAM (Lafleche & Albert, 1995; Rudolph et al., 2006; Zheng et al., 2012). For example, work by Lafleche and Albert (1995) demonstrated that older adults with AD as assessed using the MMSE, performed significantly worse on EF tasks including the self-ordering task, Hukok logical matrices task, TMT, and controlled oral word associated test. Previous research has demonstrated that patients with amnestic MCI as measured using the MoCA have decreased overall EF measured using a 2-back task, keep track task, stop-signal task, and more-odd shifting task (Zheng et al., 2012). Moreover, previous research has demonstrated that lower EF is associated with postoperative delirium. For instance, work by Smith, Attix, Weldon, Greene and Monk (2009) demonstrated that performance on the Stroop color word task is predictive of postoperative delirium, assessed through a chart review and/or the CAM. Rudolph et al. (2006) also demonstrated that CAM positive patients had lower EF based on performance on verbal fluency, category fluency, Hopkins Verbal Learning Test learning, and backward recounting of days (Rudolph et al., 2006). Further research by Greene et al. (2009) observed
lower performance on measures of EF using tasks such as the digit symbol test, symbol search test, and TMT part B in patients with postoperative delirium (Greene et al., 2009). This further highlights the motivation to investigate the relationship between which EF tasks are related to performance on the serious game.

In Chapter 5, I investigated the relationship between additional game performance measures such as false alarm rate and a combined performance metric, with performance on standard cognitive assessments in older adults in a clinical setting. As an extension of this work, I investigated these game performance measures (e.g., false alarm rate and combined performance metric) in a healthy sample against measures of inhibitory EF. Tasks used to measure EF are often administered on a computer with a monitor, keyboard and mouse in a quiet setting. Due to feasibility constraints, the ability to administer EF tasks in a clinical setting such as an ED was not possible. Thus, I carried out an experimental study investigating which EF tasks are related to game performance with a healthy sample in a controlled, laboratory setting.

The study reported in this chapter extends the earlier assessments of predictive validity and test-retest reliability, with an analysis of construct validity. Construct validity was assessed in terms of detailed relationships of game performance parameters with different types of inhibition ability. I also sought to identify which game parameters and performance measures provide the most sensitive estimate of the inhibition ability that the game measures. Additionally, based on previous research findings, I tested specific hypotheses concerning how different game parameters interact with each other in determining how difficult the game is to play.

6.2 Pilot Study

Prior to carrying out the main study, I carried out a pilot study on the serious game with six healthy, younger participants to determine what character appearance duration would likely lead to game play that was somewhat challenging, but not overly difficult. Six different character appearance durations were evaluated: 500 msec, 600 msec, 700 msec, 800 msec, 900 msec, and 1000 msec. Three different target to distractor frequencies were evaluated: 0.6, 0.7 and 0.8.

A two-way ANOVA was carried out comparing signal detection sensitivity (d’) across different combinations of character appearance duration and target frequency. The value of d’, is a measured used to describe the difference between a signal, and signal and noise (Equation 6-2)
(Wickens, 2001). There was no significant interaction between the character appearance duration and target frequency factors. However, there was a significant main effect of character appearance duration, $F(5,78) = 15.855, p < 0.001$. Figure 6-1 shows that performance, as measured by $d'$, was influenced by character appearance duration. With character appearance durations lower than about 800 msec, the task appeared to be a speeded task where accuracy depended on the amount of time that the character was available to be seen and hit. In contrast, when the character appearance duration was longer than 800 msec, sensitivity/accuracy no longer benefitted from seeing the character for a longer period of time.

$$d' = z(\text{hits}) - z(\text{false alarms})$$

**Equation 6-1. Signal detection sensitivity ($d'$) calculation.**

![Graph showing the impact of character appearance duration on mean signal detection sensitivity ($d'$). Error bars represent +/- 1 standard error.](image)

**Figure 6-1. The impact of character appearance duration to mean signal detection sensitivity ($d'$) value.**

A two-way ANOVA was carried out comparing the median correct RT based on the six character speeds and three target frequencies. There was a significant main effect of character appearance duration, $F(5,86) = 9.852, p < 0.001$ (Figure 6-2). The median correct RT was longer with a character appearance duration of 600 msec compared to 500 msec (368 msec ± 204, $p = 0.024$). The median correct RT was longer with a character appearance duration of 800 msec compared to 700 msec (563 msec ± 004, $p = 0.034$). There was no significant interaction effect between character appearance duration and target frequency.
As can be seen in Figure 6-2, median correct RT was quicker for the shorter character appearance durations which should not be surprising since there was less time to hit the target (e.g., mole) while it was still showing (i.e., the shorter median correct RT for the shorter character appearance durations is a necessary response to the task demands). The pattern of results in Figure 6-2 show that by the time character appearance duration increased to 800 msec, the speed of responding was no longer being driven as much by the character appearance duration. The data are plotted for the three different target frequencies in Figure 6-2 and it can be seen that the same general trend of having a speeded task up until around 800 msec holds for all three target frequencies. There was no significant main effect of the target frequency.

**Figure 6-2. Interaction plot comparing character appearance duration to mean of median correct RT. The different lines represent target frequency.**

A two-way ANOVA was carried out comparing the miss rate based on the six character appearance durations and three target frequency levels. There was a significant main effect of character appearance duration, $F(5,90)=22.643, p<0.001$ (Figure 6-3). The miss rate was lower with a character appearance duration of 600 msec compared to 500 msec (57 msec ± 14, $p<0.001$). Neither the main effect of the target frequency nor the interaction effect of target frequency was significant. The highest miss rate occurred with a 500 msec character appearance duration as this time was too quick for participants to consistently identify the target and hit it in time. As might be expected, the next highest miss rate was for a character speed of 600 msec. The trends in the increasing median correct RT (Figure 6-2) and decreasing miss rate (Figure
6-3), as character speed increases, suggests that there is a speed-accuracy trade-off observed in the data.

![Mean Miss Rate Based on Character Appearance Duration](image)

**Figure 6-3.** Interaction plot comparing character speed to mean probability of a miss. The different lines represent target frequency.

### 6.2.1 Game Parameter Settings

The choice of game parameter combinations used in the main part of the study was guided by previous research findings concerning the properties of relevant parameters. In this section, I briefly summarize the thinking behind the parameter selections made.

In target identification tasks (e.g., oddball tasks or Go/No-Go discrimination tasks), correct RT by healthy adults ranges between 300 and 600 msec (Madden et al., 2004; Thorpe, Fize, & Marlot, 1996). In detection response time tasks, movement time has been observed to range between 150 to 200 msec (Wyrick Spirduso, 1975). In earlier chronometric research, (e.g., (Posner, 1978)) a subtractive/additive approach was used to calculate the time required for different processing stages in RT tasks such as stimulus discrimination, and motor execution (Miller & Low, 2001). Using an additive approach, Posner (1978), demonstrated that the required RT in a task that requires both target identification and movement should range between an absolute minimum of 450 msec (300 msec for target identification + 150 msec for movement time) and a maximum of 800 msec or so where the task is transitioning to an unspeeded mode. This RT range is consistent with the findings from the pilot study that the whack-a-mole task
transitioned from a speeded to unspeeded task at a character appearance duration of around 0.8 sec. In the main experiment, this appearance time for the character appearance duration was contrasted with a higher value of 1200 msec, which represented a largely unspeeded task.

Based on my earlier research with older adults in an ED, median correct RT was typically faster with smaller grid sizes than larger grid sizes. In Tong and Chignell (2014), I observed strong correlations between performance on the game, measured using a combined performance metric, with inhibition ability in both grid sizes of 2x2 and 3x3. In this study, I used a 3x3 grid size to make the game more interesting. I also thought that there would be less likelihood of observing floor effects (where the task is not sufficiently difficult to detect cognitive differences) using the 3x3 grid, in a population of university students (who might be less likely to have cognitive impairments).

The experiment contrasted two target frequencies, 60% and 85%. With the 60% rate, targets and distractors are close to equi-probable and the participant will tend to look for the possibility of a distractor on every trial, which will encourage slower and more accurate performance. With the 85% target rate, distractors occur relatively infrequently and the participant is more likely to expect/anticipate targets, thereby speeding responses, but also leading to errors when distractors do occur. The 85% distractor frequency condition used in this experiment was similar to the 83% target frequency condition used by Jones, Cho, Nystrom, Cohen and Braver (2002) in their Go/No-Go discrimination task. They demonstrated that high target frequency conditions (83% targets to 17% distractors, versus a comparison condition where these frequencies were reversed) were associated with shorter correct RT and longer incorrect RT for hitting distractors, and higher error rates for hitting distractors. Based on the results by Jones et al. (2002), I hypothesize that higher target frequencies will result in shorter RT and higher error rates, in the context of game performance in this study.

The experiment also contrasted similar and dissimilar distractors. I hypothesized that longer RT would occur with levels/blocks that had similar targets and distractors (moles and moles with hats) as compared to blocks with dissimilar targets and distractors (moles and butterflies). This expectation was based on previous research findings that more discrimination time is required to recognize and distinguish between similar features. For instance, previous work on serial search by Treisman and Gormican (1988), demonstrated increased RT with similar targets and
distractors, as time is needed to process the stimulus and determine which mapping it belongs to, either a user’s mental model of a target or a distractor. Research by Verghese and Nakayama (1994) demonstrated the impact of target discriminability on RT by examining changes in orientation, color, and spatial frequency. They observed longer responses with less discriminable targets (i.e., smaller orientation angles, less color contrast and smaller spatial frequencies). Further findings by Woods, Wyma, Yund, Herron and Reed (2015) looked at the impact of target color and shape on RT. They observed higher error rates for similar distractors compared to dissimilar distractors. The findings from a number of studies (e.g., Treisman and Gormican (1988), Verghese and Nakayama (1994), and Woods et al. (2015)) are in accordance with the hypothesis that longer RT will be observed with similar distractors as compared to dissimilar distractors.

6.2.2 Research Hypotheses

I expected that using a shorter character appearance duration creates a speeded task, where participants respond faster compared to tasks with longer character appearance durations. I anticipated that dissimilar distractors take less time to identify compared to similar distractors, which require more processing time. Thus, in game blocks with both shorter character appearance duration and dissimilar distractors, I would expect shorter RTs but higher error rates. Higher error rates would be expected based on a speed-accuracy trade-off. In distractor trials on these blocks, I expected that participants would develop a response bias towards targets and would be more likely to erroneously hit a similar distractor due to having insufficient time to make a correct decision.

Research findings by Jones et al. (2002) suggest that RT is shorter in higher target frequency conditions whereas errors are increased. Thus, I expected to see shorter RT and higher false alarms in game conditions with higher target frequencies. In shorter character appearance duration conditions, I would expected that participants would have less time to make a perceptual judgment and inhibit their pre-potent response when shown a distractor. Thus, I expected that a combination of shorter character appearance duration and higher target frequency games would result in a disproportionately large increase in false alarm rates.

I now list three hypotheses that are motivated by the preceding discussion, with the third hypothesis being based on the expectation that RT will be more sensitive to the effects of target
frequency and distractor similarity when the task is less speeded (i.e., for the longer character appearance duration).

1. If a minimum movement time above 800 msec marks the transition from a speeded to unspeeded task, and similar distractors require more processing time, then I predict that the presence of an interaction between character appearance duration with distractor type.
   a. If similar distractors are more difficult to distinguish, and a minimum RT is required for target identification, then I expect that there will be a disproportionate increase in errors (e.g., increased miss rates and increased false alarm rates) in game blocks with a shorter character appearance duration combined with similar distractors.

2. If higher target frequencies lead to response bias as well as slower and more accurate performance, then I predict that there will be an interaction of character appearance duration with target frequency.
   a. If lower target frequencies lead to more errors (e.g., misses and false alarms) and slower RT, then I expect that there will be a disproportionate increase in misses in game blocks with lower target frequency conditions and shorter character appearance duration.

3. If higher target frequencies lead to slower and more accurate performance, and identifying similar distractor types requires more processing time, then I predict that there will be an interaction between target frequency and distractor type for unspeeded (1200 msec) trials.
   a. If higher target frequency results in slower RT for correct responses, and similar distractors requires more processing time, then I predict that there will be a disproportionate increase in median correct RT in game blocks with similar distractors and higher target frequency.
   b. If higher target frequency lead to response bias and similar distractors are require more discrimination time, then I predict that there will also be a disproportionate increase in false alarms in game blocks with similar distractors combined with high target frequency.
In addition to testing these hypotheses, I was interested in which of several game performance measures (e.g., median correct RT, false alarm rate, combined performance metric) would be most strongly correlated with cognitive ability, and whether the performance measure with the strongest correlation would depend on the combination of game parameters that was used.

6.3 Research Question

In addition to the hypotheses relating to the impact of game parameters on performance that were noted in the previous section, the study reported in this chapter will address the following research question relating to the construct validity of the game as a measure of inhibition ability:

1. Is game performance correlated with general inhibitory EF, or is it associated with one or more of the specific subtypes of inhibition that have been identified?

The research question is addressed through a correlation analysis (using both Pearson’s r and Spearman’s rho), with relationships between the game performance measures and the different measures of cognitive performance (e.g., Stroop task, Go/No-Go discrimination task, antisaccade task, and Flanker task). The purpose of using four different measures of inhibition ability is that each inhibition task has task impurities that reflect other attributes asides from inhibition ability.

6.4 Methodology

6.4.1 Study Sample

Thirty participants (14 females, 16 males) between the ages of 19–35 years (mean age = 24.9 years, median age = 25.0 years) completed this study. All participants were recruited through the University of Toronto and through local ads placed on online forums (i.e., Craigslist and Kijiji).

6.4.2 Study Design

The research protocol was approved by the University of Toronto research ethics board (protocol #35586). All participants completed the experiment individually. The study was divided into two phases with the duration of the study varying between 1 and 1.5 hours for each participant. In the first phase, all participants completed four cognitive tasks using the PEBL2 software suite (The PEBL Project, 2018), on a desktop computer with a computer mouse and keyboard. The four
tasks included a color-word Stroop task, a Go/No-Go discrimination task, an anti-saccade task, and a Flanker task. Each cognitive task began with a practice block followed by the experimental test block. The implementation of each of these tasks is described in Appendix F – Section 8.6.2. Participants were instructed to take breaks between each of the four cognitive tasks. This phase of the study took approximately 30 minutes to complete. The cognitive ability tasks were measured at the beginning of the experiment to ensure the most accurate assessment of cognitive ability (prior to the onset of any experimentally induced fatigue).

Four different measures of inhibition were selected as each inhibition task has task impurities that reflect other attributes aside from inhibition ability. As discussed in the literature review (Chapter 2 – Section 2.2), different inhibition tasks such as the Stroop task measures interference control whereas the Go/No-Go discrimination task is a measure of a response inhibition. Performance on inhibitory tasks such as the Stroop task has also been investigated in patients with delirium assessed using the CAM (Smith et al., 2009). The rationale in administering four different tasks is to investigate whether game performance is associated with one or more specific subtypes of inhibition that has been identified.

6.4.2.1 Phase 1: Cognitive Task

The performance on the four cognitive ability tasks was analyzed using median correct RT and error rates for the anti-saccade, Flanker and Go/No-Go discrimination task. These error rates were obtained by pooling data within each of the eight experimental blocks (see section 6.4.2.2 for an explanation of the eight blocks used for each participant). For the Stroop task, an interference value was calculated, as will be described below. A summary of the cognitive tasks and corresponding measures can be found in Table 8-18.

In the anti-saccade task, RTs below 200 msec and above 2000 msec were removed with responses lying outside those cut-off values being either too quick to make a meaningful decision, or too slow to be representative of the time needed to make a decision (e.g., the participant may not have been paying attention). A similar minimum cut-off approach was used by Miyake et al. (2000b), who only considered RT above 200 msec for correct trials. The choice of a 2000 msec cut-off at the high end of the RT distribution was based on inspection of the RT data (histogram) and the fact that the frequency of responses dropped off greatly after 2000 msec had elapsed. In the Flanker task, RT below 200 msec were also removed. The Go/No-Go
discrimination task consisted of two blocks. The first block was a high target frequency condition (0.8), and the second block was a low target frequency condition (0.2). For block 1, RT below 200 msec and above 1000 msec were removed based on inspection of the RT distribution, with RTs above 1000 msec appearing to be outliers. For block 2, RT below 200 msec were removed again, based on review of the distribution. For block 2, RT above a certain threshold value were not removed (the maximum RT was 1445 msec), as responses on block 2 were likely to be longer due to decreased target frequency and change in the go-target. Based on the RT histogram distribution, there was a fairly large proportion RTs that were greater than 1000 msec for block 2.

Performance on the Stroop task was analyzed using the Stroop interference value as calculated by subtracting the median correct RT on congruent trials (i.e., trials where the stimulus word and colour match) from the RT on incongruent trials (i.e., trials where the stimulus word and colour do not match) (Equation 6-2).

\[
\text{stroop interference} = RT_{\text{incongruent trials}} - RT_{\text{congruent trials}}
\]

Equation 6-2. Stroop interference calculation.

For all trial types in the Stroop task, RT below 200 msec and above 3500 msec were removed based on inspection of the RT distribution. A similar data trimming approach was taken by Miyake et al. (2000b), whereby RTs less than 200 msec were trimmed on tasks with multiple trials. Miyake et al. (2000b) used a shorter upper criterion of 2000 msec for the Stroop task, which I chose not to replicate because a significant number of responses in the Stroop task in this study were above 2000 msec.

In the second phase of the study, all participants played the game on a tablet. The game was played on a 10.1-inch Samsung Galaxy Tab 4 tablet (Samsung, 2006) in portrait orientation. Participants were provided with instructions before the beginning of each game block.

6.4.2.2 Phase 2: Game Block Design

This study used a full factorial design with 8 blocks of trials, representing all possible combinations for the three factors (2 target-to-distractor ratios, x 2 types of distractors x 2 character appearance durations). The target frequency was held constant for the first four blocks,
and within these blocks, the remaining two factors (2 types of distractors x 2 character appearance durations = 4 combinations) were varied. There were 4! (i.e., 24) unique orders of these four combinations and in the first four blocks (where the first target frequency is assigned) each of the first 24 participants received a unique ordering of the four blocks (i.e., the four combinations of distractor type and character appearance duration). The remaining six participants received one of the 24 unique orders selected at random, with the constraint that none of the orderings was repeated within the final six participants. Each participant saw a unique ordering of the 4 blocks (2 types of distractors x 2 character appearance durations = 4 combinations) of trials with one target frequency, and then saw those same four blocks but in reverse order, accompanied by the other target frequency (not seen in the first four blocks) in the second half of the experiment (blocks five through eight). Half of the participants saw low target frequencies in the first four blocks of the experiment and half saw the high target frequencies first. The ordering of the 8 blocks as defined above was completed once, then repeated a second time for each participant. In total, each participant repeated each of the 8 combinations of game settings twice (8 blocks x 2). Performance on the two instances of each of the block types for each individual was averaged to create the person’s summary statistics for that block type.

6.4.2.3 Game block information

For each game, the character size was held constant at 175 px with a grid size of 3x3 (Figure 8-7). There were 60 trials per game block. Each character appeared for either 800 msec or 1200 msec long. Only one character/stimulus appeared on the game board at a time. In summary, the second phase of the study evaluated eight different variations of the game. The eight game blocks varied based on all possible combinations of three dichotomous factors:

1. Factor 1: Proportions of targets to distractors
   a. 60% target trials, 40% distractor trials
   b. 85% target trials, 15% distractor trials
2. Factor 2: Type of distractor
   a. Moles with hats (similar distractors)
   b. Butterflies (dissimilar distractors)
3. Factor 3: Character appearance duration
   a. 800 msec
6.5 Results

6.5.1 Game Performance

Performance on the game was measured using median correct RT, false alarm rates (Equation 3-1), miss rates, and a combined game performance metric (Equation 3-2).

6.5.2 Hypothesis Testing

A repeated-measures multivariate analysis of variance (MANOVA) was carried out to investigate the following three hypotheses described earlier. In the MANOVA, the following four dependent variables were used: median correct RT, false alarm rates, miss rates, and combined performance metric. The impact of three game parameter factors: character appearance duration, target frequency, and distractor type, was assessed on these measures. There was a significant (multivariate) main effect (using Pillai’s trace) of the three factors on the four outcome game measures:

1. character appearance duration, $V=0.680$, $F(4,26)=13.794$, $p<0.001$,
2. target frequency, $V=0.913$, $F(4,26)=67.839$, $p<0.001$, and
3. distractor type, $V=0.749$, $F(4,26)=19.402$, $p<0.001$.

Pillai’s trace is reported in this case as it is the most powerful MANOVA test statistic when sample sizes are small (Field, 2013). The other multivariate test statistics, (Roy’s largest root, Hotelling’s trace, and Wilks’ lambda), were also significant for the above three main effects.

6.5.2.1 Hypothesis 1: Interaction of Character Appearance Duration with Distractor Type

There was no significant interaction effect of character appearance duration and distractor type, $V=0.136$, $F(4,26)=1.026$, $p=0.412$.

6.5.2.2 Hypothesis 2: Interaction of Character Appearance Duration with Target Frequency

There was an interaction of character appearance duration with target frequency, $V=0.549$, $F(4,26)=7.924$, $p<0.001$. Univariate ANOVAs were carried out to determine how the dependent
variables differed for this interaction effect. In higher target frequency conditions with a shorter character appearance duration, the median correct RT was disproportionately faster, $F(1,29)=6.550, p=0.016$ (see Figure 6-4), and the combined performance metric was disproportionately higher, $F(1,29)=6.493, p=0.016$ (see Figure 6-6).

6.5.2.2.1 Hypothesis 2a

In the significant interaction of character appearance duration with target frequency, there were disproportionately more misses, $F(1,29)=10.350, p=0.003$ in higher target frequency conditions combined with a shorter character appearance duration (see Figure 6-5).

![Figure 6-4. Interaction plot for median correct RT based on character appearance duration, grouped by target frequency.](image)

84
Figure 6-5. Interaction boxplot for miss rates based on character appearance duration, grouped by target frequency.

Figure 6-6. Interaction boxplot for combined performance metric based on character appearance duration, grouped by target frequency.
6.5.2.3 Hypothesis 3: Interaction of Target Frequency with Distractor Type

There was a significant interaction effect of target frequency and distractor type on the game outcome measures, $V=0.486$, $F(4,26)=6.140$, $p<0.01$. Univariate ANOVAs were carried out to determine the effect of the interaction on the dependent variables.

6.5.2.3.1 Hypothesis 3a

In the significant interaction effect of target frequency and distractor type, there was an increase in median correct RT, $F(1,29)=10.092$, $p=0.004$, in game blocks with similar distractors and higher target frequency (Figure 6-7). There was a disproportionate (i.e., greater than a simple sum of the main effects of the two factors) increase in median correct RT in game blocks with similar distractors and higher target frequency.

6.5.2.3.2 Hypothesis 3b

As predicted, in the significant interaction effect of target frequency and distractor type there was a disproportionately large increase in false alarms, $F(1,29)=15.887$, $p<0.001$ (Figure 6-8), for the case of similar distractors combined with higher target frequency.

Together, Figure 6-7 and Figure 6-8 (i.e., for median correct RT and for false alarm rates) show that the impact of target frequency is much greater in the similar distractor condition, where median correct RT was almost as fast as high target frequency responses in the dissimilar distractor condition, while the corresponding false alarm rate (for high target frequency and similar distractors) was much higher than it was for the other combinations of distractor type and target frequency.
Figure 6-7. Interaction boxplot for median correct RT based on distractor type, grouped by target frequency.

Figure 6-8. Interaction boxplot for false alarm rates based on distractor type, grouped by target frequency.

6.5.3 Summary of Hypothesis Testing Results

The first hypothesis was not supported in this study (I proposed that there would be an interaction between character appearance duration with distractor type). The hypothesis was that disproportionately more errors would occur when there was a shorter character appearance
duration and similar distractor since there would be on a speed-accuracy trade-off as individuals would have less time to make a difficult judgment between a target and similar distractor. Possible reasons for the lack of interaction may be that the 800 msec duration did not speed the task sufficiently for undermining the processing of similar distractors, or alternatively, the duration of 1200 msec was not sufficiently long to make processing of similar distractors an unspeeded task.

I observed a significant interaction between target frequency and character appearance duration that supported hypothesis 2. I found that median correct RT was disproportionately higher in game conditions with high target frequency and shorter character appearance duration. Hypothesis 2a was supported as I observed much higher miss rates in game conditions with higher target frequency and shorter character appearance durations. In speeded conditions, participants were more likely to miss a target when they appeared with a shorter character appearance duration on high frequency trials, likely because they were required to both identify the stimulus and respond correctly as quickly as possible. Longer character appearance durations enabled participants to spend more time processing the stimulus before responding, thereby reducing the likelihood of a miss.

I observed a significant interaction between target frequency and distractor type that supported hypothesis 3. Hypothesis 3a was supported as I observed disproportionately longer median correct RT in conditions with a high target frequency and similar distractors. Hypothesis 3b was supported by a particularly high false alarm rate in higher target frequency and similar distractor type conditions. Participants likely had an increased pre-potent response bias in higher target frequency games along with less time to detect similar distractors, which resulted in increased false alarms.

Based on the hypothesis testing reported above (Table 6-1), game performance worsened in terms of median correct RT, and errors (e.g., miss rates and false alarm rates) tended to increase in more difficult game settings, which were characterized by games with a higher target frequency, similar distractor types, and shorter character appearance duration.

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Outcome</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Character appearance duration</td>
<td>Distractor similarity</td>
<td>Median correct RT</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Character appearance duration</td>
<td>Target frequency</td>
<td>Median correct RT</td>
<td>Yes</td>
</tr>
</tbody>
</table>
6.5.4 Cognitive Ability Performance

To understand the underlying relationships between the cognitive measures, Pearson’s \( r \) and Spearman’s \( \rho \) bivariate correlations were calculated between the four cognitive tasks with the resulting correlations shown in Table 6-2. There were no significant correlations between the Stroop interference value and the other cognitive ability measures. The median correct RT on the anti-saccade was moderately correlated with the median correct RT on the Go/No-Go discrimination task in high target frequency trials (block 1) \((r=0.378, \rho=0.425)\). However, the strongest correlation was between the error rates on the anti-saccade and Flanker tasks \((r=0.994, \rho=0.919)\), suggesting that these tasks are largely testing the same underlying construct.

There were a number of significant correlations between the Go/No-Go discrimination task measures. The median correct RT on the Go/No-Go discrimination task in high target frequency trials (block 1) was correlated with the corresponding error rate \((r=-0.645, \rho=-0.384)\), and with the median correct RT \((r=0.614, \rho=0.561)\), and error rate \((r=-0.598, \rho=-0.487)\) in the low target frequency trials (block 2). The error rate on the Go/No-Go discrimination task in high target frequency trials (block 1) was correlated with the error rate in low target frequency trials (block 2) \((r=0.666)\). The median correct RT on the Go/No-Go discrimination task in low target frequency trials (block 2) was correlated with the corresponding error rate \((r=-0.466, \rho=-0.402)\), and with the median correct RT \((r=0.444)\) on the Flanker task.

In summary, there was a high correlation between anti-saccade and flanker performance \((r>0.9)\), and strong correlations amongst the Go/No-Go discrimination measures \((r<0.6)\). There were only weak correlations between Stroop interference and the other cognitive tasks \((r<0.3)\) While there were moderate \((r=0.4)\) correlations between the Go/No-Go discrimination measures and flanker/anti-saccade), the most parsimonious interpretation of the overall pattern of correlation results is that there are three types of inhibition represented in these four tasks, one for the Stroop, one for Go/No-Go discrimination, and one for the combination of Flanker and anti-saccade tasks.
6.5.5 Structure of Cognitive Tasks

Based on inspection of the correlation analysis in Table 6-2 as well as expected relationships between the tasks based on previous research results and theories, there are three underlying factors for the four tasks that I used, which are labeled as different subtypes of inhibition below:

1. Perceptual inhibition (Flanker and anti-saccade error rates),
2. Processing inhibition (Stroop interference), and
3. Response inhibition (Go/No-Go discrimination task performance measures).

The error rate on the Flanker and anti-saccade tasks were strongly correlated with each other. These tasks both require selective attention and pro-saccades/anti-saccades (Diamond, 2013). Second, based on Table 6-2, I observed that the Stroop interference value was not strongly correlated with any other cognitive performance measure. Third and last, the Go/No-Go discrimination performance measures (median correct RT and error rate) are strongly correlated with each other (a total of four measures across the two blocks that varied according to target frequency). The correlation amongst the Go/No-Go discrimination task variables likely reflects response inhibition as described by (Diamond, 2013).

For the perceptual inhibition construct, I created a combined scale that averaged the error rate from the anti-saccade and Flanker task (“perceptual inhibition”).

<table>
<thead>
<tr>
<th>Table 6-2. Two-tailed correlations (n=30) between cognitive measures (Pearson’s r appear on the top row and Spearman’s rho are on the bottom row in italics, in each cell). Significant values greater than ±0.5 are bolded. * p&lt;0.05, ** p&lt;0.01, *** p&lt;0.001. AS = anti-saccade task, F = Flanker task, GNG = Go/No-Go discrimination task, and S = Stroop task.</th>
</tr>
</thead>
<tbody>
<tr>
<td>S Interference</td>
</tr>
<tr>
<td>S Interference</td>
</tr>
<tr>
<td>AS Median Correct RT</td>
</tr>
<tr>
<td>AS Error Rate</td>
</tr>
<tr>
<td>F Median Correct RT</td>
</tr>
<tr>
<td>F Error Rate</td>
</tr>
<tr>
<td>GNG Median Correct RT Block 1</td>
</tr>
<tr>
<td>GNG Error Rate Block 1</td>
</tr>
<tr>
<td>GNG Median Correct RT Block 2</td>
</tr>
<tr>
<td>GNG Error Rate Block 2</td>
</tr>
</tbody>
</table>
The next part of the analysis was concerned with the question of how to characterize the response inhibition factor. I focused on block 1 of the Go/No-Go discrimination task, as it has high target frequency, and response inhibition should be higher in that case. I combined the median correct RT and error rate measures using a z-score transformation (Equation 6-3).

\[
\text{Go/No-Go discrimination task combined z-score value} = -z(\text{median correct RT}) - z(\text{error rate})
\]

**Equation 6-3. Combined metric for the median correct RT and error rate on the Go/No-Go discrimination task in high target frequency trials (block 1).**

### 6.5.6 Game Performance

To gain an understanding of how the game performance measures across all eight game blocks was related, I carried out two principal component analyses (PCA). In the PCAs described below, all analyses used an orthogonal rotation (varimax criterion). The first PCA was carried out using the median correct RT on all eight blocks, and the second PCA included the false alarm rate on all eight blocks.
6.5.6.1 Median Correct RT PCA

The Kaiser-Meyer-Olkin (KMO) measure verified the sampling adequacy for the analysis, KMO=0.93 ('superb' according to Field (2013)), and all KMO values for individual items were >0.89. Bartlett’s test of sphericity, $\chi^2(28)=325.65$, $p<0.001$, indicated that correlations between items were sufficiently large for a PCA.

An initial analysis was run to obtain eigenvalues for each component in the data. Only one component had an eigenvalue over Kaiser’s criterion of 1, which explained 86.8% of the variance, and thus one component was retained in the final analysis. Table 8-19 shows the corresponding factor loadings. The game median correct RT component had a high reliability, Cronbach’s $\alpha$ of 0.98. Thus, I will be using the mean of the median correct RT across all eight game blocks in analyses reported below.

6.5.6.2 False Alarm Rate PCA

The KMO measure verified the sampling adequacy for the analysis, KMO=0.77 ('mediocre’ according to Field (2013)), and all KMO values for individual items were >0.57. Bartlett’s test of sphericity, $\chi^2(28)=156.98$, $p<0.001$, indicated that correlations between items were sufficiently large for a PCA.

An initial analysis was run to obtain eigenvalues for each component in the data. Three components had eigenvalues over Kaiser’s criterion of 1 and in combination explained 85.3% of the variance. The scree plot showed an inflection after the first component. Given the small sample size, and Kaiser’s criterion (Field, 2013), I determined that the false alarm rates across the different blocks was best explained as being due to a single underlying factor.

I then re-ran the PCA with forced extraction of one factor with the false alarm rate on the eight game blocks, with the factor loadings shown in Table 8-20. An item reliability analysis was carried out by including all false alarm rates from the eight game blocks. This component had a high reliability, with a Cronbach’s $\alpha$ of 0.81. Thus, I will be using the mean of the median false alarm rate across all eight game blocks as a measure of false alarm rate in subsequent analyses.
6.5.7 Relationship Between Game Performance and Cognitive Ability

A correlation analysis using Pearson’s $r$ and Spearman’s $rho$ was carried out between the response inhibition subscale (Equation 6-3) and measures of game performance (Table 6-3). The mean of the median game correct RT across all eight game blocks was strongly correlated with the response inhibition subscale ($r=-0.505$, $rho=-0.476$). The mean of the median game false alarm rate across all eight game blocks was strongly correlated with the response inhibition subscale ($r=0.574$, $rho=0.493$). The combined game performance metric (Equation 3-2) was particularly strongly correlated with the response inhibition subscale and accounted for close to half the variance observed in the response inhibition subscale ($r=-0.691$, $rho=-0.606$).

To compare the strength of each correlation between each game performance measure with response inhibition, Fisher’s $z$-transformations were carried out. The correlation between the mean of median game correct RT across all game blocks and response inhibition ($r=-0.505$, $rho=-0.476$) was not significantly lower than the correlation between the game combined performance metric with response inhibition ($r=-0.691$, $rho=-0.606$), $z=-1.08$, $p=0.14$. The correlation between the mean of median game false alarm rates across all game blocks and response inhibition ($r=0.574$, $rho=0.493$) was not significantly lower than the correlation between the game combined performance metric with response inhibition ($r=-0.691$, $rho=-0.606$), $z=-0.722$, $p=0.235$.

Table 6-3. Two-tailed correlations ($n=30$) between game performance measures and response inhibition assessed by the combined metric on block 1 of the Go/No-Go discrimination task (Equation 6-3). Significant values greater than +/- 0.5 are bolded. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

<table>
<thead>
<tr>
<th>Response Inhibition</th>
<th>Response Inhibition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r-values)</td>
<td>(rho-values)</td>
</tr>
<tr>
<td>Mean of Median Game Correct RT Across All Game Blocks</td>
<td>$-0.505^{**}$</td>
</tr>
<tr>
<td>$p&lt;0.001$</td>
<td>$p=0.00778$</td>
</tr>
<tr>
<td>Mean of Median Game False Alarm Rates Across All Game Blocks</td>
<td>$0.574^{**}$</td>
</tr>
<tr>
<td>$p&lt;0.001$</td>
<td>$p=0.0566$</td>
</tr>
<tr>
<td>Game Performance Metric Across All Game Blocks</td>
<td>$0.691^{**}$</td>
</tr>
<tr>
<td>$p&lt;0.001$</td>
<td>$p=0.00387$</td>
</tr>
</tbody>
</table>

As the game performance metric had the strongest relationship with response inhibition, I carried out another correlation analysis looking at the game performance metric on each game block, and the relationship with response inhibition (Table 6-4). All game blocks were significantly
correlated using the game performance metric with response inhibition ($r = -0.510$ to $0.698$, $\rho = 0.469$ to $0.599$), with game performance on block 2 having the strongest correlation ($r=0.698$, $\rho=0.582$).

Table 6-4. Two-tailed correlations between the game performance metric (Equation 3-2) for each game block and response inhibition assessed by the combined metric on block 1 of the Go/No-Go discrimination task (Equation 6-3). Significant values greater than ± 0.5 are bolded. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

<table>
<thead>
<tr>
<th>Game Performance Metric for Game Block 1</th>
<th>Response Inhibition (r-values)</th>
<th>Response Inhibition (rho-values)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.620 ***</td>
<td>0.570 **</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.001$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 2</td>
<td>0.698 ***</td>
<td>0.582 ***</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 3</td>
<td>0.510 **</td>
<td>0.455 *</td>
</tr>
<tr>
<td></td>
<td>$p=0.004$</td>
<td>$p=0.115$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 4</td>
<td>0.641 ***</td>
<td>0.599 ***</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p&lt;0.001$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 5</td>
<td>0.626 ***</td>
<td>0.537 **</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.0022$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 6</td>
<td>0.645 ***</td>
<td>0.513 **</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.0037$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 7</td>
<td>0.567 **</td>
<td>0.560 **</td>
</tr>
<tr>
<td></td>
<td>$p=0.0011$</td>
<td>$p=0.0013$</td>
</tr>
<tr>
<td>Game Performance Metric for Game Block 8</td>
<td>0.573 ***</td>
<td>0.469</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.0089$</td>
</tr>
</tbody>
</table>

6.6 Discussion

At the beginning of this research study, I proposed three hypotheses based on game parameters and the resulting game performance. The first hypothesis was that there would be an interaction between character appearance duration and distractor type. I proposed that there would be a
disproportionate increase in errors (e.g., misses and false alarms) in game blocks with a shorter character appearance duration with similar distractors. I was unable to find evidence from results of the MANOVA to support this hypothesis, and rejected it.

My second hypothesis was that there would be an interaction between character appearance duration and target frequency. I proposed that there would be disproportionate increase in errors (e.g., increased miss rates) in game blocks with lower target frequency with shorter character appearance duration. Based on the results of the MANOVA, I observed that there was an increase in miss rates, in game blocks with a lower target frequency with shorter character appearance duration. Thus, the second hypothesis is supported. The decreased trial character appearance duration combined with the time needed for the perceptual inhibition task therefore likely increased the probability of a participant missing the time frame to make the correct decision in the time available. Participants may have had a response bias where they avoid hitting the distractor at the expense of not hitting some targets. The visual design of the game may have contributed to this response bias, since false alarm errors received visual feedback (a large X over the character) whereas miss errors were not accompanied by visual feedback. The issue of whether the form of visual feedback, when making different kinds of error, influences performance or the validity of the game as a cognitive assessment tool may be an interesting topic for future research. The present findings indicate that game difficulty can be increased by using a shorter character appearance duration and higher target frequency.

My third and final hypothesis was that there would be an interaction between target frequency and distractor type. I expected that, in trials with longer (1200 msec) character appearance durations, there would be a disproportionate (i.e., greater than a simple sum of the main effects of the two factors) increase in median correct RT in game blocks with similar distractors and higher target frequency. I also expected that there would be a disproportionate increase in false alarm rates in game blocks with similar distractors combined with high target frequency. From the MANOVA, I observed a significant interaction between target frequency and distractor type, supporting Hypothesis 3. In a follow up univariate analysis, I found a disproportionate increase in correct RT in game blocks with similar distractors and higher target frequency. This finding suggests that when participants had more time to respond with the longer character appearance duration they took a disproportionately longer time to respond in higher targets frequency trials. They were able to respond in this way because the task was no longer “speeded” by the shorter
character appearance duration and they had more time to determine if the similar distractor was different from the target. There was also a disproportionate increase in false alarms rates in game blocks with similar distractors and higher target frequency. This finding suggests that participants were primed and had a pre-potent response bias to targets when the target frequency was higher, and had a more difficult time inhibiting their response when the distractor was similar to the target, which resulted in higher false alarm rates. I also note that response bias appears to be malleable depending on the experimental conditions, with greater task difficulty leading to more misses versus false alarms in some cases, and the reverse in other cases.

6.6.1 Cognitive Ability Performance

The correlation analysis performed on the cognitive ability measures generally mirrored the relationships between tasks that have been found previously in the research literature (Table 6-2). For example, there were no significant correlations between the Stroop interference value with other measures of inhibition ability. Based on previous research on the Stroop task, the Stroop interference value measures interference control, which is not part of the processing activity required in other inhibition tasks such as the Flanker, anti-saccade and Go/No-Go discrimination tasks. In comparison, there was a significant relationship between the error rates on the Flanker and anti-saccade tasks, both of which require pro-saccade/anti-saccade responses and selective attention. Furthermore, the RT on the two Go/No-Go discrimination blocks were strongly correlated with each other, which reflects previous literature suggesting that the Go/No-Go discrimination task measures response inhibition. The low correlations observed between some of the inhibition tasks (e.g., between the Stroop task and other inhibition tasks) may have been a result of task implementation, and sampling method. Moreover, the results from the correlation analysis reflect the differences in the subtypes of inhibition ability as measured through different inhibition tasks. The results from the correlation analysis demonstrates that inhibition can be measured using different tasks and to varying degrees (i.e., variability in correlation values).

Based on the results from the correlation analysis and previous literature, I created two new subscales. In the first subscale called “perceptual inhibition”, I equally-weighted the error rate on the Flanker and anti-saccade task. This subscale is meant to reflect inhibition tasks that require perceptual judgment and processing of stimuli. The second and final subscale I created was labeled “response inhibition”, which combined both the RT and false alarm rate on block 1 of the
Go/No-Go discrimination task (Equation 6-3). This subscale was created as a measure of response inhibition.

6.6.2 Game Performance

To gain an understanding of how game performance measures varied based on different game blocks/parameters, I carried out two factor analyses on (1) the median game CRT and (2) the false alarm rates across all eight game blocks, respectively. For both analyses there appeared to be one underlying factor that accounted for performance on all eight game blocks.

6.6.3 Relationship Between Game Performance and Cognitive Ability

The combined game performance metric had the strongest correlation with the response inhibition subscale ($r=-0.691$, $\rho=-0.606$), followed by the mean of the median false alarm rate across all eight game blocks ($r=-0.574$, $\rho=-0.493$), and then by the mean of the median correct RT across all eight game blocks ($r=-0.505$, $\rho=-0.476$).

Using the combined game performance metric (Equation 3-2) all game blocks had statistically significant relationships with the response inhibition subscale. This was not surprising given that the median correct RT on all game blocks loaded onto one factor in the PCA.

6.6.4 Predicting Cognitive Performance Using Game Performance

Response inhibition was best predicted by using the game performance metric on game blocks, rather than the mean of median correct RT and the mean of false alarm rates across all eight game blocks. Using the game performance metric on all eight blocks as predictors to predict the response inhibition subscale, the most significant predictor was game block 2. This game block had a longer character appearance duration, similar distractors and lower target frequency. Based on the results from the hypothesis testing, it is reasonable to expect higher false alarms and longer median correct RT on this game block given the interaction between similar distractor type and lower target frequency.

My experimental findings demonstrate that the Go/No-Go discrimination task and Stroop interference are relatively uncorrelated with each other. In contrast, the Flanker and anti-saccade tasks were highly correlated with each other. This pattern can best be understood in terms of which processing system the inhibitory control is being directed towards. In the case of the anti-
saccade task prior research (Munoz & Everling, 2004), the anti-saccade task is associated with inhibition of the frontal eye fields, and the superior colliculus, regions associated with the planning of eye movements and visual attention. Inhibition of these same regions would also seem to be required for the Flanker task. In contrast, the Stroop task requires suppression of one of two active coding systems. The ACC seems to be strongly involved in this suppression. In contrast to these first two types of inhibition (inhibition of non-task related visual planning and control; inhibition of a non-task relevant representation), Go/No-Go discrimination involves response inhibition, inhibiting the (pre-potent) response to the more frequent target.

6.6.5 Design Recommendations

I propose the following design guidelines in creating tasks to measure response inhibition.

1. Use visually similar distractor stimuli to increase the use of perceptual judgment in discriminating between similar target and distractor stimuli.

2. Provide a longer stimuli duration such as 1200 msec, to allow a participant to respond. Decreasing stimuli duration may result in a speeded-task that results in higher error rates.

3. Varying target frequency between 0.60 and 0.70 provides sufficient go and no-go trials to estimate response inhibition. A target frequency of 0.85 seems to be too high as there are fewer chances for false alarms to recur and there may be less response inhibition on some trials as participants decide to take their chances on seeing a target, since there is a high probability that they will be correct on the current trial.

4. Use the game combined performance metric that combines z-scores of RT and false alarm rate. I recommend an equal weighting of these scores for the time being, but it would be useful for future research to determine how the relative weighting of false alarm rate and RT should (or should not) change depend on the observed false alarm rate for each participant.

6.6.6 Limitations

A study sample of 30 participants was used and the present results should be replicated and extended with larger samples in future. Additionally, only four inhibition tasks were compared with game performance and it is possible that game performance shares variance with other inhibition tasks that were not assessed in this study.
6.6.7 Further Research

Future research could expand on investigating the psychometric properties of the game by focusing on game blocks that use lower target frequency combined with similar distractors. It would also be useful to know what level of target frequency is best for the game when it is used to predict response inhibition ability. In hindsight, the high target frequency of 85% used in this study may have been too high, and it would be useful to explore a range of target frequencies between 60% and 75%.

6.7 Conclusion

Overall, the findings from this research study demonstrate that median correct RT and false alarm rates across all game eight blocks are closely strongly correlated with performance on the Go/No-Go discrimination task, which measures response inhibition. I also observed that a block that used lower target frequency, similar distractors, and longer character appearance duration had the strongest correlation with response inhibition in a population of normal younger volunteers. The implications of this study suggests that in using this game for cognitive screening, a participant only needs to play a version of the game using a low target frequency, and a character appearance duration of 1200 msec, with varying degrees of distractor similar, to best predict response inhibition ability.
Chapter 7

7 Conclusion

In this concluding chapter, I briefly review and discuss the results of the four studies and their implications. I then list the contributions of this thesis, review some of the limitations of the work, and present recommendations for future work.

7.1 General Discussion of the Results and their Implications

The first clinical study described in Chapter 4 showed the potential for the whack-a-mole game to be used as a tool for delirium screening of elderly patients in hospital EDs. In that case, a threshold value of median correct game RT was used as the screening criterion. Retrospective analysis of the data showed that a median correct RT cut-off of 1130 msec for distinguishing people at risk of having delirium had a specificity and sensitivity that were both above 80%. However, I was unable to confirm this result in the two later clinical studies because people with delirium were explicitly screened out of study two in order to more accurately measure test-retest reliability, and delirium was observed in only one of the participants in the third clinical study (second clinical validation of the serious game). Thus, while the use of the serious game as a delirium screening tool is promising, more research is needed to demonstrate the generalizability of the current threshold. In addition, future research should assess if consideration of false alarm rate as well as median correct RT will improve the effectiveness of the game as a screening tool for delirium.

In the first clinical study (first clinical validation of the serious game), median correct RT was found to correlate (around $r=0.6$) with the MMSE. Correlation with the MoCA was weaker (around $r=0.35$) and appeared to be related to speed of processing (based on the fact that the partial correlation with MoCA scores was insignificant once the variance shared with the CRT task was removed). In the second clinical study reported in the second half of Chapter 4, median correct RT on the game was shown to have strong test-retest reliability (around $r=0.8$ between adjacent administrations around 8 hours apart). In the third clinical study game performance was again found to correlate with the MMSE ($r=0.5$), but the strength of the relationship increased (to $r=0.6$) when the combined metric of game performance (which included the false alarm rate as well as median correct RT) was used instead of mean correct RT.
The correlation of median correct game RT with DI was lower (around $r=0.3$) in the first clinical study (first clinical validation of the serious game), but was stronger when using the game performance metric that combined the correct RT and false alarm rate data in the third clinical study ($r=0.5$). Overall, the results of the clinical studies demonstrate the ability of the game to predict scores on cognitive assessments (predictive validity).

In the fourth study (with healthy participants), I included four cognitive tasks commonly used to measure inhibition: the Stroop task, Go/No-Go discrimination task, Flanker task, and the anti-saccade task. The pattern of correlations found in the study reflected expectations from the prior literature on inhibitory EFs, indicating three different underlying forms of inhibition (perceptual inhibition, processing inhibition, and response inhibition) represented by the Flanker/anti-saccade, Stroop, and Go/No-Go discrimination tasks, respectively. In addition, results from this study demonstrated that designing game levels that are more difficult (i.e., games with a lower target frequency and similar distractor types), result in longer median correct RT and more false alarms, compared to easier game levels. The fourth study also provided some initial insights into the impact of different game parameters on the validity of the game as a measure of response inhibition ability. The strongest prediction of response inhibition by the game performance metric was found in block 2 which had low target frequency (and a long character appearance duration and high distractor similarity). This suggests that a target frequency of 0.85 may be too high and that a lower target frequency should be used. The results also indicated that a character appearance duration of 800 msec may be too short and that a longer duration (e.g., the 1200 msec duration used as the contrasting condition in my study) should better assess response inhibition ability. With the speeding effect below 1200 msec, due to the increased random error rate, more “noise” is introduced to the ability of the game to specifically detect response inhibition EF. Since I used only one type of similar distractor in this study, I recommend that two or more levels of distractor similarity be used in future studies to provide a wider range of testing response inhibition ability.

The main finding of the fourth study (with healthy participants) was that game performance was more strongly correlated with the Go/No-Go discrimination task than with the other inhibition tasks that were tested, indicating that, as expected, performance on the serious game is a predictor of response inhibition ability. The game performance metric also provided an effective way of capturing the response inhibition component of game performance. Due to the presence
of speed accuracy trade-offs in this kind of task it is recommended that the game performance metric be used as the preferred measure for capturing game performance since it should be able to put participants who respond slower but more accurately and participants who respond faster but less accurately on a more level playing field.

7.2 Contributions

The research reported in this thesis makes the following contributions.

1. Demonstrated the feasibility of using a serious game-based assessment tool to screen for delirium in a hospital ED.
2. Demonstrated the concurrent validity of the serious game in a clinical setting with older adults.
3. Found significant correlations with the MMSE, MoCA, and DI, as well as significant differences in median correct game RT between CAM positive and CAM negative patients.
4. Demonstrated the test-retest reliability of the serious game in a clinical setting with older hospital ED patients.
5. Found that serious game performance was shown to be predictive of response inhibition ability as measured by the Go/No-Go discrimination task. Initial findings concerning the impact of different game parameters on game difficulty and the correlation of game performance with response inhibition ability were also presented.
6. Found that false alarm rate, as a supplement to median correct RT was shown to increase the predictive power of the game in both clinical and healthy samples, with a metric that combined standardized versions of median correct RT and false alarm rate providing a good account of game performance.
7. Found that iterative game design to match the requirements of clinical use, and the resulting revised versions of the serious game were shown to be usable by the vast majority of elderly hospital ED patients.
8. Designed a serious game that, (at the time of this writing), is being used in five hospital EDs across Canada.
7.3 Limitations

The limitations of each of the three clinical studies and the final study with healthy participants have been discussed in detail at the end of each of the chapters. In the section, I briefly review some of the main limitations of the work.

Much of the data collection in this thesis (three out of the four studies) was carried out in the ED setting. In this setting, it is not possible to exert the kind of experimental control that occurs in a laboratory setting. An unpredictable pool of potential participants presented themselves at the ED, and in clinical study 1 in particular the game had to be modified a number of times to better suit the requirements of newly observed subgroups of patients. In spite of the revisions, I did my best to ensure that the game remained first and foremost a test of inhibitory EF.

While the Sunnybrook Health Sciences Centre ED may be typical of North American hospital EDs in many ways, it no doubt also has some unique properties. Since the three clinical studies in this thesis all took place in the Sunnybrook ED they will need to be replicated in other EDs before I can have confidence that the results are fully generalizable to hospital EDs.

With respect to study 4, there are many inhibition tasks that could have been used. I chose to use four of the most popular, and the finding that game performance was most closely related to performance on the Go/No-Go discrimination task was consistent with the task demands of the game where the player had to hit a target and avoid distractors. The correlation of the serious game with CRT in the first clinical study 1 shows that median correct RT on the game may also reflect overall cognitive speed to some degree which may explain why I (Tong & Chignell, 2014) found a correlation between game RT and Stroop RT (Stroop interference was not used in that study). As other researchers have noted, it is difficult, if not impossible, to develop “pure” tasks that measure a single well-defined cognitive function since all tasks will require at least a minimal amount of perceptual processing, attention, memory, etc. While the game may measure a number of things to some extent, it is clear that of the abilities that I have tested thus far, the relationship of the game to response inhibition ability is strongest.
7.4 Future Work

My thesis works reveals the need for future research to demonstrate the validity of the serious game as a screening tool for delirium and to provide stronger guidance on what game performance cut-offs to use when screening for delirium.

The clinical studies in this thesis showed that game performance was also predictive of the commonly used dementia threshold (<24) on the MMSE, with the game scores showing a significant difference between those scoring above and below this cut-off. It would be useful to explore the possible role of the game in detecting dementia in future research.

While I did my clinical studies in a hospital ED, it would be useful to study use of the serious game in long-term care (LTC) where there are known to be high rates of dementia and delirium. Since cognitive assessment is typically done only infrequently in LTC, research should be carried out to assess the possibility viability of the serious game as a cognitive assessment tool in a variety of settings including LTC and community care. One useful property of the game in this regard is that it is non-verbal, so that once a person learns how to play the game, communication difficulties are not a problem. Since there may be a natural tendency for staff in LTC to underestimate the cognitive status of people with communication difficulties, the use of the serious game would allow people to demonstrate more accurately their level of cognitive function.

One final direction for future research noted here is that different variants of the game can be developed for specific cognitive tasks. For instance, by putting numbers on moles one can turn the game into an n-back task. For example, is the number on this mole the same as the number on the mole that popped up just previously (e.g., 1-back) or is the number the same as the one on the mole before last (e.g., 2-back)?

7.5 Final Words

In this dissertation, I showed how one serious game that assesses the overall EF of inhibition can be predictive of a number of different clinical assessment tools, and can also be enjoyable for older adults to play in a clinical context. It is important that serious games for dementia are validated through scientific research such as that conducted in this dissertation, and that
promoters and businesses do not simply claim that a game or suite of games is valid based on face validity or some intended association with laboratory tasks or scientific theories.

I also think that cognitive assessment needs to move to a more game-like approach to increase the range of people that are assessed and to increase the frequency of assessment for each person, while potentially reducing the cost of such assessments. Using batteries of cognitive tests that are not game like creates a kind of selection bias where only highly motivated people will use them, since they will not be enjoyable or seem useful to most people.

Ideally, cognitive assessment should not only answer general questions about whether or not a person has a major condition such as delirium or dementia, but should also be diagnostic of the status of different cognitive abilities that a person has. Performance on the serious game studied in this dissertation appears to be predictive of inhibition ability. But many other serious games could be developed to assess a variety of factors associated with cognitive decline in ageing, while keeping individuals active and stimulated, thereby potentially slowing cognitive decline.

The “whack-a-mole” task was found to assess response inhibition ability in particular, a major element of cognitive control necessary in carrying out activities of daily living and other types of task. This opens the way for more discriminative testing, using serious games, that can assess profiles of ability across a range of cognitive functions.
References


https://doi.org/10.1037/neu0000001


Chignell, M., Tong, T., Mizobuchi, S., Delange, T., Ho, W., & Walmsley, W. (2015). Combining Multiple Measures into a Single Figure of Merit. In *7th International Conference on Advances in Information Technology*. Bangkok, Thailand.


Fernandez-Duque, D., & Black, S. E. (2006). Attentional Networks in Normal Aging and
Alzheimer’s Disease. https://doi.org/10.1037/0894-4105.20.2.133


disease patients. In *2012 ICME International Conference on Complex Medical Engineering (CME)* (pp. 215–217). IEEE. https://doi.org/10.1109/ICCME.2012.6275655


Noreen, S., & MacLeod, M. D. (2015). What Do We Really Know about Cognitive Inhibition? Task Demands and Inhibitory Effects across a Range of Memory and Behavioural Tasks. PloS One, 10(8), e0134951. https://doi.org/10.1371/journal.pone.0134951


https://doi.org/10.1016/S0197-4580(97)00109-7


https://doi.org/10.1073/PNAS.87.1.256


Appendices

8 Appendices

8.1 Appendix A. Literature Review on Computerized Cognitive Assessments.

One approach to address challenges associated with paper-and-pencil based assessments is to digitize existing cognitive screening tools, and another method is to create new screening tools. This section begins with a review of translated cognitive assessments into digital formats and then an overview of existing computerized assessment tools. This section ends with a discussion of the advantages and disadvantages of computerized cognitive assessments.

8.1.1 Translated Computerized Assessment Tools

This subsection will review paper-and-pencil based assessments that have been translated into a digital format.

Based on the earlier review of the paper-and-pencil based MMSE, MoCA and CAM, the only MoCA is the only assessment that has a digital analog. A digital version of the MoCA was released in May 2017, which requires an iOS devices and an active Internet connection. At the time of this writing (January 8, 2018), it is currently only available in English and requires a monthly subscription. In addition, no information has been provided on the validation and reliability of the electronic version of the MoCA.

8.1.2 Computerized Assessment Tools

This subsection will review computerized assessments tools that are a mix of novel assessment tools and digitized paper-and-pencil tests.

8.1.2.1 Computerized Screening Tools for Healthy Adults

A battery of digital screening tools have been developed for use by organizations such as the Department of Defense and the Federal Aviation Administration for use by healthy adults such as the Automated Neuropsychological Assessment Metrics (ANAM), and the CogScreen –
Aeromedical Edition (CogScreen-AE). These assessments were originally designed to assess the cognitive fitness of military personnel and aviation pilots by examining components such as information processing, reaction RT and spatial discrimination.

Other computerized assessments have been developed for use by healthy adults, which include the Cambridge Neuropsychological Test Battery (CANTAB), CNS Vital Signs and NIH Toolbox, CogTest Online, CogState, and Cognitive Stability Index (CSI). Many of these computerized assessments aim to evaluate the cognitive domains of working memory, attention, visuospatial memory, cognitive flexibility, and language.

Although many of these computerized assessments were not explicitly designed to screen for signs of cognitive impairment, tools such as ANAM, CNS Vital Signs, Computerized Neuropsychological Test Battery (CNTB), CogState, and CSI have been shown to differentiate healthy and non-healthy adults (i.e., adults with MCI, and dementia). These computerized assessments are either administered on a computer (e.g., ANAM) or touchscreen tablet (e.g., CANTAB) that can either be self-administered (e.g., CNS Vital Signs), or require administration by a trained administrator (e.g., CNTB, CSI and CST).

8.1.2.1.1 Automated Neuropsychological Assessment Metrics (ANAM)

The Automated Neuropsychological Assessment Metrics (ANAM) is a digital library of 30 assessments that was originally developed for use by the Department of Defense (Wild et al., 2008). This assessment suite is administered on a computer with a monitor, keyboard, and mouse by a trained examiner. Performance on the ANAM is evaluated based on accuracy (percent correct) and throughput/efficiency (number of correct responses per minute). A principal component analysis (PCA) of the accuracy, throughput/efficiency and RT performance on the ANAM was carried out to assess the underlying constructs that ANAM evaluates (Kabat, Kane, Jefferson, & Dipino, 2001). Correlation analysis between performance on the ANAM with traditional measures of cognitive status were also carried out. The PCA revealed three factors: processing speed/efficiency, retention/memory, and working memory, and the correlation analysis revealed significant relationships between ANAM subtests and traditional neuropsychological tests. Moreover, a subset of six of the 30 assessments have been used in dementia screening, which include (1) simple and (2) choice reaction times, (3) matching to sample, (4) continuous performance test, (5) a Sternberg six-letter memory task, and (6) spatial
discrimination. Research by Levinson, Reeves, Watson and Harrison (2005) used a subset of the ANAM to compare performance on the ANAM against the MMSE, and Geriatric Depression Scale (Yesavage et al., 1982) with eight participants with Alzheimer’s disease, and eight age-matched controls. Using a discriminant function analysis of the accuracy and throughput/efficiency scores, their analysis correctly classified 93.8% and 100% of the participants, respectively. Their findings revealed that throughput/efficiency was a more sensitive measure of cognitive deficit compared to accuracy. Overall, the study by (Levinson et al., 2005), suggested that a subset of the ANAM test battery can be performed by older adults Alzheimer’s disease, and that a subset of the ANAM is sensitive with a small sample population in detecting cognitive changes in Alzheimer’s disease. Overall, the ANAM is a digital test battery that has been used for assessing cognitive status in older, cognitively impaired adults.

8.1.2.1.2 CogScreen – Aeromedical Edition (CogScreen-AE)

CogScreen – Aeromedical Edition (CogScreen-AE) is a computerized cognitive assessment suite adapted for use by the United States Air Force (Carretta & Ree, 1999). It is designed to assess the domains of information processing, attention, memory, reaction time, and simultaneous information processing. CogScreen-AE is intended to screen for changes or declines in cognition in aviation pilots that may be at risk of impairing flying ability (Kay, 1995). A factor analysis using varimax rotation on the 28 CogScreen-AE scores revealed nine factors (Kay, 1995).

8.1.2.1.3 Cambridge Neuropsychological Test Automated Battery (CANTAB)

The Cambridge Neuropsychological Test Automated Battery (CANTAB) is a self-administered set of 25 cognitive assessments that is completed on a touch screen. It is designed to evaluate three cognitive domains: working memory, attention, and visuo-spatial memory. Normative data is available for ages 8 to 80 (De Luca et al., 2003; Robbins et al., 1994). Robbins et al. (1994) carried out a study with 787 healthy adults (age 55 to 88 years) using the CANTAB. A factor analysis with 771 individuals with MMSE scores identified two factors: memory and learning, and speed of response in memory and attention (Robbins et al., 1994). The CANTAB was found to have good test-retest reliability (for a sample of 162 participants with an average of 4 weeks between tests, interclass correlation = 0.75–0.80). Practice effects have been observed, which vary based on the participant’s ability and the task difficulty (Lowe & Rabbitt, 1998). CANTAB
has also been used to identify early signs of memory impairment in Alzheimer’s disease, specifically the paired associate learning and delayed matching subtests (De Jager, Milwain, & Budge, 2002; Fowler, Saling, Conway, Semple, & Louis, 1995; Fowler, Saling, Conway, Semple, & Louis, 1997; Fowler, Saling, Conway, Semple, & Louis, 2002).

8.1.2.1.4 CNS Vital Signs

CNS Vital Signs is a computer-based battery of cognitive assessments that consists of seven tests that assess five cognitive domains: memory, psychomotor speed, reaction time, cognitive flexibility, and complex attention. It is designed for cognitive screening rather than diagnosis. CNS Vital Signs is designed to be self-administered on a computer with a monitor using a keyboard and mouse. The entire test battery is estimated to take 30 minutes to complete. Research has been carried out using CNS Vital Signs to detect MCI, dementia, post-concussion syndrome, and severe TBI, ADHD and depression (Gualtieri & Johnson, 2006).

8.1.2.1.5 Computerized Neuropsychological Test Battery (CNTB)

The Computerized Neuropsychological Test Battery (CNTB) is a digital test battery that consists of 11 subtests that assess motor speed, information processing, attention, verbal and spatial memory, language, and spatial abilities. Administration of the CNTB requires a trained administrator and cannot be self-administered. Responses from the test participant are recorded by the administrator through key presses, pointing or spoken responses. The CNTB was developed to provide a measure of cognitive status across a range of AD severity (Veroff, Bodick, Offen, Sramek, & Cutler, 1998). The CNTB has two alternate versions that are highly correlated with each other (r=0.9) and has been evaluated for test-retest reliability (r=0.9) (Veroff et al., 1998).

8.1.2.1.6 NIH Toolbox

The NIH Toolbox consists of over 100 assessments for cognition, emotion, motor, and sensation, that are either performance-based or self-reported (National Institutes of Health, 2012). It was released in 2015 for iOS devices (which does that require an active Internet connection), and is available in English and Spanish. It has normative data for individuals between the ages of 3 to 85 years (Akshoomoff et al., 2014). Tests in the NIH Toolbox can be self-administered or administered by a proctored or both. The estimated time to complete each test varies between 3
minutes to 31 minutes. None of the tests in the cognition battery require the purchase of supplementary resources unlike in motor domain, which requires items such as a grip dynamometer. Test-retest reliability for tests separated by 2 weeks with 89 participants has been reported ($r=0.86–0.92$), along with internal consistency with 268 participants (Cronbach’s $\alpha = 0.84$), and convergent validity ($r=0.78–0.90$), and discriminant validity ($r=0.19–0.39$) (Heaton et al., 2014).

8.1.2.1.7 CogState

CogState is a digital battery that consists of eight subtasks in the form of card games. CogState is administered on a computer, and is designed to take 15 to 20 minutes to complete. Test participants are asked to respond by pressing the “k” key for yes and “d” key for no (Westerman, Darby, Maruff, & Collie, 2001). CogState was developed for repeat testing and has been evaluated for practice effects (Falleti, Maruff, Collie, & Darby, 2006; Fredrickson et al., 2010). Research by Falleti et al. (2006) demonstrated that performance on test-retest of intervals 10 minutes and 1-week offer no cognitive improvement after the second assessment. In both 10-minute and 1-week test-retest intervals, the highest practice effects were observed between the first and second assessments. A study by Darby, Maruff, Collie and McStephen (2002) investigated the ability to detect MCI in patients with repeated assessment of CogState over three hours with four assessments periods. They found that patients with mild MCI demonstrated less practice effects compared to healthy controls by the third and fourth administration of CogState. In another study by Maruff et al. (2004), the continuous learning task subtest from CogState along with the MMSE and CANTAB were administered eight times over a 12-month period in healthy controls and participants with MCI. At the end of 12 months, there was a significant decline in the CogState continuous learning task performance but no observed changes detected by the CANTAB. These findings suggest that CogState is sensitive to subtle changes in memory.

8.1.2.1.8 Cognitive Stability Index (CSI)

The Cognitive Stability Index (CSI) is a web-based test battery that consists of 10 tests that evaluate four domains: memory, attention, response speed, and processing speed. The battery is administered by a test administrator on a computer, and input by the subject is provided using a keyboard. CSI is designed to take 25 to 30 minutes to complete (Erlanger et al., 2002). The concurrent validity was established by comparing performance on the four factors on the CSI
against existing neuropsychological tests, and the test-retest reliability was assessed across two sessions. Computerized Screening Tools for Adults with Cognitive Impairment

In the previous section, some of the computerized assessments such as the CANTAB, CNS Vital Signs, and CogState have been used with adults with cognitive impairments; they were not designed for use by this sample population. The focus of this section will be on digital screening tools designed specifically for use by adults with cognitive conditions such as MCI and dementia. These tools include the Cognitive Screening Test (CST), Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment, Alzheimer’s Disease Assessment Scale-Cognitive Subscale, BrainCare, MicroCog, MCI Screen, and the Cognitive Stability Index and Cognitive Screening Test. Another category of computerized assessments have been designed for use by adults on drugs such as the Cognitive Drug Research Computerized Assessment System.

8.1.2.1.9 Cognitive Screening Test (CST)

The Cognitive Screening Test (CST) is a shorter version of the CSI that is targeted for dementia screening. Compared to the four factors in the CSI, the CST focuses on three skills: keyboard skills, learning and memory, and EF. Lichtenberg et al. (2006) demonstrated that the CST had a sensitivity of 80% and a specificity of 87% compared to the MMSE which had a sensitivity of 37.7%, specificity of 97.6%. They also observed an 83% concordance rate with geriatrician consensus diagnosis in categorizing 102 patients in their study as having dementia, MCI or no cognitive impairment.

8.1.2.1.10 The Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment (CANS-MCI)

The Computer-Administered Neuropsychological Screen for Mild Cognitive Impairment (CANS-MCI) a digital test battery designed to screen for mild cognitive impairment, that evaluates the domains of language, memory, and EF. It is designed to take 30 minutes to complete and is self-administered in a clinical environment. The initial setup of the assessment takes a few minutes and requires an administrator to set up a patient in front of a proprietary setup (computer with a touch screen and speakers). The CANS-MCI is scored and reported using a web-based system. This system has been evaluated with 310 community dwelling older adults (age 51 to 93 years) (Tornatore, Hill, Laboff, & McGann, 2005). The CANS-MCI has been
evaluated for internal consistency of the subtests using alpha coefficients, and test-retest reliability using Pearson correlations and paired \( t \)-tests. The concurrent validation was assessed using Pearson correlations comparing performance on the CANS-MCI with neuropsychological tests. (Tornatore et al., 2005) also carried out exploratory and confirmatory factor analyses on the scale items and revealed three factors: memory, language/spatial fluency, and executive function/mental control.

### 8.1.2.1.11 Cognitive Drug Research Computerized Assessment System (COGDRAS)

The Cognitive Drug Research Computerized Assessment System (COGDRAS) was originally designed to measure change in drug effects on cognition but has been adapted to measure changes in cognitive in patients with dementia (COGDRAS-D). COGDRAS contains eight subtests that are administered on a computer by a trained administrator. There are two response buttons: Yes and No. The test-retest reliability of the COGDRAS-D was evaluated with 26 older adults and the concurrent validity has been established by comparing performance on the COGDRAS-D with the MMSE (Simpson, Surmon, Wesnes, & Wilcock, 1991). Research by Mohr, Walker, Randolph, Sampson and Mendis (1996) has shown that the COGDRAS-D is more sensitive in distinguishing between patients with Alzheimer’s and Huntington’s disease compared to other test batteries such as the ADAS, and MMSE. In addition, this battery has been shown to be differentiate between different dementias in patients with Alzheimer’s disease compared to those with Lewy body dementia (Ballard et al., 2001). Performance on the COGDRAS has been able to differentiate adults with delirium compared to adults with no delirium (Lowery, Wesnes, & Ballard, 2007). Compared to other the digital Poon-Baro-Wens test battery, the COGDRAS-D was not as sensitive in providing diagnostic information (De Lepeleire, Heyrman, Baro, & Buntinx, 2005).

### 8.1.2.1.12 MCI Screen

The MCI Screen is a designed to evaluate memory, executive function and language in 10-minutes (Medical Care Corporation, 1997). It is administered through a trained administrator either in person or over the telephone, which accounts for the lack of visual subtests. The MCI Screen is based on the CERAD 10-word recall task, which consists of a 10-word list that is shown three times with immediate recall following each presentation. In this task, there is a self-
evaluation of memory, and a distraction task. In the distraction task, there is a triadic comparison of animals shown before delayed free and cued recall trials. In the last task, there is a free recall of animal names from the distractor task. Performance on the MCI Screen is generated at the end of the test. The two primary result options are “normal” or “impaired”. In a study by Trenkle, Shankle and Azen (2007), the MMSE, Clock Drawing Test and the MCI Screen were compared in detecting Alzheimer’s disease and related dementias in 254 adults over 65 years. They found that the MCI Screen was more sensitive compared to the MMSE and Clock Drawing Test in detecting changes in cognitive impairment in their sample. The MCI Screen test-retest reliability has been evaluated in 30 adults (normal to moderate dementia) in two sessions separated by 6 months (Trenkle et al., 2007). During the first session, a neuropsychologist administered the paper-and-pencil based CERAD wordlist and then in the second session, the neuropsychologist administered the MCI Screen.

8.1.2.1.13 MicroCog

MicroCog was originally designed to screen for cognitive impairment in older physicians. It was formally called Assessment of Cognitive Skills, and was developed by an insurance company to screen for malpractice risk (Powell et al., 1993). The MicroCog test battery consists of 18 subtests and 12 short form subtests, which evaluates five domains: attention/mental control, memory, reasoning/calculation, spatial processing, and reaction time. MicroCog is a self-administered software suite that can be completed on a computer with a monitor, keyboard, and mouse. Responses are recorded using keyboard keys. It is designed to take 1-hour to complete for individuals with no cognitive impairments, and may take longer than 1-hour for individuals with impairments.

MicroCog, when it was formally known as Assessment of Cognitive Skills, was shown to distinguish between health controls and patients with mild cognitive impairment (Green, Green, Harrison, & Kutner, 1994). Research by Elwood (2001), found that MicroCog was able to differentiate older physicians with dementia compared to healthy controls with a sensitivity of 0.98 and specificity of 0.83. They also found that MicroCog performance was not affected by patients with depression.

The validity of different subtests from MicroCog has been evaluated. For example, the short form General Cognitive Functioning score has been compared against the Full Scale IQ on the
WAIS-III (Johnson, Rust, & Rust, 2003). In Johnson et al.'s (2003) study, 15 participants completed the MicroCog short form General Cognitive Functioning subtest and 15 participants completed the WAIS-III short form. They carried out a correlation analysis between the two scores from both groups which resulted in a positive correlation ($r=0.52$, $n=30$). In contrast, Helmes and Miller (2006) observed weak correlations between performance on MicroCog and the Wechsler Memory Scale-III with a sample of 33 older adults between the ages of 60 to 89 years, living in the community.

8.1.2.1.14 BrainCare (Neurotrax)

BrainCare (Neurotax), formerly known as Mindstreams, is a test battery designed to screen for MCI in elderly adults using nine subtests: verbal memory, nonverbal memory, Go/No-Go response inhibition, Stroop interference, problem solving, visual spatial imagery, verbal rhyming, verbal naming, staged information processing speed, finger tapping, and visuomotor planning. Some of the subtests are translations of paper-and-pencil based tests such as the Stroop interference subtest from the original traditional Stroop Test. Mindsreams is administered on a computer with a monitor, keyboard and mouse. Responses are recorded using a mouse and number pad on a keyboard. It is designed to be completed within 45 to 60 minutes.

Performance on the memory, visual spatial, and verbal subtests in Mindstreams have been shown to distinguish between adults with MCI compared to healthy controls and adults with mild Alzheimer’s disease (Dwolatzky et al., 2003, 2004). Research by Doniger et al. (2006) extends the work by Dwolatzky et al. (2003, 2004) by demonstrating that Mindstreams is able to differentiate between healthy controls and adults with mild AD in the presence of symptoms of depression.

8.1.2.1.15 CogTest Online

CogTest Online, formally known as CogTest, is an online cognitive assessment battery that is self-administered on a computer, which can be completed within 40-50 minutes. It measures the domains of attention, cognitive flexibility, memory, mental speed, and social cognition (Cogtest Online, 2017). The validity and test-retest reliability of CogTest has been evaluated with 75 adults between the ages of 20 to 60 years (Barua, P., Bilder, R., Small, A. Sharma, 2005). Correlations between CogTest and corresponding paper-and-pencil equivalents ranged between $r = 0.30$ to $0.70$ and the test-retest reliability over up to 4 weeks ranged between 0.4 and 0.9.
8.1.2.1.16 Alzheimer's Disease Assessment Scale-cognition sub-scale (ADAS-Cog)

The Alzheimer’s Disease Assessment Scale-cognition sub-scale (ADAS-Cog) was designed to screen for dementia following intervention therapies, that measures the domains of: memory, language, praxis and orientation (Rosen, Mohs, & Davis, 1984; Skinner et al., 2012). There is another version of the ADAS-Cog that was modified by Mohs et al. (1997) to include more items.

The ADAS-Cog has also been used to screen for cognitive changes in MCI with limited results (Beaton, Bombardier, Katz, & Wright, 2001; Benge, Balsis, Geraci, Massman, & Doody, 2009; Husted, Cook, Farewell, & Gladman, 2000; Kirshner & Guyatt, 1985; Mohs et al., 1997; Sano et al., 2011).

8.1.3 Discussion (Advantages and Disadvantages)

The computerization of cognitive assessments offers many advantages over paper-and-pencil analogs. For example, computerized assessments can record ambient information such as date and time, and precise information such as RT and response accuracy (Wild et al., 2008). They can also offer potential resource and cost savings in terms of reducing the need to be administered in conventional clinical settings with limited, trained test administrators. In addition, the digital format of computerized assessments may also enable an easier method to create randomized or alternate versions of a test (Gualtieri & Johnson, 2006).

Compared to paper-and-pencil tests, data generated from the computerized assessments do not need be transcribed into a digital format. This ease of use may facilitate easier data sharing between a user and healthcare professionals.

Computerized assessments also have the potential be self-administered or administered with minimal assistance from a care provided such as a family member, which enables users to complete assessments in their own homes, outside of conventional clinical settings such as hospitals, which may be inaccessible. The ability to administer computerized assessments that require little to no assistance on readily-available technologies (such as tablets and computers) enables care providers to suggest to a user to complete an assessment at any time. This ease of
access may also encourage users to repeat the assessments over time, which provides more information about the user’s cognitive status.

Although there are many advantages to use computerized assessments, it is important to consider the disadvantages. The validation and reliability of digitized tests should be considered as well as the usability of the applications. The computerization of cognitive assessments may potentially be somewhat uninteresting tasks on a computer to complete. Although the validation and test-retest reliability of the tests were established, the satisfaction and enjoyment level of the patient/user was not considered. Some computerized assessments (i.e., NIH Toolbox) require a subscription fee, which may deter its use.

Existing computerized assessments have focused on the measurement of EF and cognitive domains such as response inhibition, working memory, and cognitive flexibility using tasks such as the Stroop and Go/No-Go discrimination task (e.g., BrainCare) with both healthy and cognitively impaired adults. This research provides a basis for designing serious games for cognitive screening using components from similar tasks. By combining the advantages of computerized assessments (i.e., ability to record RT and accuracy with precise measurement, providing alternate versions of assessments) and avoiding disadvantages (i.e., validation and reliability issues) into a serious game, there would be the benefits of using such a platform for cognitive screening.

8.1.4 Literature Review on Other EF (Updating and Shifting).

8.1.4.1 Updating Ability and Tasks

Updating ability is an EF required in the intake of new information and revision of existing information in working memory. It is essential in activities such as driving and cooking. Updating ability can be measured using psychological tasks such as the n-back task (Kirchner, 1958), letter memory task (Morris & Jones, 1990), and keep track task (Yntema, 1963).

In a n-back task, the participant is provided with a sequence of stimuli. The task is to indicate when the current stimulus is the same as one from a defined number, “n”, of steps earlier in the sequence. For example, in a 2-back task, (the value of n is 2), the sequence gulf-echo-gulf would be a match as the word gulf appears in the first and third positions. Performance on the n-back task requires the coordination of various processes including the simultaneous storage and
processing of the stimuli (Jonides et al., 1997; Kane & Engle, 2002), as well as updating of the working memory. Performance also relies on reallocation and matching of stimuli in a sequential order (Oberauer, 2005). Research by Finley, Benjamin and McCarley (2014), and Jaeggi, Schmid, Buschkuehl and Perrig (2008) have demonstrated that increasing values of \( n \) are related to increases in error rate and RT, as well as increased mental workload (Herff et al., 2014).

Similar findings using fMRI, and positron emission tomography (PET) have shown that higher levels of activation (e.g., as measured by percentage change in blood flow) increased approximately linearly in the dorsolateral prefrontal cortex with increasing values of \( n \) in healthy adults, suggesting that the \( n \)-back task can be used as a method in quantifying processing complexity (Drobyshevsky, Baumann, & Schneider, 2006; Jonides et al., 1997).

The letter memory task (Morris & Jones, 1990) is another example of an updating task. In this task, a participant is provided with one letter at a time, and is instructed to repeat the last \( n \) letters in the sequence aloud to memorize it. After the last letter is provided, the participant is asked to recall the last \( n \)-number of letters (ranging from 3 to 6) in the series. The total sequence length varies from 4, 6, 8, and 10 in the original implementation by (Morris & Jones, 1990). A key difference between the \( n \)-back and letter memory task, is that the letter memory task requires continuous updating after each list of letters is provided, whereas the \( n \)-back task requires updating after each new item is provided (Dahlin, Nyberg, Bäckman, & Neely, 2008). The difference between the tasks have been shown to produce differences in functional brain activation (Dahlin, Neely, et al., 2008).

Another example of an updating task is the keep-track task (Yntema, 1963). In this task, a participant is shown a set of categories (i.e., animals, countries, colours, and sports, etc.). In the second set, the participant is shown a set of words from each category. The participant is then asked to recall the last word in each category. The proportion of correct responses is often used as a measure of performance on this task (Sylvain-Roy, Lungu, & Belleville, 2015). The keep track task has been shown to be related to working memory span (Engle, Laughlin, Tuholski, & Conway, 1999).

### 8.1.4.2 Shifting Ability and Tasks

Shifting ability is needed in switching between tasks or between different components of one task (Monsell, 1996). Examples of tasks used to measure shifting include the Wisconsin Card
Sort task (WCST) (Heaton, 1981), which has been referred to as a “gold standard of executive function tests” by Swanson (2005). Additional tests for measuring shifting ability include the color-shape task (Miyake, Emerson, Padilla, & Ahn, 2004), and TMT (War Department Adjutant General’s Office, 1944).

The WCST is used to measure shifting ability where four stimulus cards are presented to the participant (Butler, Retzlaff, & Vanderploeg, 1991). The objects on the cards can differ in color, quantity, and shape (Heaton, 1981). The participant is shown an additional card and is asked to choose which one of the four original cards conforms to the same category as the additional card. As the classification rule is not provided to the participant, they have to guess the rule. They do this based on the feedback provided to them (“correct” or “incorrect”), after they choose one of the four cards to match with the additional card. Performance is based on a participant’s perseveration errors, which is the amount of erroneous responses made after switching to a new rule despite the provided feedback (Head, Kennedy, Rodrigue, & Raz, 2009).

The color-shape task is an example of a shifting task that requires a participant to identify either the colour or shape of a stimulus (Miyake et al., 2004). A cue appears above the stimulus, either “C” (or the word cue “COLOR”) for colour or “S” (or the word cue “SHAPE”) for shape, that indicates which feature of the stimulus to identify. Miyake et al. (2004) demonstrated that the letter cue condition (i.e., “C” or “S”) increases the cost of task switching on articulatory suppression compared to a control condition (e.g., no dual-task). Articulatory suppression refers to inner speech that assists individuals in goal retrieval and activation.

The TMT is another example of a shifting task, which contains two parts: A and B (War Department Adjutant General’s Office, 1944). Part A requires visuospatial abilities to complete, whereas part B requires shifting and task-switching ability to complete (Corrigan & Hinkeldey, 1987; Periáñez et al., 2009). In part A of the TMT (referred to as TMT-A), a participant is presented with circles that contain letters that are randomly arranged. The objective is to connect these circles in alphabetical order (i.e., A, B, C). In part B of the TMT (referred to as TMT-B), a participant is presented with circles that contain either letters or numbers that are randomly arranged. The objective is to connect the circles in alternating numeric-alpha order (i.e., 1, A, 2, B). TMT-B takes longer to complete than TMT-A due to the added complexity (Mortimore,
2001; Strauss, Sherman, Spreen, & Spreen, 2006). Performance on the TMT is often measured using time to completion, and error rates (Ashendorf et al., 2008).

### 8.1.4.3 Updating Tasks and Age-Related Performance

Age-related performance on updating has been investigated using the n-back, letter memory, and keep track task. For example, performance on the n-back task in healthy adults has shown that older adults have higher error rates and longer RT. This disadvantage increases as the value of n increases, compared to younger adults (Droby shevsky et al., 2006). Thus, declines in n-back performance in healthy adults have shown that performance is age-related and associated with changes in the prefrontal cortex (Mattay et al., 2006), which suggests that this task can be used to assess updating ability across the lifespan.

Research by Dahlin, Nyberg, et al. (2008) investigated the effect of training on updating memory using the letter memory and keep-track task in young and old adults. They observed that both young and old adults that received training demonstrated significant improvements on the letter memory task (but not with the keep-track task), although larger improvements were observed in the younger adult group. Moreover, older adults did not demonstrate any transfer effects on the keep-track task, which is consistent with previous literature showing little-to-no transfer effects of training updating ability (Ball et al., 2002; Derwinger, Stigsdotter Neely, Persson, Hill, & Backman, 2003; Neely & Bakman, 1993). Further research is required to support the use of the letter memory and keep-track tasks as measures of updating ability across the lifespan.

### 8.1.4.4 Shifting Tasks and Age-Related Performance

Age-related performance on EF shifting tasks have been observed to decrease with age on tasks such as the WCST (Axelrod & Henry, 1992; Beatty, 1993; Daigneault, Braun, & Whitaker, 1992; Fristoe, Salthouse, & Woodard, 1997; Haaland, Vranes, Goodwin, & Garry, 1987; Parkin & Walter, 1991), and TMT (Davies, 1968; Giovagnoli et al., 1996; Gordon, 1978; Goul & Brown, 1970; Kennedy, 1981; Lindsey & Coppinger, 1969). Research on age-related performance on the WCST suggests that healthy, older adults demonstrate higher perseveration errors compared to younger adults (Axelrod & Henry, 1992; Beatty, 1993; Daigneault et al., 1992; Fristoe et al., 1997; Haaland et al., 1987; Parkin & Walter, 1991). Head et al. (2009) demonstrated using fMRI studies with healthy adults that increases in preservation errors are associated with the integrity of the prefrontal cortex and changes in associated processes in this
area, which supports the notion that the prefrontal cortex is critical for EF. Moreover, age-related performance has been observed on the TMT, whereby older adults demonstrate longer time to completion than younger adults (Tombaugh, 2004). Error frequencies on the TMT have also been shown to be age-related but not as strongly as time to completion (Ashendorf et al., 2008).
8.2 Appendix B. Literature Review Supporting Materials.

Table 8-1. Comparative summary of standard cognitive assessments. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MMSE</th>
<th>MoCA</th>
<th>CAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screens for Administration time</td>
<td>5 to 10 minutes</td>
<td>10 minutes</td>
<td>5 to 10 minutes</td>
</tr>
<tr>
<td>Administration method</td>
<td>Dementia</td>
<td>MCI</td>
<td>Delirium</td>
</tr>
<tr>
<td>Administration method Pencil-and-paper Interview style</td>
<td>Pencil-and-paper Interview style</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-2. Comparison between traditional paper-and-pencil cognitive assessments and the use of serious games for cognitive screening.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Paper-based assessments</th>
<th>Serious games</th>
</tr>
</thead>
<tbody>
<tr>
<td>Administration method</td>
<td>Trained administrator</td>
<td>Self or minimal assistance</td>
</tr>
<tr>
<td>Administration bias potential</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Equipment</td>
<td>Paper, pencil</td>
<td>Tablets, computers</td>
</tr>
<tr>
<td>Repeatability</td>
<td>Limited repeatability—</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>not necessarily if alternate forms are available</td>
<td></td>
</tr>
<tr>
<td>Multiple variations</td>
<td>Few or none</td>
<td>Yes, can be randomized</td>
</tr>
<tr>
<td>Motivation/Entertainment</td>
<td>Low</td>
<td>High, if target users enjoy playing the game</td>
</tr>
<tr>
<td>Validation</td>
<td>Available</td>
<td>Variable depending on the serious game</td>
</tr>
</tbody>
</table>
Appendix C. Supporting Materials for the Concurrent Validation Study in Assessing Cognitive Status in Elderly Adults with Incident Delirium.

8.3.1 Supporting Methodology Materials

Table 8-3. Detailed information on the different game variants played by patients.

<table>
<thead>
<tr>
<th>Game Number</th>
<th>Grid Size</th>
<th>Target Size (px)</th>
<th>Possible Stimuli</th>
<th>Number of Patients That Played This Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2x2</td>
<td>250</td>
<td>• Moles • Butterflies</td>
<td>195</td>
</tr>
<tr>
<td>2</td>
<td>3x3</td>
<td>200</td>
<td>• Moles • Butterflies</td>
<td>203</td>
</tr>
<tr>
<td>3</td>
<td>3x3</td>
<td>175</td>
<td>• Moles • Butterflies</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>5x3</td>
<td>175</td>
<td>• Moles • Butterflies</td>
<td>189</td>
</tr>
<tr>
<td>5</td>
<td>3x3</td>
<td>250</td>
<td>• Moles • Butterflies</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>5x3</td>
<td>200</td>
<td>• Moles • Butterflies</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5x3</td>
<td>175</td>
<td>• Moles • Butterflies</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>5x3</td>
<td>175</td>
<td>• Moles • Raccoons</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>2x2</td>
<td>250</td>
<td>• Moles • Butterflies</td>
<td>19</td>
</tr>
<tr>
<td>10</td>
<td>2x2</td>
<td>175</td>
<td>• Moles • Butterflies</td>
<td>166</td>
</tr>
<tr>
<td>11</td>
<td>2x2</td>
<td>175</td>
<td>• Moles • Raccoons</td>
<td>76</td>
</tr>
<tr>
<td>12</td>
<td>2x2</td>
<td>175</td>
<td>• Moles • Butterflies</td>
<td>33</td>
</tr>
<tr>
<td>13</td>
<td>5x3</td>
<td>175</td>
<td>• Moles • Raccoons</td>
<td>122</td>
</tr>
<tr>
<td>14</td>
<td>5x3</td>
<td>200</td>
<td>• Moles • Butterflies</td>
<td>8</td>
</tr>
</tbody>
</table>

8.3.1.1 Description of Additional Cognitive Assessment Measures

This section describes the following cognitive assessment measures used in this study including the DI, RASS, DVT, and CRT task.

8.3.1.2 Delirium Index

The Delirium Index (DI) is a scale adapted from the CAM, which assesses the severity of delirium. It was designed for use by non-clinicians such as RAs to assess a patient based on observations that do not require additional information from the patient’s care providers such as
their family members and healthcare staff such as nurses (McCusker et al., 1998). The scale consists of seven questions and the total score ranges from 0 to 21, where a score of 0 indicates no impairment and higher scores suggest higher delirium severity. The DI is meant to be used in conjunction with the MMSE, where at least five questions on the MMSE constitute the observational basis. Moreover, McCusker et al. (1998) originally reported the inter-rater reliability of the DI as 0.78 between RAs, and 0.88 between RAs and geriatric psychiatrists. The convergent validity of the DI was assessed by comparing DI and Delirium Rating Scale scores ($r=0.60–0.70$). A copy of the DI is included in Appendix L.

### 8.3.1.3 Richmond Agitation-Sedation Scale

The Richmond Agitation-Sedation Scale (RASS) was developed to assess a patient’s level of sedation and agitation based on an administrator’s (e.g., nurse, physician) observations (Sessler et al., 2002). The RASS is a 10-level scale with scores ranging from -5 (unarousable) to 4 (combative). A value of 0 suggests that a patient is alert and calm (Sessler et al., 2002). (Sessler et al., 2002) assessed the inter-rater reliability of the RASS with two physicians, two nurses, and one pharmacist ($r=0.956$). Sessler et al. (2002) also evaluated the validity of the RASS by comparing it to the Ramsay sedation scale ($r=–0.78$), and the Sedation Agitation Scale ($r=0.78$).

### 8.3.1.4 Digit Vigilance Test

The Digit Vigilance Test (DVT) is paper-and-pencil task designed to measure sustained attention and psychomotor speed (Kelland Manchester et al., 1996; Lewis & Rennick, 1979). In this task, patients are presented with a series of numbers and are asked to cross out a specific number as quickly as possible, that appears randomly on two-pages with numbers arranged on 59 rows with 35 single digits. The administrator observes the patient on the first five rows to evaluate whether feedback should be provided to the patient. Performance on the DVT is measured using total completion time and errors of omission. Moreover, the DVT has a test-retest reliability of $r=0.91$ over a one-week interval (Kelland Manchester et al., 1996).

### 8.3.1.5 Choice Reaction Time Task

In the choice reaction time (CRT) task, a patient is presented with one of two possible stimuli (e.g., the letter L or R), and is asked to identify the stimulus as quickly as possible (Logan & Cowan, 1984).
8.3.2 Supporting Results Materials

Table 8-4. Summary of study sample demographics and cognitive assessment scores. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.

<table>
<thead>
<tr>
<th>Variable a</th>
<th>Males (n=80)</th>
<th>Females (n=66)</th>
<th>Total (n=146)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>Mean (SD) / Median (IQR)</td>
<td>Range</td>
<td>Mean (SD) / Median (IQR)</td>
</tr>
<tr>
<td>MMSE</td>
<td>80.6 (6.3) / 80.6 (5.7)</td>
<td>70-94</td>
<td>80.6 (6.0) / 70-94</td>
</tr>
<tr>
<td>MoCA</td>
<td>28.2 (1.5) / 27.7 (2.2)</td>
<td>25-30</td>
<td>26.7 (3.9) / 9-30</td>
</tr>
<tr>
<td>CAM</td>
<td>24.5 (2.6) / 23.2 (3.8)</td>
<td>8-30</td>
<td>23.2 (4.6) / 15-30</td>
</tr>
<tr>
<td>DI</td>
<td>0.1 (0.3) / 0.1 (0.3)</td>
<td>0-1</td>
<td>0.1 (0.3) / 0-1</td>
</tr>
<tr>
<td>RASS</td>
<td>–0.1 (0.4) / –0.1 (0.4)</td>
<td>–2 to 1</td>
<td>–0.1 (0.3) / –2 to 1</td>
</tr>
<tr>
<td>DVT</td>
<td>97.5 (5.7) / 98.7 (4.0)</td>
<td>81-103</td>
<td>97.5 (3.9) / 92-103</td>
</tr>
<tr>
<td>CRT RT (msec) a</td>
<td>1200 (300) / 1200 (500)</td>
<td>870-1980</td>
<td>1200 (400) / 780-3230</td>
</tr>
<tr>
<td>Game Correct Target Offset (px) a</td>
<td>337.5 (65.8) / 317.1 (71.7)</td>
<td>14.0 – 448.0</td>
<td>321.0 (68.4) / 13.0 – 448.0</td>
</tr>
</tbody>
</table>

a For the CRT RT and game RT rows, the median (IQR) is reported.

Table 8-5. Correlations comparing game performance to the standard cognitive assessments. MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, DVT = Digit Vigilance Test, and CRT = Choice Reaction Time.

<table>
<thead>
<tr>
<th></th>
<th>MMSE</th>
<th>MoCA</th>
<th>CAM</th>
<th>DI a</th>
<th>RASS a</th>
<th>DVT</th>
<th>CRT RT</th>
<th>CRT Target Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Correct Target Offset</td>
<td>–0.104</td>
<td>–0.042</td>
<td>0.071</td>
<td>0.048</td>
<td>–0.108</td>
<td>0.432</td>
<td>–0.053</td>
<td>0.004</td>
</tr>
<tr>
<td>p</td>
<td>0.22</td>
<td>0.67</td>
<td>0.40</td>
<td>0.46</td>
<td>0.12</td>
<td>0.008</td>
<td>0.64</td>
<td>0.97</td>
</tr>
</tbody>
</table>

160
Correlations involving the CAM were calculated using point-biserial correlations. Correlations involving the DI and RASS (and not involving the CAM) were assessed using Spearman’s rho. All other correlations were calculated using Pearson’s r.

Table 8-6. Partial correlations that control for CRT reaction time on game performance and standard cognitive assessments. CRT = Choice Reaction Time, MMSE = Mini-Mental State Examination, MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, DI = Delirium Index, RASS = Richmond Agitation-Sedation Scale, and DVT = Digit Vigilance Test.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Control for CRT RT</th>
<th>Game Median Target Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>rho</td>
</tr>
<tr>
<td>MMSE</td>
<td>–0.024</td>
<td>0.84</td>
</tr>
<tr>
<td>MoCA</td>
<td>0.160</td>
<td>0.19</td>
</tr>
<tr>
<td>CAM</td>
<td>–0.112</td>
<td>0.33</td>
</tr>
<tr>
<td>DI</td>
<td>0.066</td>
<td>0.56</td>
</tr>
<tr>
<td>RASS</td>
<td>–0.088</td>
<td>0.44</td>
</tr>
<tr>
<td>DVT</td>
<td>0.440</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 8-7. Mann-Whitney U test results comparing cognitive assessment performance based on the absence (≥24) or presence (<24) of dementia as assessed by the MMSE. MMSE = Mini-Mental State Examination.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MMSE &lt;24</th>
<th>MMSE ≥24</th>
<th>U</th>
<th>P</th>
<th>z</th>
<th>r</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Mean (SE)</td>
<td>n</td>
<td>Mean (SE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game Correct Target Offset</td>
<td>18</td>
<td>330.4 (17.5)</td>
<td>122</td>
<td>319.5 (6.2)</td>
<td>7846.0</td>
<td>0.5</td>
<td>-0.64</td>
</tr>
</tbody>
</table>

Table 8-8. Mann-Whitney U test results comparing game performance based on the absence (≥23) or presence (<23) of cognitive impairment as assessed by the MoCA. MoCA = Montreal Cognitive Assessment.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>MoCA, &lt;23</th>
<th>MoCA ≥23</th>
<th>U</th>
<th>P</th>
<th>z</th>
<th>r</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Mean (SE)</td>
<td>n</td>
<td>Mean (SE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Game Correct Target Offset</td>
<td>38</td>
<td>324.0 (7.7)</td>
<td>67</td>
<td>324.9 (5.2)</td>
<td>1201.5</td>
<td>0.6</td>
<td>-0.5</td>
</tr>
</tbody>
</table>
Table 8-9. Mann-Whitney $U$ test results comparing cognitive assessment performance based on the absence (CAM negative) or presence (CAM positive) of delirium as assessed by the CAM. CAM = Confusion Assessment Method.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>CAM Negative</th>
<th>CAM Positive</th>
<th>U</th>
<th>P</th>
<th>z</th>
<th>r</th>
<th>IQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Correct Target Offset</td>
<td>129 (6.1)</td>
<td>12 (14.9)</td>
<td>716.0</td>
<td>0.7</td>
<td>-0.4</td>
<td>0.07</td>
<td>46.5</td>
</tr>
</tbody>
</table>
8.4 Appendix D. Supporting Materials for the Test-Retest Reliability Study in Assessing Cognitive Status in Elderly Adults with Incident Delirium.

Table 8-10. Demographics of the sample study.

<table>
<thead>
<tr>
<th>Baseline features</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean age (years) (SD)</td>
<td>81.1</td>
</tr>
<tr>
<td>Female (n)</td>
<td>61</td>
</tr>
<tr>
<td>Male (n)</td>
<td>53</td>
</tr>
<tr>
<td>Mean length of stay in the ED (hours) (SD)</td>
<td>16.3 (9.0)</td>
</tr>
</tbody>
</table>

Figure 8-1. Barplot illustrating the number of assessments completed at each session.
Figure 8-2. Scatterplots depicting the serious game median correct RT between: (A) initial enrolment compared to follow up session 1 ($r=0.927$, $p<0.001$), (B) follow up session 1 compared with 2 ($r=0.833$, $p=0.001$), and (C) follow up session 2 compared with 3 ($r=0.547$, $p=0.0656$).

Table 8-11. Relationships between sessions on serious game median correct RT, was determined using two-tailed Pearson’s $r$ correlations.

<table>
<thead>
<tr>
<th></th>
<th>Initial enrolment</th>
<th>Follow up 1</th>
<th>Follow up 2</th>
<th>Follow up 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial enrolment</td>
<td>1</td>
<td>0.886 ***</td>
<td>0.631 **</td>
<td>0.864 **</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.012$</td>
<td>$p=0.003$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$n=47$</td>
<td>$n=23$</td>
<td>$n=15$</td>
</tr>
<tr>
<td>Follow up 1</td>
<td></td>
<td>1</td>
<td>0.834 **</td>
<td>0.810 **</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$p=0.001$</td>
<td>$p=0.0014$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$n=19$</td>
<td>$n=13$</td>
</tr>
<tr>
<td>Follow up 2</td>
<td></td>
<td></td>
<td>1</td>
<td>0.553</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$p=0.062$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$n=14$</td>
</tr>
<tr>
<td>Follow up 3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Table 8-12. Relationships between sessions on serious game median target offset for correct trials, was determined using two-tailed Pearson’s $r$ correlations.

<table>
<thead>
<tr>
<th></th>
<th>Initial enrolment</th>
<th>Follow up 1</th>
<th>Follow up 2</th>
<th>Follow up 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial enrolment</td>
<td>1</td>
<td>0.822 ***</td>
<td>0.547 *</td>
<td>0.150</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.035$</td>
<td>$p=0.64$</td>
<td>$n=47$</td>
</tr>
<tr>
<td></td>
<td>$n=47$</td>
<td>$n=19$</td>
<td>$n=15$</td>
<td></td>
</tr>
<tr>
<td>Follow up 1</td>
<td>0.818 **</td>
<td>0.273</td>
<td>0.150</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.002$</td>
<td>$p=0.39$</td>
<td>$p=0.64$</td>
<td>$n=19$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=13$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow up 2</td>
<td>1</td>
<td>0.492</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.10$</td>
<td></td>
<td></td>
<td>$n=14$</td>
</tr>
<tr>
<td>Follow up 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-13. Relationships between sessions on serious game median correct RT, was determined using two-tailed Spearman’s $r$ correlations.

<table>
<thead>
<tr>
<th></th>
<th>Initial enrolment</th>
<th>Follow up 1</th>
<th>Follow up 2</th>
<th>Follow up 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial enrolment</td>
<td>1</td>
<td>0.804 **</td>
<td>0.800 **</td>
<td>0.510 **</td>
</tr>
<tr>
<td></td>
<td>$p&lt;0.001$</td>
<td>$p=0.003$</td>
<td>$p=0.090$</td>
<td>$n=47$</td>
</tr>
<tr>
<td></td>
<td>$n=47$</td>
<td>$n=23$</td>
<td>$n=15$</td>
<td></td>
</tr>
<tr>
<td>Follow up 1</td>
<td>1</td>
<td>0.818 **</td>
<td>0.503</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.002$</td>
<td>$p=0.095$</td>
<td></td>
<td>$n=19$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=13$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow up 2</td>
<td>1</td>
<td>0.483</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.11$</td>
<td></td>
<td></td>
<td>$n=14$</td>
</tr>
<tr>
<td>Follow up 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-14. Relationships between sessions on serious game median target offset for correct trials, was determined using two-tailed Spearman’s $r$ correlations.

<table>
<thead>
<tr>
<th></th>
<th>Initial enrolment</th>
<th>Follow up 1</th>
<th>Follow up 2</th>
<th>Follow up 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial enrolment</td>
<td>1</td>
<td>0.730 **</td>
<td>0.329</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td>$p=0.009$</td>
<td>$p=0.23$</td>
<td>$p=0.44$</td>
<td>$n=47$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=13$</td>
<td>$n=14$</td>
<td></td>
</tr>
<tr>
<td>Follow up 1</td>
<td>0.696 *</td>
<td>0.503</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.0039$</td>
<td>$p=0.095$</td>
<td></td>
<td>$n=19$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=13$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow up 2</td>
<td>1</td>
<td>0.769 **</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p=0.0034$</td>
<td></td>
<td></td>
<td>$n=14$</td>
</tr>
<tr>
<td>Follow up 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 8-3. Scatterplots depicting the serious game median target offset for correct trials between: (A) initial enrolment compared to follow up session 1 ($r=0.822 \ p<0.001$), (B) follow up session 1 compared with 2 ($r=0.818, \ p=0.002$), and (C) follow up session 2 compared with 3 ($r=0.492, \ p=0.105$).
Figure 8-4. Bar chart illustrating the mean of median correct target offset (px) between initial enrolment compared to follow up session 3. Error bars are 95% CI.
8.5 Appendix E. Supporting Materials for Re-Assessing the Concurrent Validity of the Serious Game in Elderly Adults

Table 8-15. Detailed information on the different game variants played by patients.

<table>
<thead>
<tr>
<th>Game Number</th>
<th>Grid Size</th>
<th>Target Size (px)</th>
<th>Stimuli</th>
<th>Level Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2x2</td>
<td>250</td>
<td>• Moles</td>
<td>Beginner</td>
</tr>
<tr>
<td>1</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Intermediate</td>
</tr>
<tr>
<td>1</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Hard</td>
</tr>
<tr>
<td>1</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Harder</td>
</tr>
<tr>
<td>2</td>
<td>3x3</td>
<td>250</td>
<td>• Moles</td>
<td>Beginner</td>
</tr>
<tr>
<td>2</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Intermediate (same as above in version 1)</td>
</tr>
<tr>
<td>2</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Hard (same as above in version 1)</td>
</tr>
<tr>
<td>2</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Harder (same as above in version 1)</td>
</tr>
<tr>
<td>3</td>
<td>3x3</td>
<td>250</td>
<td>• Moles</td>
<td>Beginner</td>
</tr>
<tr>
<td>3</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Intermediate (same as above in version 1)</td>
</tr>
<tr>
<td>3</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Hard (same as above in version 1)</td>
</tr>
<tr>
<td>3</td>
<td>5x3</td>
<td>175</td>
<td>• Moles</td>
<td>Harder (same as above in version 1)</td>
</tr>
</tbody>
</table>
Table 8-16. Summary of study sample demographics and cognitive assessment scores. MMSE = Mini-Mental State Examination, CAM = Confusion Assessment Method, and DI = Delirium Index.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Males (n=82)</th>
<th>Females (n=88)</th>
<th>Total (n=170)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD) / Median (IQR)</td>
<td>Range</td>
<td>Mean (SD) / Median (IQR)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>76.7 (8.2)</td>
<td>65 – 98</td>
<td>78.7 (8.7)</td>
</tr>
<tr>
<td>MMSE</td>
<td>26.5 (3.8)</td>
<td>12 – 30</td>
<td>26.7 (3.7)</td>
</tr>
<tr>
<td>CAM</td>
<td>0.03 (0.2)</td>
<td>0 – 1</td>
<td>0.0 (0.0)</td>
</tr>
<tr>
<td>DI</td>
<td>1.5 (2.4)</td>
<td>0 – 12</td>
<td>1.5 (2.5)</td>
</tr>
<tr>
<td>Game Correct RT (msec) a</td>
<td>1140 (300)</td>
<td>500 – 2200</td>
<td>1180 (400)</td>
</tr>
<tr>
<td>Game False Alarm Rate</td>
<td>0.088 (0.1)</td>
<td>0.009 – 0.6</td>
<td>0.096 (0.1)</td>
</tr>
<tr>
<td>Game Combined Performance Metric</td>
<td>-0.12 (1.3)</td>
<td>-5.1 – 1.8</td>
<td>-0.20 (1.4)</td>
</tr>
</tbody>
</table>

a For the game correct RT row, the median (IQR) is reported.

b There were 79 patients who did not have any false alarms, and thus their false alarm rate could not be calculated.

Table 8-17. Spearman’s rho correlations comparing game performance to the standard cognitive assessments. The correlation coefficient is listed on the first line, followed by the p-value and n. * p<0.05, ** p<0.01, *** p<0.001. MMSE = Mini-Mental State Examination, CAM = Confusion Assessment Method, and DI = Delirium Index.

<table>
<thead>
<tr>
<th></th>
<th>Game Correct RT</th>
<th>Game False Alarm Rate</th>
<th>Game Combined Performance Metric</th>
<th>MMSE</th>
<th>CAM</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Correct RT</td>
<td>1</td>
<td>0.151 *</td>
<td>-0.833 ***</td>
<td>-0.494 ***</td>
<td>-0.00841</td>
<td>0.367 ***</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p=0.156</td>
<td>p=0.001</td>
<td>p=0.001</td>
<td>p=0.925</td>
<td>p=0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n=90</td>
<td>n=131</td>
<td>n=83</td>
<td>n=128</td>
<td>n=118</td>
</tr>
<tr>
<td>Game False Alarm Rate</td>
<td>1</td>
<td>-0.5874 ***</td>
<td>-0.560 ***</td>
<td>0.0992</td>
<td>0.328 **</td>
<td>p=0.0016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p=0.001</td>
<td>p=0.001</td>
<td>p=0.358</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>n=90</td>
<td>n=56</td>
<td>n=88</td>
<td>n=80</td>
<td>n=80</td>
</tr>
<tr>
<td>Game Combined</td>
<td>1</td>
<td>0.582 ***</td>
<td>-0.6844</td>
<td>-0.06844</td>
<td>-0.506 **</td>
<td>p=0.0030</td>
</tr>
<tr>
<td>Performance Metric</td>
<td></td>
<td>p&lt;0.001</td>
<td></td>
<td>p=0.443</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>n=83</td>
<td></td>
<td>n=128</td>
<td>n=118</td>
<td></td>
</tr>
<tr>
<td>MMSE</td>
<td>1</td>
<td></td>
<td>-0.143</td>
<td>-0.659 ***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p=0.181</td>
<td>p=0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n=89</td>
<td>n=85</td>
<td>n=85</td>
<td></td>
</tr>
<tr>
<td>CAM</td>
<td>1</td>
<td></td>
<td>0.209 *</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p=0.0130</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>n=140</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DI
Figure 8-5. Scatterplot illustrating the differences on game correct RT performance based on MMSE score ($\geq 24$ = normal cognitive function or possible dementia; $< 24$ = signs of dementia). MMSE = Mini-Mental State Examination.

Figure 8-6. Scatterplot illustrating the differences on game false alarm rate performance based on MMSE score ($\geq 24$ = normal cognitive function or possible dementia; $< 24$ = signs of dementia). MMSE = Mini-Mental State Examination.
8.6 Appendix F. Supporting Materials for the Construct Validation of a Serious Game for Assessment of Executive Functioning.

8.6.1 Experimental Materials.

8.6.1.1 Recruitment Flyer

Call for Participation in a Human Factors Research Project

Screening
We are seeking people who:

• age 18 years or older;
• can read texts on a computer monitor display and handheld tablet;
• have no self-reported difficulty in hearing or understanding English instructions;
• have no self-reported problem distinguishing between red, blue, purple, green, yellow, orange, brown and gray;
• have no self-reported history or current cognitive impairments (ex. dementia, mild cognitive impairment, delirium)

Please do not sign up to participate in this study unless you meet these criteria.

Experiment
For those volunteers who are invited to participate in the experiment (based on their screening results) the details of the experiment are as follows:

Period: from January 1, 2018 to January 1, 2019
Place: 8th floor in the Bahen building at St. George campus of University of Toronto
We are studying how different cognitive tasks compare to a tablet-based game measuring similar metrics.

Participants will participate in the experiment individually. First, participants will perform cognitive capacity tasks presented on a computer. Each cognitive task will take only a few minutes. Next, participants will be asked to play a game on a tablet. The entire experiment should finish within 1 hour and 30 minutes. Participants will be compensated $20.00 CAD for taking part in this study.

If you are interested in participating, please complete the screening questionnaire (accessible through the QR code above): http://givebuttons.com/research/recruitment or send an email to tiffany.tong@mail.utoronto.ca. Questions are also welcome to the same address.

Thank you,

Tiffany Tong
PhD Candidate
Department of Mechanical and Industrial Engineering, University of Toronto

8.6.1.2 Consent Form

Client Information Sheet and Informed Consent Form for the Study: Examining The Relationship Between Serious Game Parameters and Inhibition Ability

Dear Sir/Madam: Date: ______________________
Thank you for your interest in this research project. This letter has been created to provide you with the information needed so that you may decide whether you would like to participate in this study. Participation is voluntary and you are free to withdraw or stop at any time.

If at any point you feel as though any of the following details are unclear, or if you have any other questions, comments, or concerns, please feel free to contact me using the contact information at the end of this letter. If you decide that you would like to participate, please date, and sign the last page of this letter then return one copy to me and keep the other for your reference. If you do not wish to participate, there is no need to return the form. Please note, you may request a copy of our final study if desired.

The long-term goal of our research is to use digital, mobile-based games/cognitive assessments to detect changes in cognition. This study aims to understand the effectiveness of computerized assessments.

The design of the experiment is as follows. All participants will participate in the experiment individually, one at a time. Each person will be asked to perform a series of target identification tasks presented on a computer. Next, a tablet with a game will be presented to the participant and the experimenter will introduce the game. The participant will play the game on the tablet.

The participant may decline to answer any questions.

The data we collect will be anonymized and kept in a secure office. No personal or identifying information will be included in written reports or presentations, and your confidentiality and privacy will be respected at all times. Any data and information received will be kept confidential. Any study reports and presentations will have all personal identifiers removed. Data and participant information will be kept in my possession or stored in a locked office accessible only by me and the other investigators. Electronic information will be password protected. All data will be securely stored until January 1, 2022. All data will be destroyed after January 2, 2022.

The experiment should take 1 hour and 30 minutes to complete and will be held in the Bahan building at 40 St. George Street on the University of Toronto St. George Campus. There is a $20.00 compensation for participating in the study.

As mentioned previously, if you have any questions, you may contact me at tiffany.tong@mail.utoronto.ca or 416-978-7581. Alternatively, you may call the Office of Research Ethics at ethics.review@utoronto.ca or 416-946-3273.

Thank you for your consideration,
Tiffany Tong
PhD Candidate, University of Toronto

To be completed by the participant:

☐ I have read this consent form and I understand the research and what is expected of me.
☐ I understand that:
  • I am free to withdraw before or anytime during the study without the need to give any explanation.
  • I am free to elect to skip parts of the study without the need to give any explanation.
☐ I agree to participate in this study.

If I do not wish to participate in the research, I can just keep the form without signing it.

_____________________________  ____________________________  ____________________________
(Participant Signature)        (Participant Name, please print)        (Date)

_____________________________  ____________________________
(Investigator Signature)       (Investigator Name, please print)

_____________________________
(Date)
8.6.1.3 Demographics Questionnaire

Please answer the following questions to the best of your ability. If at any point you feel uncomfortable answering a question feel free to leave it blank and move on to the next one. If you ever require assistance or if you no longer wish to participate in the study, please let one of the researchers know so that they can accommodate your requests.

Sex:
- [ ] Male
- [ ] Female
- [ ] Prefer not to answer

What is your age?

What is your primary/preferred language?

What is your predominant hand?
- [ ] Right
- [ ] Left
- [ ] Ambidextrous (both hands)

How did you find out about this research study?
- [ ] Poster around the University of Toronto
- [ ] E-mail mailing list
- [ ] Online posting (i.e., Kijiji, Craigslist)
- [ ] Other (please specify)

Which of the following best describes the highest level of education that you have completed?
- [ ] Did not attend school
- [ ] Primary School
- [ ] Some Secondary School (High School)
- [ ] Secondary School (High School)
- [ ] Some Undergraduate College/University
- [ ] Undergraduate College/University
- [ ] Some Graduate College/University
- [ ] Graduate College/University
- [ ] Prefer not to answer
- [ ] Other (please specify)

If you are currently in registered in a post-secondary institution, please indicate your level:
- [ ] Undergraduate
- [ ] Graduate
- [ ] Other (please specify)

If you are currently in registered in a post-secondary institution, please indicate your faculty and department:
- Faculty:
- Department:

**Computer Skills**
Approximately, how often do you use a computer?
- [ ] Never
- [ ] Once per month
- [ ] Once every two week
- [ ] Several times a week
- [ ] Daily

**Portable Device Skills**
Approximately, how often do you use a touch-based device (ex. computer, tablet, smartphone)?
- [ ] Never
- [ ] Once per month
- [ ] Once every two week
- [ ] Several times a week
What type of touch-based devices do you use most frequently (please check all that apply)?

☐ Smartphone (ex. iPhone)
☐ Media player (ex. iPod)
☐ Tablet (ex. iPad, Android tablet)
☐ Laptop
☐ Computer monitor
☐ Gaming console
☐ Other (please specify)

Do you play games on your mobile phone?

☐ Yes  ☐ No  ☐ Prefer not to answer

If yes, how often do you play games on your mobile phone?

☐ Never  ☐ Once per month  ☐ Once every two week  ☐ Several times a week  ☐ Daily

8.6.1.4 Participant Compensation Receipt

Date (yyyy/mm/dd): _____________________________

I, _____________________________, have received $_____ CAD for participating in the research study “Examining The Relationship Between Serious Game Parameters and Inhibition Ability“.

Participant Signature
Participant Name (please print)
Investigator Signature
Investigator Name (please print)

8.6.2 Inhibition Task Implementation

In the Stroop Test, there were six color words (‘black’, ‘white’, ‘yellow’, ‘orange’, ‘purple’, and ‘green’) presented in one of the six same font colors individually and at random (Stroop, 1935). There are 36 possible word-font color combinations. On each trial, three-color names (response alternatives) were presented in black at the bottom of the display. The participant's task was to respond with the color in which the stimulus word was written, by pressing a corresponding key. The three response alternatives will be mapped to the left arrow key, down arrow key, and right arrow key, respectively.

In the Go/No-Go Discrimination Task: In this task, a two-by-two square grid was presented to participants. In this grid, there were three of the cells contain a symbol (e.g. an asterisk, *), and one cell contains a stimulus (e.g. the letter P for the go target, and the letter R for the no-go target). The objective was to respond only to the go-target (e.g. letter P) when it appeared in the two-by-two grid using the shift key on the keyboard (Yechiam et al., 2006). Feedback was
provided (“correct” or “incorrect”) once the participant provides a response. The data collected from this task includes response accuracy (e.g. correct or incorrect) and response time. There was a high target frequency block, followed by a low target frequency block.

In the Anti-saccade task, a square appeared on either the left or right side of the screen. Next, a second square appeared on the opposite side of the screen containing an arrow, which was either facing up, down, left or right. Shortly after the arrow appeared, it disappeared and was replaced by a small-hashed square. The participant’s task was to indicate the direction that the arrow was pointing. The four responses alternatives were mapped to the up arrow key, down arrow key, left arrow key, and right arrow key.

In the Flanker task, an arrow as shown in the center of the screen, either pointing to the left or right, and flanked by other arrows. In congruent trials, the flanking arrows pointed in the same direction as the center target. In contrast, in incongruent trials, the flanking arrows pointed in the opposite direction as the center target.

8.6.3 Supporting Methodology Materials

<table>
<thead>
<tr>
<th>Cognitive Task</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Saccade Task</td>
<td>• Median CRT</td>
</tr>
<tr>
<td></td>
<td>• Error rate</td>
</tr>
<tr>
<td>Flanker Task</td>
<td>• Median CRT</td>
</tr>
<tr>
<td></td>
<td>• Error rate</td>
</tr>
<tr>
<td>Go/No-Go Discrimination Task</td>
<td>• Median CRT on high target frequency trials (block 1)</td>
</tr>
<tr>
<td></td>
<td>• Error rate on high target frequency trials (block 1)</td>
</tr>
<tr>
<td></td>
<td>• Median CRT on low target frequency trials (block 2)</td>
</tr>
<tr>
<td></td>
<td>• Error rate on low target frequency trials (block 2)</td>
</tr>
<tr>
<td>Stroop Task</td>
<td>• Stroop interference value</td>
</tr>
</tbody>
</table>
8.6.4 Supporting Results Materials.

Table 8-19. Summary of exploratory factor analysis results for the game median correct RT data ($n=30$).

<table>
<thead>
<tr>
<th>Item</th>
<th>Component 1: Game Median Correct RT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median CRT Block 6</td>
<td>0.965</td>
</tr>
<tr>
<td>Median CRT Block 7</td>
<td>0.961</td>
</tr>
<tr>
<td>Median CRT Block 8</td>
<td>0.949</td>
</tr>
<tr>
<td>Median CRT Block 5</td>
<td>0.945</td>
</tr>
<tr>
<td>Median CRT Block 2</td>
<td>0.933</td>
</tr>
<tr>
<td>Median CRT Block 4</td>
<td>0.922</td>
</tr>
<tr>
<td>Median CRT Block 3</td>
<td>0.887</td>
</tr>
<tr>
<td>Median CRT Block 1</td>
<td>0.886</td>
</tr>
<tr>
<td>Eigenvalue Value</td>
<td>6.94</td>
</tr>
<tr>
<td>% Variance</td>
<td>86.78</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 8-20. Summary of exploratory factor analysis results for the game false alarm rate data ($n=30$).

<table>
<thead>
<tr>
<th>Item</th>
<th>Component 1: False Alarm Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>False Alarm Rate Block 2</td>
<td>0.906</td>
</tr>
</tbody>
</table>
Figure 8-8. Diagram illustrating the relationship between different measures of inhibition.
8.7 Appendix G. Literature Review on Response Time as a Measure of Executive Function.

The earliest use of RT as a measure in experimental psychological studies can be traced back to work by Donders in the 1860s on differences in simple versus choice RT (Miller & Low, 2001). This research led to his investigation on mental chronometry, the measurement of speed of mental process. This area was further explored by Sir Francis Galton whose work focused on RT differences in physical and behavioral variables from an anthropometric context.

8.7.1 Executive Function

RT variability in terms of intra-individual variability (IIV), differences in an individual’s performance at different time points, has been associated to changes in EF. Variability can also be examined based on between-tasks variability, where a participant is evaluated on multiple tasks at one occasion and between-occasions, where a participant is evaluated on one task at multiple occasions.

In general, a greater variability in RT IIV is often associated with poorer performance in certain cognitive domains as well as abilities. For example, work by Bunce, Tzur, Ramchurn, Gain and Bond (2008) looked at simple reaction time (SRT) and CRT correlations on EF tasks for shifting (assessed using a verbal fluency task), updating (assessed using an n-back task), and inhibition (assessed using a Stroop Task). They found that individuals with greater variability in their SRT and CRT had lower performance on switching and inhibition. Their results also revealed that individuals with higher CRT variability had lower updating performance. Further work by Hultsch, MacDonald, Hunter, Levy-Bencheton and Strauss (2000) looked at RT IIV on four different RT-based tasks, which included tasks on: CRT, SRT, lexical decision RT, and semantic decision RT. They found that variability was positively correlated with mean RT and negatively correlated with accuracy. They observed that participants with slower RT more errors. Moreover, research carried out by Burton, Strauss, Hultsch and Hunter (2009) examined at RT IIV in relation to one’s ability to perform activities of daily living such as shopping using the telephone. Examples tasks included calculating the cost of a taxi fare and determining the caloric value of
one serving of cereal. They found that individuals with less variability had better performance on these tasks.

Age-related decline in RT IIIV has also been shown in previous work, where it has been observed that older adults have larger variability in RT (Hultsch, MacDonald, & Dixon, 2002; Van der Elst, Van Boxtel, Van Breukelen, & Jolles, 2006). Looking within older adults, there also exists variability where the oldest old have larger RT variability compared to the younger old (Dixon et al., 2007).

### 8.7.2 Brain Injury

Greater and inconsistent variability in RT IIIV has also been associated with patients with brain damage. Bruhn and Parsons (1971) compared patients with and without brain damage, and matched patients based on age and education. They reported that patients with brain damage demonstrated greater RT variability, as measured using individual standard deviation, compared to the control group. This team proposed that this observed result may be due to general slower rather than brain damage. Moreover, the location and severity of brain injury has also been investigated to determine their role in RT IIIV. For example, Stuss, Murphy, Binns and Alexander (2003) looked at traumatic brain injury patients with injuries in the following four locations: left dorsolateral frontal, right dorsolateral frontal, inferior medial, and superior medial. Stuss et al. (2003) found no differences in RT variability in patients with injuries to their inferior medial or superior medial regions. In contrasts, significant differences were observed in patients with injury to their left or right dorsolateral frontal regions compared to control groups. These results led Stuss et al. (2003) to suggest that the frontal regions may play an important role in RT IIIV. Moreover, research on the severity of brain injuries in relation to variability in RT IIIV has yet to show any significant relationships (Hetherington, Stuss, & Finlayson, 1996).

Related work on the effect of traumatic brain injury on RT IIIV has also looked into the time since injury. Further work by Hetherington et al. (1996) compared two groups of traumatic brain injury patients at 5 and 10 years post-injury to a control group of patients with no brain injury. They found that their group of patients at 5 years post-injury had larger RT IIIV compared to both the control and group of patients 10-years post injury. Additional work by Collins and Long (1996) looked at mean and IIIV SRT in two patients groups that had traumatic brain injury and no cognitive impairment, and a second group of patients with traumatic brain injury that had
cognitive impairment, and then compared both groups to a healthy control group. In this study, the authors found that patients with traumatic brain injury, regardless of having a cognitive impairment, had in general, slower and more variable RTs on a SRT and Go/No-Go discrimination task, compared to the control group. They also observed that mean RT was better able to discriminate between traumatic brain injury patients with a cognitive impairment compared to the control, in contrast, RT IIV was able to discriminate between non-impaired traumatic brain injury patients compared to the control. The findings from this study may suggest that RT IIV is a more sensitive measure of cognitive impairment that may be missed by a standard neuropsychological task such as the go/no-go discrimination task.

8.7.3 Dementia

The relationship between RT IIV and cognitive impairments such as dementia (specifically AD), mild cognitive impairment, and delirium has also been investigated. In general, impaired, older adult patients typically demonstrate larger variability in their RT. In particular, an increase in RT IIV is often associated with dementia patients (Walker, Ayre, Cummings, et al., 2000; Walker, Ayre, Perry, et al., 2000). As dementia is a broad term that covers many different types including AD, and Dementia with Lewy Bodies (DLB), it is important to characterize the relationships between RT IIV based on the study sample. For example, work by Walker, Ayre, Cummings, et al. (2000), and Walker, Ayre, Perry, et al. (2000) compared RT IIV in AD and DLB compared to a healthy control group. While they found that both dementia groups had more variable RT IIV compared to the control, they also found that the DLB group had greater SRT and CRT IIV compared to the AD group. This work highlights the differences in RT IIV between different dementia subtypes. Their findings also suggested that a RT-based task be used for dementia screening in patients with DLB. In addition, work by Murtha, Cismaru, Waechter and Chertkow (2002) looked at differences in RT IIV in AD and frontal lobe dementia (FLD) patients to a healthy control group. They found that AD patients had greater RT IIV compared to frontal lobe dementia (FLD) patients. Their findings were non-significant, which may be attributed to a small sample group of only five FLD and 8 AD patients. Further work with FLD patients are necessary to investigate the role of the frontal lobes in RT IIV.
8.7.4 Mild Cognitive Impairment

Similar to the trend observed between increased RT IIV and dementia patients, the relationship is seen with mild cognitive impairment. Work by Dixon et al. (2007) examined the RT IIV in patients over 64 on CRT and SRT tasks. They found that impaired patients with MCI had higher SRT IIV compared to non-impaired adults. In contrast, there was no difference in CRT IIV in either impaired, and non-impaired groups. They also demonstrated that SRT IIV was a better predictor for the presence or absence of MCI over the use of mean RT. In a second sample, Dixon et al. (2007) looked at patients with mild and moderate MCI as well as a healthy control group. They observed greater variability in SRT IIV and CRT IIV in the moderate MCI group compared to the mild MCI and control groups. They also reported that CRT was significantly higher in the mild MCI group compared to controls. In addition, they also demonstrated that both SRT IIV and CRT IIV were strong predictors between no impairment and moderate MCI, and between mild and moderate MCI compared to using mean RT. Overall, these two samples studied by Dixon et al. (2007) suggest that patients with MCI demonstrate larger differences in SRT IIV and CRT IIV.

Another study by Christensen et al. (2005) looked at the relationship between cognitive impairment and RT IIV in older adults between the ages of 60 and 64 years. They demonstrated the same trend, where SRT IIV and CRT IIV were greater in impaired individuals compared to non-impaired individuals. Through further investigation, they looked at the relationship between RT variability in each of the subgroups (age-related cognitive decline, age-related memory impairment, and MCI) of cognitively impaired individuals. Based on this subgroup analysis, the observed difference in IIV was no longer significant in individuals with MCI compared to healthy participants.

Moreover, another study by Strauss, Bielak, Bunce, Hunter and Hultsch (2007) confirms earlier findings on the use of RT IIV as a stronger measure compared to mean RT in predicting MCI status. This study examined older adults over 64 years on their performance on RT-based tasks. They found that MCI patients had greater variability in CRT performance compared to non-impaired adults. They also found that RT variability was a strong predictor of impairment group membership compared to mean RT.
8.7.5 Using RT IIV as a Predictor of Cognitive Ability and Performance

As previously mentioned, the use of RT IIV has the ability to distinguish between patient groups with and without cognitive impairment as well brain injury. For example, research by Bruhn and Parsons (1971) demonstrated that differences in RT IIV could distinguish between healthy patients and those with brain injuries. In addition, research by Hultsch et al. (2000) was able to discriminate between patients with and without dementia using both SRT IIV and CRT IIV, which were better measures than using mean RT. The use of RT IIV has also been used as an indicator to predict cognitive decline (Lövdén, Li, Shing, & Lindenberger, 2007) and progression from healthy to impaired states. For instance, work by Cherbuin, Sachdev and Anstey (2010) has demonstrated that RT IIV can predict an individual’s progression from a healthy, non-impaired state to MCI. The use of CRT IIV plus one standard deviation also been shown to strongly predict the onset of delirium in postoperative patients following a knee or hip replacement as shown in work by Lowery et al. (2007). In their study, their CRT task required patients to respond by pressing either a “NO” or “YES” button on a screen based the stimulus.

8.7.6 Measures of RT IIV and Methodological Issues

RT II can be used as an indicator of an individual’s performance (Hultsch et al., 2002; Nesselroade & Salthouse, 2004; Rabbitt, Osman, Moore, & Stollery, 2001). As previously seen, RT IIV can be measured using different methods. Two of the most common ways use intra-individual standard deviation (ISD) and intra-individual variance (IVar). Both ISD and Ivar are calculated using an individual’s own mean RT across all trials. Limitations surrounding the use of both of these measures of IIV include the issue that RT is typically positively skewed. In addition, RT IIV is also sensitive to outliers since it is calculated across all trials. One common way to control for outliers is to trim the data by excluding very short and very long trials, which may be a result of accidental or delayed responses. Previous work by Dykiert, Der, Starr and Deary (2012), and Yan, Thomas and Stelmach (1998) have used a cut-off value of 150 msec as a minimum, and an upper bound of 3 standard deviations above the mean (Hultsch et al., 2002; Williams, Hultsch, Strauss, Hunter, & Tannock, 2005).

However, an issue with this method is it is difficult to differentiate between intentional versus abnormal trials. Another method to address methodological issues with using RT IIV is to use inter-quartile ranges of RT data. One approach is to exclude data in the 25th and 75th percentiles,
which leaves only 50% of the data for analysis. A similar method by Adam, Rétey, Khatami and Landolt (2006) uses data between the 10th and 90th percentiles. An issue with selecting data based on percentile is that RT then becomes trimmed and does not account for what occurs on two extremes of either responding too quickly or slowly.

In using RT IIV as a measure of performance, it is important to control for mean RT. There are multiple approaches such as creating a coefficient of variation, using mean RT as a covariate, and carrying out a whole distribution analysis. In the first approach, a coefficient of variation (CV) is calculated for individuals by dividing their RT ISD by one’s intra-individual mean. When a participant’s IIV increases in proportion to their RT mean, there should be no difference CV between individuals. In the case where there exist differences in CV, then it can be assumed that differences in RT IIV are not due to an individual’s difference in their mean RT. The CV is often used due to its ease in calculation and simplicity. This approach is however, not without its limitations, such that it may confound unsystematic and systematic variation. For instance, trial-to-trial RT performance may improve due to practice, which would result in a decrease in RT for one block.

A second method for controlling for mean RT involves including mean RT as a covariate. This method allows researchers to model the relationship between mean and ISD, however it does not control for systematic sources of variation such as fatigue and practice effects across a block of trials. Moreover, another method is to using regression modeling with purified residuals. In this approach, RT is regressed against potential confounders such as age, sex, and trial number, in an attempt to remove systematic variation. The resulting residuals are often referred to as purified residuals as they are independent of confounding effects. These purified residuals are then used to calculate an individual’s RT ISD. Potential issues with this approach is that predicted values when controlling for confounding variables such as age and sex will result in the same predicted values for individuals with the same age and sex.

The fourth and final method to control for mean RT is to use a whole distribution analysis. This procedure involves fitting a convolution of a Gaussian and exponential distribution to the RT data. This method renders three variables, a mean, and standard deviation for the normal component, and a tau for the exponential component. The mean provides information on the
average performance, the standard deviation represents variability, and tau reflects the distributed tail of the distribution.
8.8 Appendix H. Assessing Cognitive Status in Elderly Adults with Prevalent Delirium Study.

8.8.1 Assessing Cognitive Status in Elderly Adults with Prevalent Delirium

This chapter describes the evaluation of the serious game in screening for prevalent delirium in a sample of elderly adults selected from a hospital ED. The chapter begins with the motivation for the study, followed by the study design, results, and discussion of the findings.

8.8.1.1 Motivation

In Chapter 4, the serious game was evaluated in older adults using emergency services that were present in a hospital ED for a minimum of eight hours, as the goal was to screen for incident delirium, which is delirium that occur during a patient’s hospital stay (Anderson, 2005). In contrast, this chapter will focus on evaluating the concurrent validity of the game-based assessment tool with older adults with prevalent delirium, which is delirium that is present on admission to a hospital (Anderson, 2005).

The prevalence of delirium in a hospital ED has been estimated to range between 10 to 17% (Elie et al., 2000; Hustey & Meldon, 2002; Hustey et al., 2000; Kakuma et al., 2003; Lewis et al., 1995). The occurrence of delirium can lead to adverse outcomes such as functional decline, and increased risk for mortality. In North America, delirium in the ED is unrecognized in 75 to 83% of patients (Hustey & Meldon, 2002; Hustey, Meldon, Smith, & Lex, 2003; Hustey et al., 2000). One possible explanation for the lack of delirium recognition includes a clinician’s gap in knowledge on a patient’s cognitive status over time (Stiell, Forster, Stiell, & van Walraven, 2003). In a fast-paced and time-pressured environment such as a hospital ED, the ability to carry out delirium assessment repeatedly is challenging and not often operationally feasible due to limited resources.

The potential to recognize delirium upon admission could assist patients to receive the proper intervention strategies to prevent further deterioration and minimize adverse outcomes. This provides the motivation to assess the ability of the serious game in a hospital ED to screen for prevalent delirium in older adults.
8.8.1.2 Research Question

The research question that this chapter will address is as follows:

1. How well does performance on the serious game played by older adults in a clinical setting predict the presence of delirium as determined using standard tests such as the CAM, and DI? How well does game performance agree with other methods of cognitive assessment in a clinical setting (i.e., MoCA).

The research question was explored by examining correlations (using both Pearson’s $r$ and Spearman’s $\rho$) between measures of cognitive status (e.g., MoCA, CAM, DI) with game performance (median correct RT).

8.8.1.3 Methodology

8.8.1.3.1 Software Design and Development

The serious game-based cognitive screening tool was implemented on a tablet along with a battery of digitized cognitive, functional, and frailty assessments. This software was available in both French and English. The entire application, including the serious game, was developed in the programming language C# using the Unity engine. The change in software systems from Java to C# was for ease of implementation. The motivation to change the game platform was to create a game that could be used on other platforms in addition to Android.

The serious game was originally designed with three levels that varied based on the game board size, type of stimuli as well as the number of targets and distractors that could appear on the game board at once. The targets were displayed using 175 px.

After piloting the game, two of the three levels were revised to increase their difficulty by increasing the number of targets on the game board at once. All variants of the game can be found in Appendix I. Following this, a practice level was added, which was then revised to increase the practice difficulty. In the final iteration of the game, there were three levels and one practice level (Appendix I). The data analysis excluded all practice data.
Compared to the version of the serious game used in Chapter 4, the levels are comparatively harder due to the presence of multiple targets on the game board at once. The targets were also set to a smaller target size of 175 px.

8.8.1.3.2 Study Protocol

This study included adults over 65 years of age who called 911 for paramedic services, or who presented to a participating ED. Participating EDs included the following four Canadian EDs: (a) Sunnybrook Health Sciences (Ontario), (b) L’hopital L’Enfant Jesus (Quebec), (c) Nova Scotia Emergency Health Services (Nova Scotia), and (d) Ottawa Civic and Ottawa General Hospital (Ontario) recruited patients. Clinical RAs recruited eligible patients in the waiting room of participating EDs. Patients who arrived by ambulance were recruited on stretchers or in the ambulance off-load area while awaiting transfer to an ED bed. A research ethics cover sheet along with the protocol from Sunnybrook Health Sciences Centre (protocol #166-2015) was also approved by the research ethics board at the University of Toronto (protocol #32025).

Patients were approached by RAs and asked to play the serious game on the tablet. After using the device and software, patients were asked to complete a system usability scale (SUS) (Brooke, 1996), a modified questionnaire used to examine history of previous technology use (Tierney et al., 2014), a single ease questionnaire, and were asked for feedback.

8.8.1.3.3 Patient Sample Population

The patient inclusion criteria for this study included adults 65 years of age and older who called 911 for paramedic services or who presented to an ED. Exclusion criteria included patients who: 1) lived in a full care nursing home; 2) had a critical illness rendering them unable to communicate or provide consent; 3) had known moderate to severe cognitive impairment assessed using the MoCA (Nasreddine et al., 2005); 4) had visual impairment that makes them unable to use the tablet; or 5) had other communication difficulties preventing use of the tablet. Patients who met the inclusion criteria and were not excluded, were invited to participate, and could decline for any reason, such as being critically ill.

8.8.1.3.4 Requirements Analysis

I conducted a requirements analysis with a physician and a team of clinical RAs at the participating institutions who worked in emergency services with elderly adults. Insight on the
needs of patients and caregivers were provided by these informants, based on their previous experience, and on their experience with the tablet technology in this study.

8.8.1.3.5 Usability Study

The research team held weekly online and phone meetings to review and discuss ongoing usability feedback gathered from the SUS, modified technology questionnaire, and single ease questionnaire.

8.8.1.3.6 Apparatus

The tablets used in this study were Samsung Galaxy Tab 4 tablets with 10.1-inch screens manufactured by Samsung (Samsung, 2016).

8.8.1.3.7 Statistical Analysis

A similar statistical analysis approach used in Chapter 4 was used in this study to analyze the cognitive data and game results (Chapter 4 – Section 4.1.5). My interest was in correlations as a measure of the effect size of the underlying relationship between game performance and the cognitive assessments, but I used nonparametric correlation measures for some of the comparisons that involved categorical or narrow ordinal scales. Correlations between the dichotomous CAM and the other measures were assessed using point-biserial correlations (Field, 2013). Correlations involving the DI (and not involving the CAM) were assessed using Spearman’s rho because the DI uses a small number of ordered categories. The remaining comparisons were done using Pearson’s r correlations.

The game data was analyzed using median correct RT. The target offset was calculated for correct trials only and can be found in Appendix I, but will not be presented in the results.

8.8.1.4 Results

8.8.1.4.1 Description of Sample Population

330 patients were enrolled in the study (124 males, 109 females, 97 unreported), between 65 to 94 years (mean age = 75.7 years, standard deviation = 7.6).

There were a number of participants in the sample with low MoCA scores (down to 0). There were 126 participants who were negative for the CAM and only 3 participants who were positive
(a positive result on the CAM suggests that the patient has delirium). Moreover, the DI scores ranged from 0 to 10 (the score indicates the severity of delirium). The overall median correct RT on the serious game was 1200 msec (IQR 500). A summary of the scores on the cognitive assessments can be found in Table 8-16.

8.8.1.4.2 Requirements Analysis

The requirements analysis provided information about technologies to use, different use-cases, and scenarios. It provided information on the diverse needs of the patient population that emergency services personnel such as clinicians, paramedics, and RA work with.

Based on the requirements analysis, the decision to use tablets was selected as a medium for the serious game. The tablet device provided a lightweight and portable form-factor that is well suited to busy environments such as emergency care. The tablets could be used with capacitive styluses, which were useful for patients who were frail or had decreased dexterity skills. The tablets were also familiar to the paramedics on the research team who had previous experience using tablet-based technologies to record patient information.

8.8.1.4.3 Usability

Usability feedback indicated that patients and caregivers were interested in using technology as part of their health care experience. However, participants commented that the “tablet was too heavy”, and that “it’s heavy, make it lighter, or have a book stand”. Another emergent theme related to the use of technology for delivering assessments. On the negative side, feedback included “it would be confusing for those not familiar with technology”, “right direction but adopting new idea might be difficult”, and “I’m an old fashioned guy not into tech”. In contrast, positive feedback regarding the use of technology included “it’s great”, “innovative and good”, “very interesting and helpful”, and “would like to see more of this in the near future”.

8.8.1.4.4 SUS

264 patients completed the SUS. The mean SUS score was 76.4 (standard deviation = 18.9), and the range was from 22.5 to 100. There were 92 (34.7 %) scores below the acceptable usability score of 68 (Brooke, 1996). However, the interactive, computerized tool was usable by most elderly emergency patients.
### 8.8.1.4.5 Modified Technology History Questionnaire

The modified technology history questionnaire contained five questions regarding patient use and experience with computer, smartphones, and tablet technology (Table 8-23). 60.9% indicated that they have a computer in their home. Another 56.8% indicated that they use a computer daily. Moreover, 41.7% indicated that they have a smartphone/tablet. Of those that owned a smartphone/tablet, 42.0% indicated that they play games on their device. In the same subset, 46.7% of patients indicated that they used their device daily. Over half of the users indicated that they were comfortable with using a computer, tablet, or other electronic device.

### 8.8.1.4.6 Comparison between Game and Cognitive Performance

The concurrent validity of the game was compared to performance on the MoCA, CAM, and DI. Game performance data was analyzed using median correct RT. A Pearson’s $r$ (Table 8-21) and Spearman’s $\rho$ (Table 8-26) correlation analysis was carried out between measures of cognitive status (e.g., MoCA), delirium (e.g., CAM) and delirium severity (e.g., DI) and game performance. The Pearson’s $r$ correlations revealed that the game median correct RT results were correlated with the MoCA and DI but not with the CAM.

As a follow-up to the Pearson’s $r$ correlation analysis in Table 8-21, I carried out the same analysis using Spearman’s $\rho$ correlations (Table 8-26). The significant correlations between the MoCA and DI with the game median correct RT, were also observed to be significant using Spearman’s $\rho$.

#### Table 8-21. Two-tailed Pearson’s $r$ correlation between game performance and cognitive measures. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. MMSE = Mini-Mental State Examination, CAM = Confusion Assessment Method, and DI = Delirium Index.

<table>
<thead>
<tr>
<th></th>
<th>Game Median Correct RT</th>
<th>MoCA</th>
<th>CAM $\rho$</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game Median Correct RT</td>
<td>1.000</td>
<td>-0.442 ***</td>
<td>0.137</td>
<td>0.201 **</td>
</tr>
<tr>
<td></td>
<td>$n=259$</td>
<td>$p&lt;0.001$</td>
<td>$p=0.13$</td>
<td>$p=0.002$</td>
</tr>
<tr>
<td>MoCA</td>
<td>1.000</td>
<td>-0.103</td>
<td>-0.486 ***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n=237$</td>
<td>$p=0.27$</td>
<td>$p&lt;0.001$</td>
<td>$n=222$</td>
</tr>
<tr>
<td>CAM</td>
<td>1.000</td>
<td></td>
<td>0.387 ***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$n=129$</td>
<td>$p&lt;0.001$</td>
<td>$n=129$</td>
<td></td>
</tr>
<tr>
<td>DI</td>
<td></td>
<td></td>
<td></td>
<td>1.000</td>
</tr>
</tbody>
</table>

190
Correlations involving the CAM were calculated using point-biserial correlations. Correlations involving the DI (and not involving the CAM) were assessed using Spearman’s rho. All other correlations were calculated using Pearson’s r.

8.8.1.5 Discussion

This study demonstrated the use of a serious game to screen for prevalent delirium in patients accessing emergency services. Patients found the device and software easy to use, and they suggested many new features that they would like to see in the next iteration.

8.8.1.5.1 SUS

I obtained a relatively high SUS score (76.4), suggesting that interactive tablets may be a feasible medium for delivering a battery of assessments that are easy to use, for elderly adults without an extensive technology background. This finding suggests that tablet-based technology is a possible medium for cognitive screening.

8.8.1.5.2 Modified Technology History Questionnaire

Based on findings from the modified technology history questionnaire, the study sample appears to be more familiar with older computer technology compared to new smartphones and tablets technology. In this sample, 42% of smartphone/tablet owners already play games on their device and 46% of users use their device on a daily basis. This demonstrates that older adults have adopted this technology into their daily lives and some are already used to playing mobile games. This finding is of particular interest, as older adults may be more open to playing game-based cognitive assessments. In addition, as touch-screen devices such as smartphones and tablets become more prevalent with current generations, it is likely that more adults will be familiar with this technology and more accepting in the future. Furthermore, over half of the sample patients indicated that they were comfortable using a computer, tablet, or other electric device. This suggests that technology-based screening is a viable platform for aging adults.

8.8.1.5.3 Game Performance

Based on the game settings, more difficult levels resulted in longer median correct RT. The difficulty of each level varied based on the number of simultaneous targets and non-targets on
the game board as well as the game board size. More difficult levels had a larger game board (e.g., a 4x4 game board) and a maximum of six stimuli that could appear at once on the game board compared to the easiest level with a smaller game board (e.g., a 3x3 game board) and a maximum of three stimuli that could appear simultaneously on the game board.

The game median correct RT was longer (1200 msec) in this version of the game compared to the version used with patients with delayed cognitive assessment as described in Chapter 4, where the game median correct RT was 900 msec. This comparison suggests that the presence of multiple characters on the game board at once may have been too difficult for participants. In considering that 41.7% of patients had experience with a smartphone/tablet and 41.7% of this sample played games on their device, this technology experience could have provided a benefit to patients in this study when interacting with the serious game. Despite the longer median correct RT, it suggests that the game used in this study was still difficult even though some patients had experience with technology and games.

8.8.1.5.4 Comparison between Game and Cognitive Performance

The results from the correlation analysis carried out between game performance with the MoCA, CAM, and DI, revealed significant correlations between the game median correct RT with the MoCA and DI but not with the CAM. In contrast to the results Chapter 4, the game median correct RT was not correlated with the CAM in this study. One explanation for the lack of correlation between the game median correct RT with the CAM is that there were only 3 CAM positive patients in this sample population, and 21 patients with a DI score of 4 and above. Moreover, the lower correlations between the game with the MoCA and DI, may have been due to the sample distribution of there being too few patients with delirium as assessed using the CAM. Another explanation is that there may be that the version of the game used in this study was much more difficult as multiple stimuli could appear simultaneously. Another reason could be that the simultaneous appearance of multiple stimuli required participants to recall the rules of the game in their working memory. For example, the participant would need to recall which stimuli to hit versus which stimuli they need to inhibit their response to, thus leading to the longer observed median correct RT. Furthermore, the results do not support concurrent validity of the serious game with measures of delirium as measured using the CAM. Based on these findings, it is difficult to assess the relationship between the CAM and game-based assessment.
However, the concurrent validity of the game as a measure of general cognitive status was demonstrated through the correlations between game performance and the MoCA and as a measure of delirium severity using the DI.

8.8.1.5.5  Recommendations

Based on the findings of this clinical study, I suggest the following design recommendations in designing a serious game for cognitive screening to be used in a clinical setting by older adults.

- Design a game with settings that can be adjusted to vary the difficulty.
- Evaluate the serious game with target users to receive feedback on game difficulty.
- Present one stimulus at a time since this may reduce the difficulty of the task to minimize frustration with the game for older adults.

8.8.1.6  Conclusion

In evaluating the game-based screening tool in emergency contexts, I demonstrated that the use of tablet-based screening technology is feasible and usable, which reconfirms previous findings in using this technology in an ED as described in Chapter 4. Moreover, in evaluating the serious game in emergency contexts to screen for prevalent delirium, I was able to establish validity between the game and cognitive performance (with the MoCA) and DI. The concurrent validity of the game with the CAM was not established, potentially due to increased difficulty settings in the game and a low number of patients with delirium in the sample. Further research should be carried out to validate the game-based assessment to screen for prevalent delirium with emergency patients using the same variant of the game used to screen for incident delirium.
8.9 Appendix I. Supporting Materials for Assessing Cognitive Status in Elderly Adults with Prevalent Delirium Study

Table 8-22. Summary of study sample demographics and cognitive assessment scores. MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, and DI = Delirium Index.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Males (n=124)</th>
<th>Females (n=109)</th>
<th>Total (n=330) b</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SD) / Median (IQR)</td>
<td>Range</td>
<td>Mean (SD) / Median (IQR)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>74.5 (7.5)</td>
<td>65 – 94</td>
<td>77.0 (7.5)</td>
</tr>
<tr>
<td>MoCA</td>
<td>23.5 (5.7)</td>
<td>0 – 30</td>
<td>22.25 (6.1)</td>
</tr>
<tr>
<td>CAM</td>
<td>0.02 (0.1)</td>
<td>0 – 1</td>
<td>0.04 (0.2)</td>
</tr>
<tr>
<td>DI</td>
<td>0.7 (1.3)</td>
<td>0 – 6</td>
<td>1.2 (2.0)</td>
</tr>
<tr>
<td>Game Median Correct RT (msec) a</td>
<td>1002 (500)</td>
<td>700 – 2200</td>
<td>1300 (400)</td>
</tr>
<tr>
<td>Game Target Offset (px)</td>
<td>28.88 (7.2)</td>
<td>13.4 – 56.8</td>
<td>31.3 (8.6)</td>
</tr>
</tbody>
</table>

a For the game RT rows, the median (IQR) is reported.

b The sex for 109 patients were not reported.
Table 8-23. Modified technology questionnaire questions and responses.

<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. Do you own your own computer or have one in your home?</td>
<td>No: 102</td>
</tr>
<tr>
<td></td>
<td>Yes: 201</td>
</tr>
<tr>
<td>Q2. How often do you use a computer?</td>
<td>Daily: 138</td>
</tr>
<tr>
<td></td>
<td>Weekly: 25</td>
</tr>
<tr>
<td></td>
<td>Monthly: 4</td>
</tr>
<tr>
<td></td>
<td>Rarely: 20</td>
</tr>
<tr>
<td></td>
<td>Never: 56</td>
</tr>
<tr>
<td>Q3. Do you own your own tablet or have one in your home?</td>
<td>No: 176</td>
</tr>
<tr>
<td></td>
<td>Yes: 126</td>
</tr>
<tr>
<td>Q3a. If yes, do you play games on your tablet/smartphone?</td>
<td>No: 91</td>
</tr>
<tr>
<td></td>
<td>Yes: 66</td>
</tr>
<tr>
<td>Q3b. If yes, how often do you use a tablet/smartphone?</td>
<td>Daily: 77</td>
</tr>
<tr>
<td></td>
<td>Weekly: 25</td>
</tr>
<tr>
<td></td>
<td>Monthly: 4</td>
</tr>
<tr>
<td></td>
<td>Rarely: 14</td>
</tr>
<tr>
<td></td>
<td>Never: 45</td>
</tr>
<tr>
<td>Q4. Have you ever used an Automated Banking Machine (ABM)?</td>
<td>No: 61</td>
</tr>
<tr>
<td></td>
<td>Yes: 243</td>
</tr>
<tr>
<td>Q4a. If yes, how long have you been using them?</td>
<td>Within the last month: 6</td>
</tr>
<tr>
<td></td>
<td>6 months - 1 year: 3</td>
</tr>
<tr>
<td></td>
<td>1-2 years: 2</td>
</tr>
<tr>
<td></td>
<td>2-5 years: 9</td>
</tr>
<tr>
<td></td>
<td>Greater than or equal to 5 years: 229</td>
</tr>
<tr>
<td>Q5. I am very comfortable using a computer, tablet or other electronic</td>
<td>Strongly agree: 116</td>
</tr>
<tr>
<td>devices.</td>
<td>Agree: 65</td>
</tr>
<tr>
<td></td>
<td>Neither agree nor disagree: 35</td>
</tr>
<tr>
<td></td>
<td>Disagree: 34</td>
</tr>
<tr>
<td></td>
<td>Strongly disagree: 48</td>
</tr>
<tr>
<td></td>
<td>NA's: 5</td>
</tr>
</tbody>
</table>

8.9.1 Game variants used in the validation study.

Table 8-24. Game settings for the different versions of the game. The target size for all levels was held constant at 175 px.

<table>
<thead>
<tr>
<th>Version</th>
<th>Game Name</th>
<th>Grid Size</th>
<th>Number of Moles (Target)</th>
<th>Number of Raccoons (Target)</th>
<th>Number of Butterflies (distractor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A (Beginner) a</td>
<td>3 by 3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Intermediate</td>
<td>3 by 3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Hard</td>
<td>4 by 4</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>B (Intermediate) a</td>
<td>3 by 3</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>C (Hard) a</td>
<td>4 by 4</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>Practice</td>
<td>4 by 4</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>D (Practice) a</td>
<td>3 by 3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

a Final version of this level.

Table 8-25. Two-tailed Pearson’s r correlation between game performance and cognitive measures. a Correlations involving the CAM were calculated using point-biserial
correlations. * \( p<0.05 \), * * \( p<0.01 \), * *** \( p<0.001 \). MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, and DI = Delirium Index.

<table>
<thead>
<tr>
<th>Game Correct Target Offset</th>
<th>Game Median Correct RT</th>
<th>MoCA ( p&lt;0.01 )</th>
<th>CAM ( * p&lt;0.05 )</th>
<th>DI ( p&lt;0.001 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n=259 )</td>
<td>( n=235 )</td>
<td>( n=127 )</td>
<td>( n=234 )</td>
</tr>
</tbody>
</table>

\( a \) Correlations involving the CAM were calculated using point-biserial correlations. Correlations involving the DI (and not involving the CAM) were assessed using Spearman’s \( \rho \). All other correlations were calculated using Pearson’s \( r \).

Table 8-26. Two-tailed Spearman’s \( \rho \) correlation between game performance and cognitive measures. * \( p<0.05 \), * * \( p<0.01 \), * *** \( p<0.001 \). MoCA = Montreal Cognitive Assessment, CAM = Confusion Assessment Method, and DI = Delirium Index.

<table>
<thead>
<tr>
<th>Game Median Correct RT</th>
<th>Game Correct Target Offset</th>
<th>MoCA ( p=0.0023 )</th>
<th>CAM ( * ) ( p=0.079 )</th>
<th>DI ( p=0.001 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n=259 )</td>
<td>( n=259 )</td>
<td>( n=235 )</td>
<td>( n=119 )</td>
<td>( n=222 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Game Correct Target Offset</th>
<th>Game Median Correct RT</th>
<th>MoCA ( p=0.001 )</th>
<th>CAM ( * ) ( p=0.001 )</th>
<th>DI ( p=0.001 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n=259 )</td>
<td>( n=235 )</td>
<td></td>
<td>( n=127 )</td>
<td>( n=234 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MoCA</th>
<th>Game Correct Target Offset</th>
<th>Game Median Correct RT</th>
<th>MoCA ( p=0.001 )</th>
<th>CAM ( * ) ( p=0.001 )</th>
<th>DI ( p=0.001 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n=237 )</td>
<td>( n=235 )</td>
<td>( n=237 )</td>
<td></td>
<td>( n=119 )</td>
<td>( n=222 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CAM</th>
<th>Game Correct Target Offset</th>
<th>Game Median Correct RT</th>
<th>MoCA ( p=0.001 )</th>
<th>CAM ( * ) ( p=0.001 )</th>
<th>DI ( p=0.001 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n=237 )</td>
<td>( n=235 )</td>
<td>( n=237 )</td>
<td></td>
<td>( n=119 )</td>
<td>( n=222 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DI</th>
<th>Game Correct Target Offset</th>
<th>Game Median Correct RT</th>
<th>MoCA ( p=0.001 )</th>
<th>CAM ( * ) ( p=0.001 )</th>
<th>DI ( p=0.001 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n=236 )</td>
<td>( n=235 )</td>
<td>( n=237 )</td>
<td></td>
<td>( n=119 )</td>
<td>( n=222 )</td>
</tr>
</tbody>
</table>
8.10 Appendix J. Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting Study.

8.10.1 Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting

This chapter examines the potential for increasing the sensitivity of the game as a cognitive measure by having the person walk while playing the serious game so as to create additional cognitive load. A secondary objective is to explore how varying the different game parameters in the game are related to measures of EF and to see how feasible it was to play the game in a dual-task (DT) walking context. The chapter begins with the motivation for the study, followed by the study design, results, and discussion of the findings.

8.10.1.1 Motivation

Patients with cognitive impairments may have higher risk of falls and injuries due to decreased EF abilities (and thus diminished high level control of gait) and decreased motivation to remain physically active. EF are important in tasks such as walking, balance, and obstacle avoidance (Hausdorff, Yogev, Springer, Simon, & Giladi, 2005; Persad et al., 1995). Thus, understanding how executive functioning is related to both cognitive status and walking in daily livings is important, especially in scenarios where a secondary task is being performed while walking, such as avoiding an obstacle to answer a telephone.

Disturbances in gait are common across the dementia spectrum and may be detected early at the stage of MCI (Montero-Odasso et al., 2014). In particular, stride time variability, a measure of the reliability of lower limb movements that depend on cortical control (Lord, Howe, Greenland, Simpson, & Rochester, 2011), has been shown to be low among cognitively healthy older adults (Beauchet, Dubost, Herrmann, & Kressig, 2005; Grabiner, Biswas, & Grabiner, 2001) and high among individuals with MCI or mid-stage AD (IJmker & Lamoth, 2012). In the MCI population where gait changes may be less evident, it has been suggested that assessing gait while dual-tasking may provide insight into the evaluation of EF (Montero-Odasso et al., 2009; Montero-Odasso, Verghese, Beauchet, & Hausdorff, 2012). Dual-tasking is a motor-divided attention task
that requires individuals to walk while doing a cognitively demanding task, which may unmask gait disturbances that are only evident under cognitive stress (Montero-Odasso et al., 2014).

It has been demonstrated that task selection and standardization of both the cognitive and walking task can impact the effect of DT performance (Menant, Schoene, Sarofim, & Lord, 2014). Howcroft, Kofman, Lemaire and McIlroy (2016) suggest that DT gait assessment may be improved by addressing task standardization, selection, and quantification of cognitive task performance. The research study described in this chapter sought to test whether playing the serious game while walking provided a more accurate assessment of cognitive status than playing the game as a single task (while sitting).

8.10.1.2 Research Question

This chapter will address the following research question:

1. Does use of the game in a DT setting (while walking) improve its sensitivity as a measure of EF relative to performing the serious game alone (without walking)?

This research question was addressed by carrying out a study comparing performance on the serious game while sitting down in a single task condition to playing the game while walking in a DT condition. A correlation analysis was carried out comparing game performance in single and DT conditions to performance on cognitive tasks carried out while sitting. The DT cost on each game level due to walking was calculated in terms of stride variability and stride time. The DT cost for each level was compared to cognitive performance using a correlation analysis.

8.11 Methodology

8.11.1 Recruitment

Participants were recruited from the University of Toronto and through online advertising using sites such as Craigslist.ca and Kijiji.ca.

8.11.1.1 Protocol

8.11.1.1.1 Study Design

The research protocol was approved by the University of Toronto research ethics board (protocol #33430). All participants completed the experiment individually. Participants were first asked to
complete a demographics questionnaire, which included questions on sex; age; native tongue; level of education; and computer skills. There were three phases in the study, which were counter-balanced across participants: (1) psychological tasks, (2) playing the serious game-based cognitive assessment in a single task condition, and (3) DT assessment (gait analysis and game-based cognitive assessment).

In the first phase, participants were asked to complete three psychological tasks on a computer. These tasks included the: (1) Stroop task (a measure of inhibition and general executive function) (Miyake & Friedman, 2012), (2) n-back task (a measure of updating), and (3) WCST (a measure of shifting) (Heaton, 1981). Each of these tasks measures different EF (Miyake & Friedman, 2012). These tasks were administered using the Psych Toolkit available from http://psychtoolkit.com on a computer. Data analysis focuses on game performance and DTW performance compared to the median correct RT and error rates on the three cognitive tasks. A particular emphasis was made on game performance in single and DTW to the performance on the Stroop task, as it is hypothesized to be a general measure of EF.

The gait analysis tool (Blumenthal, 2017) was an Android-based mobile application that was administered on a Samsung Galaxy S4 smartphone (Samsung, 2006). The device was affixed to the participant’s waist using a smartphone holder (Figure 8-11). The application used the device's embedded accelerometer to identify steps from movement along the Y (vertical) axis.

The game-based cognitive assessment was completed on an Asus ZenPhone2 smartphone with a 5.5-inch screen (ASUSTeK Computer Inc., 2015). There were six levels of the game, which lasted 1 minute and 30 seconds each. A description of each of the levels is described below:

- Level 1: single stimulus game with moles as targets
- Level 2: single stimulus game with moles as targets and butterflies as non-targets (dissimilar distractor condition)
- Level 3: single stimulus game with moles as targets and moles with hats as non-targets (similar distractor condition)
- Level 4: multiple stimuli game with moles as targets
- Level 5: multiple stimuli game with moles as targets and butterflies as non-targets (dissimilar distractor condition)
• Level 6: multiple stimuli game with moles as targets and moles with hats as non-targets (similar distractor condition)

In single stimulus games (levels 1, 2 and 3), only one character would appear on the game board at a time. In multiple stimuli games (levels 4, 5, and 6), three characters (one target and two non-targets) would appear on the game board at a time. The objective was to hit the target character (or one of the target characters) as quickly as possible.

Participants were asked to play through a tutorial level of the game first, which lasted for 1 min. Next participants were asked to play through all six levels of the game, which appeared in a separately randomized order for each participant.

8.11.1.1.2 Statistical Analysis

The cognitive data and serious game results were non-normally distributed based on visual inspection of the data. The RT data for the cognitive task and game were positively skewed, thus the median RT was used for the cognitive task and game. As my interest was in correlations as a measure of the effect size of the underlying relationship between game performance and the cognitive assessments, I used correlations between the cognitive tasks and game measures using Pearson’s $r$ correlations, which were followed up with Spearman’s $\rho$ correlations.

8.11.1.2 Results

8.11.1.2.1 Description of Sample

Twenty participants (9 females, 11 males), between the ages of 18-39 years (mean = 22.9 years, standard deviation = 5.8), completed this study.

8.11.1.2.2 Game-Based Cognitive Assessment Parameters That Influence Cognitive Ability in a Single-Task Condition

In the following analyses of the role of targets and non-targets, and the impact of the number of stimuli, measures of cognitive ability were: the Stroop effect was calculated as the difference between median correct RT in incongruent and congruent trials. Other measures included the perseveration error rate on the WCST, and error rate the $n$-back task. Performance on the game was calculated as the median correct RT.
In the following section, the data from the game played while sitting; in a single task setting will be analyzed. First, to assess the relationships between game and cognitive performance, the game performance measure (median correct RT) will be compared to performance on the cognitive measures using both Pearson's $r$ and Spearman's $\rho$ correlations. Second, performance between levels with a single stimulus versus multiple stimuli will be carried out using paired $t$-tests. These analyses will investigate whether the number of stimuli has an influence on game performance. Third and finally, game performance within single stimulus levels will be carried out using ANOVAs to determine differences between single stimulus games. This analysis will be repeated for multiple stimuli levels.

Table 8-27 shows the two-tailed Pearson’s $r$ correlations between median correct RT for each game level played in a single task condition, and the cognitive measures. The purpose of this analysis was to assess the ability of the game at each level to predict EF ability. There were no statistically significant correlations between any of the game levels with the Stroop task (Table 8-27). Level 1 had a significant relationship with the error rate on the WCST and $n$-back task.

As a follow up, a Spearman’s $\rho$ correlation analysis was carried out comparing the median correct RT on each level while sitting to the cognitive measures (Table 8-34). The same Pearson’s $r$ correlations observed with the median correct RT on game level 1 was also observed with Spearman’s $\rho$. Using Spearman’s $\rho$, level 2 was correlated with the perseveration error rate on the WCST.

Table 8-27. Two-tailed Pearson’s $r$ correlation analysis between the median correct RT game-based cognitive assessment played by while sitting with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Effect</th>
<th>WCST Error Rate</th>
<th>$n$-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.165</td>
<td>0.536</td>
<td>0.469</td>
</tr>
<tr>
<td></td>
<td>$p=0.4807$</td>
<td>$p=0.0181$</td>
<td>$p=0.0372$</td>
</tr>
<tr>
<td>2</td>
<td>0.269</td>
<td>0.385</td>
<td>0.0058</td>
</tr>
<tr>
<td></td>
<td>$p=0.252$</td>
<td>$p=0.104$</td>
<td>$p=0.981$</td>
</tr>
<tr>
<td>3</td>
<td>-0.0915</td>
<td>0.277</td>
<td>0.280</td>
</tr>
<tr>
<td></td>
<td>$p=0.701$</td>
<td>$p=0.250$</td>
<td>$p=0.232$</td>
</tr>
<tr>
<td>4</td>
<td>-0.0574</td>
<td>0.160</td>
<td>-0.0312</td>
</tr>
<tr>
<td></td>
<td>$p=0.810$</td>
<td>$p=0.514$</td>
<td>$p=0.896$</td>
</tr>
<tr>
<td>5</td>
<td>-0.0935</td>
<td>0.235</td>
<td>0.170</td>
</tr>
<tr>
<td></td>
<td>$p=0.695$</td>
<td>$p=0.332$</td>
<td>$p=0.475$</td>
</tr>
<tr>
<td>6</td>
<td>-0.187</td>
<td>0.184</td>
<td>-0.213</td>
</tr>
<tr>
<td></td>
<td>$p=0.430$</td>
<td>$p=0.451$</td>
<td>$p=0.368$</td>
</tr>
</tbody>
</table>
8.11.1.2.2.1 Comparison between Single and Multiple Character Game Levels

A paired, \( t \)-test was carried out to determine the differences in median correct RT based on game levels with single (levels 1, 2, and 3) versus multiple characters (levels 4, 5, and 6) (Figure 8-9). There was a statistically significant difference between game levels with single versus multiple characters, \( t(20)=9.53, p<0.001 \). Median correct RT was longer in single character levels (mean = 509 msec, standard deviation = 0.036) compared to multiple character levels (mean = 392 msec, standard deviation = 0.060). Inspection of Figure 8-9 suggests that the speed advantage for multiple characters was present in levels 4 and 5 but not in level 6. Thus, rather than being confusing, multiple characters likely let people respond to the closest character. Note that level 6 had the similar distractor (moles with hats).

Figure 8-9. Boxplot of median correct RT per game level.

8.11.1.2.2.2 Comparison within Game Levels

An ANOVA was carried out to determine the differences in median correct RT in single character levels (levels 1, 2, and 3). There was a statistically significant difference between the three game levels, \( F(2,60)=24.89, p<0.001 \). Post-hoc pairwise Tukey tests revealed that median correct RT was faster in level 1 (mean = 484 msec, standard deviation = 0.030) compared to level 2 (mean = 556 msec, standard deviation = 0.054, \( p<0.001 \)), and level 3 (mean = 598 msec, standard deviation = 0.068, \( p=0.0098 \)). As expected, the presence of a distractor made the task
more difficult, and a similar distractor (level 3) slowed RT more than a dissimilar distractor (level 2).

An ANOVA was used to assess the differences in median correct RT in multiple character levels (levels 4, 5, and 6). There was a significant overall difference between the three game levels, $F(2,60)=222.1, p<0.001$. Post-hoc pairwise Tukey tests revealed that median correct RT was faster in level 4 (mean = 220 msec, standard deviation = 0.032) compared to level 5 (mean = 451 msec, standard deviation = 0.049, $p<0.001$), and level 6 (mean = 581 msec, standard deviation = 0.077, $p<0.001$). Similar to single character levels, the presence of a distractor made the task more difficult, and a similar distractor (level 6) slowed RT more than a dissimilar distractor (level 5).

8.11.1.2.2.3 Comparison between Single to Multiple Character Levels

Paired $t$-tests were carried out to compare median correct RT between the corresponding pairs of single and multiple character levels. For example, levels 1 and 4 were compared, level 2 and 5 were compared, and levels 3 and 6 were compared. Between game levels 1 and 4, which only had targets, the median correct RT on level 4 was significantly faster than level 1, $t(20)=28.943$, $p<0.001$. Next, between game levels 2 and 5, which had dissimilar distractors, the median correct RT on level 5 was significantly faster than level 2, $t(20)=9.34$, $p<0.001$. Finally, there was no difference in median correct RT between game levels 3 and 6, which had similar distractors.

8.11.2 Cognitive Task Performance While Walking

The analyses reported in the preceding section were then repeated for the data where the participants were walking while they played the game. First, to assess the relationships between game and cognitive performance, the game performance measure (median correct RT) will be compared to performance on the cognitive measures using both a Pearson's $r$ and Spearman's rho correlation analysis. Second, performance between levels with a single stimulus versus multiple stimuli will be carried out using paired $t$-tests. These analyses will investigate whether the number of stimuli has an influence on game performance. Third and finally, game performance within single and multiple stimuli levels will be carried out using ANOVAs to determine differences between single stimulus/multiple stimuli games.
A Pearson’s $r$ correlation analysis between the median correct RT on the game played while walking with the cognitive measures was carried out (Table 8-28). This analysis revealed that level 4 on the game was significantly correlated with the perseveration error rate on the WCST. Moreover, the Pearson’s $r$ correlation observed with level 4 on the game with the error rate on the WCST was also observed with the Spearman’s $\rho$ analysis (Table 8-36).

Table 8-28. Two-tailed Pearson’s $r$ correlation analysis between the median correct RT on game-based cognitive assessment while walking with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Effect</th>
<th>WCST Error Rate</th>
<th>$n$-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0492</td>
<td>0.233</td>
<td>0.00333</td>
</tr>
<tr>
<td></td>
<td>$p=0.837$</td>
<td>$p=0.336$</td>
<td>$p=0.989$</td>
</tr>
<tr>
<td>2</td>
<td>0.288</td>
<td>0.287</td>
<td>-0.140</td>
</tr>
<tr>
<td></td>
<td>$p=0.218$</td>
<td>$p=0.233$</td>
<td>$p=0.556$</td>
</tr>
<tr>
<td>3</td>
<td>0.312</td>
<td>0.0913</td>
<td>-0.338</td>
</tr>
<tr>
<td></td>
<td>$p=0.180$</td>
<td>$p=0.710$</td>
<td>$p=0.145$</td>
</tr>
<tr>
<td>4</td>
<td>0.202</td>
<td>0.656 **</td>
<td>-0.023</td>
</tr>
<tr>
<td></td>
<td>$p=0.394$</td>
<td>$p=0.00230$</td>
<td>$p=0.922$</td>
</tr>
<tr>
<td>5</td>
<td>0.0226</td>
<td>0.0472</td>
<td>0.0933</td>
</tr>
<tr>
<td></td>
<td>$p=0.925$</td>
<td>$p=0.848$</td>
<td>$p=0.696$</td>
</tr>
<tr>
<td>6</td>
<td>0.240</td>
<td>0.165</td>
<td>-0.291</td>
</tr>
<tr>
<td></td>
<td>$p=0.308$</td>
<td>$p=0.500$</td>
<td>$p=0.214$</td>
</tr>
</tbody>
</table>

8.11.2.1.1.1 Comparison between Single and Multiple Character Game Levels

A paired, $t$-test was carried to determine the differences in median correct RT based on game levels with single (levels 1, 2, and 3) versus multiple characters (levels 4, 5, and 6) (Figure 8-10). There was a statistically significant difference between game levels with single versus multiple characters, $t(21)=12.356$, $p<0.001$. Median correct RT was longer in single character levels (mean = 564 msec, standard deviation = 0.069) compared to multiple character levels (mean = 412 msec, standard deviation = 0.059). Inspection of Figure 8-10 suggests that the speed advantage for multiple characters was present in levels 4 and 5 but not in level 6. Thus, rather than being confusing, multiple characters likely let people respond to the closest character. Note that level 6 had the similar distractor (moles with hats). A similar pattern of results was obtained in the single-task condition.
8.11.2.1.1.2 Comparison within Game Levels

An ANOVA was carried to determine the differences in median correct RT in single character levels (levels 1, 2, and 3) only while walking. There was a statistically significant difference between the three game levels, $F(2,63)=13.62$, $p<0.001$. Post-hoc pairwise Tukey tests revealed that median correct RT was faster in level 1 (mean = 536 msec, standard deviation = 0.059) compared to level 2 (mean = 596 msec, standard deviation = 0.068, $p = 0.035$), and level 3 (mean = 658 msec, standard deviation = 0.098, $p<0.001$). As expected, the presence of a distractor made the task more difficult, and a similar distractor (level 3) slowed RT more than a dissimilar distractor (level 2).

An ANOVA was used to assess the differences in median correct RT in multiple character levels (levels 4, 5, and 6) only while walking. There was a significant overall difference between the three game levels only, $F(2,63)=121.1$, $p<0.001$. Post-hoc pairwise Tukey tests revealed that median correct RT was faster in level 4 (mean = 248 msec, standard deviation = 0.061) compared to level 5 (mean = 469 msec, standard deviation = 0.056, $p<0.001$), and level 6 (mean = 568 msec, standard deviation = 0.087, $p<0.001$). Similar to single character games, the presence of a distractor made the task more difficult, and a similar distractor (level 6) slowed RT more than a dissimilar distractor (level 5).
8.11.2.1.1.3 Comparison between Single to Multiple Character Levels

Paired t-tests were then carried out to compare median correct RT between the corresponding pairs of single and multiple character levels while walking. For example, levels 1 and 4 were compared, level 2 and 5 were compared, and levels 3 and 6 were compared. Between game levels 1 and 4, the median correct RT on level 4 was significantly faster than level 1, $t(21)=18.98$, $p<0.001$. Next, between game levels 2 and 5, the median correct RT on level 5 was significantly faster than level 2, $t(20)=14.236$, $p<0.001$. Finally, between game levels 3 and 6, the median correct RT on level 6 was significantly faster than level 3, $t(21)=5.460$, $p<0.001$.

8.11.2.1.2 Dual-Task Cognitive Assessment

A DT cost calculation (Equation 8-1) was carried out on stride variability (standard deviation) and mean stride time for each game level. The purpose of this analysis was to investigate changes in gait performance due to playing the game-based cognitive assessment as an additional task. A two-tailed Pearson’s $r$ correlation analysis was then carried out comparing the DT value for each level with performance on the cognitive tasks (Table 8-22, and Table 8-30).

$$DT\ cost = \frac{\text{dual\ task\ value} - \text{single\ task\ value}}{\text{single\ task\ value}} \times 100$$

Equation 8-1. Dual-task cost calculation.

A correlation analysis using Pearson’s $r$ was carried out comparing the DT cost on game performance using stride variability to cognitive measures (Table 8-22, Table 8-30). A second correlation analysis using Pearson’s $r$ was carried out comparing the DT cost on game performance using stride time to cognitive measures (Table 8-29). The value of $n$ is listed for each correlation. Software issues with the gait tool resulted in data collection errors for some users.

There were no significant correlations, between the DT cost on game performance using stride variability with any of the cognitive measures (Table 8-22). As a follow up, Spearman’s $\rho$ correlations were also carried out between the DT cost on game performance using stride variability to cognitive measures (Table 8-37). As with the Pearson’s $r$ correlations, there were
also no observed significant correlations between the DT cost on game performance using stride variability with any of the cognitive measures.

Table 8-29. Two-tailed Pearson’s $r$ correlation analysis between performance measures on the cognitive task and DT cost of playing the serious game using stride variability. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.00939</td>
<td>0.392</td>
<td>-0.247</td>
</tr>
<tr>
<td>$p=0.970$</td>
<td></td>
<td>$p=0.103$</td>
<td>$p=0.309$</td>
</tr>
<tr>
<td>$n=19$</td>
<td></td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>2</td>
<td>-0.0894</td>
<td>0.0806</td>
<td>-0.077</td>
</tr>
<tr>
<td>$p=0.716$</td>
<td></td>
<td>$p=0.742$</td>
<td>$p=0.754$</td>
</tr>
<tr>
<td>$n=19$</td>
<td></td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>3</td>
<td>-0.127</td>
<td>0.376</td>
<td>-0.138</td>
</tr>
<tr>
<td>$p=0.603$</td>
<td></td>
<td>$p=0.120$</td>
<td>$p=0.573$</td>
</tr>
<tr>
<td>$n=19$</td>
<td></td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>4</td>
<td>-0.0418</td>
<td>0.115</td>
<td>-0.193</td>
</tr>
<tr>
<td>$p=0.869$</td>
<td></td>
<td>$p=0.652$</td>
<td>$p=0.442$</td>
</tr>
<tr>
<td>$n=18$</td>
<td></td>
<td>$n=17$</td>
<td>$n=18$</td>
</tr>
<tr>
<td>5</td>
<td>-0.145</td>
<td>0.217</td>
<td>0.101</td>
</tr>
<tr>
<td>$p=0.567$</td>
<td></td>
<td>$p=0.402$</td>
<td>$p=0.690$</td>
</tr>
<tr>
<td>$n=18$</td>
<td></td>
<td>$n=17$</td>
<td>$n=18$</td>
</tr>
<tr>
<td>6</td>
<td>-0.168</td>
<td>0.146</td>
<td>-0.177</td>
</tr>
<tr>
<td>$p=0.504$</td>
<td></td>
<td>$p=0.571$</td>
<td>$p=0.482$</td>
</tr>
<tr>
<td>$n=18$</td>
<td></td>
<td>$n=17$</td>
<td>$n=18$</td>
</tr>
</tbody>
</table>

Based on the Pearson’s $r$ correlation analysis between the DT cost of playing the game using stride time, there were no observed significant correlations with any of the cognitive measures (Table 8-30). As a follow up, Spearman’s $\rho$ correlations were also carried out between the DT cost on game performance using stride time to cognitive measures (Table 8-38). As with the Pearson’s $r$ correlations, there were also no observed significant correlations between the DT cost on game performance using stride time with any of the cognitive measures.

Table 8-30. Two-tailed Pearson’s $r$ correlation analysis between performance measures on the cognitive task and DT cost of playing the serious game using stride time. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.0189</td>
<td>0.258</td>
<td>0.006</td>
</tr>
<tr>
<td>$p=0.939$</td>
<td></td>
<td>$p=0.302$</td>
<td>$p=0.980$</td>
</tr>
<tr>
<td>$n=19$</td>
<td></td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>2</td>
<td>0.149</td>
<td>0.272</td>
<td>0.204</td>
</tr>
<tr>
<td>$p=0.544$</td>
<td></td>
<td>$p=0.275$</td>
<td>$p=0.403$</td>
</tr>
<tr>
<td>$n=19$</td>
<td></td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>3</td>
<td>-0.357</td>
<td>0.393</td>
<td>0.002</td>
</tr>
<tr>
<td>$p=0.133$</td>
<td></td>
<td>$p=0.107$</td>
<td>$p=0.993$</td>
</tr>
<tr>
<td>$n=19$</td>
<td></td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
</tbody>
</table>
An ANOVA was carried out comparing the DT cost of stride variability and stride time across all six game levels. There was no statistically significant difference of the DT of stride variability and stride time costs between game levels.

8.11.2.2 Discussion

The purpose of the exploratory research carried out in this study was to inform the redesign of a serious game for cognitive assessment. In my previous research, I had developed a game that correlates well with Stroop task performance for able-bodied participants and that also correlates quite well ($r = 0.5$ to $0.6$) with the MMSE for elderly ED patients (Tong, Chignell, Tierney, & Lee, 2016). The serious game that I developed represents one region within a potentially very large design space defined by the interplay between numbers of different parameters some of which are listed below:

- Number and layout of holes on the game board
- Type of characters/images used as targets and non-targets
- Presence or absence of distractors
- Degree of visual similarity between targets and distractors
- Number of targets used (e.g., moles, raccoons, etc.)
- Number of non-targets used
- Number of targets/non-targets that can appear at the same time
- How quickly characters pop up and how long they stay up before going back down

I designed an experiment to explore a portion of this design space defined by the type of distractor (none, dissimilar, similar) and by the distinction between single and multiple targets. In addition to exploring the image of distractor type and number of simultaneous targets, I also wanted to see if DT assessment (i.e., walking while playing the serious game) would provide a
more sensitive test of cognitive status, motivated by previous research results showing that the act of walking is both affected by cognitive load and contributes to cognitive load.

Since this was design-oriented research, I used a sample of 20 participants and looked for relatively large effect sizes that I would then seek to replicate in a follow up experiment. I found that the median correct RT on game level 1, which only had targets, provided the strongest correlation with the error rate on the WCST in the single task condition (playing the serious game while sitting down). This suggests that performance on the serious game may require shifting ability. One possibility may be that performance on the Stroop task requires interference control whereas response inhibition is required on the serious game when considering median correct RT. Thus, the Stroop task and serious game may be measuring different forms of inhibition (Diamond, 2013).

In a DT condition, the median correct RT game level 4 was correlated with the error rate on the WCST. Both game levels 1 (in the single task) and level 4 (in the DT condition), only contain targets, and were both correlated with the error rate on the WCST. This further suggests that performance on the game requires shifting ability.

In both single and DT conditions, game performance was not correlated with Stroop task performance. As mentioned earlier, it is possible that both the serious game and the Stroop task require different forms of EF and inhibition. This raises the question of what the psychometric properties of the game are, and which forms of EF and inhibition it is related to.

The DT cost of stride variability and DT cost of stride time were not significantly correlated with any of the cognitive measures. Based on these results, it is possible that using the game as a secondary task in a DTW study may not be a good prediction of cognitive performance. Another explanation may be that the secondary task should be carefully designed to result in observable results by having an impact on a participant both cognitively and physically. Another explanation may be that the cognitive measures used are not related to performance on the serious game. For example, the use of the Go/No-Go discrimination task could be explored in future studies as a measure of inhibition instead of, or in addition to using the Stroop task.

In a single task condition, I observed that game performance on single target games was generally slower compared to multiple target games. One potential reason may be that a serial
versus parallel processing effect was observed (Treisman, 1986). In single target games, I suggest that a serial processing effect was observed compared to parallel processing, which was needed in multiple target games.

I observed that game performance in a single task condition was fastest in level 4 compared to all other levels. One explanation for the observed RT is that this game level demonstrates a Fitts’ Law effect, whereby less time is taken to reach closer targets (Fitts, 1954). As there are only targets in this game level, the time taken to respond to each target may be less compared to levels where a discrimination task between a target, and non-target is required. In addition, game performance was faster in levels with distractors in multiple character games (levels 5 and 6) compared to single character games (levels 2 and 3), potentially due to a “pop out” effect (Treisman, 1986), whereby the presence of distractors influenced the time taken to identify the target stimuli (e.g., mole). Moreover, in a single task condition, the presence of multiple stimuli created a speeded effect on game performance while the presence of similar distractors slowed game performance and increased the number of false alarms.

8.11.2.2.1 Recommendations

Based on the findings of this study, I suggest the following design recommendations in designing a serious game for cognitive screening to be used in a single and DT condition.

- To create a speeded task, multiple stimuli can be used.
- Select a visually similar distractor to increase the probability of a false alarm.

8.11.2.2.2 Limitations

This study was carried out using only one gait analysis tool and one game-based cognitive assessment in a laboratory setting with healthy adults. The gait analysis tool required an active Internet connection to record data. In the DT portion of the study, there were cases were data was lost due to a weak Internet connection. In future research, the data could be stored locally on the device to minimize the possibility of data loss.

In future research, it would beneficial to study other tools available. In addition, the motivation for this work was in screening for risk of cognitive impairment through cognitive and functional/gait assessment in older adults. In this current research, I focused on first evaluating the two tools with a healthy, young adult sample to evaluate the gait assessment tool’s usability.
and feasibility. Future research should focus on re-evaluating these two tools with healthy, older adults, and other adults either at risk or with cognitive impairments.

Suggested future research will focus on integrating the gait analysis tool with the game-based cognitive assessment to enable both functional and cognitive measures to be measured in one application deployable on a single mobile device compared to the two devices used in this study.

8.11.2.3 Conclusions

The goals of this study were to investigate the role of different game parameters in assessing cognitive status, and how performance on the serious game in a DTW scenario would assess cognitive status in healthy, young adults. First, I have shown that performance on the serious game is not related to the Stroop task, and may require a different form of inhibition. Second, I have demonstrated that there is no relationship between the DT cost of stride variability and stride time while walking and playing a game-based assessment.
8.12 Appendix K. Supporting Materials for Assessing Cognitive Status Using a Serious Game While Walking in a Dual-Task Setting Study.

8.12.1 Experimental Materials.

8.12.1.1 Recruitment Flyer

**Call for Participation in a Human Factors Research Project**

**Screening**

We are seeking people who:

- age 18 years or older;
- can read texts on a computer monitor display and handheld tablet;
- have no self-reported difficulty in hearing or understanding English instructions;
- have no self-reported problem distinguishing between red, blue, purple, green, yellow, orange, brown and gray;
- have no self-reported history or current cognitive impairments (ex. dementia, mild cognitive impairment, delirium)

Please do not sign up to participate in this study unless you meet these criteria.

**Experiment**

For those volunteers who are invited to participate in the experiment (based on their screening results) the details of the experiment are as follows:

*Period:* from September 1, 2016 to September 1, 2017
*Place:* 8th floor in the Bahen building at St. George campus of University of Toronto

We are studying how different cognitive tasks compare to a tablet-based game and gait analysis tool measuring similar metrics.

Participants will participate in the experiment individually. First, they will play a game on a tablet and then use a gait analysis tool. Then they will perform cognitive capacity tasks presented on a computer. Each task will take only a few minutes, and the entire experiment should finish within one hour. Participants will be compensated $10.00 for taking part in this study. Participants may be invited to participate in the same study at a later date, and will compensated $10.00.

If you are interested in participating, please send an email to tiffany.tong@mail.utoronto.ca. Questions are also welcome to the same address.

Thank you,
Tiffany Tong
PhD Candidate
Department of Mechanical and Industrial Engineering, University of Toronto

8.12.1.2 Consent Form

**Client Information Sheet and Informed Consent Form for the Study: Examining Game-Based and Gait-Related Cognitive Assessments**

Dear Sir/Madam: 

Date: 
Thank you for your interest in this research project. This letter has been created to provide you with the information needed so that you may decide whether you would like to participate in this study. Participation is voluntary and you are free to withdraw or stop at any time.

If at any point you feel as though any of the following details are unclear, or if you have any other questions, comments, or concerns, please feel free to contact me using the contact information at the end of this letter. If you decide that you would like to participate, please date and sign the last page of this letter then return one copy to me and keep the other for your reference. If you do not wish to participate there is no need to return the form. Please note, you may request a copy of our final study if desired.

The long-term goal of our research is to use digital, mobile-based games/cognitive assessments to detect changes in cognition. This study aims to understand the effectiveness of computerized assessments.

The design of the experiment is as follows. All participants will participate in the experiment individually, one at a time. First, a tablet with a game will be presented to the participant and the experimenter will introduce the game. The participant will play the game on the tablet. Next, each participant will be asked to use the gait analysis tool on a mobile phone. Next, each person will be asked to perform a series of tasks presented on a computer. These include target identification, color identification, card sorting, and an n-back task.

The participant may decline to answer any of these questions.

The data we collect will be anonymized and kept in a secure office. No personal or identifying information will be included in written reports or presentations, and your confidentiality and privacy will be respected at all times. Any data and information received will be kept confidential. Any study reports and presentations will have all personal identifiers removed. Data and participant information will be kept in my possession or stored in a locked office accessible only by me and the other investigators. Electronic information will be password protected. All data will be securely stored until September 31 2021. All data will be destroyed after September 31 2021.

The experiment should take 1 hour and will be held in the Bahen building at 40 St. George Street on the University of Toronto St. George Campus (near College Street and Spadina Avenue). There is a $10.00 compensation for participating in the first part of the study. You may be invited to participate in the same study at a later date, and will be compensated $10.00.

As mentioned previously, if you have any questions, you may contact me at tiffany.tong@mail.utoronto.ca or 416-978-7581. Alternatively, you may call the Office of Research Ethics at ethics.review@utoronto.ca or 416-946-3273.

Thank you for your consideration,

Tiffany Tong
PhD Candidate, University of Toronto

To be completed by participants:

☐ I have read this consent form and I understand the research and what is expected of me.
☐ I understand that:
  • I am free to withdraw before or anytime during the study without the need to give any explanation
  • I am free to elect to skip parts of the study without the need to give any explanation
☐ I agree to participate in this study.

If I do not wish to participate in the research, I can just keep the form without signing it.

______________________  (Signature)
______________________  (Name, please print)
______________________  (Date)
______________________  (Investigator)
______________________  (Name, please print)
______________________  (Date)
### 8.12.1.3 Demographics Questionnaire

**Demographic and Computer Skill Questionnaire**

Please answer the following questions to the best of your ability. If at any point you feel uncomfortable answering a question feel free to leave it blank and move on to the next one. If you ever require assistance or if you no longer wish to participate in the study, please let one of the researchers know so that they can accommodate your requests.

- **Sex:**
  - [ ] Male  
  - [ ] Female

- **What is your age? _______**

- **Is English your native tongue/first-language? (Yes/No)**

- **Which of the following best describes the highest level of education that you have completed?**
  - [ ] Did not attend school  
  - [ ] Primary School  
  - [ ] Secondary School (High School)  
  - [ ] Some Undergraduate College/University  
  - [ ] Undergraduate College/University  
  - [ ] Some Graduate College/University  
  - [ ] Graduate College/University

- **Please indicate if you currently attend either of the following institutions:**
  - [ ] University of Toronto  
  - [ ] Other (please specify) ______

---

**General Technology Questionnaire**

**Computer Skills**

Approximately how often do you use a computer?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
<td>Once per month</td>
<td>Once every two weeks</td>
<td>Several times a week</td>
<td>Daily</td>
</tr>
</tbody>
</table>

**Portable Device Skills**

Approximately how often do you use a touch-based device (ex. computer, tablet, smartphone)?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
<td>Once per month</td>
<td>Once every two weeks</td>
<td>Several times a week</td>
<td>Daily</td>
</tr>
</tbody>
</table>

- **What type of touch-based devices do you use most frequently?**
  - [ ] Smartphone (ex. iPhone, Blackberry)  
  - [ ] Media player (ex. iPod)  
  - [ ] Tablet (ex. iPad, Android tablet)  
  - [ ] Laptop  
  - [ ] Computer monitor

---

### 8.12.1.4 Participant Compensation Receipt

**Study:**

Exercising Game-Based and Gait-Related Cognitive Assessments

**Date:**

________________________________________________

**Participant Name:**

________________________________________________

**Participant’s Signature:**

________________________________________________

**Compensation Amount:**

$ ..................................................

**Paid by:**

________________________________________________

**Gift Card Code:**

________________________________________________
8.12.2 Supporting Materials.

8.12.2.1 Literature Review

8.12.2.1.1 Dual-Task Assessment of Cognitive Status

In this section, I review past research relating to gait and cognitive assessment, with particular interest in the use of serious games, and dual-task (DT) methods.

Patients with cognitive impairments, such as AD, may have higher risk of falls and injuries due to decreased EF abilities (and thus diminished high level control of gait) and decreased motivation to remain physically active. EF are important in tasks such as walking, balance, and obstacle avoidance (Hausdorff et al., 2005; Persad et al., 1995).

8.12.2.1.2 Movement and Cognition Interplay

Growing awareness of links between cognition and motor performance has fostered approaches that examine the beneficial effects of physical exercise on cognition, and/or benefits to gait and movement of cognitive exercise.

Szturm et al. (2013) compared the effect of performing a cognitive task while sitting, versus walking on a treadmill. They found a significant decrease in performance on the cognitive task when walking. Earlier work by Mizobuchi, Chignell and Newton (2005) compared the impact of sitting versus walking on text input on a handheld device. They observed faster text entry while sitting compared to walking. They demonstrated that text entry in a DT while walking (DTW) condition decreased text entry speed and increased error rates. Mirelman et al. (2014) carried out a study with young, healthy adults where participants were asked to complete an arithmetical task (counting down by seven) in a DTW condition. They found supporting evidence that gait is adversely affected by cognitive load and reduced EF. Overall, these studies support decreased performance in DTW conditions.

Pichierri, Murer and de Bruin (2012) compared a group that used a dancing game as an intervention, in addition to strength training and exercise, with a control group who received only strength and exercise training. Their findings suggested that playing the dancing game improved foot placement accuracy and walking performance, in single and DT conditions, while also improving resistance to falling. Other research involving games for DTW, has been carried out
by (Bruin, Reith, Dörflinger, & Murer, 2011), which investigated the use of a physical-computer game on DTW performance with assisted living residents. This game consisted of a monitor showing dance steps, which the user had to respond to using a dance pad. They observed a significant decrease in the DT cost of stride time and velocity (i.e., faster walking speed) for the experimental group that used the computer game, after the intervention, relative to their earlier stride time and velocity. Further work by van het Reve et al. (2014) examined the ability of cognitive training games to improve DTW performance, in adults living in homes for the aged, or living in the community. The control group received only a strength and balance training program whereas the intervention group also received cognitive training via games. They concluded that cognitive training is beneficial in improving balance and reducing risk of falls. Furthermore, research by Heiden and Lajoie (2010) investigated the effect of a game-based biofeedback training program in community-dwelling adults, versus a control group that only received exercise training. This game consisted of a user standing on two pressure sensors to control a virtual game of tennis (pong). The goal was for the participant to shift their weight to move the virtual paddle. They observed improved balance in older adults that played the game compared to adults in the control group. Overall, previous research findings suggest that cognitive-motor tasks (e.g., a serious game with a physical component) may be useful for more sensitive measures of cognitive status as well as for longer-term training of gait and cognition.

8.12.2.1.3 Dual-Task Walking in Older Adults

In healthy adults, performing a DTW has been shown to impact walking speed (Beauchet et al., 2005; Hollman, Kovash, Kubik, & Linbo, 2007; Lindenberger, Marsiske, & Baltes, 2000; J. Verghese et al., 2007), and stronger effects have been observed in older adults (Hollman et al., 2007; Lindenberger et al., 2000; Theill, Martin, Schumacher, Bridenbaugh, & Kressig, 2011). Changes in gait measures such as decreased walking speed (Camicioli, Howieson, Lehman, & Kaye, 1997; Persad, Jones, Ashton-Miller, Alexander, & Giordani, 2008), increased stride time (Muir et al., 2012), and increased stride time variability (Manuel Montero-Odasso et al., 2009; Muir et al., 2012; Sheridan, Solomont, Kowall, & Hausdorff, 2003) have been observed in adults with cognitive impairments. Differences in gait performance under DT conditions reveal significant changes compared to gait performance under single task conditions (Manuel Montero-Odasso et al., 2009; Muir et al., 2012), suggesting that measuring DTW may be more a sensitive measure of cognitive status, especially in older adults with cognitive deficits.
Other research suggests a relationship between DTW and falls risk. For example, research by Yamada et al. (2011) looked at the use of the “Basic Step” and “Skip Slalom” games from the Nintendo Wii Fit program in a DTW study with community-dwelling adults. They grouped their sample into two groups of healthy adults with no cognitive impairments: fallers and non-fallers. They observed significant correlations between the “Basic Step” game with (1) DTW lag walking time ($r=-0.547$) and with (2) performance on the Timed Up and Go task ($r=-0.688$).

Further research by Szturm, Betker, Moussavi, Desai and Goodman (2011) evaluated the ability of an interactive exercise game to improve balance in older hospitalized adults. The intervention group was asked to play a video game in addition to an exercise program. The control group only received the exercise program. They observed significant improvement in post-treatment compared to pre-treatment balance (but not gait) measures in both the control and intervention group, however greater improvements were observed in the intervention group.

8.12.2.1.4 Dual-Task Cognitive Assessment

Previous research on the reliability of a serious game played in a DTW condition has been carried out in adults with cognitive deficits (Tarnanas, Papagiannopoulos, et al., 2015). Research by Tarnanas, Papagiannopoulos, et al. (2015) investigated the use of a virtual reality (VR) serious game in combination with a DTW analysis as a reliable measure of detecting MCI in elderly adults. They demonstrated that serious game performance in a DTW condition was a more reliable measure of MCI compared to playing the serious game in a single task condition. Tarnanas, Papagiannopoulos, et al.'s (2015) earlier research focused on the validity of their serious game to predict MCI status in older adults in a single task condition only and not DTW (Tarnanas, Tsolaki, Nef, Müri, & Mosimann, 2014).

A DTW concept for detecting cognitive impairment eliciting medium to high cognitive control, called the Walking Stroop Carpet, was demonstrated by Perrochon, Kemoun, Watelain and Berthoz (2013), but to my knowledge, this is the first study reporting on a complex everyday activities serious game employing DTW, which might indicate prodromal AD.
8.12.3 Supporting Methodology Materials.

Figure 8-11. Illustration of how the smartphone was affixed to the individual during the gait analysis portion of the study using a smartphone holder.

8.12.4 Supporting Results Materials.

Table 8-31. Two-tailed Pearson’s $r$ correlation analysis between the median target offset game-based cognitive assessment played by while sitting with cognitive measures. $n=20$ for the correlations with all cognitive tasks except for the WCST, where $n=19$. * $p<0.05$, ** $p<0.01$, *** $p<0.001$. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>$n$-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0681</td>
<td>0.0651</td>
<td>0.420</td>
</tr>
<tr>
<td></td>
<td>$p=0.775$</td>
<td>$p=0.791$</td>
<td>$p=0.065$</td>
</tr>
<tr>
<td>2</td>
<td>-0.459$^*$</td>
<td>-0.0964</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>$p=0.0416$</td>
<td>$p=0.695$</td>
<td>$p=0.366$</td>
</tr>
<tr>
<td>3</td>
<td>-0.404</td>
<td>0.235</td>
<td>0.187</td>
</tr>
<tr>
<td></td>
<td>$p=0.00774$</td>
<td>$p=0.332$</td>
<td>$p=0.430$</td>
</tr>
<tr>
<td>4</td>
<td>-0.236</td>
<td>0.247</td>
<td>0.188</td>
</tr>
<tr>
<td></td>
<td>$p=0.317$</td>
<td>$p=0.307$</td>
<td>$p=0.428$</td>
</tr>
<tr>
<td>5</td>
<td>-0.583$^{**}$</td>
<td>0.173</td>
<td>0.315</td>
</tr>
<tr>
<td></td>
<td>$p=0.007$</td>
<td>$p=0.478$</td>
<td>$p=0.176$</td>
</tr>
<tr>
<td>6</td>
<td>-0.141</td>
<td>0.385</td>
<td>0.559$^{**}$</td>
</tr>
</tbody>
</table>
Table 8-32. Two-tailed Spearman’s rho correlation analysis between the median correct RT game-based cognitive assessment played by while sitting with cognitive measures. \( n=20 \) for the correlations with all cognitive tasks except for the WCST, where \( n=19 \). * \( p<0.05 \), ** \( p<0.01 \), *** \( p<0.001 \). WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.248</td>
<td>0.516</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td>( p=0.293 )</td>
<td>( p=0.0238 )</td>
<td>( p=0.033 )</td>
</tr>
<tr>
<td>2</td>
<td>-0.269</td>
<td>0.506</td>
<td>-0.073</td>
</tr>
<tr>
<td></td>
<td>( p=0.251 )</td>
<td>( p=0.0272 )</td>
<td>( p=0.759 )</td>
</tr>
<tr>
<td>3</td>
<td>0.217</td>
<td>0.317</td>
<td>0.286</td>
</tr>
<tr>
<td></td>
<td>( p=0.359 )</td>
<td>( p=0.186 )</td>
<td>( p=0.221 )</td>
</tr>
<tr>
<td>4</td>
<td>-0.0369</td>
<td>0.203</td>
<td>-0.051</td>
</tr>
<tr>
<td></td>
<td>( p=0.877 )</td>
<td>( p=0.405 )</td>
<td>( p=0.830 )</td>
</tr>
<tr>
<td>5</td>
<td>0.0820</td>
<td>0.391</td>
<td>0.181</td>
</tr>
<tr>
<td></td>
<td>( p=0.731 )</td>
<td>( p=0.0978 )</td>
<td>( p=0.446 )</td>
</tr>
<tr>
<td>6</td>
<td>0.216</td>
<td>0.250</td>
<td>-0.272</td>
</tr>
<tr>
<td></td>
<td>( p=0.361 )</td>
<td>( p=0.303 )</td>
<td>( p=0.246 )</td>
</tr>
</tbody>
</table>

Table 8-33. Two-tailed Spearman’s rho correlation analysis between the median target offset game-based cognitive assessment played by while sitting with cognitive measures. \( n=20 \) for the correlations with all cognitive tasks except for the WCST, where \( n=19 \). * \( p<0.05 \), ** \( p<0.01 \), *** \( p<0.001 \). WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.00903</td>
<td>-0.0511</td>
<td>0.353</td>
</tr>
<tr>
<td></td>
<td>( p=0.970 )</td>
<td>( p=0.835 )</td>
<td>( p=0.126 )</td>
</tr>
<tr>
<td>2</td>
<td>-0.365</td>
<td>0.128</td>
<td>0.159</td>
</tr>
<tr>
<td></td>
<td>( p=0.114 )</td>
<td>( p=0.602 )</td>
<td>( p=0.502 )</td>
</tr>
<tr>
<td>3</td>
<td>-0.295</td>
<td>0.294</td>
<td>0.297</td>
</tr>
<tr>
<td></td>
<td>( p=0.207 )</td>
<td>( p=0.222 )</td>
<td>( p=0.203 )</td>
</tr>
<tr>
<td>4</td>
<td>-0.168</td>
<td>0.205</td>
<td>0.276</td>
</tr>
<tr>
<td></td>
<td>( p=0.478 )</td>
<td>( p=0.400 )</td>
<td>( p=0.239 )</td>
</tr>
<tr>
<td>5</td>
<td>-0.591 **</td>
<td>0.168</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td>( p=0.0061 )</td>
<td>( p=0.493 )</td>
<td>( p=0.164 )</td>
</tr>
<tr>
<td>6</td>
<td>-0.160</td>
<td>0.507 *</td>
<td>0.534 **</td>
</tr>
<tr>
<td></td>
<td>( p=0.500 )</td>
<td>( p=0.0268 )</td>
<td>( p=0.015 )</td>
</tr>
</tbody>
</table>

Table 8-34. Two-tailed Pearson’s \( r \) correlation analysis between the median correct target offset on game-based cognitive assessment while walking with cognitive measures. \( n=20 \) for
the correlations with all cognitive tasks except for the WCST, where n=19. * p<0.05, ** p<0.01, *** p<0.001. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.396</td>
<td>0.101</td>
<td>0.522</td>
</tr>
<tr>
<td></td>
<td>p=0.0839</td>
<td>p=0.678</td>
<td>p=0.018</td>
</tr>
<tr>
<td>2</td>
<td>0.345</td>
<td>0.0913</td>
<td>0.472</td>
</tr>
<tr>
<td></td>
<td>p=0.136</td>
<td>p=0.710</td>
<td>p=0.036</td>
</tr>
<tr>
<td>3</td>
<td>0.433</td>
<td>0.250</td>
<td>0.718</td>
</tr>
<tr>
<td></td>
<td>p=0.0566</td>
<td>p=0.303</td>
<td>p=0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.362</td>
<td>0.341</td>
<td>0.382</td>
</tr>
<tr>
<td></td>
<td>p=0.117</td>
<td>p=0.152</td>
<td>p=0.097</td>
</tr>
<tr>
<td>5</td>
<td>0.354</td>
<td>0.171</td>
<td>0.468</td>
</tr>
<tr>
<td></td>
<td>p=0.126</td>
<td>p=0.485</td>
<td>p=0.037</td>
</tr>
<tr>
<td>6</td>
<td>0.515</td>
<td>0.102</td>
<td>0.442</td>
</tr>
<tr>
<td></td>
<td>p=0.0202</td>
<td>p=0.678</td>
<td>p=0.051</td>
</tr>
</tbody>
</table>

Table 8-35. Two-tailed Spearman’s rho correlation analysis (n=20) between the median correct RT on game-based cognitive assessment while walking with cognitive measures. * p<0.05, ** p<0.01, *** p<0.001. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.145</td>
<td>0.393</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>p=0.541</td>
<td>p=0.0959</td>
<td>p=0.740</td>
</tr>
<tr>
<td>2</td>
<td>-0.265</td>
<td>0.479</td>
<td>-0.154</td>
</tr>
<tr>
<td></td>
<td>p=0.258</td>
<td>p=0.0381</td>
<td>p=0.518</td>
</tr>
<tr>
<td>3</td>
<td>-0.236</td>
<td>0.410</td>
<td>-0.230</td>
</tr>
<tr>
<td></td>
<td>p=0.318</td>
<td>p=0.081</td>
<td>p=0.329</td>
</tr>
<tr>
<td>4</td>
<td>-0.233</td>
<td>0.656</td>
<td>-0.119</td>
</tr>
<tr>
<td></td>
<td>p=0.322</td>
<td>p=0.0022</td>
<td>p=0.617</td>
</tr>
<tr>
<td>5</td>
<td>-0.0135</td>
<td>0.348</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>p=0.955</td>
<td>p=0.144</td>
<td>p=0.764</td>
</tr>
<tr>
<td>6</td>
<td>-0.250</td>
<td>0.355</td>
<td>-0.160</td>
</tr>
<tr>
<td></td>
<td>p=0.289</td>
<td>p=0.136</td>
<td>p=0.500</td>
</tr>
</tbody>
</table>

Table 8-36. Two-tailed Spearman’s rho correlation analysis (n=20) between the median correct target offset on game-based cognitive assessment while walking with cognitive measures. * p<0.05, ** p<0.01, *** p<0.001. WCST = Wisconsin-card sorting task.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.449</td>
<td>0.0648</td>
<td>0.446</td>
</tr>
<tr>
<td></td>
<td>p=0.0407</td>
<td>p=0.792</td>
<td>p=0.049</td>
</tr>
<tr>
<td>2</td>
<td>0.321</td>
<td>0.0216</td>
<td>0.458</td>
</tr>
<tr>
<td></td>
<td>p=0.167</td>
<td>p=0.930</td>
<td>p=0.042</td>
</tr>
<tr>
<td>3</td>
<td>0.472  *</td>
<td>0.314</td>
<td>0.713</td>
</tr>
<tr>
<td></td>
<td>p=0.0355</td>
<td>p=0.190</td>
<td>p=0.000</td>
</tr>
<tr>
<td>4</td>
<td>0.364</td>
<td>0.187</td>
<td>0.341</td>
</tr>
<tr>
<td></td>
<td>p=0.114</td>
<td>p=0.443</td>
<td>p=0.141</td>
</tr>
</tbody>
</table>
### Table 8-37. Two-tailed Spearman’s rho correlation analysis between performance measures on the cognitive task and dual-task (DT) cost of playing the serious game using stride variability. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.126</td>
<td>0.0519</td>
<td>-0.256</td>
</tr>
<tr>
<td></td>
<td>$p=0.609$</td>
<td>$p=0.838$</td>
<td>$p=0.293$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>2</td>
<td>-0.131</td>
<td>-0.119</td>
<td>-0.0960</td>
</tr>
<tr>
<td></td>
<td>$p=0.593$</td>
<td>$p=0.637$</td>
<td>$p=0.696$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>3</td>
<td>-0.0729</td>
<td>0.0166</td>
<td>-0.198</td>
</tr>
<tr>
<td></td>
<td>$p=0.767$</td>
<td>$p=0.948$</td>
<td>$p=0.416$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>4</td>
<td>-0.0610</td>
<td>0.0702</td>
<td>-0.144</td>
</tr>
<tr>
<td></td>
<td>$p=0.810$</td>
<td>$p=0.789$</td>
<td>$p=0.566$</td>
</tr>
<tr>
<td></td>
<td>$n=18$</td>
<td>$n=17$</td>
<td>$n=18$</td>
</tr>
<tr>
<td>5</td>
<td>0.0310</td>
<td>0.0259</td>
<td>-0.0549</td>
</tr>
<tr>
<td></td>
<td>$p=0.968$</td>
<td>$p=0.288$</td>
<td>$p=0.652$</td>
</tr>
<tr>
<td></td>
<td>$n=18$</td>
<td>$n=17$</td>
<td>$n=18$</td>
</tr>
<tr>
<td>6</td>
<td>-0.168</td>
<td>0.146</td>
<td>-0.177</td>
</tr>
<tr>
<td></td>
<td>$p=0.903$</td>
<td>$p=0.922$</td>
<td>$p=0.829$</td>
</tr>
<tr>
<td></td>
<td>$n=18$</td>
<td>$n=17$</td>
<td>$n=18$</td>
</tr>
</tbody>
</table>

### Table 8-38. Two-tailed Spearman’s rho correlation analysis ($n=18$) between performance measures on the cognitive task and dual-task (DT) cost of playing the serious game using stride time. * $p<0.05$, ** $p<0.01$, *** $p<0.001$.

<table>
<thead>
<tr>
<th>Game Level</th>
<th>Stroop Interference Effect</th>
<th>WCST Error Rate</th>
<th>n-back Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.129</td>
<td>0.270</td>
<td>-0.142</td>
</tr>
<tr>
<td></td>
<td>$p=0.599$</td>
<td>$p=0.279$</td>
<td>$p=0.563$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>2</td>
<td>0.271</td>
<td>0.583</td>
<td>-0.0845</td>
</tr>
<tr>
<td></td>
<td>$p=0.261$</td>
<td>$p=0.0111$</td>
<td>$p=0.731$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td>3</td>
<td>-0.0975</td>
<td>0.469</td>
<td>0.0332</td>
</tr>
<tr>
<td></td>
<td>$p=0.691$</td>
<td>$p=0.0495$</td>
<td>$p=0.344$</td>
</tr>
<tr>
<td></td>
<td>$n=19$</td>
<td>$n=18$</td>
<td>$n=19$</td>
</tr>
<tr>
<td></td>
<td>p</td>
<td>q</td>
<td>r</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>4</td>
<td>-0.261</td>
<td>0.473</td>
<td>0.0332</td>
</tr>
<tr>
<td></td>
<td>p=0.295</td>
<td>p=0.0552</td>
<td>p=0.896</td>
</tr>
<tr>
<td></td>
<td>n=18</td>
<td>n=17</td>
<td>n=18</td>
</tr>
<tr>
<td>5</td>
<td>-0.329</td>
<td>0.490</td>
<td>0.00207</td>
</tr>
<tr>
<td></td>
<td>p=0.183</td>
<td>p=0.0458</td>
<td>p=0.994</td>
</tr>
<tr>
<td></td>
<td>n=18</td>
<td>n=17</td>
<td>n=18</td>
</tr>
<tr>
<td>6</td>
<td>-0.332</td>
<td>0.462</td>
<td>-0.124</td>
</tr>
<tr>
<td></td>
<td>p=0.179</td>
<td>p=0.0620</td>
<td>p=0.623</td>
</tr>
<tr>
<td></td>
<td>n=18</td>
<td>n=17</td>
<td>n=18</td>
</tr>
</tbody>
</table>
8.13 Appendix L. Delirium Index.

The Delirium Index (DI) is an instrument for the measurement of severity of symptoms of delirium that is based solely upon observation of the individual patient, without additional information from family members, nursing staff, or the patient's medical chart. The DI was designed to be used in conjunction with the MMSE; at least the first five questions of the MMSE constitute the basis of observation. Additional questions may be necessary for scoring certain symptoms as noted.

1 Inattention
   0 Attentive.
   1 Generally attentive but makes at least one error in spelling “WORLD” backwards.
   2 Questions can generally be answered but subject is distractible and at times has difficulty keeping track of questions. May have some difficulty shifting attention to new questions, or questions may have to be repeated several times.
   3 Unresponsive or totally unable to keep track of or answer questions. Has great difficulty in focusing attention and is often distracted by irrelevant stimuli.
   9 Cannot assess.

2 Disorganized thinking
   0 Responses are logical, coherent, and relevant.
   1 Responses are vague or unclear.
   2 Thought is occasionally illogical, incoherent, or irrelevant.
   3 Unresponsive or thought is fragmented, illogical, incoherent, and irrelevant.
   9 Cannot assess.

3 Altered level of consciousness
   0 Normal.
   1 Hypervigilant or hypovigilant (glassy eyed, decreased reaction to questions).
   2 Drowsy/sleepy. Responds only to simple, loud questions.
   3 Unresponsive or comatose.

4 Disorientation (additional questions on age, birth date, and birthplace may be used)
   0 Knows today's date (±1 day) and the name of the hospital.
   1 Does not know today's date (±1 day) or does not know the name of the hospital.
   2 Does not know the month or year or does not know that is in the hospital.
   3 Unresponsive or does not know name or birth date.
   9 Cannot assess.

5 Memory impairment (Additional questions may be asked on how long patient has been in the hospital, circumstances of admission.)
   0 Recalls three words or details of hospitalization
   1 Cannot recall one of the words or has difficulty recalling details of the hospitalization.
   2 Cannot recall two of the three words or recalls few details of the hospitalization.
   3 Unresponsive or cannot recall any of the three words or any details of the hospitalization.
   9 Cannot assess.

6 Perceptual disturbances (Patient is asked whether she/he has had any unusual experiences or has seen or heard things that other people do not see or hear. If yes, she/he is asked whether these occur during the daytime or at night and how frequently. Patient is also observed for any evidence of disordered perception.)
   0 Unresponsive, no perceptual disturbances observed, cannot assess.
   1 Misinterprets stimuli (for example, interpreting a door closing as a gunshot).
   2 Occasional nonthreatening hallucinations.
   3 Frequent, threatening hallucinations.

7 Motor disturbances
   0 Normal.
   1 Responds well to questions but moves frequently or is lethargic/sluggish.
   2 Moves continuously (and may be restrained) or very slow with little spontaneous movement.
   3 Agitated, difficult to control (restraints are required) or no voluntary movement.

Scoring:
Total score is sum of seven item scores.
If questions 1, 2, 4, or 5 are checked “9,” replace 9 by the score of item 3.