How does practice affect working memory? The efficacy of adaptive-difficulty working memory training programs.

by

James M. Stone

BSc (Psychology), Lancaster University (2009)
MSc (Psychological Research Methods), Lancaster University (2010)
MRes (Applied Social Statistics), Lancaster University (2011)

A thesis submitted for the degree of
Ph.D.
at
Lancaster University
September 2015

How does practice affect working memory? The efficacy of adaptive-difficulty working memory training programs.

James M. Stone

Abstract

Working memory refers to a mental 'workbench' whereby new or goal relevant information is held in a readily accessible state in order to achieve success with cognitive problems. Working memory has been shown to be relevant to individual differences in many aspect of cognition including fluid intelligence, while also identified as a core deficit in cognitive developmental disorders. Therefore the possibility that working memory capacity is trainable, and that such interventions produce generalisable cognitive benefits is highly noteworthy. Following these initial claims from early studies using an adaptive working memory training intervention, commercial products have been developed and marketed, based on the premise that working memory can indeed be trained and lead to transfer to a wider range of cognitive abilities.

A close examination of the literature suggests that these claims are based on a combination of mixed generalisable results and often a lack of evidence for genuine working memory improvement. In some examples the analyses testing for working memory improvement fail to show such effects while in others the chosen assessment tasks are too similar to the trained tasks to be evidence for general working memory improvement. The potential for improved fluid intelligence and amelioration of deficits leading to developmental issues due to a working memory training intervention is of clear practical importance. However, for such arguments to be made convincingly it is of critical importance that there is increased understanding of the effects of working memory training on the construct itself. Thus far the effects seen in the literature have not proved to be robust and therefore the mechanisms of any proposed effect need to be examined to illuminate the conceptual and empirical benefits of such training interventions, whilst also establishing their robustness and reliability. In addition, there is reason to believe that tighter methodological designs are needed to make the domain more credible with particular emphasis on suitable active control groups.

Therefore this thesis pursued a series of studies assessing the efficacy of adaptive difficulty WM training interventions with an emphasis on the impact on the working memory construct (near-transfer). Thus, the assessment of potential improvements in working memory formed the core analyses within this thesis. Therefore, in addition to the three working memory training studies, this thesis will also addressed various methods of assessing working memory performance from typical behavioural assessments, so as to inform the methods utilised in the pre-post changes in the training studies. This was achieved by means of analysing alternative scoring methods, assessment protocols, and modelling the difficulty of different list length trials using a Rasch model in working memory tasks.

Each training study utilised a randomised pre-post (experiment one also incorporated a follow-up phase) intervention design where the control group completed a demanding regime of tasks with minimal stress on working memory. In each experiment participants completed a battery of tasks prior to the onset of the training phase that lasted 5-6 weeks. During the training phase an intervention group would complete the specified training regime while an active control group would complete different tasks. Following the conclusion of the training phase the initial battery of tasks was re-administered (experiment one also incorporated a follow-up phase). Additionally, all software used was developed specifically for this project including control group tasks therefore maintaining consistency in the 'look and feel' between the two groups.

In experiment one (N = 55) an intervention using a variety of working memory based tasks (Working Memory Period, Memory Updating, Colour Corsi, and Stroop) was assessed and compared with an active control intervention in children aged 9-10. No improvements were found in the individual tasks measuring verbal and visuo-spatial working memory, nor were there training effects. In experiment two (N = 76) children aged 9-11 participated in three working memory interventions with each consisting of a single task (Working Memory Period, Colour Corsi, and N-Back). No improvements were seen in any training group beyond what was also seen in an active control group in composite measures of verbal and visuo-spatial working memory, or processing speed. A final working memory training experiment (N = 55) replicated the results from experiment two (Dual N-Back replacing standard N-Back) using a sample of healthy adults.

These data suggest, in typically developing children and adults, that adaptive working memory training interventions may not improve working memory functions. These results cast doubt on the potential for such interventions to produce improved performance in a wider range of higher order cognitive abilities. Potential reasons for the existence of a number of positive results in the literature are considered including ineffective controls, publication bias, and potential false positives as a consequence of multiple comparisons - with regards the number of tasks used in a pre-post battery in addition to the number of measures one can extract from each task compounded by the possibility of analysing these data using multiple methods.

Declaration

I declare that the work presented in this thesis is my own. None of the data or material contained in this thesis has been submitted for previous or simultaneous consideration for a degree at this or any other institution.

Date:

Signature:

Acknowledgements

I would like to thank my supervisor John Towse for his mentorship throughout this project. John was equally supportive as he was insightful in the many discussions we have had throughout the course of completing this project. I couldn't have asked for a better supervisor and I will be forever grateful I had the opportunity to work with him.

Two of the presented experiments would not have been possible without the generous willingness to participate shown by eight classes across four schools in the North West of England. I thank the pupils and staff of these schools for welcoming myself and being willing to participate in somewhat intensive studies.

Finally a big thank you to my family especially my Mother, Sister, and Grandparents. Without your support this would never have been possible.

Contents

1 Literature Review **15**

1.1 Introduction . . . 15

1.2 The assessment of memory . . . 16

1.2.1 Early methods and role in intelligence testing . . . 16

1.2.2 Memory measurement in the wider context . . . 19

1.3 From tools to theory . . . 22

1.3.1 Models of short-term memory . . . 22

1.3.2 Influence of the WM concept on memory assessment . . . 24

1.4 Cognitive Training . . . 26

1.4.1 Early attempts to improve cognitive ability . . . 26

1.4.2 WM as a trainable construct? . . . 29

1.4.3 Further successes and range of application of WMT . . . 34

1.4.4 Commercialisation of WMT . . . 38

1.5 Is WMT effective? . . . 39

1.5.1 Classification of types of effects . . . 39

1.5.2 Efficacy of WMT - A critical evaluation . . . 42

1.6 Does adaptive-difficulty working memory training actually improve working memory ability? . . . 51

1.7 Summary and thesis objectives . . . 55

2 Working Memory Training Developmental Study One **58**

2.1 Working Memory Training - Study One . . . 58

2.2 Method . . . 64

2.2.1 Participants . . . 64

2.2.2 Design and materials . . . 64

2.2.3 Procedure . . . 76

2.3 Results . . . 78

2.3.1 Data Hygiene . . . 78

2.3.2 Practice Effects (RQ1) . . . 80

2.3.3 Transfer Effects (RQ2/3 . . . 87

2.3.4 Assessing the active control group (RQ5) . . . 94

2.3.5 Post-Hoc Power Analyses . . . 96

2.4 Discussion . . . 97

2.4.1 P1: Performance Change on the trained tasks . . . 97

2.4.2 Generalisable benefits of the intervention (Transfer Effects) . . . 98

2.4.3 Potential impact of administration environment . . . 100

2.4.4 Sustained attention of control group 101

3 Task Validation and Scoring Comparisons 102

3.1 Introduction 102

3.1.1 Outline of the goals of this chapter 105

3.2 Method 108

3.2.1 Participants 108

3.2.2 Design and Materials 108

3.2.3 Procedure 110

3.3 Results 111

3.3.1 Working Memory Period 111

3.3.2 Operation Span 118

3.3.3 Colour Corsi 119

3.3.4 Arrow Span 119

3.3.5 Scoring comparisons and correlational analyses 121

3.4 Using the Rasch Model To Assess WM Task Items 126

3.5 Discussion 131

3.5.1 WMP Properties 131

3.5.2 Arrow Span Properties 132

3.5.3 Scoring Methods 133

4 Working Memory Training Developmental Study Two 137

4.1 Introduction 137

4.2 Method 140

4.2.1 Participants 140

4.2.2 Design and Materials 141

4.2.3 Procedure 151

4.3 Results 152

4.3.1 Comparison of excluded participants and retained participants at T1152

4.3.2 Correlation/PC Analysis of T1 data 152

4.3.3 Practice Effects 157

4.3.4 Transfer Effects 164

4.4 Discussion 170

5 Working Memory Training Adult Study 175

5.1 Introduction 175

5.2 Method 178

5.2.1 Participants 178

5.2.2 Design and materials 179

5.2.3 Procedure 185

5.3 Results 186

5.3.1 Correlation/PC Analysis of T1 data 186

5.3.2 Practice Effects 193

5.3.3 Transfer Effects 196

5.4 Discussion 206

6 General Discussion 210
6.1 Brief Overview 210
6.2 Does Practice Lead To Enhanced Working Memory? 213
6.3 Practical and Theoretical Implications 220
6.4 Strengths and Limitations 223
6.5 Methodological Implications 227
6.6 Future Work 229
6.7 Conclusions 232

References 234

A Working Memory Training Developmental Study One 252
A.1 Ethics Statement 252
A.2 Method 252
A.2.1 Mental Arithmetic Stimuli 252
A.3 Data Clean Up 254
A.3.1 WMP2 254
A.3.2 Stroop 255
A.4 Practice Effects 256
A.4.1 WMP 256
A.4.2 WMP2 256
A.4.3 Colour Corsi 257
A.4.4 Memory Update 257
A.4.5 Stroop 258

B Working Memory Training Developmental Study Two 259
B.1 Method 259
B.1.1 Software Images 259
B.1.2 Mental Arithmetic Stimuli 261
B.2 Practice Effects 263
B.2.1 WMP 263
B.2.2 Colour Corsi 263
B.2.3 Spatial N-Back 263

C Working Memory Training Adult Study 265
C.1 Method 265
C.1.1 Software Images 265
C.2 Practice Effects 268
C.2.1 WMP 268
C.2.2 Colour Corsi 269
C.2.3 Dual N-Back 269

List of Tables

2.1 Mean and standard deviation values for pre, post, and follow-up near transfer measures. Accuracy = proportion correct, RT = response time in seconds. 88
2.2 Mean and standard deviation values for pre, post, and follow-up far transfer measures. Accuracy = proportion correct, Response Time = seconds . . . 90
2.3 Mean number of correct responses for each block of Mental Arithmetic questions. 91
2.4 Overview of Omnibus Mancova results for transfer measures; ($[Post, Follow-up] \sim Pre + group$); Λ = Wilk's Lambda statistic, F statistic, and associated p-value. 92
2.5 Overview of subsequent Ancova analyses and effect sizes for near-transfer measures; d = Cohen's d, g = Hedges' g, BF = Bayes Factor. 93
2.6 Overview of subsequent Ancova analyses and effect sizes for far-transfer measures; d = Cohen's d, g = Hedges' g, BF = Bayes Factor. 94

3.1 Descriptive statistics for the Working Memory Period task. 112
3.2 Model fit values for the various combinations of fixed effects applied to the MLM analysis of WMP response times. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion. 115
3.3 Parameter estimates for fixed effects for the MLM analysis of WMP response times (log scale). 117
3.4 Descriptive statistics for the Operation Span task. 118
3.5 Descriptive statistics for the Colour Corsi task. 119
3.6 Descriptive statistics for the Arrow Span task. 120
3.7 Correlation matrix consisting of each scoring method for all tasks administered. Variable names take the form task.method. The tasks are: os = Operation Span, as = Arrow Span, cc = Colour Corsi, wmp = Working Memory Period. The scoring methods are: span = the maximum list length or level the participant was successful, fta = full trial accuracy score, nc = total number of correct individual memoranda, t = t-score, nc with correction for transposition errors. 125
3.8 Descriptive Statistics for four scoring methods for each task. The tasks are: os = Operation Span, as = Arrow Span, cc = Colour Corsi, wmp = Working Memory Period. The scoring methods are: span = the maximum list length or level the participant was successful, fta = full trial accuracy score, nc = total number of correct individual memoranda, t = t-score, nc with correction for transposition errors. 127

3.9 Rasch Model fit statistics for the Operation Span test. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion. 129
3.10 Parameter coefficients for Operation Span Items. The $p(x = 1|z = 0)$ column indicates probability of success on that item for a participant of average ability. 131

4.1 Correlation Matrix of each primary dependent variable; Digit = Digit Span t-score, Op = Operation Span t-score, Reading = Reading Span t-score, Matrix = Matrix Span t-score, Symm = Symmetry Span t-score, IT = Inspection Time mean SOA, OOO = Odd One Out decision time, M-Arith = Total correct operations across all blocks 153
4.2 Variance explained by the extracted principal components (initial unrotated solution) 154
4.3 Loadings matrix for the 3-PC promax solution 155
4.4 Variance explained by the PCs (memory span only - initial solution) 156
4.5 Loadings matrix for the 2-PC promax solution (memory span only) 156
4.6 Pre- and post-training performance on each of the transfer tasks for the three training groups and active control. Values represent means with standard deviation in parenthesis. Tasks; CC = Colour Corsi, WMP = Working Memory Period, SNB = Spatial N-Back 163
4.7 Summary of performance on the processing phase of complex span tasks at pre- and post-training. Accuracy = proportion correct, RT = response time in milliseconds 167
4.8 Summary of Ancova results plus the effect size (Hedges' g) for pairwise comparisons comparing the adjusted means for each training group to the active control group. 168
4.9 Summary of the time by group interaction fixed effect for each depdendent variable of interest. The likelihood ratio test compares the fit of the model with and without the interaction effect where a significant result indicates improved model fit 169

5.1 Correlation Matrix of each primary dependent variable; OOO = Odd One Out decision time, IT = Inspection Time mean SOA, M-Arith = Total correct operations across all blocks, F-Recall = Total correctly recalled items across all FR trials, Digit = Digit Span t-score, Op = Operation Span t-score, Matrix = Matrix Span t-score, Symm = Symmetry Span t-score, Rotation = Rotation Span t-score, Reading = Reading Span t-score.187
5.2 Variance explained by the Principal Components from the initial unrotated solution with all variables included 189
5.3 Loadings matrix for the 4-PC promax solution (loadings below 0.3 suppressed) 189
5.4 Loadings matrix for the 3-PC promax solution (loadings below 0.3 suppressed) 190
5.5 Loadings matrix for the 2-PC promax solution with only memory span measures included (loadings below 0.3 suppressed) 191

5.6 Correlation matrix of processing speed measures from complex span tasks and primary dependent variables from processing speed tasks 192
5.7 Summary of ANCOVA results plus the effect size (Hedges' g) for pairwise comparisons comparing the adjusted means for each training group to the active control group. Tasks; CC = Colour Corsi, DNB = Dual N-Back, WMP = Working Memory Period. Scores; FTA = full-trial accuracy, T = transposition score (number of indivdidual items correct plus partial credit for transpositions) 199
5.8 Summary of the time by group interaction fixed effect for each dependent variable. The likelihood ratio tests the significance of the effect. FTA = full-trial accuracy, T = transposition score (number of indivdidual items correct plus partial credit for transpositions) 200
5.9 Pre- and post-training performance on the memory span transfer tasks for the three training groups and active control. Values represent means with standard deviation in parenthesis. 201
5.10 Summary of processing elements of complex span tasks at pre- and post-training 202
5.11 Pre- and post-training performance on the non-span transfer tasks for the three training groups and active control. Values represent means with standard deviation in parenthesis. IT = Inspection Time Mean SOA (ms), OOO = Odd One Out Decision Time (ms), F-Recall = Free Recall Total Items Recalled, M-Arith = Mental Arithmetic Total Correct Operations. . 203

A.1 Question sets for each block in the Mental Arithmetic task used in the WMT Developmental One study 253

B.1 Question sets for each block in the Mental Arithmetic task used in the WMT Developmental Two study 262

List of Figures

2.1 Working Memory Period (WMP) Task 66
2.2 Illustration of the memory updating task 68
2.3 Illustration of the Colour Corsi task 69
2.4 Illustration of the Operation Span task used in experiment one. 74
2.5 Overview of experiment design 78
2.6 Distribution of participants between school classes and experimental groups 79
2.7 Parameter estimates with 95% confidence intervals for the effect of session number on primary indicators of performance per training task 83
2.8 Operation RT (ms) deviation from baseline as a function of session by serial position; Top - SP1, Mid - SP2, Bottom - SP3. Points indicate parameter estimates (adjusted mean) with 95 % confidence intervals. 85
2.9 Stroop effect deviation from baseline as estimated by the GLMM with 95 % confidence intervals. 86
2.10 Key boxplots for the near- and far-transfer measures. MS = Matrix Span, OS = Operation Span, SS = Silly Sentence, MA = Mental Arithmetic. . . 89
2.11 Assessing attentiveness to control tasks with measures of trials completed and accuracy. Parameter estimates with 95% confidence interval from GLMM analysis. 95

3.1 Mean response time (ms) for each level of WMP operation (bars represent 95% confidence interval). 113
3.2 Mean response time (ms) per serial position on the WMP task (bars represent 95% confidence interval). 114
3.3 Mean response time for WMP operations at each serial position grouped by level of operation. 116
3.4 This chart displays how many participants would be successful at higher list lengths than where traditional span terminating algorithms would end administration. These values represent the number of times the values for participants differ on the termination span variable with the max span (1) and max span (2) variables. AS = Arrow Span, CC = Colour Corsi, OS = Operation Span, WMP = Working Memory Period 121
3.5 Distribution properties of scoring methods for Operation Span (raw sample, sample density plot, normal density plot) 123
3.6 Distribution properties of scoring methods for Working Memory Period (raw sample, sample density plot, normal density plot) 123

3.7 Distribution properties of scoring methods for Arrow Span (raw sample, sample density plot, normal density plot) 124
3.8 Distribution properties of scoring methods for Colour Corsi (raw sample, sample density plot, normal density plot) 124
3.9 Operation Span - Item Characteristic Curves. 128
3.10 Reading Span - Item Characteristic Curves. 129
3.11 Symmetry Span - Item Characteristic Curves. 130

4.1 Flow diagram highlighting participant involvement throughout WMT study 2. 141
4.2 Illustration of the Colour Corsi task 143
4.3 Illustration of the Working Memory Period (WMP) Task 144
4.4 Illustration of the Symmetry Span task 147
4.5 Illustration of the Operation Span task 148
4.6 Illustration of the Reading Span task 149
4.7 Illustration of the Odd One Out task 150
4.8 Illustration of the Inspection Time task 150
4.9 Scree plot for initial PCA solution 154
4.10 Scatterplot showing the relationship between number of 'sessions' completed and the number of trials this resulted in for each person 158
4.11 This figure highlights the number of participants who reached each stage of the training for each of the groups. 159
4.12 CC; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks . . 160
4.13 WMP; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks . . 161
4.14 N-Back; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks . . 162
4.15 Pre- and Post-Training scores for the verbal and visuo-spatial WM PC scores as determined by the coefficients derived from pre-training PCA analysis with standard errors. 165

5.1 Flow diagram highlighting participant involvement throughout the WMT adult study 179
5.2 Illustration of the Rotation Span task 181
5.3 Illustration of the Dual N-Back task 184
5.4 Scree plot highlighting the eigenvalues for extracted PCs in the initial solution (Adult - WMT3) 188
5.5 Scree plot for initial PCA solution using the span subset data 191
5.6 Number of participants who reached each stage of the training for each of the groups. 193

5.7 CC; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks . . 194
5.8 WMP; - Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks . . 195
5.9 Dual N-Back - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number 196
5.10 Standardised Pre- and post-training scores for the verbal (top) and visuospatial (bottom) composite measures as determined by the coefficients derived from the PC analysis with standard errors 198
5.11 Scatterplots showing the observed mean training improvement against difference scores on the PCs (top - verbal; bottom - spatial) 205

A.1 Distribution of the proportion of correct operations; WMP (top) and WMP2 (bottom) 254
A.2 WMPmean level spaghetti plot 256
A.3 WMP2 mean level spaghetti plot 256
A.4 CC mean level spaghetti plot 257
A.5 MU mean ISI spaghetti plot 257
A.6 Stroop mean level spaghetti plot 258

B.1 Screenshot showing the 'look and feel' of the Spatial N-Back task used in the second developmental training study. This highlights the status panel changes that were altered from that used in the adult study in an effort to make the program look more colourful and appealing to the much younger sample. 260
B.2 Screenshot that shows the 'Stacker' game that was included in the second developmental study as a reward for completing a training session. Participants were given two minutes to play Stacker at the end of each completed session. 260
B.3 WMP Training; Top - Performance over all training sessions for each participant as measured by the mean level achieved at each session, Bottom - Split into trial blocks 263
B.4 CC Training; Top - Performance over actual logged training sessions for each participant as measured by the mean level achieved, Bottom - Split into trial blocks 264
B.5 N-Back Training; Top - Performance over all training sessions for each participant as measured by the mean level (n) achieved at each session, Bottom - Split into trial blocks 264

C.1 Screenshot showing the look and feel of the Colour Corsi task as used in the adult WMT study. 266
C.2 Screenshot showing the look and feel of the Working Memory Period task as used in the adult WMT study. 266

C.3 Screenshot showing the look and feel of the dual N-Back task as used in the adult WMT study. The microphone graphic was not visible, it is included in the diagram to show that as grids were presented so were letters in the auditory domain. 267
C.4 Screenshot showing the look and feel of the Sudoku (one example of the Active Control tasks) task as used in the adult WMT study. 267
C.5 WMP (Adult Study); Top - Performance aggregate per training 'session' for each participant as measured by the mean level achieved, Bottom - Split into trial blocks 268
C.6 CC (Adult Study); Top - Performance aggregate per training 'session' for each participant as measured by the mean level achieved, Bottom - Split into trial blocks 269
C.7 Dual N-Back Training (Adult Study) - Performance over all training sessions for each participant as measured by the mean level (n) achieved at each session 269

Chapter 1

Literature Review

1.1 Introduction

In this thesis I wish to explore two themes that while interconnected can be considered separate areas of study. The primary empirical questions are whether (and if so, to what extent?) a person's working memory (WM) can be improved, and in a broader sense, what are the consequences of repeated practice of WM tasks? This issue has risen to prominence in the past decade and has clear practical applications in addition to theoretical implications such as the static vs dynamic nature of the construct. The second question is methodological and focuses on how working memory is measured. Both of these topics are large areas in their own right and are worthy (as they have been) of being standalone topics. Obtaining an accurate measure of working memory ability is an objective whenever a researcher administers a WM measurement in their experimental design. However, the importance of the 'accuracy' of a WM measurement changes depending on the context one wishes to use it in. In the context of measuring improvement in working memory tasks then I believe the importance of the accuracy of those measurements to be exceptionally high, and in addition to that the importance of understanding what 'sub-processes' a given WM task is tapping. Another consideration in how working memory is measured that needs special consideration in this case is that of how one scores a participant to reflect their ability. Perhaps one needs to be aware of the sensitivity of various scoring methods and ensure they use a scoring method that is sensitive enough to exhibit any change that an intervention might cause.

These issues will be the main foci of this literature review and throughout this thesis, although in order to discuss them, other issues will be addressed too. They are also closely intertwined, so it will not be possible to consider each of them in isolation.

The structure of this review will be split into several key sections. I will discuss (briefly) the history of memory assessment tools in the research domain and discuss the development of the different paradigms that have dominated the inquiry into the structure of human memory. There will then be a discussion on how these tools influenced theory and vice versa, with emphasis on theoretical concepts that will be drawn upon in the rest of this thesis. The focus will then shift towards discussing the literature with regards to attempts to improve cognition, discussing some earlier work that can be considered a precursor to the modern attempts to improve cognition via working memory training (WMT). A thorough evaluation of the current WMT literature will be conducted where the strengths and weaknesses will be highlighted. This review will conclude with a statement on the particular research issues I will attempt to address in this thesis.

1.2 The assessment of memory

1.2.1 Early methods and role in intelligence testing

The importance of memory as a key mental faculty has long been accepted and discussed, from the philosophers of ancient Greece to present day researchers. The study of memory played a significant role in establishing Psychology as a scientific discipline in its own right. One such reason for this is the role that memory assessment took in the battery of tasks that made up intelligence tests at the beginning of the 20th century.

At this time the French government had made state education mandatory for all children. This development led to the need for an assessment tool that could be used to identify children who would need additional help in order to be successful in their education. Alfred Binet and Theodore Simon were tasked with creating such a tool, the Binet-Simon scale (Binet & Simon, 1905). The initial Binet-Simon scale consisted of 30 tasks that were appropriate for given ages. When a participant became unsuccessful on a task the participant had reached their limit and their mental age was derived as being the age that corresponded with the highest difficulty task that they were successful at. Of the 30 tasks included in this scale, 5 of them were classified by Binet and Simon as testing memory. In addition to these tasks, there were assessments of intelligence around this time that were not explicitly stated as memory assessments but what would clearly involve a memory component.

Two of the tasks in the Binet-Simon scale were based on the weight-discrimination test described by Francis Galton (Francis, 1883). In Galton's weight discrimination task participants were given three identical looking objects that differed in their weight and they were asked to rank these objects in ascending weight order. The difference in weight would be decreased until a participant was unable to accurately judge the weight of the items. This test was based on Galton's idea that a key aspect of intelligence is the acuteness of one's senses and the sensitivity with which one interacts with the external environment. This task can be seen as an early example of construct validity being a cause for concern as there are different methods (strategies) a participant can employ to tackle this problem. Depending on how a participant chose to solve the problem there are different cognitive processes potentially being assessed. For example, I may approach this task by actually assigning a weight to each object which would then allow me to rank them once I have all three of my weight judgements. This method of solving the problem would involve a perceptual process (of what Galton wanted to measure) of attempting to accurately weigh the objects, but also a memory component of keeping those judgements in mind while also remembering which object each judgement belongs to. An alternative process that might lead to success on this task could be to make a series of comparisons between two objects that allows me to infer where the other object lies in the ranking (i.e. A weighs more than B, B weighs more than C, therefore I know that A weighs more than C). This method then may or may not measure a memory component depending on whether the participant tries to remember the results from the dual comparisons or represents them externally by placement of the objects.

During the first world war the American Army invested resources into developing intelligence assessments that would quickly and effectively screen recruits and inform superiors on what role each recruit could potentially fill. The result of this was the Army Alpha and Army Beta tests (see Yerkes, 1921). These tests were informed by the existing intelligence tests such as the Binet-Simon scale which had by that time been adapted for use in America (Terman, 1916), revised version named the Stanford-Binet Intelligence Scales. The Army Alpha test used a subtest called memory span which is a version of the digit span task (in itself an extension of the 'repetition of three figures' task in the Binet-Simon scale). Additionally, the army tests had an additional individual examination that was given to recruits who scored particularly low or particularly high, as well as those where the examiners were unsure of a participants result. In this individual examination a backwards digit span task was used. Over a million recruits were assessed using these measures.

These batteries of tests that incorporate memory measurement in their pursuit of measuring the intelligence of test-takers were initially motivated by the prevailing opinions of what intelligence was, as opposed to theoretical models that had been rigorously tested. However, as Psychology as a discipline emerged the fields of memory and intelligence grew considerably and while intelligence theories are much more sophisticated now, the role of memory has not been diminished. The most recent versions of the most widely used intelligence tests at least involve one task assessing working memory. The Stanford-Binet fifth edition (Roid, 2003) computes a working memory factor as part of its overall assessment. The intelligence tests initially published by David Wechsler are very widely used today. The latest version of the Wechsler Adult Intelligence Scale (Wechsler, 2008) computes a working memory index, as does the Wechsler Intelligence Scale for Children (Wechsler, 2003). For a much more detailed review of the history of intelligence testing see Boake (2002).

1.2.2 Memory measurement in the wider context

The history of intelligence testing shows that the role of memory was always assumed to be a pivotal element of any theory of cognition/human faculty. As intelligence testing grew so did the amount of work being conducted in order to understand the human memory system and as ideas relating to our memory system evolved, so did the procedures used to assess memory performance.

Many of the current paradigms for testing memory relate to three assessment procedures developed before the turn of the 20$^{\text{th}}$ century. The building block of Ebbinghaus' study into human memory was rote learning of lists of nonsense syllables (Ebbinghaus, 1913). Ebbinghaus standardised the presentation process using a metronome to read a randomly selected list of nonsense syllables which he would then attempt to recall. Calkins (1894) introduced the paired associates learning task where the test taker is required to learn stimuli pairs and when given a stimulus (cue) they should respond with the associated item. And finally, Jacobs (1887) reported a 'memory span' test. In this test participants were given a number of stimulus items at a rate of one every half a second. Upon being given the final item in the list the participant was required to recite the items in the order in which they had been given. Jacobs started with lists of very few items giving participants two different lists (trials) at each set size. The number of items in the lists was increased every two trials. The task was scored such that the participant was given a 'span' score equal to the largest list successfully recalled. Jacobs tested participants using three different types of stimuli; nonsense syllables, numerals (digit span), and letters. Therefore the numeral version of Jacobs memory span task is the concept behind the digit span task used by Binet-Simon (although they only used lists of size 3) and in the army alpha test, as well as being a core task in the most recent versions of the WISC and WAIS.

They provided a framework to test memory processes through subtle manipulations of the various task elements. Ebbinghaus and Jacobs both commented on the heterogeneous results obtained when using different forms for the stimuli (Jacobs found an average span of 7.4 for numerals and 6.5 for letters on a sample of 41 girls aged 12). Ebbinghaus noted other variables which affected success on his task including the length of the to-be-remembered (abbreviated to TBR henceforth) list, the speed of presentation of the TBR list, the frequency of re-learning, as well as noting that even with nonsense syllables as the TBR items there were items that were more easily recalled than others (Witmer, 1935 would replace nonsense syllables with consonant trigrams e.g. 'CTN'). Stimulus properties within-group having significant effects on participants ability to successfully recall TBR item emerged such as words/letters with phonological similarity (Conrad, 1964; Conrad & Hull, 1964).

Specific task components began to be analysed; the encoding of information into memory, the storage of said information, and retrieval of the information upon request. Results from paired-associate type tasks with manipulations of the encoding (learning) phase (e.g. incidental learning procedures (Hyde & Jenkins, 1969; Johnston & Jenkins, 1971) led to Tulving and Thomson (1973) proposing the encoding specificity principle whereby the 'state' at encoding is integral to whether an item can be recalled due to the memory trace it creates.

Immediate free recall (IFR) tasks follow the memory span paradigm but typically involve longer list lengths and remove the need for the participant to maintain the serial order of the presented memoranda. Studies using an IFR paradigm were able to expand upon the existing research to provide a richer view of short-term memory. For example, Mäntylä (1986); Mäntylä and Nilsson (1988) showed the dramatic effects of varying richness of cue in an IFR paradigm. Serial position curves which show the robust effect of recency on responses dominates explanations of IFR performance (e.g. Murdock Jr, 1962; Murdock, 1965; Postman & Phillips, 1965). The differing dominance of recency and primacy in IFR and memory span tasks leads to the paradigms being seen as measuring different constructs, or simply not being used together due to the difficulty to explain performance on both with one explanation (but see Ward, Tan, & Grenfell-Essam, 2010, for a discussion of similar properties between the two).

Investigation of the forgetting process led to further manipulations. In supporting a decay hypothesis of forgetting over an interference hypotheses of forgetting, Brown (1958) adapted the memory span procedure so that in addition to reading aloud the TBR letters in the trial (which ranged from 1-4) the participant was given another 5 letters to read aloud but these were not required for recall. Peterson and Peterson (1959) asked participants to count backwards in threes/fours for various durations after they had been presented the TBR array (spelled out consonant trigram). The introduction of 'distractor' elements to the memory tasks extended to processing throughout the presentation phase (e.g. Murdock, 1965) and processing between TBR items (e.g. Tzeng, 1973).

Craik and Watkins (1973) designed a clever method of delivering the words that were to form the TBR array for a free recall trial. To systematically manipulate the amount of rehearsal an item received they asked participants to listen to a list of words and keep in mind the most recent word that started with a particular letter. For example, if 'C' is the letter to watch for and the list of words was "broom, crow, teach, ball, crust, tent, draw" the participant would need to write down the word "crust". A series of lists (27, each with 21 words) like this would provide a TBR array that the participant was then surprisingly asked to recall. This allowed Craik and Watkins to assign a value to each word in the TBR array that was equal to the amount of non-critical list items that followed the target before the end of the list. This value would be a good indicator of the amount of subvocal rehearsal a particular item received during the encoding phase. This clever design is the precursor to the 'updating' category of modern working memory tasks.

1.3 From tools to theory

1.3.1 Models of short-term memory

The preceding section focused on the historical development of tasks to emphasise the historical significance of such methods, and the striking continuity in task implementation. For a more in depth review of many of the studies in this era of memory research see Bower (2000).

Theoretical models ought to be able to account for key memory phenomena, therefore any model, should explain, for example; the role of proactive/retroactive interference (e.g. Keppel & Underwood, 1962; Postman, 1971), the role of rehearsal (e.g. Brown, 1958; Peterson & Peterson, 1959; Craik & Watkins, 1973), chunking (Miller, 1956; Simon, 1974), the variation in span due to type of stimuli and modality (e.g. Hyde & Jenkins, 1969; Conrad & Hull, 1964). There were a number of models put forward such as the 'mechanical model for human attention and immediate memory' (Broadbent, 1957), the levels of processing framework (Craik & Lockhart, 1972), and the very detailed and influential multi store model (Atkinson & Shiffrin, 1968). However, it is the model proposed by Alan Baddeley and Graham Hitch that seemed to provide the biggest spark in the field of short-term memory (Baddeley & Hitch, 1974; Baddeley, 1986, 2000).

Baddeley and Hitch (1974) combined results from 10 of their own experiments with results from the literature to begin formulating their multi-component model ('working memory-LTS system' originally). The Baddeley and Hitch model postulates a system that at its core is a mental "work space" of limited capacity. The resources available to this workspace can be divided between storage of information and processing demands. A phonemic rehearsal buffer is available to the system that is itself subject to a capacity limit (later termed the phonological loop). Initially Baddeley and Hitch defined the phonological store in some detail while also suggesting there should be an analogous buffer for visual information citing Brooks (1968). Brooks had shown that encoded spatial information was more easily disrupted by subsequent spatial information than verbal information. The visual buffer was later detailed and termed the visuospatial sketchpad (Baddeley & Lieberman, 1980; Baddeley, 1986). Despite the relative success of their specific model, perhaps the greatest contribution is the outline of short-term memory as a working memory concept that is now such a strong theoretical construct that the idea of a working memory system appears to be more commonly examined than a short-term memory system. The hypothesised WM construct has become an essential component in explaining human cognition - ”At present, working memory capacity is the best predictor for intelligence that has yet been derived from theories and research on human cognition” Süß, Oberauer, Wittmann, Wilhelm, and Schulze (2002, p. 284).

The impact and importance of the concept of working memory has been immense in terms of generating empirical work by providing a framework for hypothesis testing. But beyond that, the importance of WM becomes apparent when one assesses the array of higher order cognitive functions that have been shown to be dependent upon a WM system. Working memory ability has been shown to reliably correlate with other cognitive abilities such as; fluid intelligence (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002), Arithmetic (McLean & Hitch, 1999), ability to prevent mind wandering during tasks requiring focus (Kane, Conway, Miura, & Colflesh, 2007), attention (Kane & Engle, 2003), general learning disabilities (Alloway, 2009), and many more. In addition to its prominence within cognitive research there are a wide variety of other disciplines incorporating WM ability in to their research programs and assessing the impact of this cognitive system on their respective fields. Some examples of topics that have seen measures of WM used as a predictor include depression (Arnett et al., 1999), learning computer languages (Shute, 1991), life event stress (Klein & Boals, 2001), regulating emotion (Kleider, Parrott, & King, 2010), and multitasking (Bühner, König, Pick, & Krumm, 2006; Hambrick, Oswald, Darowski, Rench, & Brou, 2010).

Given the widespread acknowledgement of working memory as useful theoretical construct and the range of other behaviours/abilities that appear to be in some part associated with it, the task of understanding exactly what processes underlie the construct takes on even greater importance. But also, if working memory predicts so many other important abilities such as fluid intelligence, then what would be the impact on wider cognitive abilities is working memory capacity can be raised?

1.3.2 Influence of the WM concept on memory assessment

Baddeley and Hitch (1974) report experiments using a variety of tasks where the theme is such that there is storage of items combined with a task that includes a processing demand (Baddeley's reasoning task, comprehension tasks, articulatory suppression). For example, in experiment 3 participants were required to complete a reasoning task (based on Baddeley, 1968). In addition to a control condition there were three experimental conditions with varying degrees of an articulatory suppression task to complete concurrently. The first experimental condition involved articulating the word 'the' repeatedly, the second condition required participants to count from one to six repeatedly, while the final condition gave the participant a random 6-digit sequence at the start of each question on the reasoning task to recite. The response times on the reasoning task showed that from the control condition through the increasingly demanding articulatory suppression techniques the time taken to respond to the reasoning task trials increased. The random 6-digit condition showed significantly increased RT compared to the other conditions. Baddeley and Hitch summarised the result as such "the trade-off between reasoning speed and additional storage load suggests that the interference occurs within a limited capacity 'work space' which can be flexibly allocated either to storage or to processing". I see that link or 'trade-off' between the resources allocated to maintenance of the stored items and the resources allocated to contending with additional demands as being the essence of the working memory concept.

A group of tasks that are designed to capture the conceptual requirements of simultaneous processing and memory operations thought to be inherent to working memory functioning are now known as complex span tasks. Daneman and Carpenter (1980) developed the reading span task which elegantly combines a traditional memory span task for word lists with concurrent processing that will compete for the limited resources available. In the reading span task participants are given a series of sentences to read aloud. The last word in each sentence needs to be remembered for recall at the end of the trial. In Daneman and Carpenter's procedure a participant would be given three trials at each set size (number of sentences and hence number of TBR words) from two through six. If a participant was unable to respond correctly to all three trials at a given set size then the test was terminated at that point. The highest set size where a participant was able to correctly recall the TBR array on at least two of the three trials was taken as a measure of that persons reading span. In addition to administering their reading span task, Daneman and Carpenter also tested their participants using a traditional memory span task for words, a reading comprehension test that provided measures based on comprehension of facts and also pronoun references, and obtained their verbal SAT scores. They found that the reading span score correlated with all three measures of reading comprehension (.72, .90, and .59 for the fact questions, pronoun reference questions, and verbal SAT respectively). In contrast the word span scores provided modest correlations with the reading comprehension measures (all $p < .05$). The additional demands of the sentence processing and the inability to engage in rehearsal in the reading span task produce a span measure that relates to higher order cognitive functions to a striking degree.

Complex span tasks follow the paradigm of storage demands combined with additional processing requirements. The form of the to-be-remembered (TBR) items and the processing task can take various forms. Turner and Engle (1989) introduced the operation span task with two versions that differed in the target stimuli. The processing part of the operation span task involves presenting a mathematical operation (e.g. '$(6/2) + 2 = 5$') to which the participant must assess whether or not the printed answer is correct. In the 'Operations Word' version each operation was followed by the presentation of a word and these words made up the TBR array. Another version used in their experiment was 'Operations Digit' where the participant was required to recall the numbers that were given as answers to the operations (regardless of whether the operation was true or false). Other variants of a verbal complex span task exist such as the counting span task (Case, Kurland, & Goldberg, 1982)which was designed to be appropriate for a wide developmental population. Alongside verbal complex span tasks, a number of visuospatial complex span tasks have developed over time. For example, Shah and Miyake (1996) introduced the 'rotation span' task. This combined a processing phase which involved mentally rotating letters and judging whether or not they were regular or mirror images with a storage phase that presented arrows in varying orientations and lengths. The symmetry span task (Kane et al., 2004) uses grid locations in a 5x5 matrix as the storage units while the processing phase requires judgements on the symmetry of a pattern filled in an 8x8 matrix.

1.4 Cognitive Training

1.4.1 Early attempts to improve cognitive ability

Jacobs (1887) noted that the simple memory span measure he obtained from his participants appeared to correlate with the 'forms' the students were in (i.e. the scholastic measures used to rank the children and placed them in groups based on ability). The top 10 (scholastically) boys aged 12 had an average memory span of 9.1, the next 10 scored 8.3, while the third 10 scored 7.9. Alfred Binet introduced a series of exercises that he termed 'mental orthopaedics' that he derived to help those children who scored low on intelligence measures as he was an early proponent of the idea that we are not given an intelligence level based on some pre-determined biological factor but that through practice our faculties could be improved.

Given the acknowledgement of the limited capacity nature of STM and the relationship of this capacity to higher order cognition it follows that researchers would be interested in whether this capacity can be increased. Early attempts to assess the fluidity of memory limits focused on what strategies might affect the capacity limits that were obtained i.e. mnemonic strategies. Processes such as rehearsal, organisation of TBR material, attributing/acknowledging meaning of items, associations between items were are examples of the types of strategies that were studied (e.g. Flavell, Friedrichs, & Hoyt, 1970; Bower, 1970; Brown, Campione, Bray, & Wilcox, 1973; Butterfield, Wambold, & Belmont, 1973). These studies were particularly focused on participants who had some form of intellectual development disorder (IDD henceforth).

For example Brown and Barclay (1976) split their sample of children with an IDD into three training groups; label, anticipation, and rehearsal. Participants were given two training sessions over two days. The task used was a 'recall-readiness' task (Flavell et al., 1970) whereby a sequence of images that represented an object were presented to participants but behind an occluder. Participants could remove the occluder from any image to view it as many times as they liked and for as long as they liked providing that only one image was viewed at a time. After a pre-training phase where baselines were measured the training began. In the label training the participants were instructed to reveal the items in the correct serial order and apply a label to each item in turn, and to repeat these four times (compulsory, they could do more if they wished). The anticipation and rehearsal groups were also asked to reveal the trial items in serial position and to do this four times (again, a minimum of four time). The first run through was identical whereby participants were told to apply labels to the items. For the remaining three sets of exposures the anticipation group were instructed to attempt to recall what was behind each occluder before exposing it to verify or update their belief. The rehearsal group were trained to employ a strategy of rehearsing the items in groups of three. The authors gave a test version of the recall-readiness task immediately post-training, one day after, and two weeks after. They found that the label training group did not show any significant gains at any time point while the rehearsal and anticipation groups showed a significant improvement in performance immediately post-training but this benefit appeared to decrease over the subsequent testing sessions. Additionally they had an older and younger group (means of 144 months and 116 months) and the regression back to baseline performance was not as pronounced for the older group.

A limitation to these types of strategies is that they tend to be relatively specific to certain types of information in certain learning/recall contexts. While the processes that affect performance on a specific task can be very interesting in terms of understanding the sub-processes required to perform said task, it is of little concern if the goal is to produce a broad reaching cognitive improvement.

There were much more involved intervention programs administered and assessed for their effect on intellectual development. One significant example is the Carolina Abecedarian project. The Abecedarian project was a vast intervention experiment involving the identification of 'at risk' children from birth. At risk in this instance means that the family met a number of criteria relating to socio-economic and psychometric properties. Children from low-income families are more likely to struggle scholastically (Jensen, 1969). The intervention program that the selected children were involved in included a day care program that ran each weekday for 50 weeks of the year. Children began attending the day care program between 6 and 12 weeks of age. Children were kept in the 'infant' program until they were signed off by the teachers to move to the 'toddler' program (13-15 months typically) where they remained until 3 years of age. The scope of the curriculum employed in the project is too vast to describe in full detail, but generally the activities that the children participated in emphasised social, emotional, and cognitive areas of development (such as language). Participants were given individualised programs of activities based upon teacher/researcher judgements. The early intervention continued until the children were 5 years old. The control group in this project were given all the nutritional and medical support that the experimental group received but did not attend the day care program.

The results of such an ambitious project are spread over a significant number of research papers so I will highlight just two. Ramey and Haskins (1981) is one of the earlier outputs from the project which outlines results from the age-appropriate standardised IQ measures through the first three years of the program. The results showed that the experimental group continued on a typical developmental trajectory in performance while the control group seemed to show decline between certain time points (specifically between 12-18 months). Muennig et al. (2011) is a recent publication from the project which focuses on adult health and behaviour issues in a follow-up study for those who participated in the project as infants. The results showed that at age 21 the participants who had received the early intervention showed improved health and participated in healthier behaviours.

The overall results of the Abecedarian project seemed to suggest that the children who were given the early intervention showed increased IQ scores although the benefit was smaller as age increased (participants were also tested at age 15 and 21), had higher scholastic attainment scores, and completed more years of education. It is important to note that the size of the IQ difference was very modest at follow-up time points *(4 points)*. Arthur Jensen (1998) in his book The G Factor concluded after a review of attempts to improve IQ including the Abecedarian project that:

”Anything less than very early and intensive intervention, including medical and nutritional advantages, during the preschool years (and also prenatally), is probably inadequate to cause a lasting increase in the child’s level of g” - page 344.

1.4.2 WM as a trainable construct?

While working memory as a concept has been embraced in cognition research, the specific theories regarding the workings of such a system are contested. A unifying element however comes in the form of the notion of a limit (often a ’capacity limit’), whether this be limits to the number of representations that can be held in the buffers of the multi component model (Baddeley & Hitch, 1974; Baddeley, 2000) or the number of activated memories in the focus of attention (Cowan, 1995, 1999) to name but two examples. The concept of WM appears to provide a framework that underpins such a wide range of cognitive functions that as variation occurs in the limits of WM, so should the variation in the higher order functions to at least some extent. The question then becomes to what extent is this system dynamic; are the underlying processes/functions that combine to make up what we call working memory fixed based on biological factors or can they be strengthened, made faster, more efficient, quite simply: improved?

Introduction of adaptive-difficulty WM training interventions

Torkel Klingberg began a research program focused specifically on trying to 'train' ADHD participants on working memory based tasks. There is a need to define the notion of training here and the way in which it differs from the studies that preceded it (see Abikoff, 1991, for a review of ADHD intervention studies up to that time point). The cognitive training studies that Abikoff (1991) discusses are paradigms whereby participants are taught a method/strategy/technique in order to improve their level of performance on cognitive measures, which then may or may not transfer to other situations. In these studies the focus is on the particular method/strategy/technique, the participants are trained to use that. In contrast, the focus starting with the work of Klingberg and colleagues in their 'working memory training' studies involves a 'reinforcement by repetition' paradigm.

A useful analogy here can come from the process of training for a physical competition such as running. A coach can advise a runner on different techniques that may facilitate improvements in performance such as body position, stride distance, pacing, and equipment to use. This involves trying to make the most of the 'athletic engine' someone has. A key part of training may also involve specifying the aerobic and anaerobic conditioning the athlete undertakes, developing muscular power and endurance for example. This attempts to improve the athletes 'athletic engine'. This type of training will improve performance through repetition.

Klingberg, Forssberg, and Westerberg (2002) conducted a study aiming to assess the potential impact of a behavioural WM based training intervention for children diagnosed with ADHD. The sample comprised 7-15 year-olds and were split into a training intervention group and a control group. The experiment was pre-post design with a battery of tasks administered prior to a training phase and then being re-administered after a training phase. The evaluation battery consisted of five tasks; matrix span, span board, Stroop, Raven's Coloured Progressive Matrices (RCPM), and a choice reaction time task.

The matrix span task was a computerised task presenting participants with a 4x4 grid where a sequence of the segments are highlighted that the participant must recall in order at the end of the presentation phase. The span board task as described is the same as the Corsi-Block tapping task (Corsi, 1972) but with 10 blocks rather than 9. The wooden blocks are arranged in front of the participants and the experimenter taps one block at a time at a rate of 1 per second. The participant needs to recall the sequence of taps in the correct order. The Stroop task used was the standard form of the test (Stroop, 1935) whereby words are presented in varying hues and participants are required to respond to the colour of the text ignoring the actual words. The RCPM task is designed to measure reasoning ability while the choice reaction task was a simple test of reaction time in various conditions (with/without warning and with/without choices).

The training regime devised by Klingberg and colleagues involved four tasks that were all included in each training session. The first was a matrix span task, second was backwards digit span, thirdly the letter span task, and finally a choice reaction time task with an inhibition element (asked not to respond when the cue was a certain colour). A training session consisted of 30 trials of each of the four tasks where difficulty was adjusted based on performance. Participants were asked to train daily (weekdays) over the 5-6 week training period which resulted in a mean of 24.3 training sessions completed.

The control group in this experiment were asked to complete a training regimen that was a 'low-dose/placebo' regimen. In their version of the training tasks they only had to complete 10 trials per task and the number of TBR stimuli was set at 2 for all matrix span and backwards digit span trials, and 3 for letter-span trials. Klingberg et al. (2002) suggested that the group who received the adaptive 'high-dose' training group improved (over and above the control group) on the matrix span task, span board task, Stroop accuracy, and RCPM. The effects of the training on the post-training measure of matrix span and span board are likely to reflect task-specific improvements given the matrix span task was a trained task, and span board is essentially a non-computerised version of the same task. However, the additional benefit of the adaptive training on the Stroop measure and RCPM suggested a potential generalised benefit that had not been observed to that point. There was also a second experiment noted in the 2002 paper with just 4 healthy adults forming the participant pool. The sample size in the second experiment clearly prohibits much generalisation to be made from the results obtained but it is worth mentioning that the adults showed significant improvements post-training on the same measures as the ADHD children in experiment one (Raven's Advanced Progressive Matrices (RAPM) was used instead of RCPM but otherwise the same).

This work was followed up with an fMRI study (Olesen, Westerberg, & Klingberg, 2004) which due to the neuroimaging aspect (discussed below) consisted of a small sample (three healthy adult participants in the experimental group). The participants in experiment one followed the same training regimen as outlined in Klingberg et al. (2002) but without Stroop and choice reaction time tasks being in the training regimen. The behavioural data was compared to a control group (passive; n = 11) and the training appeared to generalise again to span board, Stroop, and RAPM. Klingberg et al. (2005) conducted a more direct follow up experiment. They tested 44 children diagnosed with ADHD (age range 7-12 years; n = 20 in treatment group) using a similar methodology to the 2002 work. The training regimen was now called Robomemo and unfortunately was not documented to the same level of detail. The participants completed 90 trials per training session. The training software is described as including "visuospatial WM tasks (remembering the positions of objects in a 4 x 4 grid as well as verbal tasks (remembering phonemes, letters, or digits)". This likely means a matrix span task combined with a simple memory span task where the stimuli changed occasionally between digits, letters and phonemes. It would seem plausible to presume that the split between spatial and verbal based training tasks was 50/50. The control group in this experiment was also an active control group but the difficulty of the trials remained static at a low level (2-3 items TBR). The control group however were not a 'low-dose' comparison as they completed 90 trials at each training session also. An additional facet to this study is that the evaluation battery was administered at a 3 month follow up time point in addition to the standard pre/post-training. The results showed the same pattern at post-training as Klingberg et al. (2002) suggesting the adaptive training group showed significantly greater gains than the non-adaptive control group on span board, digit span, Stroop, and RCPM. The results of the 3-month follow up showed that the effect on span board and digit span remained but the Stroop and RCPM effect was no longer statistically significant (p-values of .07 and .12 respectively).

These results served as a catalyst for research in to WM training methods. After they believed they had demonstrated that CogMed produced generalised benefits to WM (Klingberg et al., 2002), a neuroimaging approach followed. Olesen et al. (2004) conducted two experiments using functional magnetic resonance imaging to collect brain scan data on participants before, during, and after the 5-week training period. In experiment one, three participants were scanned twice prior to the onset of training and once following the intervention. In this experiment the intervention consisted of a 'visuospatial WM task', backwards digit span, and letter span. While being scanned participants completed a visuospatial WM task (matrix span variant with recognition probe responses) and a control task (when 'locations' were highlighted they stayed highlighted and participants simply needed to 'click them away' in random order). with regards to the behavioural results, compared to a passive control group($n = 11$) the three experimental participants showed transfer to Span board, Stroop, and RAPM. The imaging results pointed towards increased activity in the right middle frontal gyrus, right inferior parietal cortex, and bilaterally in the intraparietal cortex when participants engaged in the visuospatial WM (VSWM) task at t_2 compared to t_1. The second experiment had eight participants in the experimental (training/scan) group. This time the intervention involved three visuospatial based tasks from CogMed (Grid, 3D Grid, Grid rotation). The matrix span task completed in the scanner involved recall of the entire sequence in serial order as opposed to the recognition probes. The behavioural data showed transfer to Stroop over and above the control group but not on Digit span or Span board. The imaging results showed an increase in activity in parietal and prefrontal cortices as well as in the thalamic and caudate nuclei. There were some specific differences in the imaging results between the studies, experiment one showed more prefrontal activity increase in the right hemisphere while experiment two showed the increased activation in the left hemisphere. This could be explained by the differing interventions and/or the different demands of the matrix span variant during scanning between the two experiments.

This was direct evidence that a physiological change was brought about due to the short-term intensive WM training. This became the rationale for why the behavioural intervention was likely to be successful in subsequent publications from this research group (e.g. Klingberg et al., 2005; Westerberg et al., 2007). Furthermore, other studies have been published which add to the story regarding physiological consequences of adaptive WM training such as; Rueda, Rothbart, McCandliss, Saccomanno, and Posner (2005) who used electrophysiological data to suggest that training intervention produced similar patterns of change to what natural development would show, Langer, von Bastian, Wirz, Oberauer, and Jäncke (2013) who concluded ”working memory training shifted brain network characteristics in the direction of high performers”, and Dahlin, Neely, Larsson, Bäckman, and Nyberg (2008) who found that transfer effects after an updating based intervention were mediated by overlapping brain activation (striatal regions).

1.4.3 Further successes and range of application of WMT

The evidence for training and transfer in the work of Klingberg and colleagues led to an increase in the amount of work focused on the possibilities of working memory training (WMT). Some striking results came from the work of Susanne Jaeggi and colleagues (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008). They found evidence to suggest that a working memory training regimen consisting of only one task was able to increase fluid intelligence (Gf) in a sample of healthy adults (mean age = 25.6 years). The training task used by Jaeggi et al. (2008) was a very demanding dual n-back task. The n-back task was first introduced by Kirchner (1958) and can be categorised as a continuous performance task (CPT). Participants are required to attend to a stimuli stream and respond when the current stimulus matches the stimulus N items ago. Jaeggi et al. (2003) introduced a dual-task paradigm version of the n-back whereby they presented two streams of stimuli simultaneously and success required the participants to attend to each stream and respond appropriately when a match to the item N presentations ago appeared. The version used to train participants WM in the 2008 paper presented a visuospatial stream in the form of a sequence of locations (8 possible locations surrounding a central marker). In addition, a verbal stimuli stream was presented in the form of a sequence of consonants that were presented auditorily. Jaeggi and colleagues asked their participants to complete either 8, 12, 17, or 19 days of training on the dual n-back. They were able to show an overall effect of training on Gf scores (via shortened versions of RAPM and BOMAT) based on a comparison to a control group. In addition, the size of the benefit appeared proportional to the amount of training days completed. Further success of a CPT based training task was conducted by Zhao, Wang, Liu, and Zhou (2011) who found that in a sample of typically developing children aged 9-11 the training group showed significant gains on Raven's Standard Progressive Matrices (RSPM) post-training compared to a control group.

Other successful studies emerged using the adaptive training program developed by Klingberg. Holmes, Gathercole, and Dunning (2009) showed the adaptive training led to enhanced WM performance as measured by untrained tasks (at post-training as compared to a control group). Interestingly the evaluation battery in this study also included scholastic measures including a mathematical reasoning subtest from the WOND. This measure showed no added benefit from the training at the post-training phase but at a 6-month follow up phase the authors claim a training effect had emerged. A follow up

(Holmes et al., 2010) testing a sample of children with an ADHD diagnosis was reported to show the significant effect of the adaptive training on untrained WM tasks, an effect which was maintained at a 6 month follow up phase. And also Thorell, Lindqvist, Bergman Nutley, Bohlin, and Klingberg (2009) showed in children (typically developing) as young as 4 (age 4-5 sample) the adaptive training led to enhanced WM performance as well as a transfer to improved attentional measures. Incidentally, the transfer didn't work the other way (they also had a group who trained on an attentional based task battery) seeming to offer even more weight to the potential importance of working memory training.

A wide variety of WMT studies have now been conducted that vary in terms of the specifics of the training tasks, the population of interest, the potential transfer (generalisability) effects, and durations of training. Schmiedek, Lövdén, and Lindenberger (2010) is a study that stands on its own with regards to the amount of training the participants completed. In this experiment the participants completed 100 training sessions, the duration of which was one hour per session. Participants formed two groups; young-adults aged 20-31, and older-adults aged 65-80. The training task battery consisted of multiple tasks representing the following constructs; WM, episodic memory, and perceptual speed. Participants would complete every task at each training session. Transfer was assessed using multiple tasks per construct. The authors differentiate between a 'WM Near' and 'WM Far' factor. The near factor was comprised of animal span, 3-back numerical, and a spatial memory updating task. The WM-Far factor included reading span, counting span, and rotation span. In addition, a reasoning and episodic memory factor was also constructed from constituent tasks for analysis of transfer. For both the younger and older age groups significant transfer was observed for the WM-Near factor but not for WM-Far. Even though only three out of the 12 tasks that made up the training battery were designated as WM tasks, the total training time on these WM tasks would exceed the majority of published studies due to the number of training sessions completed by participants. The effect of that WM training as well as training on other facets of cognition was able to produce transfer to tasks that were very similar to the trained tasks but unable to show transfer to measures of the same construct that were substantially different in paradigm. Additionally, the young adult group showed significant transfer to episodic memory and reasoning but the older group did not (although individual task analysis showed transfer to Raven's).

Further variation in the types of WM training studied is highlighted by the work of Dahlin, Nyberg, Bäckman, and Neely (2008) who adopted a running memory training task paradigm involving the ongoing maintenance of the last 4 stimuli in a stream. They found far-transfer to episodic memory in a young-adult group but failed to show near-transfer to digit span and computation span tasks, they did however show near-transfer in a 3-back numerical task. This is an example of the effects of training on WM failing to be robust. The near-transfer was only seen on the 'nearest' transfer measure and not on the STM/WM measures that had a differing paradigm. Incidentally, the older-adult group failed to show any transfer in this study. Complex span tasks have also been utilised as training tasks by Chein and Morrison (2010) . This is a rare training regime which may be quite surprising given that complex span tasks make such a large part of the WM assessment 'toolbox' and are generally seen as the go-to tasks for measurements of WM ability. The participants trained on reading span (verbal WM) and symmetry span (spatial WM) tasks. They showed transfer to a temporary memory composite with a very large effect size (D = 1.42). However, the temporary memory composite was made up of scores on identical tasks to the training tasks in addition to simple span versions of said tasks. They also report far transfer to Stroop and reading comprehension (Nelson-Denny reading test) but fail to find far transfer for ETS reasoning battery or RAPM. Additionally, the study only included a passive control group.

1.4.4 Commercialisation of WMT

Working memory training is increasingly being deployed as part of the recent boom in commercial 'brain training' products. These commercial products have a range of guises such as online subscription models, tablet/mobile apps, and specialist computerised software. For example the training program initially developed by Klingberg and colleagues in the research studies above, was commercialised, launching in Sweden in 2003. This occurred even though the materials are essentially based on existing psychological tasks. Subsequently, CogMed was acquired by Pearson (to be incorporated into its clinical assessment portfolio) from Karolinska Development AB, the commercial arm of the research institute and "The price was less than $100 million". The following extract illustrates the extent of costs facing the end user:

"The U.S. is CogMed's fastest-growing market, according to SharpBrains. Pearson sells licenses for the software to between 500 and 1,000 psychologists in the U.S., for prices of about $100-$200 per patient, practitioners said. They said they charge patients between $650 and $2,250 for CogMed training, including testing and supervision services".

The importance of product commercialisation is that it generates a potential conflict of interest for researchers. Those working for the company are, other things being equal, invested in a particular research outcome (positive far-transfer effects) as well as in protecting the status of the findings that support the product. In addition, those supported by the product, who are contracted to carry out evaluative research or aided by discounted fees etc. may not remain unbiased in their attitudes towards training effectiveness.

It is important to note that CogMed it not the only commercial product in this field. CogMed is of particular interest in the scientific literature due its foundation in Klingberg and colleagues early research and the number of published studies that cite CogMed as the training tool. In addition to the huge selection of 'brain training' apps in the Apple app store and Android's Play store there are other products similar to CogMed that are aimed at a less general audience. One such product that deviates from the prototype is

Meemo which is built on the premise of WMT but is not computerised and is marketed as a 'whole class working memory programme'.

1.5 Is WMT effective?

The research discussed in the previous sections suggests that working memory training is a promising behavioural intervention particularly perhaps for those with IDDs where low working memory ability is a significant symptom, but also for the typically developing child and healthy adults. It seems plausible that for children with reduced working memory, there is a potential-to-achievement gap that training might help bridge. Nonetheless, while the results of this research are promising there is a need to interpret the results of these studies cautiously.

1.5.1 Classification of types of effects

The effects that are of interest in studies involving WMT interventions can be organised in to three distinct categories. For the purposes of intervention studies it is important to differentiate between practice effects, near-transfer effects, and far-transfer effects. These terms have been mentioned already when discussing the literature but at this point it is important to discuss and analyse these in more detail.

Practice Effects When a task is repeated multiple times there is a certain level of improvement that would be expected based on the fact the task is being repeated. The size of this improvement will differ between tasks and also between individuals. These effects include classic test-retest issues i.e. a participant might be better on the second attempt at a task due to increased familiarity of the demands of the task, or perhaps a lower level of anxiety regarding the task due to previous exposure. A practice effect can be classified as an effect that improves subsequent scores on the same task but remains task-specific. For example, take the classic reading span task, Towse, Cowan, Horton, and Whytock (2008) provide data for 9 and 11 year old groups completing the reading span task. The authors further split each age group into a descending/ascending trial sequence condition giving four groups. Three of these four groups showed overall reading span improvement from t_1 to t_2 where the temporal gap was no more than 10 weeks. These test-retest practice effects may also be indicative of maturation effects when a developmental population is the focus of a study (e.g. Hitch, Towse, & Hutton, 2001).

Ericcson, Chase, and Faloon (1980) describe the performance of a participant (S.F.) over a long period of practising the standard digit span task. S.F. practised the digit span in the laboratory for one hour per session, between 3-5 times per week, for one and a half years. This resulted in 230 hours of lab practice. The results show that over this period of time the memory span of S.F. steadily improved from seven initially to 80 at the end of the study. However in their concluding paragraph the authors note ”These data suggest that the reliable working capacity of short-term memory is about three or four units, as Broadbent has recently argued, and that it is not possible to increase the capacity of short-term memory with extended practice” (p. 1182). The reason for this is that they were able to show that all the improvements that S.F. made on the digit span task were a result of development of mnemonic strategies and then the refinement of said strategies. Through verbal reports by S.F. and in analysis of recall timings it was clear that strategy use was the driving force of improvement as opposed to an improvement in the actual capacity of short-term working memory. The clearest evidence that S.F. was entirely dependent upon the strategies developed for the impressive performance on digit span was evident when the stimuli were altered. After three months of training on digit span, S.F. was given a letter span task. Based on the progression plot shown S.F. was recalling approximately 20 digits successfully at this time, but when letter stimuli were used instead, performance dropped down to baseline levels (6 letters). This is an example of a lack of ’transfer’ and these type of effects will be discussed next.

Transfer Effect If a training procedure is to be practically meaningful, it needs to show a generalised effect to some degree. To measure whether or not an intervention has had effects that generalise beyond the specifics of the trained task the researcher must look for transfer effects. A transfer effect is observed when a task that is different from the training task shows improvement as a result of the training. Clearly, one has to be cautious with regards to explicitly claiming that an improvement is a direct consequence of the training they received, but as the methodological rigour of a study is improved so one might have greater confidence that the training may be the cause of the observed effect. Transfer effects come in two varieties; near-transfer and far-transfer. A near-transfer effect is when training on a task that primarily taxes construct X produces a significant improvement in a different task that is also supposed to primarily tax construct X. A far-transfer effect requires training on a task that primarily taxes construct X to produce a significant improvement in a task that primarily taxes construct Y, where construct Y is known to depend on or be associated with construct X.

An example of a near-transfer effect is training on a matrix span task producing a significant improvement in rotation span (see Kane et al., 2004) scores. These two tasks are highly related as they both depend heavily on short-term retention of visuospatial material. However, it is unlikely that a strategy that was developed for aiding performance in the matrix span task would also be beneficial when completing the rotation span task due to the difference in properties of each stimulus. An example of far-transfer would be training on a digit span task produces a significant improvement in a Raven's Matrices task. Raven's Matrices tasks are well established as measures of nonverbal reasoning and working memory processes and executive functions have been shown to be predictive of performance on Raven's tasks (e.g. Carpenter, Just, & Shell, 1990).

The application of WMT research outside of academia has, as already noted, produced a large commercial market for 'brain-training' products/services that are based on improving working memory. The success of these programs should be assessed on the weight of evidence for the existence of genuine transfer effects in WMT research. Practice effects may be interesting conceptually and enrich our understanding of the mechanisms that support encoding, maintenance, and recall. Nonetheless, transfer effects are of much more general value in that they represent the practical application (e.g. as a proposed intervention to improve the development of those with IDDs where low working memory is a critical symptom). In addition, they allow theoretical development with respect to the static/dynamic nature of capacity. The publication of studies describing near and far transfer effects are of significance beyond the academic community. Some of the claims of transfer need to be taken cautiously however. For example, a common near-transfer effect reported is from CogMed training to a span board task (e.g. Klingberg et al., 2002, 2005). The visuospatial tasks in CogMed training are matrix span tasks, and the span board task is a Corsi blocks task (Corsi, 1972). Therefore, the training task (matrix span) is a computerised version of the evaluation task (span board). The differences in presentation format (computer screen versus real world 3D blocks) are a visceral difference but are the mental processes required to succeed or strategies that aid performance the same? If the answer to those questions is yes then it is difficult to class this as even a near-transfer effect rather than a practice effect. This conceptual point forms the foundation of the critical issues with the current state of the current body of WMT evidence will now be expanded upon.

1.5.2 Efficacy of WMT - A critical evaluation

Zach Shipstead, Thomas Redick, and Randall Engle have published two 'reviews' (Shipstead, Redick, & Engle, 2010, 2012) of working memory training studies where they discuss key methodological issues in such studies. In their 2010 review there are two general points of emphasis discussed in relation to WMT studies. First, the inclusion of a control group is not necessarily enough. The extent to which the control group is treated as similarly as possible to the experimental group is of paramount importance. The authors cite the known tendency for participant's task performance to change based on knowledge of being observed (e.g. McCarney et al., 2007). And yet the quality of control groups can be criticised in a significant number of WMT studies. Second, task purity is important. Care is needed with respect to the evaluation battery (pre-/post-training tasks). They note that ”the most prominent threat to this type of generalisability is an assumption that the results of single tasks (e.g. Raven’s Progressive Matrices, Stroop task) can be interpreted as pure measures of abstract hypothetical constructs (e.g. Gf, attention)” (p. 253). The vast majority of studies up to that point had only considered single task performance as measures of transfer of training effects. No one task measures a desired construct and nothing else. Even tasks that are known to have high construct validity have unexplained variance components suggesting that there are other factors that are contributing to performance. Therefore, conclusions based on single task transfer should be considered tentatively. Single task measures of transfer can be indicative of a change but leave many questions and uncertainty as to the nature of the change.

Shipstead et al. (2012) built upon the issues they raised in their 2010 review by reinforcing those points while adding two further broad points on WMT studies. Shipstead and colleague’s discuss the ”conflation of working memory with short-term memory” and discuss the additional methodological issues that arise when subjective measures are included in the evaluation battery of WMT studies. The distinction between ”short-term memory” and working memory is important. It is highlighted that the tasks that form the CogMed battery are simple span tasks and therefore could be considered short-term memory tasks rather than working memory tasks. Perhaps more generally the problem is that the cognitive processes that are required for success in the selected tasks are not discussed in detail. And furthermore, the studies tend to operate without a clear theoretical framework of working memory in mind. For example CogMed training tasks are simple span tasks, there are a number of variations but the mechanics of each one is that a series of stimuli is presented at a steady rate and the participant must respond with the items in correct serial order to be successful. These memory tasks could be operationalised as training domain specific storage components of working memory (if subscribed to a model of WM with domain-specificity storage such as the multi-component model of Baddeley, 1986) or a domain-general component of the WM system (such as the focus of attention spotlight Cowan, 1999). Shipstead and colleagues final note regarding the use of subjective measures is of particular relevance to the WMT studies with atypical populations as they often include at least one measure of behaviour but this is an aspect of these studies I have chosen not to discuss in detail.

These reviews put the WMT studies under the spotlight at a time when it was clearly needed. The lack of methodological rigour in the successful studies to that date is rather surprising given that one of the clear problems with the quality of the research was an issue as fundamental as quality control groups in an intervention study. The control groups used can be classified as such; no control, passive control, active control. A passive control group is tested on the evaluation battery at the same times as the experimental group but participates in no other way in the experiment. An active control group is one which is given something to do during the training phase and this can vary in terms of how it compares to the actual intervention. To quantify some of the issues drawn on by Shipstead and colleagues I will simply collate the information that is presented across numerous tables in the (2012) review. With regards to control group usage, of the 37 studies considered (across all populations; children, young-adult, older-adult) the breakdown of type of control group used was; None = 5, Non-Adaptive Span = 7, Contact = 14, Games = 2, Other = 9. The 'other' category includes studies where control group data from previous studies was used, general trivia, and other construct training such as perceptual speed tasks. All but a few working memory training studies use single task measurements as the only tests of transfer. Greater discussion of this point will follow with regards to the 'pattern' of transfer.

A further conclusion drawn from Shipstead et al. (2010) was that "there is little doubt that adaptive span training consistently improves performance on tasks which measure simple retention of short lists" (p. 268). This statement comes from the need for each study to show that WM has been improved as a result of the intervention so that the narrative leading to far-transfer effects is coherent. Showing the WM system has improved in some way (almost always measured by 'span') can be demonstrated by the presence of near-transfer effects whereby the WM training task leads to improvements on a different task paradigm that is known to be a valid WM measure. I would argue that often this demonstration is missing in the WMT literature and is in fact replaced by a demonstration of practice effects.

Meta-Analyses

A further review first published online in 2012 took a meta-analytic approach to reviewing the WMT literature (Melby-Lervåg & Hulme, 2013). In their meta-analysis 23 studies were included resulting in 30 overall training group comparisons and included clinical, typically developing, and adult samples. This paper marked the first review including a meta-analysis where there were a sufficient number of published WMT studies to allow reasonable conclusions to be drawn. Studies were included that met certain criteria, namely that the training group participated in a working memory based intervention, include a control group (no control group studies were excluded but passive control groups were included), include outcome measures including WM tests, reading ability, arithmetic, attention (Stroop), or standardized tests of nonverbal or verbal ability. Moderator variables were included in the analyses to address a number of issues; age (grouped; young children, older children, young adults, older adults), training dose (low/high), control group type (active/passive), learner status (clinical/typical), and intervention type (CogMed, JungleMemory, Cognifit, N-Back, other).

There are a number of important results to take from this meta-analysis regarding both near- and far-transfer. There was a small overall effect of WMT on nonverbal ability ($d = 0.19[0.03, 0.37]$). However, the moderator variables show an important effect of the type of control group used. Studies using untreated control groups show higher effect sizes [0.23,0.56] compared to those using treated control groups [-0.24.0.22]. A similar sized effect emerged for Stroop performance ($d = 0.32[0.11, 0.53]$) but no moderator variables emerged as significant predictors. There were however only 10 effect sizes in this analysis compared with 22 for the nonverbal ability analysis. The outcomes for arithmetic, reading, or verbal ability suggested no overall effect of WMT.

As for near transfer effects, the authors found that transfer to verbal WM (d = 0.79 [0.5,1.09]) and VSWM (d = 0.52 [0.32,0.72]) was significant at post-training with n of 21 and 18 respectively. There are some interesting points to note regarding the moderator variables in these analyses. The only significant moderator variable for verbal WM was age where pairwise comparisons showed that younger children benefited significantly more than older children. For VSWM the only significant moderator variable was intervention type where CogMed produced increased benefits compared to the each alternative. This means that training dose which was coded as low (¡= 8 total hours) or high (¿= 9 hours) did not affect the magnitude of improvement, nor did the 'learner type' where experiments were coded as either sampling from a clinical or observed low WM group or 'unselected'. Also, rather interesting when compared with the far-transfer result is that the difference in type of control group was not significant for these effects.

Therefore, this influential meta-analysis offers support for the notion that WMT has some impact on the WM system post-training. However, there is little support for the generalisation of these effects to other cognitive abilities because it is important to accept the importance of an active (treated) control group in such studies. As Melby-Lervåg and Hulme (2013) allude to in their discussion of the findings: the importance of a treated control group to attempt to account for extraneous factors affecting the outcome is particularly significant as 'Hawthorne' effects can account for up to a 0.3 standard deviation increase (Clark & Sugrue, 1991). Thus, the claims of various commercial outlets that WMT interventions can produce significant generalised cognitive benefits appear unsupported at this stage.

There is however some optimism to be taken from the Melby-Lervåg and Hulme (2013) meta-analysis in that the overall effect sizes for near-transfer effects were significant and of an important magnitude. Interestingly the heterogeneity between the observed effect sizes for the near-transfer measures was significant suggesting a wide range in the observations whereas the far-transfer results had smaller non-significant heterogeneity.

This leads to a critical observation regarding the near-transfer results which warrants further discussion. Given the wide range of effect sizes observed for verbal WM and VSWM and the accompanying range of tasks used as either training or outcome measures,; do these issues mediate the differing effect sizes. Let us assess some of the largest effect sizes included in the near-transfer analyses in terms of the training and outcome measures used in each.

The largest of the verbal WM effect sizes was a Cohen's d of 2.38 extracted from Holmes et al. (2009). The training regime in this study was CogMed which includes a number of verbal and visuospatial tasks but that all are described as variants of traditional matrix/digit/word span type tasks. The single task used to assess verbal WM as an outcome was the counting recall task. In this task participants were required to count a visual array of dots and remember the answer to each array for recall after the final array was presented. Therefore while the training and outcome tasks are clearly not identical there is a high degree of similarity. The memoranda to store and recall are likely to be the same in the outcome task as in some of the variants of verbal training ta sks in CogMed. Another large effect size in the verbal WM analysis was provided by Borella, Carretti, Riboldi, and De Beni (2010) where $d = 2.09$. The verbal WM outcome was backwards digit span and the training tasks were variants on the Categorization Working Memory Task (CWMS; Borella, Carretti, and De Beni (2008)) which all involved the maintenance of presented lists of words for recall either in serial or shuffled order. Here, the similarity of the transfer and training tasks are identical in terms of procedure. The differing stimuli (words/digits) do however offer some control over the utilisation of developed strategies at the training phase to aid performance in the outcome measure. With regards VSWM two of the largest effect sizes were provided based on CogMed training. Bergman Nutley et al. (2011) provided a Cohen's d of 1.55 and the outcome measure was the 'grid task' (matrix span). Klingberg et al. (2002) provided a d of 1.66 based on a span board outcome measure. Both of these outcome measures are near identical to the types of visuospatial training tasks used in the CogMed program. The span board task alters the environment from computerised to real world blocks but is mechanically the same.

This leads to the supposition that when near-transfer is assessed there is a scale of how 'near' a transfer is, and the nearer it is the higher the probability of observing a significant effect. Therefore, it could be argued that Melby-Lervåg and Hulme (2013) adopt a lenient set of criteria for inclusion which inflated the near-transfer effects. As a counter example to highlight that these results were not based solely on 'very-near' transfer effects Thorell et al. (2009) provided Cohen's *d* values of 1.09 and 1.06 for both their training groups transfer to verbal WM. The training groups in each comparison were a purely visuospatial CogMed regime, and an inhibitory control regime. Thus providing an example of large effect sizes for near-transfer effects that cannot be attributed to the degree of similarity between the trained and outcome tasks.

Re-evaluation of WMT success

At this point let's re-evaluate the 'successful' studies discussed previously based on the methodological aspects discussed above and the overall pattern of transfer effects obtained.

With regards to the inclusion of a suitable control group, Klingberg et al. (2002) used an active control group as the members were asked to complete as many 'training' sessions as the experimental group but they completed a non-adaptive version of the training tasks and were only required to complete 1/3 as many trials. This will have resulted in substantially less time-on-task with fewer and quicker trials. Klingberg et al. (2005) addressed this issue by having the control group complete as many trials as the experimental group. Holmes et al. (2009) included an active control group for the pre and post training measures but only tested the experimental group at the 6-month follow-up phase. Their subsequent study (Holmes et al., 2010) had no control group. Thorell et al. (2009) used a passive control group paired to the CogMed training intervention group as well as an active control group paired to their inhibition training intervention group.

The control groups were then combined for analysis due to not showing any differences in performance. Jaeggi et al. (2008) used passive controls for comparison while Zhao et al. (2011) required their control group to play irrelevant video games while the training group engaged with the intervention.

Klingberg et al. (2002) report near-transfer from CogMed training intervention to matrix span and span board (measures of VS-STM) and this provides the evidence for the claim that the intervention has increased WM capacity. The CogMed training regime is described as ”visuospatial WM tasks (remembering the position of objects in a 4 x 4 grid as well as verbal tasks (remembering phonemes, letters, or digits)” (Klingberg et al., 2005, page 79). There is essentially no difference between a number of the training tasks in the intervention and the tasks used to assess near-transfer. Any potential strategy that a participant may have been able to develop to aid performance on the training tasks would also be applicable to the transfer measure thus mediating a difference between baseline and post-training performance. In addition, as control participants completed non-adaptive versions of the training tasks they would never have reached a level of difficulty where there was an incentive to develop strategies. Thus, this type of near-transfer measure is not valid to show a genuine improvement in WM function. Klingberg et al. (2005) report near-transfer to digit span and span board tasks after CogMed intervention. Both of these studies showed far-transfer to attention (Stroop) and nonverbal reasoning (RCPM) but with no valid measures of whether any improvement in core WM function was actually obtained. Therefore the explanation of such transfer becomes based on assumptions that WM was improved or that transfer occurred via some other mechanism of change.

Holmes et al. (2009) were able to demonstrate near-transfer from CogMed training to measures of working memory sufficiently different from the training task (verbal and visuospatial WM latent factors) but failed to find far-transfer to the WASI, reading subtest of the WORD, and reasoning subtest of the WOND. But they were able to show transfer to a more ecological task based on remembering sequences of instructions. Holmes et al. (2010) reported near-transfer to WM based on administering the AWMA pre- and post-training but as noted above there was no control group in this study.

One of the more interesting examples of transfer assessments in the cohort of studies I discussed above as 'success stories' comes from the work of Susanne Jaeggi and colleagues. The training regime employed by these researchers involves n-back variants and they have used complex span measures as indicators of near-transfer to demonstrate WM improvement. In this instance the training and the transfer task are presumed to be heavily reliant on the WM system but clearly operate on very different paradigms. This could be described as the furthest near-transfer assessment. The remembering component of n-back is recognition judgements (does this stimulus match the stimulus *n* items ago?) as opposed to serial recall. Jaeggi et al. (2008) show a striking far-transfer effect of dual n-back training to Gf but fail to show any near-transfer to reading span. Incidentally near-transfer to digit span was observed. Additional, in a subsequent study Jaeggi, Studer-Luethi, et al. (2010) showed no transfer of n-back training to an operation span task. Zhao et al. (2011) also used an updating type task to train WM but included no near-transfer measures at all.

A potential moderating variable in analysis of transfer effects may be the degree to which certain populations are likely to engage in strategies that enhance performance on the assessed WM tasks, and potential variability between training tasks in encouraging participants to develop such strategies. There are age differences in the likelihood to spontaneous engage with strategies to aid performance when completing WM tasks (Flavell et al., 1970) with older children more likely to use mnemonic strategies. Additionally there is variation between adults in engagement of mnemonic devices and this variation is linked to variation in overall performance (Dunlosky & Kane, 2007). Studies involving instructing participants to engage in mnemonic strategies (primarily but not exclusively, rehearsal) have further highlighted the importance of strategy use and the importance of such 'mental algorithms' in determining performance. Turley-Ames and Whitfield (2003) saw that strategy instruction improved span scores while others have more intensively trained strategy usage with positive results (McNamara & Scott, 2001;

St Clair-Thompson, Stevens, Hunt, & Bolder, 2010). Peng and Fuchs (2015) examined verbal WM training with and without strategy instruction and found comparable improvement between groups on untrained WM and reading comprehension measures but very different practice effect profiles for trained tasks. Thus it is evident that the way in which participants approach the training they are given can be influential on performance on the trained task. It is likely that developed strategies throughout training may be deployed in measures of near-transfer if the transfer assessment is a task where the developed strategy can be utilised. In light of these findings and assessment of the wider literature it seems that strategy identification, experimentation, and successful utilisation is a likely moderator of practice and near-transfer effects observed. Dunning and Holmes (2014) through open-ended interviews pre- and post-training were able to show that those who were given adaptive WM training were more likely to engage with chunking/grouping mnemonic strategies. While these results suggest a benefit to engaging in adaptive WM training it is qualitatively different from the often suggested generalised neural network benefits. Additionally, strategies can be task- specific and therefore rarely lead to generalised improvements, therefore if engaging useful strategies is the actual consequence of undergoing an adaptive WM training regime then it may be more efficient for all groups to replace that extensive intervention period with strategy-based tuition/intervention (Carretti, Borella, & De Beni, 2007; St Clair-Thompson et al., 2010).

1.6 Does adaptive-difficulty working memory training actually improve working memory ability?

It is understandable why there is optimism amongst researchers to study behavioural interventions that could lead to widespread cognitive enhancements. The rise of the working memory concept as a cornerstone of cognition seems to be a very viable candidate for a targeted training program. As noted previously there are clear theoretical, practical, and commercial implications for the outcome of research in this area. The neurophysiological literature of WMT seems to paint a promising picture showing that brain activity not only increased in cortical areas known to be important for WM (e.g. Olesen et al., 2004) but that in some instances the actual network of activation may shift (Langer et al., 2013). However, we have fallen very short of providing a clear and/or robust sequence of behavioural effects that allow for any reasoned conclusion regarding the efficacy of such interventions. Beyond the clear methodological limitations of a significant amount of the research within the field, there are clear conceptual gaps in what we know.

The previous section of this review was titled 'Is WMT effective?' and the title of this section seems to be a restatement of that question. However, they can be interpreted as asking different questions where the previous section title is the question asked in WMT studies and the evidence supporting the study conclusion is based on whether or not far-transfer effects emerge from that study. The question of whether a WMT intervention actually improves WM ability is rarely given a thorough investigation. Studies often suggest a WMT program improved WM performance by referring to performance improvement on the trained tasks from the early training sessions compared to the final training sessions. Alternatively, it may be demonstrated by a transfer effect to a task that is almost identical to at least one of the trained tasks. As discussed, these methods are insufficient for claiming a generalised improvement in WM ability. It is my opinion that the unclear picture of the success/failure of WMT interventions is largely to do with the focus on demonstrating far-transfer effects when conducting studies. It is easy to understand why a researcher would set out to demonstrate a far-transfer effect of the intervention they have chosen to investigate. If one is able to suggest that their training group improved skills such as nonverbal reasoning it is going to generate far more attention. There is also the added benefit in terms of terminology where it may catch the public eye due to being able to suggest an improvement in 'intelligence'. However, this has manifested itself into a problem because of our lack of understanding of what exactly can and does change from a practice/near-transfer perspective and under what conditions. The basic rationale for far-transfer to a construct such as g is that one practices at the limits of ones WM ability (almost always measured as span), this leads to an improvement in WM functioning at some point, thus a more powerful WM system leads to improvements on the numerous other facets of cognition that we know relies on WM to some extent. Therefore the most parsimonious explanation is completely dependent upon exhibiting improved working memory ability post-training. As a caveat, this improvement needs to be shown as a near-transfer effect and not a practice effect i.e. as discussed previously the measurement of WM improvement needs to use a WM task(s) that are sufficiently different from the trained tasks that any strategic advantage a participant has developed through repeated testing can be of no benefit to the criterion task. This demonstration of improved WM is often not actually tested or is tested insufficiently. This means that when far-transfer is seen it is rarely actually demonstrable that the improved scores are mediated by improved WM function. This means that researchers are left with far-transfer effects where they do not have an evidence based explanation for the mechanism of change.

All of this means that we must conclude that despite a large number of published studies in the area of WMT we have still yet to answer the question regarding whether the adaptive difficulty WM interventions actually improve WM. Therefore, what I propose is an approach where the focus of WMT studies shifts to near-transfer effects. More specifically; a focus on identifying near-transfer effects that are robust i.e. training task to transfer task success replicated sufficiently and secondly adequate investigation of the mechanisms underlying such change. Tests to demonstrate that interventions have successfully improved WM function in the WMT literature have generally used span measures (and more often than not, simple span measures (e.g. Klingberg et al., 2002, 2005; Borella et al., 2010; Brehmer, Westerberg, & Bäckman, 2012)). However, focusing only on span is likely not the best approach both in terms of detecting improvements and understanding what those improvements represent. Firstly, from a methodological perspective a span measurement may not be sensitive enough to capture improvements below a threshold. Some examples of reported mean values for span in various samples are; 2.35 (0.48) for adult reading span (Friedman & Miyake, 2005), 3.45 (1.02) for children's (mean age 8 years 7 months) operation span (Towse, Hitch, Hamilton, Peacock, & Hutton, 2005), 2.81 (0.79) and 3.86 (0.79) for children (grade 3) and adults respectively for counting span (Cowan et al., 2005). These values suggest that the process of improving from one span level to another is a very significant cognitive advancement. There are likely improvements to be made in ones WM functioning that don't add up to a span level improvement simply due to the size of the feat and a preoccupation with span outcome measurements could lead to missing smaller levels of improvement.

A span measurement doesn't necessary represent the fixed capacity of a cognitive system but the overall result of a cocktail of cognitive functions that contribute to performance on the given task. There are multiple phases involved in conducting a task of STM/WM and at each of these phases there are processes that are at work that will affect success on that trial, such as encoding, maintenance, and recall. Although different viewpoints on the nature of working memory may lead to further classification of phases. Different theories put the focus on different processes that are integral to WM performance and thus the 'span' of any individual. For example, proponents of inhibition based accounts of WM (Lustig, May, & Hasher, 2001; Bunting, 2006) would be interested in the ability to resist interference and how this ability may shift as a result of a WMT intervention. Researchers have used other outcome measures from span tasks to uncover theoretically relevant phenomena that would not be revealed through analysis of 'span' alone. For example research assessing differences in the temporal profiles of responses (e.g. Cowan et al., 2003; Towse et al., 2008). This research shows that much more can be inferred regarding WM function by assessing multiple outcome measures rather than letting a span measure account for all the individual differences.

As a final note on the concept of span, there are two clear ways of expressing how we conceptualise span. For some, span may reflect an independent module that is part of the WM system that is a fixed capacity and plays the role of container where other processes are responsible for adding to the container (encoding) and taking from the container (retrieval). In this conceptualisation the processes responsible for specific processes such as executive functions that are known to be integral to WM performance may falter and result in poor performance but that is not reflective of a reduction in span but an efficiency problem. Therefore in this case, the improvement in the independent storage module that is a container of fixed size is the critical element that one would need to improve in order to see improvement in WM performance above a previous limit. Now, an alternative and perhaps increasingly posited conceptualisation would be one where the notion of a storage container whose size is independent of the related EF and WM sub-processes is untenable. Instead, a span measure is a product of these processes working together. In this conceptualisation there is no independence in the storage module, it essentially does not exist as a module, instead span is a quantification of the number of representations maintained in a readily accessible state that these processes are able to juggle. Therefore, in this conceptualisation there is no storage module that needs expanding for generalised improvement of the WM system. Instead, the sub processes that work together to make up the WM system are the modules that need to be improved. In the storage independent module conceptualisation a genuine improvement in the sub-processes would lead to increased efficiency (i.e. fewer mistakes on sub-span trials) but not to improvements in the overall span. What we end up with as a measure in each conceptualisation is a number of representations that are held in WM but the underlying idea of what that number represents is significantly different in each.

1.7 Summary and thesis objectives

In this review I have attempted to offer a brief overview of the development of the most often used tools used by researchers in short-term/working memory measurement and how they have influenced research/theory in turn. Research in this field eventually led to the very influential multi-modal model of working memory (Baddeley & Hitch, 1974; Baddeley, 1986, 2000). The surge of experimental research investigating working memory consistently suggested that working memory performance was predictive of many higher order measures of cognitive ability. The emergence of working memory as a keystone of cognition and a growing understanding of brain plasticity led to WM being identified as a viable candidate as a trainable construct that may lead to widespread generalisable benefits. Along the way I have alluded to methodological components of assessing and scoring WM that are pertinent to assessing potential improvement in the construct.

The current state of the WMT literature is such that it makes it difficult to make any firm assertions either way with respect to the efficacy of WMT interventions such as CogMed. While there are now quite numerous randomised pre-post assessments of such interventions there are a medley of factors that make some different from others such as population group, task selection, dose, intensity, methods of analysis, and control group quality. The choice of tasks as both intervention and transfer-assessment is an issue widespread amongst WMT studies. As Shipstead et al. (2012) noted there is a tendency to use simple span tasks in both settings and define them as WM tasks. Simple span tasks associate with higher order cognitive tasks (Gf) when they are tested at higher span levels (5-6+) (Unsworth & Engle, 2007b). The criticism is not that simple span tasks should not be used as training tasks, given the adaptive difficulty setup of the training, participants are going to be completing these tasks at a level that requires them to engage in maintenance, search, reactivation from inactive memory, or other such executive processes that one believes underlie WM performance. However, there is a general lack of interest paid to the details of training tasks and the actual cognitive processes that are being trained and the relationship of these processes to the ones that underlie the transfer measures in the evaluation battery. There is a general lack of a role for traditional complex span tasks in WMT studies as a training (but see Chein & Morrison, 2010) or evaluation measure (exceptions include Jaeggi et al., 2008; Jaeggi, Studer-Luethi, et al., 2010). Given the role that complex span tasks have played in the development of working memory as a theoretical construct and in establishing links between working memory and a plethora of other cognitive abilities, the lack of inclusion in most WMT studies is surprising. We have also seen the potentially important role that including active (treated) control groups as opposed to passive (untreated) control groups, described by Shipstead et al. (2010) and described as well as demonstrated in the meta-analysis conducted by Melby-Lervåg and Hulme (2013).

Therefore in this thesis we set out to conduct a number of studies that assess the efficacy of adaptive difficulty WMT interventions. The focus of this work will be on whether or not the selected interventions produce a generalised improvement in WM ability. We will assess a variety of WM based interventions using a variety of WM-centric transfer measures with a particular focus on complex span tasks. The populations of interest will be non-clinical and we hope to recruit for studies based on a developmental sample (young children) and also adults. This participant pool will result in a range of WM abilities being recruited and thus allow for individual difference analysis on whether such factors are significant in predicting WM improvement. The pattern of results in the literature makes it difficult to predict exactly which near transfer measures will yield significant results. Where they are found, the effect will be dissected in detail to assess the link between changing processes in the training tasks and changing processes in the outcome measure which will help to identify what the mechanisms of change are when WM appears to improve. This is the first step to characterising the changes in the WM system as a result of such interventions which could lead to a greater understanding of WM processes and lead to replicable findings.

In addition to pursuing these empirical questions we will use the opportunity to discuss WM assessment and scoring in a wider context as the considerations we make regarding our near-transfer WM assessments will also speak to WM testing generally.

Chapter 2

Working Memory Training Developmental Study One

2.1 Working Memory Training - Study One

In the literature review I concluded with some statements on the current state of the working memory training literature. I showed that the results obtained are difficult to use to form an evidential base for the support or to refute the efficacy of working memory training as an intervention. The reasons for this include the array of variables that one must consider when evaluating the WMT literature such as population of study, training paradigm tested, evaluation (pre-post) tasks used, training dose, use of adequate controls, and analysis methods used to draw inference. I discussed these issues in relation to positive and negative results seen in the literature both in the near- and far-transfer domains. I also discussed the commercialisation of WMT interventions and noted how these products/services are marketed to all demographics. As an example, the MeeMo intervention is marketed as a successful intervention for improving the WM for Key Stage 2 pupils as evidenced by the following quote "MeeMo is a targeted programme for Key Stage 2 that improves EVERY child's Working Memory and capacity to learn across ALL subjects" (archived at http://web.archive.org/web/20150702152949/http://www.risingstarsuk.com/series/meemo/). As the majority of WMT studies that have utilised a developing sample have focused on a 'selected sample' based on clinical diagnoses or selected based on pre-screening (i.e. low WM) it was decided that any addition to the WMT literature consisting of a typically developing sample would be of value. In addition to this point, the expansion of the 'brain-training' commercialisation appears to be expanding into educational settings whereby it would be used as a 'whole class' intervention. MeeMo is an example of an intervention that is 'whole class' by design but given the expansion of such WMT interventions into educational settings it easy to see the potential for computerised versions of WMT being incorporated in the same way. This is most readily seen if one envisions the situation where a school decides that, as MeeMo claims, WMT will be beneficial to all children so they decide to implement a WMT programme into the curriculum. The only practical way of doing so, due to time and financial resources, would be in a group setting. Therefore this study represents an attempt to assess potential benefits of an adaptive-difficulty computerised WMT intervention that is administered in a group setting.

The primary point of emphasis extracted from a review of the literature was that despite a significant number of studies, some of which showing dramatic transfer effects; we are no closer to understanding the actual effects of these interventions on the WM system (i.e. near-transfer). It is clear that to make sense of observed far-transfer one needs to understand the changes in the trained construct. This leads to two points of emphasis for the design of the study here and in the rest of this thesis. The first of these points is that we will focus on near-transfer evaluation tasks and where far-transfer tasks are used they will be 'not-so-far' transfer tasks. The second point is that the practice effects on the trained tasks will be assessed in some detail. Generally, the practice effects are not subject to the same level of critical assessment as the transfer assessments in WMT studies. The prototypical usage of practice effects in the WMT literature is to use a simple analysis that shows an improvement was actually made on the trained tasks such as t-tests on early vs. late training sessions. By subjecting the practice effects to a more critical level of assessment and combining this with near-transfer assessments we can build a larger evidence base of actual change occurring in the WM system.

For our first study we decided to use a battery of training tasks as opposed to a single task. This decision was made based on a number of factors. Firstly, simply to be able to document the practice effects on multiple tasks rather than a single task in this group setting. Secondly, a variety of tasks will help keep the children motivated throughout the training phase of the study. Any repeated task is likely to present motivational issues as enthusiasm gives way to fatigue. This is especially of concern given the age group we are testing (9-11 years). Through multiple-task training sessions it is hoped that a session moves on to a different task prior to the onset of fatigue/boredom. The method section will outline a number of other steps we took to attempt to maintain motivation in our sample.

In any WMT study the selection of transfer measures is of great importance. Due to restrictions on the amount of testing time allowed from the participating schools we were only able to include four tasks in the pre-post evaluation battery. Therefore to provide a variety of measures we selected measures that varied in terms of domain (visuospatial/verbal), memory requirements (STM/WM/LTM), and also a non-memory based transfer task. In terms of near transfer we will use a verbal-based WM task (operation span) and a visuospatial-based WM/STM task (matrix span) that shares properties with the WM tasks in the training battery but are not identical. These tasks will provide measures of ‘span’ amongst other outcome variables. In addition, the operation span task will provide operational speed and accuracy.

Two measures designed to assess ’further’ transfer will be the ’silly sentence’ task and a speeded mental arithmetic task. These tasks are selected to assess transfer at a more basic level than studies have tended to so far. Often, researchers are looking directly at the effects of a training intervention on higher order cognitive faculties such as fluid intelligence. Partly this is due to the headline generation if evidence is found that an intervention improves measures of such a construct. However, the far transfer results within the literature is very mixed and one wonders if researchers may be better served attempting to 'bridge the gap' of explaining how WM training would improve Gf. If WMT improves Gf then what is it about the WM system that has been improved that may have led to such a result? Perhaps the trained WM system is processing information at a faster rate, perhaps links between WM and other cognitive constructs such as LTM have been strengthened and made more efficient, or perhaps WMC has genuinely been increased. For these reasons we have selected two measures to assess potential transfer that are lower level cognitive assessments. The silly sentence task involves processing verbal information and interacting with information stored in long-term memory to judge the accuracy of the sentence. The speeded mental arithmetic task will assess general arithmetic skills as well as speed of processing. If WMT does not yield lower order cognitive improvements on tasks such as these then it would make it very difficult to explain how it may improve general fluid intelligence.

To summarise and formally state the questions of interest:

- RQ1. How much, if at all, do participants improve on the training tasks? How much training is required before measures of performance reach asymptote? These questions are important because they offer insight into working memory functioning in and of themselves. For example finding that different tasks have different trajectories and then comparing the relative contributions of each to any transfer effects. Assessing improvement on the trained tasks and answering questions regarding why some may improve more than others (i.e. which ones can be attributed to strategy usage vs. increased resources) will be informative for the generation of WMT packages. It is predicted that participants will show improvement on all training tasks but the extent of improvement will vary between tasks.

- RQ2. If the adaptive difficulty training regime we have outlined for this study is successful then transfer should be seen to the untrained tasks. This will take the form of an increase in post-training scores for the experimental (training) group over and above any change seen in the active control group. We are agnostic regarding predicting the outcome due to the weight of evidence presented in previous work failing to reach a general consensus, in addition to the level of methodological concern over the evidential base.

- RQ3. We were fortunate to be able to arrange to revisit the participating schools 6-months after the completion of the post-training phase. Therefore we will be able to administer the pre-post battery again at this phase. If near or far transfer is observed at post-training, are these effects robust beyond the termination of training? (Melby-Lervåg & Hulme, 2013) includes long-term follow-up effects in their meta-analysis and find that the effects do not stand up to the decays of time. However, the number of studies assessing these effects was very low in each analysis. If transfer is observed at the post-training phase and this improvement is a genuine improvement in cognititve systems then these improvements should be present at the long-term follow-up phase.

There are also additional aspects that we would like to pay attention to in this study. A number of these relate to the methodological issues that are an integral part of such research. For example, in taking the laboratory to the classroom and testing in groups it is necessary to be aware of the potential impact this can have on the collected data. Group testing and conducting the study in the classroom presents an opportunity to test a training intervention as it may actually be deployed in the curriculum if positive results were to be consistently found. In this study pre-post measures will be conducted in small groups (4-6 children at a time) as it is more important that each pre/post measurement is as accurate as possible. The training sessions will be conducted in full class sessions. It would be foolish to suggest that this isn't going to impact on performance on some of the training sessions for some children but the impact of lapsed concentration on any one training trial is small compared to the impact of an inattentive trial on the pre/post measures.

What should the 'training' regime constitute for an active control group in WMT studies? We've seen that many studies have problematic control groups; whether it is absent, has no contact throughout the training phase, or completes exercises where it is likely that the Ps understand they are not being tested like some of their peers. The reported experiments with active control groups often either use irrelevant computer games, or more rigorously the non-adaptive versions of their adaptive training tasks. An ideal control program would match the experimental program in every way but the aspect of the program that is under experimental investigation. In this instance that is the taxing of WM system/processes. Therefore a control program should have participants spending the same amount of time engaged in a computerised task, have the same amount of interaction with teachers/experimenters and be similarly engaged with the computerised tasks as their experimental counterparts. We will select a range of computerised cognitive tasks for the control group that attempt to meet this criterion. Taking genuine tasks used in cognitive research but that involve no WM component. More specifically, there will be no memory storage component to the tasks. Prohibiting any type of process that can be suggested to relate to WM performance is a very difficult task.

- RQ4. Consider the possible effects of administration environment. Are there any indications that these issues lead to very different data than what would be obtained conducting all sessions in the lab.

- RQ5. Assess the control group's data for signs of continued attention and motivation as the experiment progresses. The control group's engagement with the training can be assessed by ensuring number of trials completed in the time-frame and accuracy do not significantly decrease as a function of training session number.

2.2 Method

2.2.1 Participants

Participants consisted of 55 children from two different schools from the North-West of England (drawing upon a single year 5 class in each case). One year 5 class from each school participated. All children were within the standard year-5 range (10-11 years). Each school was sent £50 book tokens to the two classes who participated. No child was completely omitted from the study for any reason but a number of children excluded from particular analyses (due to session absence). Whenever this occurred it is highlighted and explained further.

2.2.2 Design and materials

The experiment is a randomised controlled trial (RCT) that consists of four phases; baseline phase, training phase, post-training phase, 6-month follow-up. The training phase was administered over a five-week period and during this phase participants either participated in sessions consisting of computerised cognitive tasks (working memory training (WMT) group or non-working memory training (control) group). Participants were randomly assigned to either condition with the caveat that a similar proportion of participants within each school were in each group. To measure any transfer of training a battery of tasks was completed at each of the non-training phases (baseline, post-training, 6-month follow-up).

All of the computerised tasks used in this experiment were programmed for this task by the experimenter using the Python programming language and the PsychoPy library (Peirce, 2007).

Training Phase

A battery of five tasks made up the task pool for the training group. Each task would be worked on for 5-minutes before switching to a new task, where one session consisted of three of the five tasks. The participants would train on three of the tasks in each session. Task difficulty varied based on performance levels in each session, the adaptive mechanisms are explained below. The active control group had three tasks to complete that were chosen so as to place little or no stress on working memory.

Working Memory Training (WMT)

Working Memory Period The working memory period (WMP) training task (Towse et al., 2005) requires participants to store and maintain information while concurrently engaging in mental arithmetic operations. Therefore, it can be seen as similar to traditional complex span measures. However, the WMP task differs from traditional complex span measures in that the manipulation of task difficulty is not through increasing the number of to-be-remembered items but in increasing the cognitive load of the operation phase.

One trial of the WMP task consists of the participant answering three mathematical operations using the number keys on the keyboard. The answers to the mathematical operations are the to-be-remembered material in the trial. After the answer to the third operation is given the participant is shown the recall screen. To input their answer the participant uses the mouse to select from an onscreen keypad. The participant is reminded that the serial order is important when inputting their response via an on screen reminder.

Participants use the number keys on the keyboard to give their response to the operations. After answering three operations they then use the mouse to select three numbers from an onscreen number pad to input their previous answers. By having separate controls for answering operations (keyboard) and recall phase (mouse) it to some extent controls for the potential to remember the pattern/location of key presses (muscle memory). Feed-

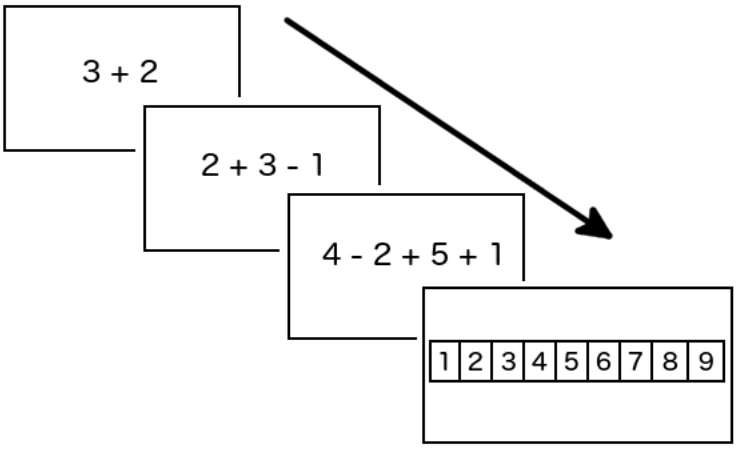


Figure 2.1. Working Memory Period (WMP) Task

back was given at the end of each trial using a green tick or a red cross image in the centre of the screen. In addition, a tally was updated and displayed after each trial keeping track of the number of correct/incorrect responses in each session for the participant.

Adaptive Difficulty: After every five trials the program assessed the performance of the individual over this epoch to determine what processing demand/duration level the following five trials should be. If four or more correct responses had been given the processing demand level would increase by one, if two or fewer correct responses were given then the processing demand level would drop by one, while three correct responses left the level unchanged. There were three levels of difficulty based on the stimuli obtained used in Towse et al. (2005). The level determined how many 'parts' there were to each operation, for example a level one operation might be "4 + 3" while a level three operation would be "3 + 4 - 2".

Working Memory Period v2 In an alternate version of the WMP task participants were shown two *letters* in sequence, *at the start of each trial* out of the possible candidates L, S, C, T, V, H, N, P, R. Then the task followed the same mechanics as the original, they were shown three operations where the amount of "parts" of the operation was determined by the current level of difficulty. After answering the three operations participants used the mouse to select the two letters in the order they recall seeing them at the start of the trial. This task therefore used different memory stimuli and implemented a slight variation on the storage/processing dynamic as all the to-be-remembered material was presented at the start of the trial followed by the processing elements.

Adaptive Difficulty: Identical to the original WMP task. The number of letters to-be-remembered was static but the demand/duration of the operations was adapted.

Memory Updating The memory updating (MU) training task involves the updating and maintenance of two items over the course of six updates. A trial began with two blue boxes, one left of centre and one right of centre. A '+' appears in both to indicate a trial is about to begin. After 500ms the +'s are replaced with the starting numbers for each box. The numbers are present for the current inter-stimulus-interval (ISI) before disappearing. A 750ms (constant) blank pause is in-between every display of numbers. After the start numbers the participant will see a small operation instruction appear in one of the boxes such as "+3" which they need to apply to the number they currently believe is "in" the box. The recall phase begins after three updates have been given for each box, at which point a "?" appears in the left-side box indicating that the participant must input the number they believe is now ''in' the box. After a response is given the "?" shifts to the right box. Once both responses are registered the boxes display either a red cross or a green tick to indicate whether or not the participant gave correct or incorrect responses (independently, one box may be correct while the other incorrect). A tally which is displayed on screen updates for each correct and incorrect response given so participants can keep track of their performance level.

Adaptive Difficulty: The difficulty is altered by increasing/decreasing the ISI and therefore changing how long the participant has to update the number in the box which has consequences for the time to rehearse/reinforce the two currently held numbers in memory. The ISI is manipulated after every trial. The starting point in a session is 1500ms, this will increase or decrease by 100ms each trial depending on the result of the

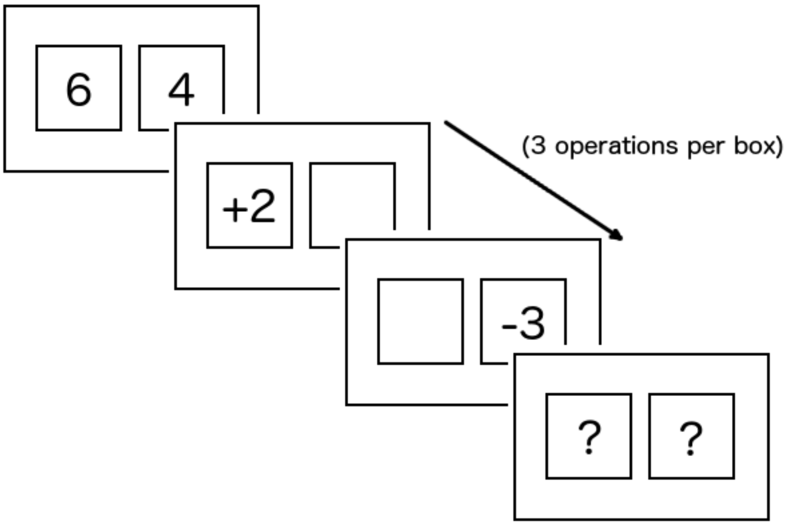


Figure 2.2. Illustration of the memory updating task

previous trial. If the participant correctly recalled both numbers then the ISI decreased, if either one was wrong then the ISI increased.

Colour Corsi (location-colour binding The Colour Corsi (CC) training task involves the storage and retrieval of multi-feature stimuli.

The CC task is similar to a matrix-span task (e.g. Kane et al. (2004)) which in turn is similar to a computerized version of the Corsi-Block task (hence the name). The difference is that rather than simply highlighting grids in sequence and the participant needing to recall said sequence after the presentation phase, instead when the grids highlight they do so in one of four colours (red, green, blue, yellow). The participant must recall the sequence in the correct order but also indicate which colour each grid was in the sequence. The recall mechanism is simply two mouse clicks per grid response, one to select a colour and the second to select the grid. Feedback was given per trial by way of a red cross or green tick image while a counter also displayed the number of correct/incorrect trials per session.

Adaptive Difficulty: The difficulty level of the CC task was assessed every three trials. If the participant had successfully remembered the colour-location sequences for two or more of the previous three trials then the level increased by one. If they had one correct

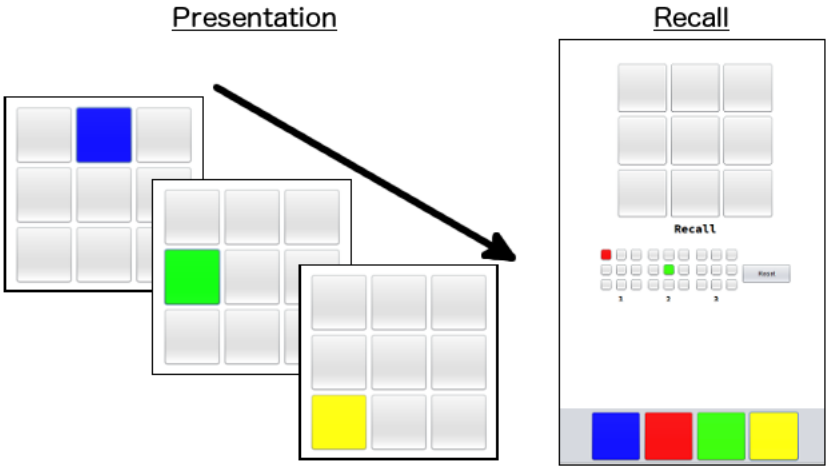


Figure 2.3. Illustration of the Colour Corsi task

response then the level remained unchanged, while if no trials in the last three were correctly recalled the level dropped by one. The level determined how many location-colour items there were in each trial, e.g. level three trials had three location-colour items to store and recall.

Stroop A computerised version of the classic Stroop task (Stroop, 1935). The Stroop task involves presenting the participant with a word such as "green" that may or may not be coloured green. Trials where the colour matches the word (the word "green" and a green font colour) are termed congruent while those where there is a mismatch (the word "green" with a blue font colour) are incongruent. The participant must respond to one aspect of the stimuli such as the actual colour of the font. We used red, green, and blue as our colour/word combinations and asked participants to respond to the colour of the word they saw using the 'R', 'G', and 'B' keys. While this task is not traditionally used as a measure of working memory there are a number of theories around the functioning of working memory that place attentional processes at the fore (e.g. Kane, Conway, Hambrick, & Engle, 2007). Past studies have shown improved performance by children on Stroop tasks after adaptive WM training (Klingberg et al., 2002, 2005) further demonstrating the dependency between memory and attention systems. Additionally, the inclusion of a task that isn't focused on tapping the WM system adds an element of variety to the training regime that may help keep participants motivated to complete the training.

Our Stroop implementation had similar feedback mechanisms to the other training tasks in that there was a counter displayed to show the participant how many they had answered correctly and incorrectly in any given session. However, in addition, given the large number of trials that can be completed in a small time frame on this task we also incorporated an extra feedback mechanism to aid motivation while completing the Stroop training. Every 15 trials the participant was shown a summary of their performance for those trials. They were shown the number out of 15 they had got correct and what their average response time was for those trials followed by a sentence "Can you beat it? Press spacebar to try".

Adaptive Difficulty: There was no obvious increment/decrement to make to adapt difficulty in the Stroop task like there is with conventional working memory tasks but it was important to try and keep the dynamic of the procedure similar to the other training tasks and two of the principal features of them throughout are feedback and adaptive difficulty. The Stroop effect itself has been shown to be stronger when the rate of congruent trials is higher (e.g. Tzelgov, Henik, & Berger, 1992). Given that the Stroop effect increases it seems reasonable to influence the rate of presentation of incongruent/congruent trials as a way of manipulating difficulty. Therefore every 20 trials the level was assessed. If the participant had given 18 or more correct answers the level would increases, 12 or fewer and it would decrease, while 13-17 correct maintained the current level. There were 5 levels in the Stroop training task and each level influenced the probability that a trial would be congruent or not. At level one the probability of the trial being congruent was .6 meaning that approximately 40% of the trials would be incongruent. The probability of a trial being congruent decreased by .1 per level increase so at level five the probability was .2 meaning that approximately 80% of the trials would be congruent.

Active-Control (AC)

Dots Task The dots task was a very simple visual search type task. A warning prompt initially appeared to warn the participant to get ready for the presentation of a new trial. A field of dots would then present itself in the middle of the screen where some dots were red and some were black. The number of each was randomly decided on each trial and either number could be between one and nine inclusive. The task for the participant was to count the number of red dots and respond using the number keys as quickly as possible.

To maintain appearances between the two groups of tasks the dots task also had feedback and motivational mechanisms. The number of correct and incorrect responses was tallied and displayed to the participant at all times. Additionally, at the conclusion of each trial they were told if the trial was correct and how quickly they had answered e.g. "Correct! You answered in 1.16 seconds". As well as the correct/incorrect tally the program always displayed to them their fastest response time for a correct trial to give them a target to beat.

Subitizing In the Subitizing task the participants were flashed a number of black dots (ranging from one to six inclusive) for a very small amount of time (starting at 155ms). The participant simply had to respond with the number of black dots they thought they saw. The amount of time the dots were present on screen was adapted with performance. A correct response decreased the presentation time by 5ms while an incorrect response increased the presentation time by 5ms.

The feedback and motivational aspects were in line with the methods used in the dots task. Participants were given instant feedback on the veracity of their answer and were shown their response time. The top left corner always showed the number correct so far, the number incorrect, and their fastest response time for a correct trial.

Number Line The number line task presented participants with a horizontal line in the centre of the screen with vertical bars at either end. Below the left vertical bar a "0" was displayed while below the right vertical line the number “1000” was displayed. The task required the participant to place a marker on the horizontal line that represented the position of a random number given to them between 1 and 999 inclusive. For example if they were asked to place a marker where 500 would be on the line then they would need to try and place their marker at exactly the midpoint of the number line. When a participant clicks on the line a small red marker appears to indicate their selected location. They are able to move this marker until they are happy with the location at which point they must click the submit button to enter that response.

After submitting a response the participant is greeted with the feedback screen. They are shown the exact number that the location they submitted represents and then told how much they missed by. For example the target number may be 456 and I may have placed my marker at number 487, in this case I have missed by 31. In the top left corner of the screen they are shown their “Lowest score so far” which represents the closest they have got in that session.

Pre, Post and 6-Month Follow-Up Assessment

Near-Transfer

Operation Span A measure of performance on this verbal complex span measure was included to assess possible near-transfer effects of this training battery. Our implementation of the operation span task follows the same procedure as the automated operation span task used by Unsworth, Heitz, Schrock, and Engle (2005). Participants were shown a mathematical operation with an answer such as “3 + 4 = 7” which they needed to calculate themselves and decide if the given answer was correct or not. Upon giving an answer using the ‘y’ or ‘n’ key they were shown one of 12 letters (L, S, C, T, V, H, N, P, R, J, K, F). This process repeated however many times the set size of that trial was set to be e.g. if it was a set size two trial then there would be two operations to evaluate and two letters to remember. After all the letters had been presented the participant was shown a recall screen which consisted of a 3 x 4 grid of the 12 possible letters and a submit button above. Using the mouse the participant input the letters they remembered in the sequence they believed they were presented before clicking the submit button. Figure 2.4 illustrates a typical trial for the Operation Span task.

In this task after each operation the box flashed a red cross or green tick image before showing the letter that the participant must remember. This was implemented to reiterate to the child participants that it was important that they attend to the operations as well as the letters.

It was decided that this task would follow a similar protocol to the training tasks in that they would start at a low level (in this case, 2) and change with performance. This is different from the traditional administration method of having a number of trials at each set size and either having a termination policy (once a participant fails a certain proportion at a given set size) or having them run through all of the trial regardless of performance. There are a number of reasons why we were reluctant to use either of these methods. Given the self-paced nature of the task these traditional methods gave no control over how long the task would last meaning for some of the participants they may have had to leave before getting to one of the other tasks in the baseline battery. If we were to administer all the trials up to a certain set size then it is very feasible that many of the children will be sat having to sit through trials they have very little chance of succeeding at leading to a loss of motivation and interest which may also impact on performance in any tasks still remaining in the battery. Finally, we were operating under very tight schedules when given the chance to work with the children in this study. Therefore participants started with three span two trials and the span size of the next three trials was decided by their performance. If they were unsuccessful at 2 or more of these trials then they would be given more span size two trials. If they were successful at two or more then the next set of trials would be at the next highest span size and so forth. The scoring of the tasks would then only take into account the first attempts at each span length the participants were able to reach in order to account for the variation in number of trials completed in

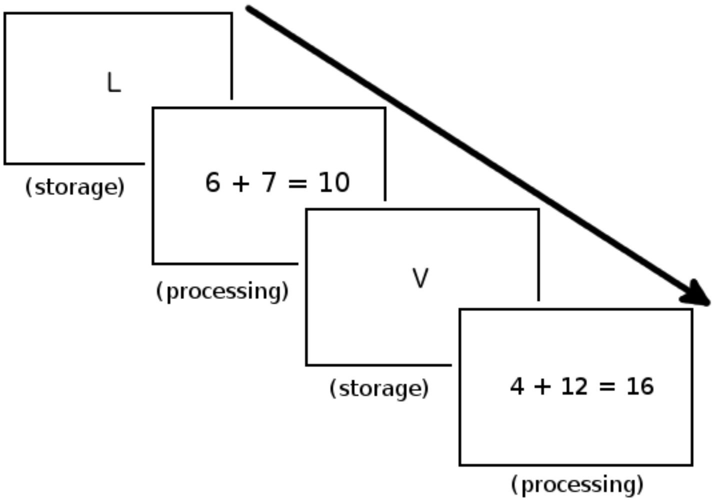


Figure 2.4. Illustration of the Operation Span task used in experiment one.

time frame.

Matrix Span The matrix span task was included as a measure of very near-transfer to spatial short term memory. The task was simply a 3 x 3 grid that would highlight a sequence one grid at a time. The participant needed to store and maintain the grid sequence and using the mouse reproduce the sequence when asked. This is essentially the same task as the Colour Corsi without the binding element. Participant was informed if they were correct or incorrect at the end of each trial. As with the operation span task we followed an adaptive administration technique rather than a fixed set of trials with/without termination.

Far-Transfer

Silly Sentence In the silly sentence task the participants were shown a sentence such as ”The sky is green” and asked to make a judgement as to whether or not the sentence was factual or not as quickly as possible. It was stressed that the participants should try and be as quick as possible but not at the cost of accuracy. At each administration the participants responded to 38 sentences where half were true sentences and half were false sentences.

The silly sentence was selected as a possible far-transfer measure as to make a quick judgement on these sentences there needs to be some interaction between the current sentence being held in working memory and long-term knowledge that allows a person to check the facts. Therefore we use this task as a simple measure of the passing of information between the WM and LTM systems.

Mental Arithmetic We also administered a speeded mental arithmetic task. The mental arithmetic task was split into six blocks where each block had a different "type" of mental arithmetic operation, they were:

- Block one - Addition (without carry) e.g. "3 + 4".
- Block two - Addition (with carry) e.g. "5 + 9".
- Block three - Subtraction (without carry) e.g. "7 - 3".
- Block four - Subtraction (with carry) e.g. "22 - 14".
- Block five - Multiplication
- Block six - Division

The participants were given one minute to answer as many operations as possible within each block. We split the blocks up as we feel it offers a number of benefits. Firstly, it allows the comparison between performance on addition/subtraction with and without the need to carry which may prove informative as carry operations should load on working memory slightly more than those without. Secondly, there is substantial variation between children of this age group in their speed of mental arithmetic and also in where they are with their maths learning. For example some children are not yet at the point that they can carry out division operations. Therefore it seemed sensible to separate the operations into blocks to allow flexibility in the analysis stage. If we had thrown all the operations into one block and set the time limit it is likely that some participants will have encountered a division/multiplication problem they could not do and perhaps stopped responding or some other behaviour which would have impacted on their overall mental arithmetic score.

The stimuli for each block was taken from a paper mental arithmetic test (need reference for Hitch paper where it was used). Each block had 30 operations so there was a maximum score of 30 (it was not expected anyone would get too close to this total in the 60 seconds time limit). The operations had three clear levels of difficulty, for example, in the subtraction without carry block the first 10 operations would be similar to ”8 - 5”, the second set of 10 operations might look like ”47 - 25”, and if a participant got to the final set of 10 trials they were greeted with operations such as ”453 - 311”.

2.2.3 Procedure

Ethical approval was obtained for the current study, details of which are available in section A.1 of the appendices. The training phase ran for 5 weeks. During this 5-week period the researcher arranged three sessions each week with each school where the researcher could go into the class and run a training session. The software (Python + PsychoPy) was installed on the school machines so that they could be used for the pre/post and training sessions. One school had a suite of Toshiba 14” laptops that we could use while the second school used 11” Compaq netbooks. For this reason the code for the programs was edited to work best on a 1024x600 resolution (as this was the netbooks native resolution). This meant that the program would run full screen on the netbooks but windowed on the larger Toshiba laptops. Participants were randomised into either the Active Control or WMT group before any baseline information was collected.

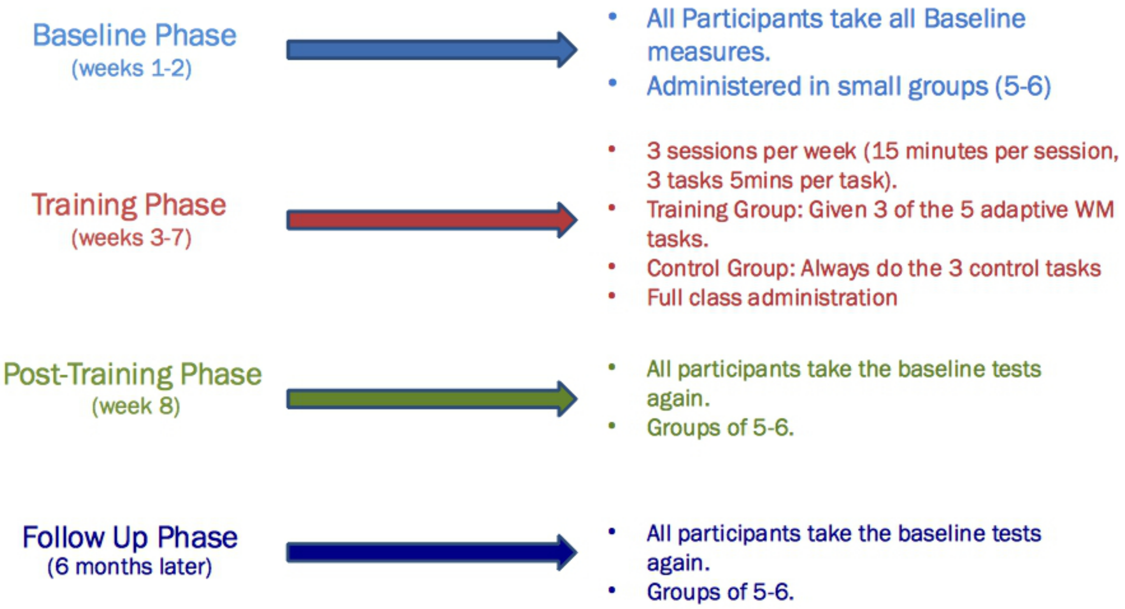
The actual time spent on the computerised tasks was 15 minutes in each of these sessions. The working memory training group was given a selection of three of the five training tasks to complete (rotated to try and ensure equal amounts of each task completed) while the active control group completed the three control tasks each session.

The training sessions were administered in a grouped classroom environment because the impact on the schools day to day running needed to be as minimal as possible. While this method presents its own set of problems (i.e. children distracting others, mixture of control/training groups in the same classroom) it also has some benefits over some methods used in other training studies. For example, many studies (e.g. von Bastian, Langer, Jäncke, & Oberauer, 2013) use self-administration at home, which in itself comes with its own set of pros and cons. The group administration allowed the researcher to be present along with at least one teacher which allowed for some control to be maintained over the conditions of the training. A consequence of this design is that blinding procedures were not possible as the researcher needed to observe and control the classroom during the training sessions. Additionally, it was possible for the participants to observe that others were not completing the same tasks as themselves.

The transfer measures (pre/post/follow-up) were administered at three different time points. The baseline (pre-training, T1) phase was two weeks prior to the onset of training (due to a one-week school holiday). The post-training phase (T2) took place in the week following the conclusion of the training phase. The follow-up phase (T3) took place in the week 6-months after the post-training phase. There were some individual data points lost due to technical faults (will be detailed in results) but the actual participant attrition between the time points was N = 55 at T1, N = 55 at T2, N = 42 at T3. The drop-off at T3 was due to a number of children moving schools in the intervening 6-month time period and a number of children were absent from school when this phase took place.

All training sessions were coordinated by the thesis author who was present at all sessions involving data collection and was supported in some (but not all) training sessions by a volunteer MSc student.

Figure 2.5 summarises the experimental procedure used in this study.



Baseline Phase
(weeks 1-2)

- All Participants take all Baseline measures.
- Administered in small groups (5-6)

Training Phase
(weeks 3-7)

- 3 sessions per week (15 minutes per session, 3 tasks 5mins per task).
- Training Group: Given 3 of the 5 adaptive WM tasks.
- Control Group: Always do the 3 control tasks
- Full class administration

Post-Training Phase
(week 8)

- All participants take the baseline tests again.
- Groups of 5-6.

Follow Up Phase
(6 months later)

- All participants take the baseline tests again.
- Groups of 5-6.

Figure 2.5. Overview of experiment design

2.3 Results

2.3.1 Data Hygiene

To address research question 4, before combining data from the two schools it was decided to collate data separately and compare the performance due to observations in differences in the school environments and the behaviour of the pupils. Additionally, one of the schools we worked with had more temperamental IT facilities and therefore on any given training session a number of computers would not work and thus some children would have to sit out of the training for that session. Separate analyses yielded no differences between the schools on training tasks or transfer measures. For brevity these analyses are not produced here, only the combined analysis. Only after it was clear that performance was comparable across both schools despite the difference in environment/situation were the data combined for analysis.

After the school comparison process was complete then the next major step in screening was to apply some quality control to the dataset. Identification of outliers and datum that are confusing are often sought out and excluded before moving on to analysis but quality control of the data is even more important here due to the whole-class training procedure. There are likely to be a number of sessions that do not reflect the participants actual ability at that given time due to possible distractions.

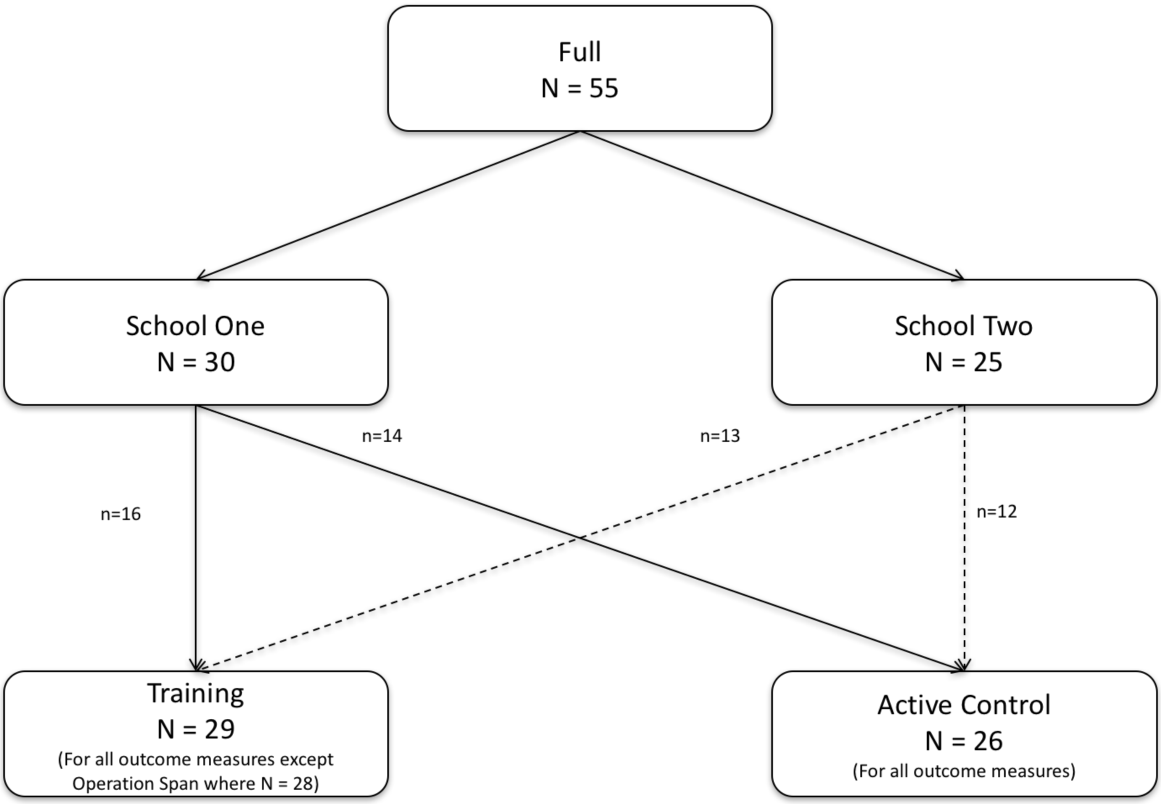


Figure 2.6. Distribution of participants between school classes and experimental groups

Figure 2.6 shows the overall participant structure in this experiment. An obvious approach would have been to designate one class as the experimental group and one class as the active control group. But this structure would have the negative consequence of introducing environment as a non-controlled extraneous variable. Therefore the decision was taken to divide each class into training and AC participants.

Memory Update One session was excluded from analyses as the participant only logged two trials for that session (mean value was 11.5).

WMP2 Exploratory plots for the WMP and WMP2 tasks revealed that on occasion children opted for the strategy during the WMP2 task to ignore the operations. To recap, the WMP2 task presents TBR stimuli prior to the operations. While the children were instructed that it was important they also tried to answer the operations correctly the task was programmed to assess accuracy based solely on the items input at the recall phase. Based on the difference in distributional properties for the proportion of operations correct between WMP and WMP2 it was clear that some sessions in the WMP2 dataset had been completed ignoring the operations so as to get to the recall screen as quickly as possible and thus be 'correct' on all trials. A number of sessions were removed from the dataset based on a cut-off value for the accuracy of operations (0.34). This cut-off represented the mean minus two standard deviations as determined by the WMP data (M = 0.78). If the majority of a participants sessions were below that cut-off then all of their WMP2 data were removed (n = 7), 13 other sessions were removed that came from 5 participants. See Figure A.1 in the appendix for an illustration of the original distribution of WMP2 operation correct proportions compared with WMP.

Stroop There are ten sessions that produced overall mean response times below 500ms. The mean of the accuracy for these trials was 0.6 compared to a mean accuracy of 0.88 for the rest of the sessions. I think this is a reasonable quick fix diagnostic for non-attended sessions and thus those ten sessions were removed from the dataset to be analysed.

WMP and CC These datasets were left intact as there was no cause for concern.

See appendix *section A.3* for full details of excluded data.

2.3.2 Practice Effects (RQ1)

With regards the data provided by participants on the training tasks there is only one method of analysis from our toolkit that is suitable and that is the generalised linear mixed model. The reason for this is because of the unbalanced nature of the data given the varying number of sessions each participant completed. When it comes to assessing the impact of the training intervention on the transfer measures there is an obvious reason to exclude participants who failed to meet a specified quantity of training. In order to assess the effects of repeated administration of a task, a participant who completes three sessions has provided three valuable data points that can be used, regardless of whether this number is less than others in the dataset. That is where the GLMM approach offers a substantial benefit over traditional techniques such as a repeated measures ANOVA, for the ANOVA we either have to reduce the number of levels in the session variable or reduce the participant pool in order to have a balanced dataset.

For each training task I would like to assess the pattern of performance over the repeated sessions to evaluate any performance change observed in these tasks. I will present some exploratory information for each of the 5 tasks followed by the results obtained from GLMM analysis. Statistical significance of the session factor can be measured by means of a likelihood ratio test comparing the null model with the addition of session as an IV. The parameter estimates obtained for the levels of the session factor will uncover the overall pattern of performance over repeated sessions.

For the same reasons that a standard repeated measures ANOVA approach is inappropriate for this analysis, general descriptive statistics are not particularly meaningful either and are therefore not reported. It is important to note that what I am suggesting here is that the parameter estimates obtained from the mixed model with session as a fixed factor will result in adjusted means for that particular dependent variable where the adjustment is based on overall participant ability (random intercept) and thus is not unduly biased by the drop-off in participant numbers as session number progresses. The reduced n at each successive sessions will be evident in the increased error bars associated with these estimates. Thus the mixed model is being used to provide descriptive information regarding the general performance change over sessions.

Primary Indicators

For each of the 5 training tasks we will look at the primary dependent variables that describe the participants overall performance. For WMP, WMP2, and CC this will be the average level that participants were operating at for the duration of that session. The mental counters task varied in difficulty based on the ISI which increased or decreased based on the performance level of the participant. And for the Stroop task we will look at the mean difficulty level the participant was operating at. Recall that the Stroop task was administered in 15-trial blocks and the prevalence of incongruent trials shifted based on performance, there were 5 levels of 'difficulty'.

Figure 2.7 shows one plot per task for the primary indicators of performance described above. I provide spaghetti plots showing every participants trajectory in the appendices (figures A.2, A.3, A.4, A.5, A.6 for WMP, WMP2, CC, MU, and Stroop respectively, the level of variability between participants is evident in these plots).

In figure 2.7 we can see that the effect of session on the primary indicators when we account for the differing levels of ability. From this figure it would seem that it is reasonable to conclude that there is little to no effect of repeated sessions on our primary indicators of performance. This is a strange finding. The following brief sections will examine the training tasks in a bit more detail but it is important to note here that the original expectation would have been to see more improvements in these tasks than has been observed. This would have meant the following section acted as an 'unpacking' of the overall effect of mean level improvement. Despite seeing little improvement beyond the first few sessions there are still some potentially interesting take aways from the practice effect data.

Working Memory Period

In the WMP task the participant needs to be successful in answering the operation as the correct answer for each of the three operations forms the memoranda. Therefore it is

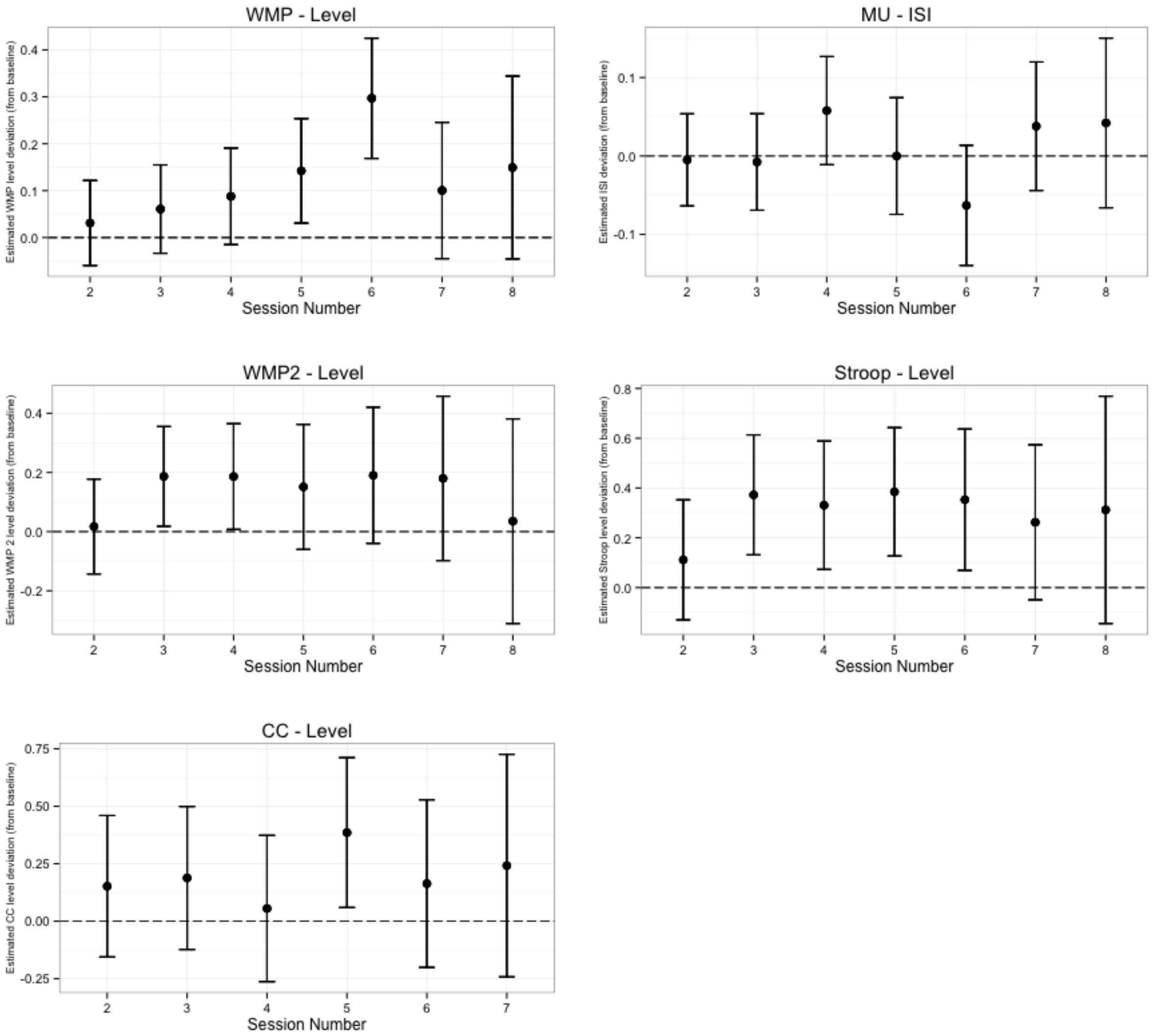


Figure 2.7. Parameter estimates with 95% confidence intervals for the effect of session number on primary indicators of performance per training task

interesting to look at how the speed of response for the operations changes as the task is repeated. The pattern of deviations for each of the three serial positions can be seen in figure 2.8. From this figure we can see that, as with most of the changes observed, the bulk of any change occurs in the initial few sessions. The RT for the first serial position shows a marginal increase before regressing to baseline. However, for serial positions 2 and 3 the increase in RT is much larger and does not show the same regression.

Stroop

We have seen that the level of Stroop trials that participants were generally operating at increased by approximately 0.4 between sessions one and three and then remained relatively stable for subsequent sessions. This may reflect a ceiling effect as the average level at session one was 4.19. The level was reactive to accuracy of trials. Generally, the accuracy of Stroop trials is not the dependent variable of interest, rather the speed of resolving conflict What may be of particular interest is the size of the Stroop effect as sessions progress (difference between RT for congruent and incongruent trials). Figure 2.9 shows the deviation of the size of the Stroop effect from the baseline (M = 156ms). This figure shows that the size of the Stroop effect decreased by 70-100ms between sessions one and four but was not reduced beyond that. Assessing if the imposed level differences affected 'difficulty' is not straight forward due to the way participants performed. An overall assessment of the Stroop effect at each difficulty level is not sensible due to the huge amount more trials conducted overall at level 5 compared to lower levels. The actual numbers were 508, 563, 4009, 4209, 19323 trials at each of the 5 difficulty levels in ascending order. The overall Stroop effect for levels three, four, and five, calculated by taking the mean of the RT of incongruent trials at that level and subtracting the mean of the congruent trials was observed as 144ms, 119ms, and 102ms respectively.

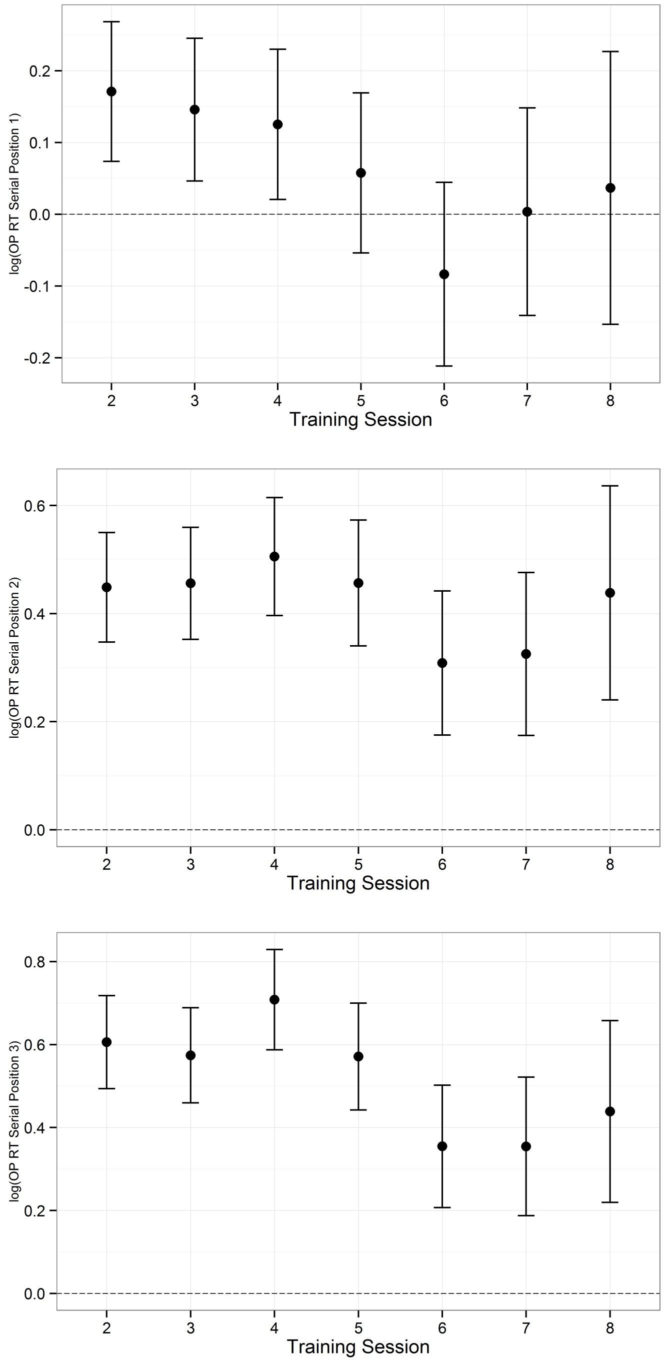


Figure 2.8. Operation RT (ms) deviation from baseline as a function of session by serial position; Top - SP1, Mid - SP2, Bottom - SP3. Points indicate parameter estimates (adjusted mean) with 95 % confidence intervals.

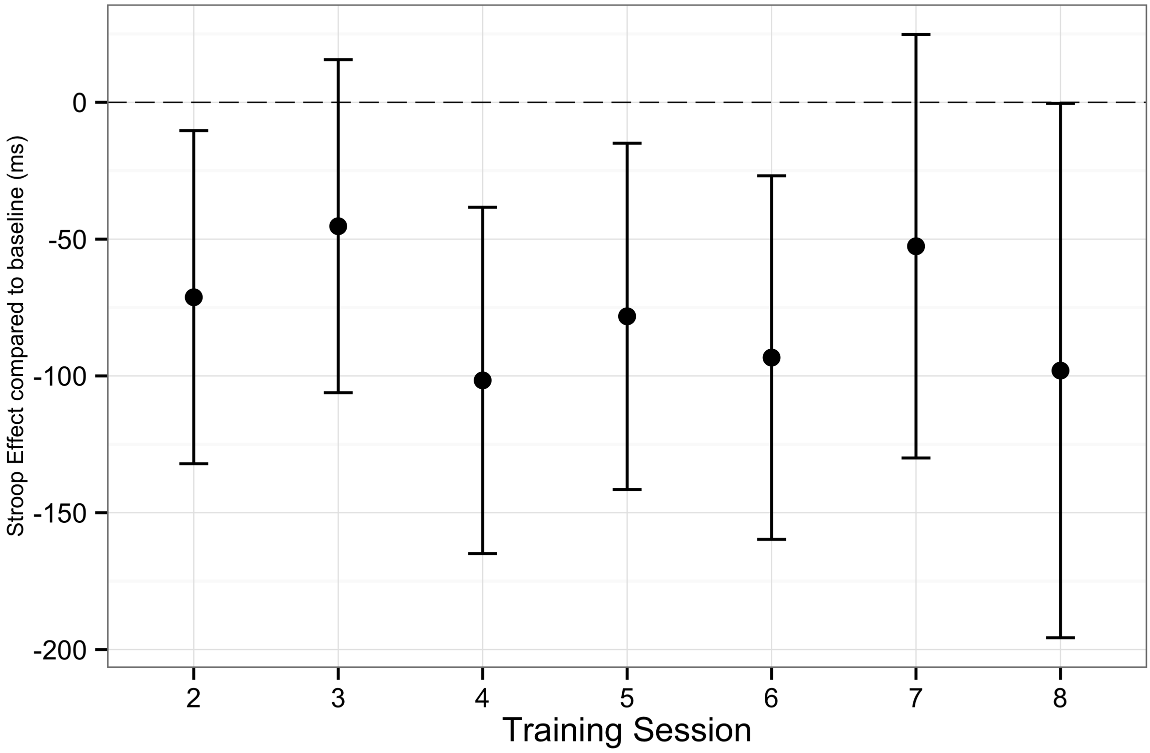


Figure 2.9. Stroop effect deviation from baseline as estimated by the GLMM with 95 % confidence intervals.

CC

The practice effects observed on the Colour Corsi task were minimal as shown in Figure 2.7.

Memory Update

Performance on the Memory Updating task was similarly unaffected by repeated training sessions (Figure 2.7).

2.3.3 Transfer Effects (RQ2/3

Participants who had completed less than 21 training tasks (7 full sessions) were excluded from the measurement of transfer as this was deemed too small an amount of training to expect any impact. This resulted in six participants being removed from the analysis stage for these measures (all from school two where IT failure was a common occurrence and hence a lower number of sessions in general compared to school one). Therefore the resultant n of the training group was 23 (control $n = 25$). One additional participant was excluded from Operation Span analyses due to lost data of their post-training session so the training group n for analyses involving OS is 22.

Exploratory analysis of transfer

Near-Transfer

Table 2.1 show basic summary statistics for the critical measures from the near-transfer tasks for both groups at each time point. It is clear to see the substantial variation between participants on many of these measures at all time points.

Matrix Span The values for the matrix span seem to support an improvement between pre-training scores and follow-up scores on each of the three measures, but for both the training and active control groups. The post-training scores provide a quirky element to

Table 2.1

Mean and standard deviation values for pre, post, and follow-up near transfer measures. Accuracy = proportion correct, RT = response time in seconds.

			Mean (sd)		
			pre	post	follow-up
Matrix Span	FTA Score	Training	27.28 (5.77)	29.11 (8.22)	31.47 (7.13)
		Control	30.0 (9.81)	27.18 (7.69)	30.52 (10.34)
	Max Span	Training	5 (.74)	5.22 (.67)	5.59 (.51)
		Control	5.28 (.68)	5.12 (.83)	5.55 (.6)
	Recall RT	Training	1.4 (.36)	1.28 (.34)	1.15 (.31)
		Control	1.34 (.32)	1.34 (.53)	1.19 (.27)
Operation Span	FTA Score	Training	10.45 (6.53)	14.73 (6.42)	17.69 (8.2)
		Control	12.12 (6.62)	15.44 (5.5)	18.86 (7.66)
	Max Span	Training	2.64 (1.22)	3.09 (.81)	3.35 (.79)
		Control	2.72 (1.06)	3.32 (.69)	3.52 (.75)
	Op Accuracy	Training	.85 (.15)	.82 (.15)	.86 (.1)
		Control	.82 (.19)	.83 (.15)	.86 (.12)
	Op RT	Training	7.06 (3.17)	5.12 (1.18)	5.38 (1.61)
		Control	6.99 (2.4)	6.03 (1.28)	5.79 (1.49)

the data in that the control group drops in performance by an average of 2.75 points on FTA score while remaining approximately static on the recall RT measure which is an averaged RT measure for time taken to give response (overall response RT is divided by set span on that trial i.e. number of clicks required to give the response). Although given the relatively large standard deviations involved these deviations in scores, both increases and decreases, are highly likely to be noise as opposed to systematic effects. In fact, the presence of a decrease of approx. 2.75 in the control FTA score should act as a warning with regards to how it may be tempting to interpret an increase of the same magnitude (regardless of p-value). Figure 2.10 shows boxplots for FTA scores for the MS task in the top-left plot which illustrates the different profiles of FTA scores which is primarily driven by the control groups post-training score.

Operation Span The values for Operation Span in Table 2.1 show a similar story. There is generally an improvement in performance across the repeated testing for both groups. Figure 2.10 illustrates the FTA scores in graphical format (bottom-left).

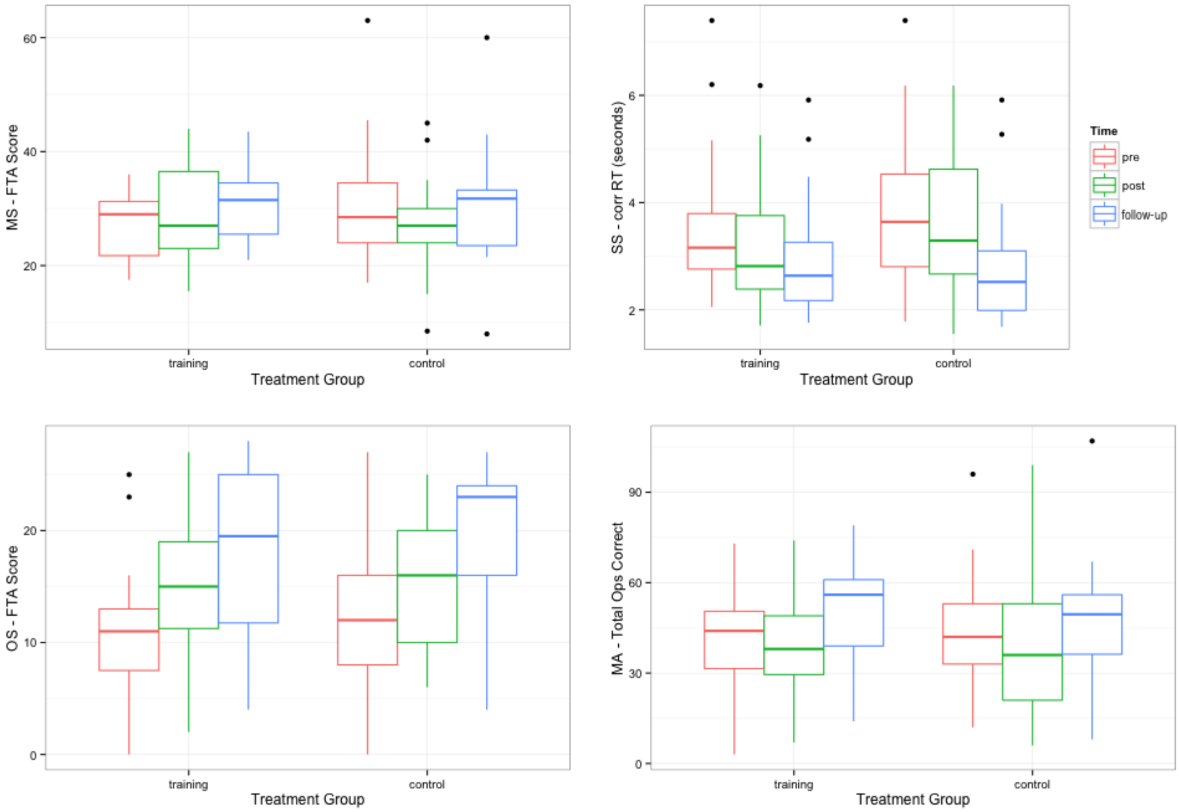


Figure 2.10. Key boxplots for the near- and far-transfer measures. MS = Matrix Span, OS = Operation Span, SS = Silly Sentence, MA = Mental Arithmetic.

Far-Transfer

Table 2.2 shows the basic summary information for the critical measures from the far transfer tasks for both groups at the various time points.

Table 2.2

Mean and standard deviation values for pre, post, and follow-up far transfer measures. Accuracy = proportion correct, Response Time = seconds

			Mean (sd)		
			pre	post	follow-up
Silly Sentence	Accuracy	Training	.93 (.07)	.96 (.04)	.97 (.04)
		Control	.95 (.09)	.94 (.1)	.96 (.04)
	Response Time	Training	3.48 (1.34)	3.21 (1.43)	2.88 (1.19)
		Control	4.04 (2.13)	3.6 (1.35)	2.94 (1.65)
Mental Arithmetic	Attempted	Training	53.04 (15.28)	65.83 (22.22)	62.24 (17.74)
		Control	61.12 (19.32)	61.28 (26.52)	58.65 (17.65)
	Accuracy	Training	.75 (.24)	.65 (.27)	.79 (.13)
		Control	.75 (.24)	.69 (.29)	.76 (.21)

Silly Sentence As one would expect the accuracy scores are extremely high to begin with and this is maintained (if not marginally improved for the training group at post/follow-up). The speed at which participants made their judgements showed reasonable improvement but this was the case for both groups. Figure 2.10 (top-right) shows boxplots summarising SS RT data.

Mental Arithmetic The values for the MA task in Table 2.2 suggest a sizeable spike in the overall number of questions attempted after the training intervention which is a measure of speed of operation. Conversely, the accuracy (proportion correct) exhibits a decrement for this group which would temper the enthusiasm of the previous statement. The overall number of trials attempted is a sensible measure of how fast participants were answering operations as all participants had 6 minutes in total (6 x 1-min blocks). Figure

2.10 (bottom-right) illustrates the somewhat strange profile of accuracy scores suggesting an overall improvement in both groups between pre and follow-up but a decrement at post-training for both. The number of correctly answered operations are shown per block, per group, in Table 2.3. Recall that the blocks differ in terms of the type of operations given to participants. Block one and two were addition blocks where the operations in the first block did not require a carry in the tens index (e.g. 3+4), while block two required carry (e.g. 4+8). This additional demand on manipulating held numerical memory representations appears to present itself in a small, but consistent across the time points, difference. The difference between subtraction blocks without carry (block 3) and with carry (block 4) appears to be much larger.

Table 2.3

Mean number of correct responses for each block of Mental Arithmetic questions.

		Mean (sd)		
		pre	post	follow-up
Block 1	Training	9.39 (3.64)	9.39 (3.49)	10.47 (3.47)
	Control	9.6 (3.93)	8.8 (3.33)	9.65 (4.59)
Block 2	Training	7.57 (2.66)	6.61 (2.9)	7.35 (3.39)
	Control	8.16 (3.26)	6.76 (3.83)	7.5 (3.65)
Block 3	Training	8.65 (4.05)	9.04 (3.71)	10.94 (9.44)
	Control	9.44 (4.77)	7.64 (5.16)	9.2 (4.98)
Block 4	Training	2.78 (3.2)	2.0 (2.83)	4.29 (3)
	Control	3.56 (3.24)	2.96 (3.52)	3.45 (2.98)
Block 5	Training	7.3 (3.5)	7.78 (3.06)	9.12 (3.94)
	Control	7.72 (3.54)	7.48 (3.8)	8.55 (4.24)
Block 6	Training	5.09 (3.86)	5.26 (4.61)	8.12 (3.97)
	Control	5.76 (3.59)	5.36 (4.1)	7.85 (4.88)

Formal analysis of transfer

Table 2.4 shows the Wilks test statistic, an approximate F value and the corresponding p-values for the effect of group in a MANCOVA analysis with the post—follow-up

Table 2.4

Overview of Omnibus Mancova results for transfer measures; ($[Post, Follow - up] \sim Pre + group$); Λ = Wilk's Lambda statistic, F statistic, and associated p-value.

		Λ	F	p
Matrix Span	FTA Score	0.91	$F(2,33) = 1.74$	.19
	Max Span	0.98	$F(2,33) = 0.31$	.73
	rt_click	0.97	$F(2,33) = 0.49$	.62
Operation Span	FTA Score	1	$F(2,33) = 0.019$	.98
	Max Span	0.99	$F(2,33) = 0.26$	.77
	Op Accuracy	0.99	$F(2,33) = 0.14$	.87
	OP RT	0.96	$F(2,33) = 0.73$	.49
Silly Sentence	Corr RT	0.93	$F(2,33) = 1.27$	.29
	Accuracy	0.92	$F(2,33) = 1.4$	.27
Mental Arithmetic	Number of Operations	0.89	$F(2,33) = 1.9$	.17
	Accuracy	0.91	$F(2,33) = 1.55$	.23

scores forming the dependent variables and the pre-training scores as a covariate. Rausch, Maxwell, and Kelley (2003) suggest that this approach is the best 'omnibus' approach to assessing if the groups differ in any way over time. As can be seen by the respective p-values there is no suggestion of a 'statistically significant' effect of group on post-training or follow-up scores. The next step in this analysis could be to 'unpack' a significant effect by carrying out the respective ANCOVA analyses on the single dependent variables (again pre-training scores as covariate). I have in fact conducted these analyses for two reasons. Firstly, the MANCOVA requires a complete dataset therefore the 11 participants we were unable to test at follow-up are not included. When assessing pre-post effects these 11 participants can be included and therefore the power of the ANCOVA with post scores as DV with pre scores as a covariate will be increased for assessing that effect. Secondly, to calculate effect sizes (Hedges' g) and Bayes Factors (BF) to assess the weight of the evidence.

Table 2.5

Overview of subsequent Ancova analyses and effect sizes for near-transfer measures; d = Cohen's d, g = Hedges' g, BF = Bayes Factor.

	Variable	Time	F	p	d	g	BF
Matrix Span	FTA Score	post	$F(1,45) = 2.5$	.12	0.4	0.39	0.71
		follow-up	$F(1,34) = 1.37$	.25	.26	.26	0.52
	Max Span	post	$F(1,45) = 0.98$	.33	0.26	0.26	0.41
		follow-up	$F(1,34) = 0.46$	.5	0.21	0.21	0.37
	rt_click	post	$F(1,45) = 0.42$	.52	0.17	0.17	0.31
		follow-up	$F(1,34) = 0.97$	.33	0.29	0.28	0.46
Operation Span	FTA Score	post	$F(1,44) = 0.0001$	.99	0	0	0.29
		follow-up	$F(1,34) = 0.013$	.91	0.03	0.03	0.33
	Max Span	post	$F(1,44) = 1.098$	.3	.29	.28	0.45
		follow-up	$F(1,34) = 0.395$	.53	0.19	0.19	0.37
	OP Accuracy	post	$F(1,44) = 0.405$	.53	0.17	0.17	0.34
		follow-up	$F(1,34) = 0.054$	.82	0.07	0.07	0.32
	OP RT	post	$F(1,44) = 2.235$	.14	0.43	0.42	0.43
		follow-up	$F(1,34) = 0.017$	.9	0.04	0.04	0.33

Near-Transfer ANCOVA results Table 2.5 summarises the results from the various ANCOVA analyses for the near-transfer dependent variables. There are no statistically significant results found. The effect size measures (Cohen's d and Hedges' g presented) range from 0 to 0.4 while the BF values are all very small and in no way support the notion that these data show positive effects of the WMT intervention on these transfer tasks.

Far-Transfer ANCOVA results Table 2.6 summarises the results for the ANCOVA analyses relating to far-transfer DVs. Again there is no evidence provided for an effect of the training intervention. The effect sizes range from 0.11 to 0.51 and the BF values are small (even the MA values are too small to legitimately support an inference to the population).

Table 2.6

Overview of subsequent Ancova analyses and effect sizes for far-transfer measures; d = Cohen's d, g = Hedges' g, BF = Bayes Factor.

	Variable	Time	*F*	*p*	*d*	*g*	BF
Silly Sentence	corr RT	post	$F(1,45) = 0.21$	.65	0.12	0.12	0.39
		follow-up	$F(1,34) = 2.579$	.12	0.37	0.36	0.65
	Accuracy	post	$F(1,45) = 1.598$	.21	0.32	0.32	0.53
		follow-up	$F(1,34) = 2.66$	.11	0.51	0.5	0.75
Mental Arithmetic	Total Correct	post	$F(1,45) = 3.52$	.07	0.28	0.27	1.13
		Follow-up	$F(1,34) = 5.71$	.02	0.44	0.43	2.52

2.3.4 Assessing the active control group (RQ5)

The three tasks given to the active control group at each training session were all self-paced, with regards that if a participant did not respond to a trial stimulus then it would not automatically move on after any given time period. This procedure lends itself to a basic but effective method of assessing if participants maintained a reasonable level of engagement with the tasks for the duration of the intervention phase. If the number of trials attempted per session remains stable (or increases) in tandem with the accuracy not decreasing significantly then it would be reasonable to conclude that participants maintained their effort levels as session number progressed.

As there is some attrition in the control dataset as already discussed regarding the training group's data (different numbers of sessions completed by Ps) then a generic table of descriptive statistics would not suffice. Therefore the same type of model used to assess the practice effects was applied to assess the impact of session number on the amount of trials completed and the most appropriate accuracy measure for each control task. The parameter estimates of the session factor on these measures are summarised in figure 2.11. The columns represent the values relating to number of trials attempted with the Dots Task on the top row, Number Line task in the middle, and Subitizing task on the bottom row. These plots show that there was no drop-off in number of trials attempted

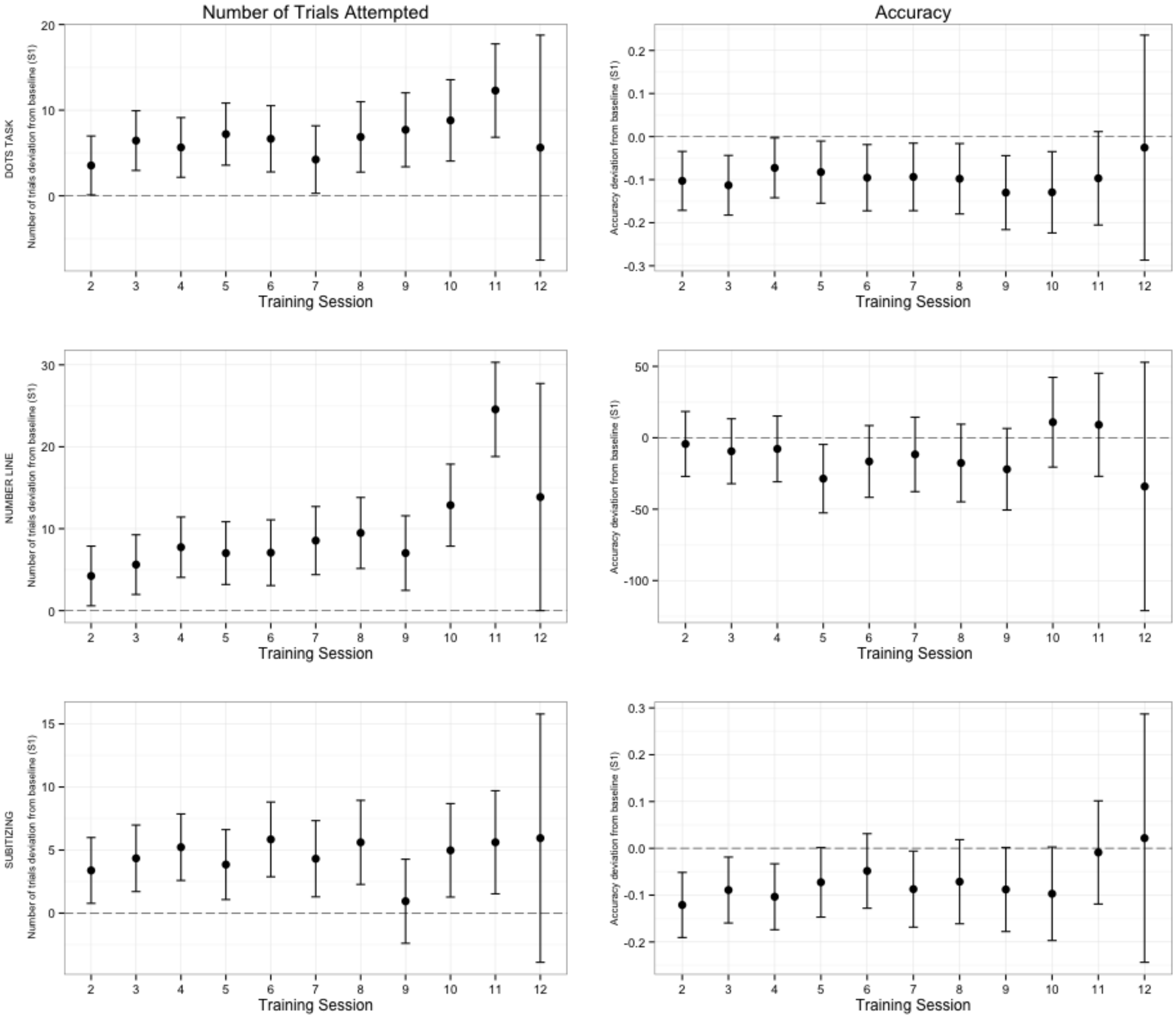


Figure 2.11. Assessing attentiveness to control tasks with measures of trials completed and accuracy. Parameter estimates with 95% confidence interval from GLMM analysis.

as the intervention phase progressed. All three tasks show a similar pattern; a slight jump in number of trials from the first to the second session and then relatively stable. The second column shows the effect of session number on a measure of accuracy. For the dots task and Subitizing task the measure of accuracy is simply the proportion of correct responses. The number line task measure of accuracy is the average deviation from the target number in each session where a lower score represents greater accuracy. We can see from these three plots (second column, figure 2.11) that accuracy does not show a pattern of decreasing over time. There is a curious drop in accuracy from session one to session two for both the dots task and the number line task, this decrement then holds over the future sessions. This is likely due to the motivational aspect provided in these tasks where in each session they were given constant feedback on accuracy and speed of response. The fastest correct response was recorded in the top right of the screen as a mark to beat. Perhaps the inverse relationship between number of trials completed and accuracy between sessions one and two is explained by the participants becoming accustomed to this mechanic. For example, perhaps in session one Ps were focused on accuracy and ensuring they were completing the task appropriately but by session two they had confidence in understanding the goals of the task and shifted focus to being as fast as possible.

2.3.5 Post-Hoc Power Analyses

Post-Hoc power analyses were computed to estimate how many participants we would have required for a satisfactory level of power (80%). Sample size estimates were produced under two scenarios; a conservative estimate based on mean effect size reported in near-transfer meta-analyses (Melby-Lervåg & Hulme, 2013) and a less strict estimate based on larger effect sizes reported by (Klingberg et al., 2005).

Mean effect size for near-transfer to verbal and visuospatial WM was 0.79 and 0.52 respectively. For 80% power we would need 27 participants in each group for verbal WM assessments and 60 participants in each group for visuospatial WM.

The effect sizes found by Klingberg and colleagues were generally of a larger magnitude for near-transfer. Klingberg et al. (2005) report a 0.93 estimate for transfer to a span-board task. This represents one of their more conservative near-transfer effect sizes. Using this value 20 participants per group would be required in order to yield 80% power.

2.4 Discussion

2.4.1 P1: Performance Change on the trained tasks

In many WMT studies the training program is outlined and the number of sessions participants actually completed is noted but then the results section moves straight to the transfer effects (e.g. Westerberg et al., 2007; Bergman Nutley et al., 2011; Holmes et al., 2009, 2010). But the change (if any) seen in the trained tasks is surely of great interest to researchers when interpreting the potential benefit of such interventions. The results we obtained in this study suggest that beyond the first few repeated sessions of the tasks we included in our battery, performance did not improve significantly. It is interesting to note that the control tasks showed a similar pattern where there were some performance shifts in the very early sessions but none beyond that. As alluded to earlier the lack of practice effects does not make any transfer effects invalid, but it does make the interpretation of such effects more difficult. The recent surge of WM training intervention studies is primarily because of the consistently found predictive power of WM 'span' on higher order cognitive tests. The notion is such that if ones WM span is such a key factor in determining higher order faculties then if it was found that the span limit for a person is not static and can be altered with repeated practice then this would have wide reaching cognitive benefits. Showing the wider reaching cognitive benefits without any significant improvement on the training tasks would therefore be a puzzling result and not readily explained by the 'simplest' explanation.

At least with regards to the primary indicators of performance level shown so far there appears to be a consensus amongst these tasks that there is a small performance increase seen in the early sessions but that this upwards trajectory quickly stalls. These results are somewhat surprising, while it would not have been sensible to presume that performance would simply continue to improve in a linear fashion over any number of repeated sessions (hence treating session as a factor rather than a continuous variable), it would have been expected that performance improved on the trained tasks more than we see. This leads to important implications with regards to the overall effectiveness of a working memory training program in that if participants are not showing improvement on the trained tasks then how reasonable is it to expect that any changes will be seen in the transfer measures, and if such change was observed in the transfer measures it is somewhat more difficult to explain.

One significant issue regarding the study we present here is that due to selecting a battery of training tasks and rotating these tasks per session combined with some scheduled sessions lost due to extraneous factors we ended up with less data than is typical of a training study where researchers may have collected data on 15 or more sessions on their training task/s. However, the results we obtain do point to the question - what would have been the benefit of further sessions? In that performance on each task seems to have stabilised by session 3/4 showing no additional improvement beyond that.

2.4.2 Generalisable benefits of the intervention (Transfer Effects)

The four tasks we selected to be transfer measures in this task were selected to provide a mixture of possible transfer benefits. Firstly, two tasks that assess near-transfer as they are primarily WM based tasks and are similar in make up to some of the tasks that form the training battery. The matrix span task is a visuospatial short-term storage task that is procedurally the same as the Colour Corsi training task without the need to bind a colour to each location. The operation span task shares similarities with the working memory period training tasks in that the short term retention of verbal information is being tested while resource-demanding additional tasks are also attended to.

Near transfer effects are commonly reported in the training literature (e.g. Klingberg et al., 2002, 2005; Ball et al., 2002; Holmes et al., 2009, 2010; Beck, Hanson, Puffenberger, Benninger, & Benninger, 2010). However, the results here suggest that there were no near-transfer effects due to our training battery. It is somewhat surprising that the training group would not show an advantage on these tasks given how closely they match up to a number of the training tasks in the battery. Perhaps the lack of improvement on the training tasks points to an inability in this group to develop mnemonic and/or other strategies to aid performance, the type of which may have transferred over to these untrained tasks. The issue of whether improvement is a product of increased resources or strategy utilisation will be picked up in more detail in the general discussion.

Additionally, there were no observed far-transfer effects either. Speed on the silly sentence task improved at each successive time point but at that same rate for both groups. Speed of mental arithmetic operations showed a somewhat erratic trend due to a decrease in both groups from pre-post. The number of successfully answered questions across the six blocks of questions presented in the Mental Arithmetic task showed trends towards a significant effect of the intervention but considering the sample sizes, effect sizes, and Bayes Factors seen it would be unreasonable to conclude that this is anything but very weak evidence.

The approach we wanted to take when tackling a WM training intervention study was to operationalise the specific processes that were involved in the training tasks as well as the transfer tasks and assess at a more fundamental level what changes might be brought about by such an intervention. However, the results obtained offered no evidence for any change at all as a result of the intervention; processing speed, efficiency, interaction with LTM, and short-term storage of memory representations were all processes assessed and none showed a significant change.

2.4.3 Potential impact of administration environment

The impact of our naturalistic environment on the study was profound from a practical and technical perspective. The reliance on computers that we had little control over, the rigidity of when sessions had to be completed, and the group administration all played a role in the amount and quality of the data acquired.

On any given scheduled training session there would be a number of computers that would fail in one way or another and led to a number of children missing out on a given session. The schools both had a similar IT system in that they had a number of laptops that was (approximately) the same number as a typical class. Therefore for every computer that we had an issue with on any given day, a child was unable to complete the training session for that day (there was no scope for making up for these sessions at other times). The reasons for computer issues were numerous but the two most significant reasons were; a) a problem from its previous use, and b) erratic compatibility with python. Reason (a) is due to the communal aspect of the laptops within the school, if the laptops had previously been used by a class and not put away properly some may be unavailable for a period of time when they are next used (e.g. frozen or dead battery due to not being plugged back in to the charging port). Issue (b) was an extremely strange issue in that on one day a specific computer would execute the python software without a hitch but the following day the very same computer would be unable to execute the software. There appeared to be no pattern to when this problem occurred. Therefore I was unable to fix it during the course of the study hence when this issue occurred the effect was another lost session for a participant.

Despite these issues, enough data was collected to conduct some meaningful analyses and to some degree assess the potential impact of the training intervention.

2.4.4 Sustained attention of control group

The integrity of the control group was of paramount importance in the design of this study. The results of the control group data validate the approach we took in selecting non-memory based but commonly used computerised tasks related to cognition. At face value, while our control group could clearly see that they had classmates who were doing different tasks it certainly didn't come across as though what they were doing was meaningless, as might be the case when one has a control group who play generic video games or answer general knowledge questions (e.g. Kerns, Eso, & Thomson, 1999; Jaeggi, Buschkuehl, Jonides, & Shah, 2011).

These data show that the participants in our control group did not attend less to later sessions compared with earlier sessions. This allows a greater degree of confidence that the control group participants were in fact an 'active control group' and can be matched to the training group for extraneous factors such as time spent participating in research, amount of time spent attending to computerised tasks, and any other related factors.

Chapter 3

Task Validation and Scoring Comparisons

3.1 Introduction

Several thesis chapters discuss complex studies that describe and interpret training studies for working memory capacity. The scale of these studies is such that it is not feasible to carry out assessments of how measures should be scaled, implemented, or related to each other. These form objectives of the current experiment and chapter.

The implementation of the Working Memory Period (WMP) task as a training tool in the previous experiment and in subsequent training studies in this thesis presents an opportunity to examine its properties in more detail. The working memory period paradigm presents an interesting addition to the investigation of WM training due to the way in which difficulty is manipulated in contrast to those manipulations of more often used span tasks. However, there are limited data available to assess the properties of the WMP task particularly with adult participants. One such important property of interest is the relationship between WMP and Operation Span performance as indicators of the convergence/divergence between the two. This has implications for the general use of a period-span task in general WM investigations and also to inform our suggestions regarding the degree to which any transfer between WMP and traditional verbal span measures falls on the near-near to far-near scale. The working memory period task was introduced by Towse et al. (2005) and was found to be related to more traditional complex span tasks (operation and reading span) via correlational analyses. However, while significant, these correlation patterns were of a modest magnitude thus highlighting the degree to which different processes are being recruited in each of these tasks.

As there are a selection of different tasks one can select as a measure of WM, there are then a set of scoring methods that one can apply to extract a measure that represents how the participant performed. Scoring methods can deviate from one another in different ways. Firstly, psychometric properties with regards to reliability, distributional properties, discrimination (the measure must be sensitive enough to tease apart different levels of participants ability). These properties are to some extent necessary of a good measure and are therefore important for all tasks/scoring methods. Additionally, there is the consideration that some scoring methods may reflect different concurrent processes that participants engage in to complete the WM task.

St Clair-Thompson and Sykes (2010) administered a battery of 5 STM/WM tasks to a group of 7-8 year old children and also obtained scholastic measurements for each child from their school for maths, reading, writing, and science. The authors were primarily interested in the difference in predictive power of absolute scoring and proportion correct methods. In absolute scoring participants are given a number of points for a trial where all to-be-remembered items were recalled in correct serial position. The number of points awarded is scaled based on the list length e.g. a fully correct trial with list length three yields three points towards the absolute score. The proportion correct method credits individual units with trials recalled in the correct serial position even if some units within the trial were not successfully recalled. The measure is derived by scoring each trial on the number of correctly recalled items divided by list length and then averaging over all trials. They found that the proportion correct scores of the STM/WM tasks often explained unique variance in the scholastic attainment scores after controlling for the absolute scores.

A later study (St Clair-Thompson, 2012) examined differences between these two scoring methods but also considering an administration manipulation, namely whether list lengths (LLs) of trials were given in ascending order or randomised. St-Clair Thompson found that when administered with LLs randomised scores on the Counting Span task were significantly different but not for the Reading Span task. More importantly however were the results suggesting that only the randomised versions of both tasks were significantly related to Raven's Advanced Progressive Matrices (RAPM), this was the case for both absolute and proportion correct scoring methods. This followed the work of Unsworth and Engle (2007b) who suggested that proportion correct methods of scoring WM tasks produced higher predictive power on criterion tasks of STM/WM as these methods benefit from using data from higher list length trials that are often not recalled in entirety, and these trials tap secondary memory due to primary memory reaching capacity. Therefore the proportion correct scoring method measures primary memory as well as a contribution from secondary memory which is absent in absolute scoring and would also be absent in a max span measurement. Thus, while measurements of WM ability are used in a large amount of research programs there is clear evidence that the assessments one makes regarding a persons or groups WM ability can be significantly influenced by decisions regarding the administration of the task (i.e. list length order, termination or not, etc.) and also by the method one chooses to produce scores on the resultant data.

These differences in outcomes based on procedural and scoring variations have been used to discuss the nature of WM functioning. For example Lustig et al. (2001) suggest that the boost in performance for longer list length trials when a descending order of LLs was used is evidence for the important role of suppressing proactive interference in tasks of STM/WM. This is due to the longer LLs being conducted when there have been few previous trials and thus a much smaller pool of potential interfering stimuli from these trials whereas generally the longer LL trials are conducted when the interference pool is at its greatest and are therefore greater influenced by the relevant interference mechanisms in such processing.

In the previous chapter the WM outcome measures considered from the transfer battery were absolute scoring and max span. As noted here, there are other potential scoring methods that have been the focus of discussion and shown to hold different properties. Therefore as well as obtaining data to further inform our conclusions based on the properties of the WMP task in adults and its relationship to other WM tasks, we will also take the opportunity to assess alternative scoring methods on those tasks and how these may be used in the future WMT studies to improve evaluation of possible WM change.

Many working memory training studies use a memory span score when assessing near-transfer (e.g. Klingberg et al., 2002; Holmes et al., 2009). These scores are often treated as an interval variable which would imply that differences of the same magnitude are equivalent across the range of possible scores. However, perhaps this assumption is not suitable when operationalising the measure in terms of memory span improvement as a result of training. Through the secondary analysis of a large dataset using the Rasch model we will quantify the difficulty of different list length items on three complex span tasks. By using the Rasch model which is in the Item Response Theory family of models we take into account the ability of each participant in the dataset and obtain more nuanced measures of the difficulty of items compared to simple proportion correct methods.

3.1.1 Outline of the goals of this chapter

1. The working memory period task was successfully used as a training task in our previous WMT experiment. However, data using a working memory period paradigm are limited, and we are not aware of data from adults. Thus by administering the WMP task to an adult sample in this study using a standardised administration we aim to profile the general pattern of performance of adults on this task. Therefore we wished to collect information on the WMP task in addition to various other indicators of WM. This will provide useful validation of the WMP task as well as allow a further investigation into the properties of the WMP task as a measure of WM ability with regards to the patterns of shared variance between WMP, OS, and visuospatial related measures. Of critical importance are the differences that emerge as a result of increasing levels in the WMP task. How analogous are level increases between OS and WMP where the mechanics of a level increase differ with regard storage and processing requirements. An increase in level in the OS task adds one extra TBR item to the array while also adding an additional processing component that thus adds two shifting components.

2. In the WMT studies that follow we are able to administer a more substantial battery of pre-post transfer measures and therefore we are able to adhere to the suggestion of Shipstead and colleagues (2010, 2012) of having multiple tasks per construct measured for transfer. Selection for pre-post tasks was based significantly on the findings of (Kane et al., 2004) in order to select related tasks that would form a suitable factor for the constructs we wished to test. As part of a visuospatial WM measure it was decided to include the rotation span task in the battery. Task development and small scale pilot work revealed that rotation span was hard and recall accuracy was weak. Yet, we lacked evidence as to why the task was especially hard. Therefore, to assess if floor effects are a significant concern if we were to use this task we administered the arrow span task which is the simple span equivalent for rotation span. The simple span equivalent was used for a number of reasons; a) quicker to administer, b) better equivalent to Colour Corsi for comparison purposes (as Colour Corsi is mechanically like a simple span task too albeit requiring the binding of features), and c) potential for comparison to rotation span data down the line to assess what makes the task difficult, the processing component or storage of more visually rich stimuli.

3. While colour-location binding tasks are not new, our specific implementation of it as a matrix span task with a colour component (requiring serial recall) is to my knowledge unique, especially within the WMT literature. Therefore the opportunity was taken to administer the Colour Corsi task as a standard span type measure with a stepwise list-length algorithm. This will increase our understanding of what a Colour Corsi based training regime would be training in addition to giving a reference to adults general ability at completing this task.

4. The procedural component of simple and complex span tasks that researchers often use of terminating administration when a participant fails to reach a specified level of performance on a given level will be evaluated. This practice has been common in the field since Daneman and Carpenter (1980) outlined their methodology in administering the reading span task where participants were given three trials at each list length where the list length would ascend provided at least one set at the current list length was successfully recalled in full. An alternative to this approach is to require participants be successful on a majority of the trials at a specific list length e.g. 2/3. The participant can then be given a score/span that is equivalent to the highest level where the majority of trials were successful and then an additional fixed amount may be added if some trials at the higher list length were successful. The assumption made by using this procedure is that a participant will not be successful at higher list length trials once they have been unsuccessful at the majority of trials at a lower list length. The degree to which this assumption holds for the various tasks administered in this study will be assessed.

To help address these points we recruited participants to complete the Working Memory Period, Arrow Span, Colour Corsi, and Operation Span tasks.

5. By applying the Rasch model to a large dataset of three complex span tasks we will assess if increases in span size can be considered equivalent, and if they cannot then the implications for using memory span measures in WM training experiments will be discussed.

3.2 Method

3.2.1 Participants

Sixty-two undergraduate students from the University of Lancaster were recruited to participate in this study in exchange for course credit. The age range was 18 - 23 with mean 19.3 years and 42 were female.

3.2.2 Design and Materials

The tasks were all developed using the JAVA programming language and were built on the framework provided by the Tatool library (von Bastian, Locher, & Ruflin, 2013). The Operation Span and Rotation Span (of which Arrow Span is based) are available as part of a published WM battery (Stone & Towse, 2015).

Operation Span The Operation Span task involves requires repeated switching between storage and processing phases. The storage phase presents a digit (10-99) in the centre of the screen for 1000ms for the participant to later recall in the correct serial position. The processing phase presents participants with a mathematical operation to verify for accuracy e.g. 4 x 7 = 28. The processing operations were randomly selected (without replacement) from a pool 100 generated operations that were pre-generated using a Matlab script. In this process each type of operation was represented equally (addition, subtraction, division, and multiplication) and within operation type the proportion of correct and incorrect operations was 50%. Every storage phase was followed by a processing phase where the number of storage-processing pairs was equal to the list length of the current trial. At this point participants were prompted to input the numbers they remembered one at a time using the keyboard. A schematic diagram outlining the task was provided via Figure 4.5 in the previous chapter. Participants completed three trials at each list length from two through six.

Working Memory Period While the WMP task involves both processing and storage components they are not represented by different phases operationally. Participants were shown a mathematical operation built using only addition and subtraction components. The answer to the operation formed the TBR memoranda for the task. Rather than adjust the number of operations to increase difficulty, analogous to the operation span, the number of operations in a trial remains static. Instead the length of the presented operations was manipulated as a function of level (task difficulty). Thus a level increase represents an increase in the processing required to generate the TBR items. Each trial consisted of four operations and thus four digits were to be recalled (this was increased from three operations used in the previous experiment to increase difficulty). Examples of operations at levels one, two, and three are $4 + 5$, $3 + 6 - 4$, $2 + 2 + 5 - 6$ respectively. See Figure 4.3 in the previous chapter for further illustration of the protocol for WMP. Participants completed three trials at each level from one through six.

Colour Corsi The Colour Corsi (CC) training task involves the storage and retrieval of colour-location representations. The CC task is a matrix span task where the grid locations are highlighted with one of four possible colours and these colours also need to be recalled alongside the grid reference. A 3x3 grid was presented in the centre of the screen. A grid would fill in one of the four possible colours for 1000ms followed by a 500ms inter stimulus interval. The number of colour-location units presented and required for storage and recall was equal to the list length of the particular trial. See Figure 4.2 for an illustration of the CC task. The participants were given three trials at each list length between two and six.

Arrow Span The stimuli in the Arrow Span task are images of arrows that are differentiated in two characteristics; length (long (300px) or short (100px)), or it can differ in its angle of rotation (0°, 45°, 90°, 135°, 180°, 225°, 270°, or 315°). Each arrow was presented for 1000ms with a 500mx ISI. After all arrows had been presented a recall screen was presented. The recall screen presented the 16 possible arrows in a 2 x 8 grid where the top row of arrows was the short arrows and the bottom row the long arrows. Participants used the mouse to select the arrows they remembered seeing in the correct order. The participants were given three trials at each list length between two and six.

3.2.3 Procedure

Testing sessions were devised to support group administration up to 6 participants at a time were tested. The tasks were completed on 21.5" iMac computers.

Participants were welcomed and shown to their place where they were given a study information sheet and an informed consent form. After each participant had read the study information sheet and signed the informed consent form the researcher began to give verbal instructions regarding the four upcoming tasks. Once started the participants were able to work at their own pace through the tasks i.e. they did not wait for all other members of the group to finish task one before moving on to task two. The onset of each task began with detailed instructions and screenshots which served as a reminder of each task demands (additional to the previous verbal instructions). An instruction was included suggesting to participants that they should raise their hand and ask for assistance from the researcher if they were still unsure what the task was going to ask of them.

The order of tasks was fixed; WMP, CC, OS, and finally AS. The order of task administration was fixed due to the group testing policy. It was decided it would likely cause distraction if participants were able to see others completing different tasks. Due to individual differences the total duration of the experiment was different for each participant. Typically the total testing time was 30 minutes but some variation was seen resulting in those who were faster/slower than 30 minutes.

3.3 Results

Seven participants did not provide data on the AS task due to time constraints. Data for a participant were excluded if for a given task if they were unable to provide any successful trials. This resulted in the exclusion of one set of WMP data and two sets of OS data. A further four sets of OS data were excluded as the accuracy of operations was below 80%. Therefore overall n for WMP, OS, CC, and AS respectively was 61, 56, 62, 55.

3.3.1 Working Memory Period

The WMP task involves two sub elements; the processing of each operation and the maintenance of the results for recall at the end of the trial. Table 3.1 shows descriptive statistics for the WMP task. An initial observation from these data is that while almost all participants correctly answer a level 6 trial (the highest level administered) there are many errors along the way, as evidenced by the mean full trial accuracy (FTA) score of 32.77 where 63 is the maximum possible. The FTA score is the absolute scoring method described previously; I use FTA for the added descriptive properties. Thus the FTA score for WMP only gives credit for trials where all four digits were correctly recalled and the number of points awarded for a successful trial is equal to the level of that trial. The 'termination span' variable reflects the average 'level/span' score that would have been attributed to participants if administration ceased when unsuccessful on more than one trial at a particular list length. Max Span (1) and Max Span (2) reflect the level/span score attributed to the participant is simply the highest level they provided a correct trial at, or the highest level they provided at least two correct answers. Taken together these scoring methods seem to show that while there is a decreasing level of accuracy as level increases, mistakes at lower levels do not necessitate that the person will fail at higher levels. Only 17 of the 61 participants were unable to successfully recall at least one level six trial despite the mean level of termination span being 2.66.

Table 3.1

Descriptive statistics for the Working Memory Period task.

Variable	Mean	Median	std. error	Min	Max
FTA Score	32.77	32.00	13.47	4.00	57.00
Max Span (1)	5.49	6.00	1.01	2.00	6.00
Max Span (2)	4.74	6.00	1.65	0.00	6.00
'Termination Span'	2.66	2.00	1.83	0.00	6.00
Proportion Correct	0.81	0.83	0.12	0.47	0.97
Level 1 Accuracy	2.31	2.00	0.81	0.00	3.00
Level 2 Accuracy	2.23	2.00	0.92	0.00	3.00
Level 3 Accuracy	1.61	2.00	0.92	0.00	3.00
Level 4 Accuracy	1.51	2.00	1.06	0.00	3.00
Level 5 Accuracy	1.38	1.00	1.04	0.00	3.00
Level 6 Accuracy	1.38	2.00	1.04	0.00	3.00

With regards to the time taken to answer operations there is a clear trend towards higher level operations taking longer to answer, see Figure 3.1. However, there is not a trend towards operations in the latter serial positions (see Figure 3.2) taking more time to answer which one may expect for a variety of reasons (picked up in the discussion). In Figure 3.2 we can see that serial position one is actually position that results in the largest RT with a seemingly equivalent value across the remaining serial positions, although some trend upwards.

To formally test the effects of level and serial position on the time taken to answer operations in the WMP task the log of the response time was used as a dependent variable in a gaussian mixed effects regression model with an added random effect for participant . The unit of analysis is trial response time and within-participant variation is accounted for by including a random-intercept effect for each participant. The fixed effects of level and serial position will be added individually, and then combined. Of particular interest given the observed pattern of descriptive statistics will be if the serial position effects differ after the effect of level is controlled for. First, two models with intercept and level as fixed

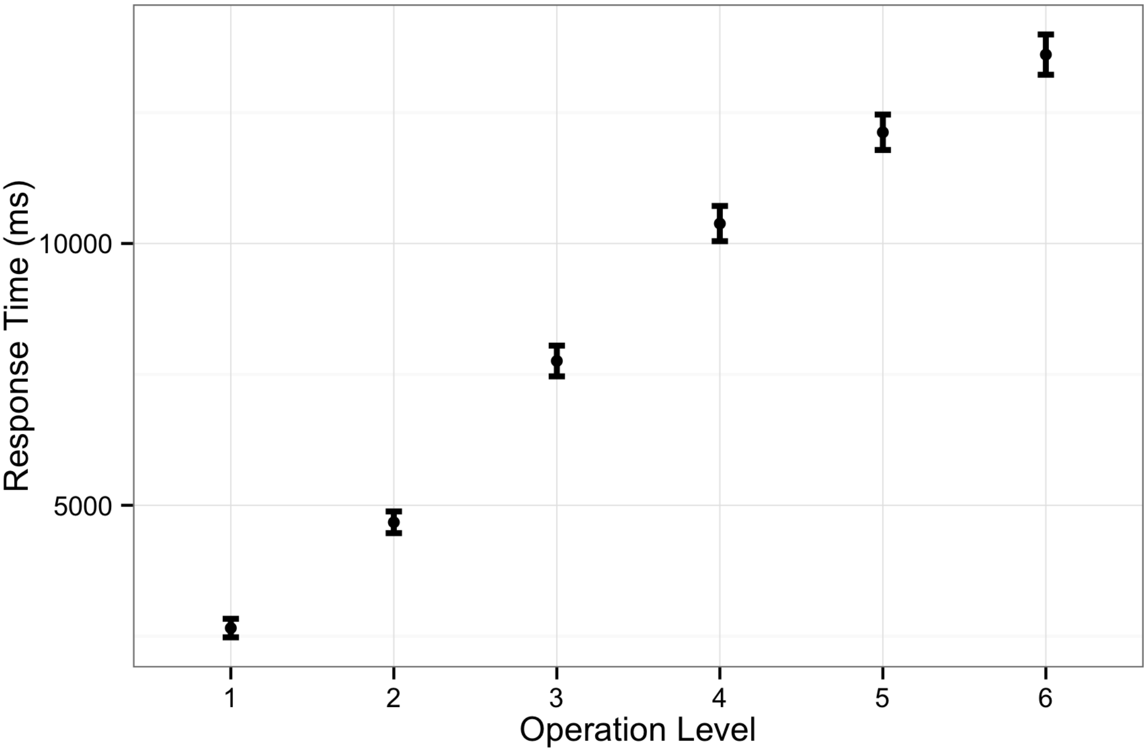


Figure 3.1. Mean response time (ms) for each level of WMP operation (bars represent 95% confidence interval).

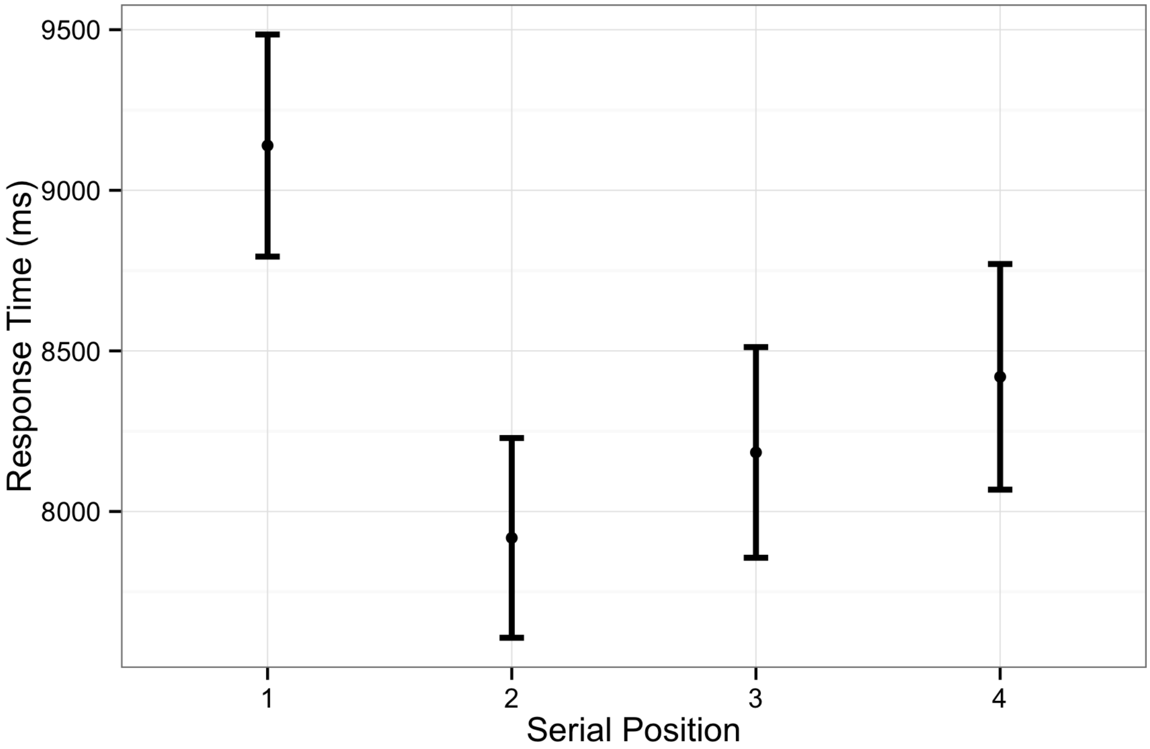


Figure 3.2. Mean response time (ms) per serial position on the WMP task (bars represent 95% confidence interval).

effects were computed - one with level as a categorical predictor and one with level as a continuous predictor. Given the observed pattern of RTs as a function of serial position (Figure 3.1) it may be reasonable and more parsimonious to treat it as a continuous predictor. As Table 3.2 shows, despite the added parameters to be estimated the model fit was substantially better with level treated as categorical, $\chi^2(4) = 408.09, p < .0001$. Serial position was only considered as a factor given the pattern of observations in Figure 3.2.

Table 3.2

Model fit values for the various combinations of fixed effects applied to the MLM analysis of WMP response times. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

Fixed Effects	AIC	BIC	Log-Likelihood
Intercept Only (Null)	10741.2	10760.3	-5367.6
Int + Level (factor)	7320.9	7371.9	-3652.5
Int + Level (continuous)	7721	7746.5	-3856.5
Int + Serial Position	10707	10745.3	-5347.5
Int + Level + Serial Position	7238.1	7308.2	-3608.1
Int + Level * Serial Position	7188.8	7354.4	-3568.4

The model fit statistics in Table 3.2 show that the addition of serial position as an explanatory variable is significant after controlling for the effects of level, $\chi^2(3) = 88.8, p < .0001$. The subsequent addition of an interaction between these variables also significantly improves the model fit, $\chi^2(15) = 79.3, p < .0001$. Table 3.3 shows the parameter estimates from the best fit model (interaction params excluded), each factor level is compared to the base category (1). The model confirms that RTs increase considerably as a function of level, as well as serial position one producing the longest RTs. Note that serial position two was set to be the comparison category for serial position and thus the non-significant values for positions three and four reflect neither differed significantly from position two. The RTs per serial position were re-plotted grouped by level in Figure 3.3.

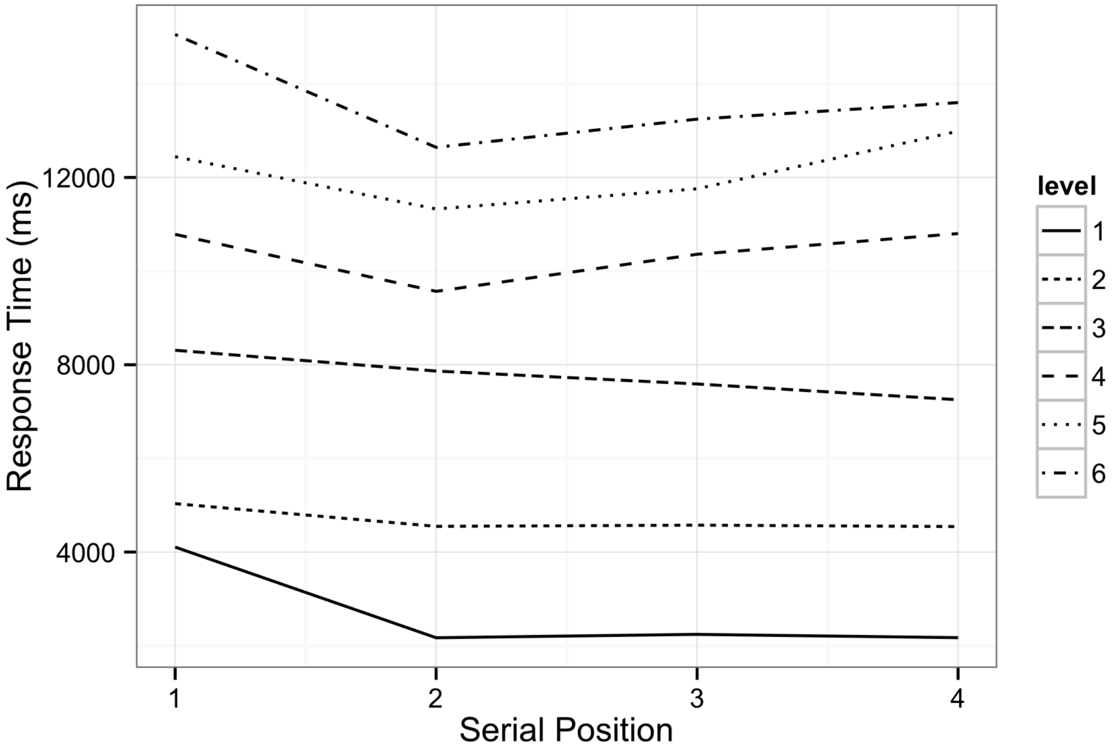


Figure 3.3. Mean response time for WMP operations at each serial position grouped by level of operation.

Table 3.3

Parameter estimates for fixed effects for the MLM analysis of WMP response times (log scale).

Fixed Effect	beta value	std. error	t-value
Intercept	7.57	0.05	159.86*
Level (2)	0.71	0.06	12.53*
Level (3)	1.22	0.06	21.53*
Level (4)	1.52	0.06	26.51*
Level (5)	1.7	0.06	29.62*
Level (6)	1.79	0.06	30.8*
S.Position (1)	0.5	0.06	8.77*
S.Position (3)	0.01	0.05	0.19
S.Position (4)	-0.05	0.06	-0.91

3.3.2 Operation Span

Table 3.4

Descriptive statistics for the Operation Span task.

Variable	Mean	Median	std. error	Min	Max
FTA Score	12.59	11.50	7.14	2.00	29.00
Max Span (1)	3.55	3.00	1.03	2.00	6.00
Max Span (2)	2.55	3.00	1.17	0.00	4.00
'Termination Span'	2.12	2.00	1.39	0.00	4.00
Proportion Correct	0.50	0.49	0.14	0.15	0.80
Operation Accuracy	0.92	0.93	0.05	0.8	1.00
Span 2 Accuracy	2.04	2.00	0.83	0.00	3.00
Span 3 Accuracy	1.64	2.00	1.03	0.00	3.00
Span 4 Accuracy	0.64	0.00	0.82	0.00	3.00
Span 5 Accuracy	0.16	0.00	0.37	0.00	1.00
Span 6 Accuracy	0.04	0.00	0.19	0.00	1.00

Descriptive statistics for the OS task are provided in Table 3.4. As noted previously the OS task and WMP share similar properties in that both are tasks involving verbal material and a combination of storage/processing. As Table 3.4 shows the effect of increasing list length (analogous to level increase in WMP) on OS trials has a more impactful consequence on the participants ability to correctly recall all TBR memoranda. The success rate on OS trials drastically decreases with ascending list length with very few participants having any success at span lengths 5 and 6 as evidenced by the span accuracy variables. These variables indicate the mean number of trials (out of 3) the participants were able to entirely recall at each list length. Additionally, the FTA scores for OS are relatively low considering the maximum score is 60.

The differences between the termination span score and the two variants on max span (see description above) are not as large as seen for the WMP data but differences still remain.

3.3.3 Colour Corsi

Table 3.5

Descriptive statistics for the Colour Corsi task.

Variable	Mean	Median	std. error	Min	Max
FTA Score	31.90	31.00	9.19	4.00	52.00
Max Span (1)	4.97	5.00	0.83	2.00	6.00
Max Span (2)	4.13	4.00	1.00	1.00	6.00
'Termination Span'	3.85	4.00	1.04	1.00	6.00
Proportion Correct	0.71	0.71	0.13	0.30	0.95
Grid Success	0.76	0.79	0.13	0.38	0.95
Colour Success	0.85	0.84	0.07	0.65	1.00
Span 1 Accuracy	2.94	3.00	0.25	2.00	3.00
Span 2 Accuracy	2.74	3.00	0.54	1.00	3.00
Span 3 Accuracy	2.66	3.00	0.65	0.00	3.00
Span 4 Accuracy	2.02	2.00	0.80	0.00	3.00
Span 5 Accuracy	1.08	1.00	0.93	0.00	3.00
Span 6 Accuracy	0.34	0.00	0.60	0.00	2.00

Table 3.5 displays descriptive statistics for the Colour Corsi task. Performance drops as list length increases as would be expected. The increased difficulty appears modest between spans one, two, and three, before increasing at the later span sizes. The gap between termination span and max span (2) is much smaller than observed for WMP and smaller than seen in the OS task. Interestingly less mistakes were made on colour selections than grid selections.

3.3.4 Arrow Span

Table 3.6 shows descriptive statistics for the Arrow Span task. These data suggest that the participants performance as measured by the FTA outcome is lower on the AS task than the CC task, $t(54) = 12.55, p < .0001, d = 1.69$. Note a dependent t-test was calculated using $n = 55$ who provided data on both tasks, and d is the standardised mean difference. This validates the concerns regarding the difficulty of the Arrow Span task as these data show that it is significantly more difficult than an alternative visuo-spatial task. There are potentially numerous factors that contribute to this which will be picked up in the discussion.

Table 3.6

Descriptive statistics for the Arrow Span task.

Variable	Mean	Median	std. error	Min	Max
FTA Score	17.58	16.00	8.18	3.00	37.00
Max Span (1)	4.45	4.00	0.94	3.00	6
Max Span (2)	3.05	3.00	1.43	0.00	6
'Termination Span'	2.02	3.00	1.73	0.00	5
Proportion Correct	0.45	0.44	0.12	0.14	0.68
Angle Success	0.54	0.53	0.11	0.27	0.79
Length Success	0.73	0.74	0.09	0.54	0.89
Span 2 Accuracy	1.84	2.00	0.98	0.00	3.00
Span 3 Accuracy	1.91	2.00	0.80	0.00	3.00
Span 4 Accuracy	1.16	1.00	1.01	0.00	3.00
Span 5 Accuracy	0.51	0.00	0.69	0.00	3.00
Span 6 Accuracy	0.16	0.00	0.42	0.00	2.00

These data do suggest that floor effects are not very likely with an adult sample as these data suggest that spans two, three, and four trials showed a reasonable degree of success. List lengths 2 and 3 do not appear to discriminate well between individuals in this sample based on the similar values for success rate, but beyond these list lengths there is a sharp reduction in success as LL increases. The length and angle success variables refer to the proportion of individual responses where the correct length/angle was recalled. It is no surprise to see that length success is higher as there were only two possible lengths versus eight possible angles of orientation. Once again there is a large difference between the 'max load' values depending on whether it was calculated based on a minimum of one or two successful trials at that LL. This clearly suggests that participants are able to successfully respond to trials of a LL beyond where the task would be terminated in a

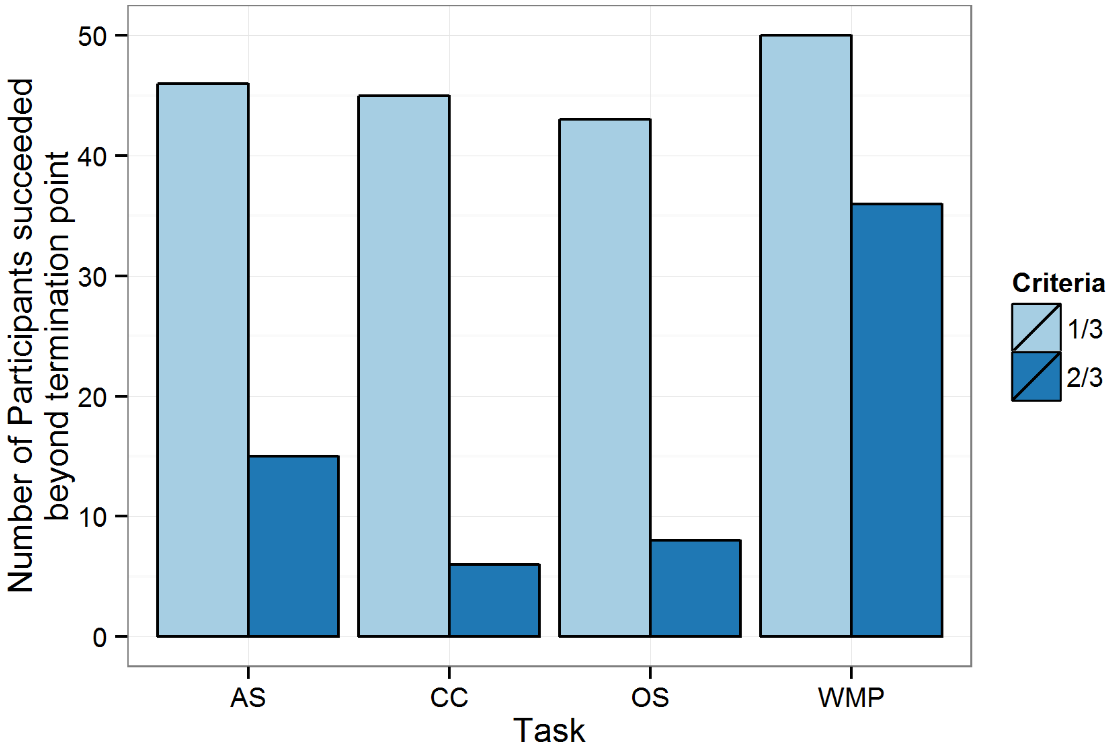


Figure 3.4. This chart displays how many participants would be successful at higher list lengths than where traditional span terminating algorithms would end administration. These values represent the number of times the values for participants differ on the termination span variable with the max span (1) and max span (2) variables. AS = Arrow Span, CC = Colour Corsi, OS = Operation Span, WMP = Working Memory Period

paradigm where task administration ceases once a certain amount of errors at a LL are made.

3.3.5 Scoring comparisons and correlational analyses

To assess the potential consequences of the terminating algorithm administration process commonly associated with span tasks I calculated the span score for each participant, for each task, under such conditions (termination span). I also calculated a span score based on the highest level/load where at least two (max span 1) out of the three trials were correct. Note this is irrespective of if an earlier list length was unsuccessful. Additionally, a span score was calculated with a more lenient criterion of just one correct trial out of three (max span 1). Figure 3.4 displays the frequency that these scoring methods produce different results. Using the most lenient criterion more than 2/3s of participants yield a higher span score. Perhaps the two trials correct criterion is more sensible to mitigate against chance success at a higher span level. Using this criterion we can see that there is generally much more agreement with the terminating span score but there is still some disagreement and that this seems to vary with task. The specific values are 6 (9.7%), 8 (13.3%), 15 (27.3%), and a huge 36 (59%) respectively for the CC, OS, AS, and WMP tasks.

Table 3.7 shows the correlations between each pair of tasks using each of three scoring methods. The correlation pattern differs between the terminating span score and the more continuous methods that use all trials to form a score. Using pure 'span' the correlation between OS and WMP is 0.16 (non-significant) whereas FTA/NC/TRANS pairs range from .41-.57 and are all significant ($p < .01$). The correlation between 'span' for CC and AS is -0.12 while the other pairwise combinations yield positive correlations between .36 and .44 (*all ps* < .01). The correlational pattern for the continuous measures yields an intuitive result where all correlations are positive but the correlations between the visuospatial tasks and between the verbal based tasks are stronger than cross-domain relationships. This is not the case for the span scores.

For each of the four tasks measured in this study I have computed three alternative scores that describe performance on the task. The full trial accuracy (FTA) score only gives a participant any credit for a particular trial if they recall all of the elements correctly (correct items and in the correct serial position) where the score given for each trial is 0 for failure or x where x is equal to the list length. Therefore this is an all-or-nothing scoring method per trial, there is no credit given if a participant was to give 4 correct items in a 5-item trial. In addition, there are two partial-credit scoring methods assessed here. The "number correct" method of scoring simply gives a point every time a participant recalls a correct item in the correct serial position, so if 4 out of 5 items in a 5-item trial are given correctly then the participant scores 4 for that trial. Finally, the "trans" scoring method builds on the number correct method by also giving partial credit where a person

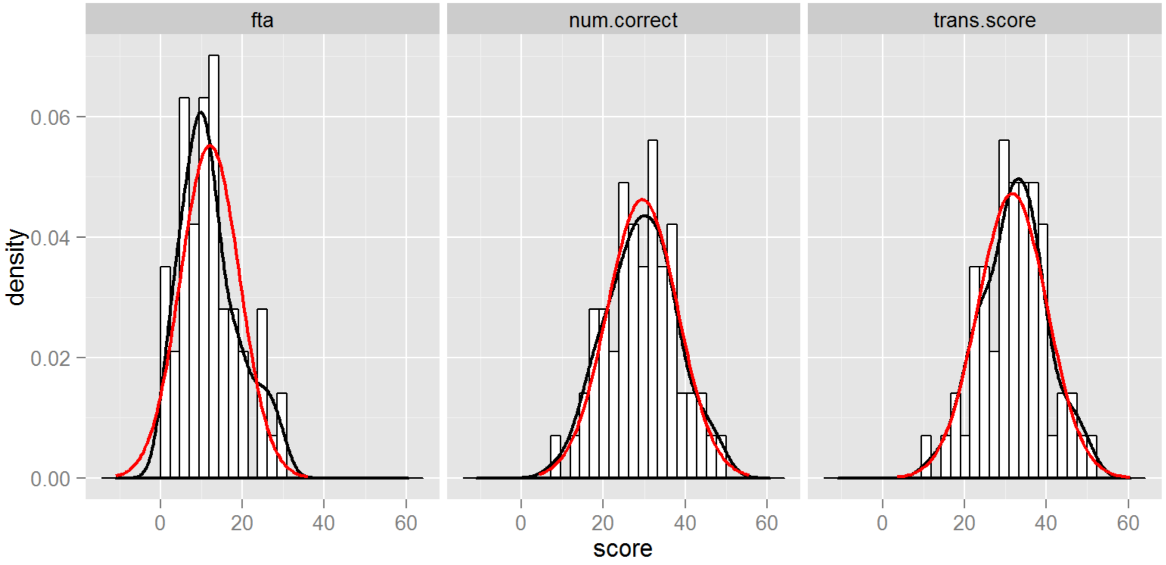


Figure 3.5. Distribution properties of scoring methods for Operation Span (raw sample, sample density plot, normal density plot)

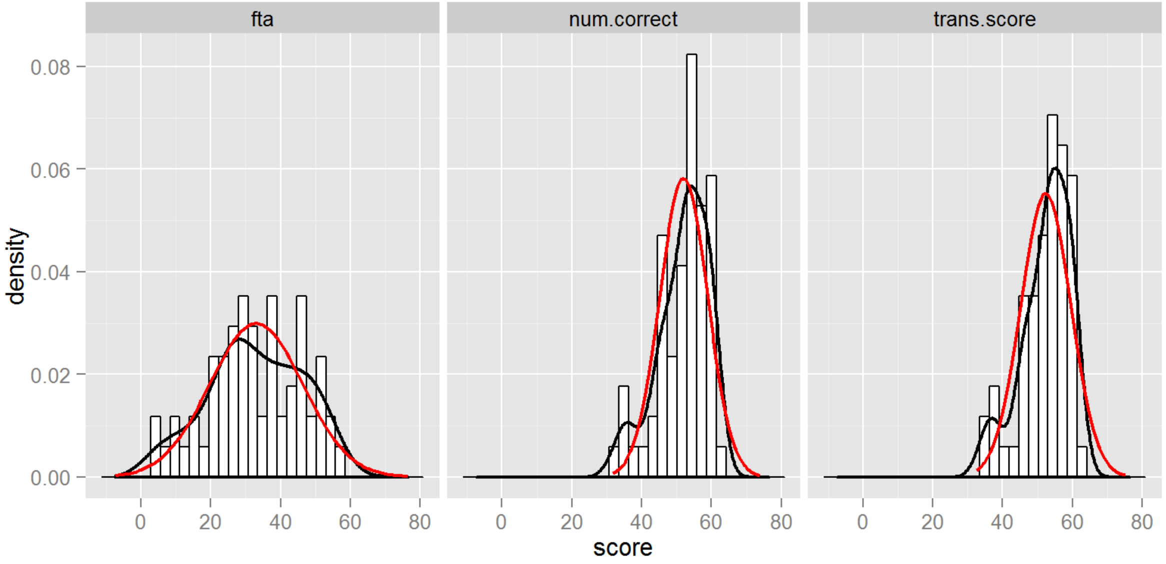


Figure 3.6. Distribution properties of scoring methods for Working Memory Period (raw sample, sample density plot, normal density plot)

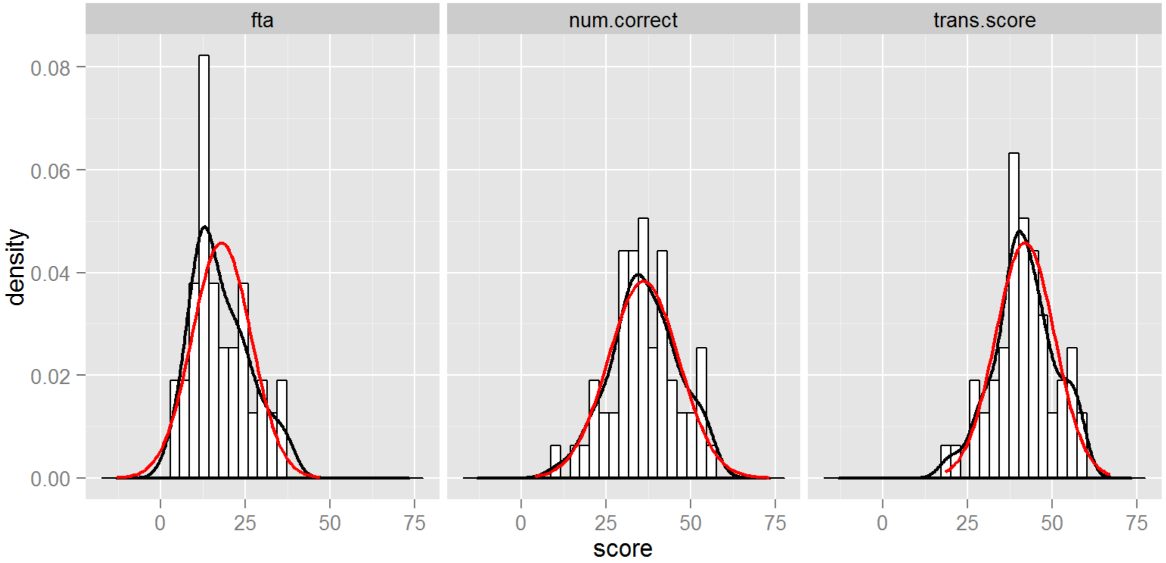


Figure 3.7. Distribution properties of scoring methods for Arrow Span (raw sample, sample density plot, normal density plot)

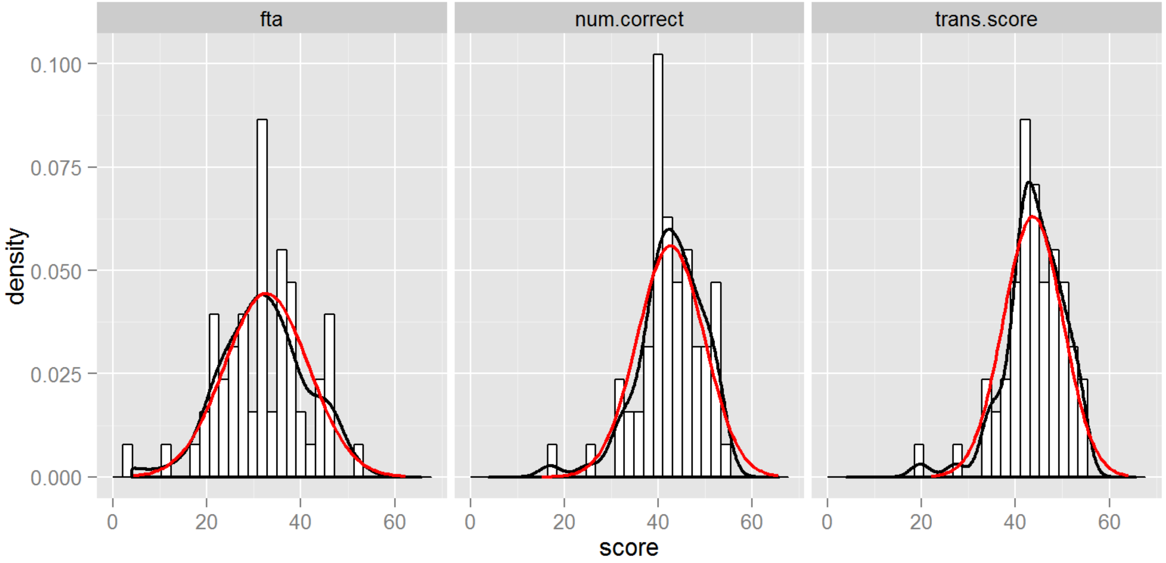


Figure 3.8. Distribution properties of scoring methods for Colour Corsi (raw sample, sample density plot, normal density plot)

Table 3.7

Correlation matrix consisting of each scoring method for all tasks administered. Variable names take the form task.method. The tasks are: os = Operation Span, as = Arrow Span, cc = Colour Corsi, wmp = Working Memory Period. The scoring methods are: span = the maximum list length or level the participant was successful, fta = full trial accuracy score, nc = total number of correct individual memoranda, t = t-score, nc with correction for transposition errors.

	os.span	os.fta	os.nc	os.t	as.span	as.fta	as.nc	as.t	cc.span	cc.fta	cc.nc	cc.t	wmp.span	wmp.fta	wmp.nc
os.fta	0.73*														
os.nc	0.69*	0.87*													
os.t	0.68*	0.87*	0.99*												
as.span	-0.07	0.14	0.15	0.20											
as.fta	0.06	0.30*	0.32*	0.38*	0.65*										
as.nc	0.19	0.36*	0.35*	0.39*	0.61*	0.82*									
as.t	0.18	0.37*	0.37*	0.41*	0.63*	0.83*	0.99*								
cc.span	0.31*	0.35*	0.30*	0.31*	-0.12	-0.04	-0.09	-0.10							
cc.fta	0.27*	0.28*	0.31*	0.31*	0.24	0.44*	0.21	0.22	0.67*						
cc.nc	0.30*	0.22	0.23	0.24	0.29*	0.45*	0.36*	0.35*	0.65*	0.85*					
cc.t	0.29*	0.23	0.25	0.26*	0.30*	0.47*	0.38*	0.38*	0.66*	0.85*	0.99*				
wmp.span	0.16	0.35*	0.28*	0.34*	0.08	0.23	0.25	0.25	0.35*	0.40*	0.36*	0.36*			
wmp.fta	0.25	0.41*	0.50*	0.54*	0.05	0.05	0.14	0.14	0.25	0.23	0.19	0.21	0.56*		
wmp.nc	0.21	0.42*	0.51*	0.56*	0.11	0.14	0.18	0.20	0.24	0.24	0.20	0.22	0.58*	0.93*	
wmp.t	0.21	0.41*	0.52*	0.57*	0.11	0.14	0.18	0.20	0.25	0.25*	0.21	0.23	0.58*	0.93*	1.00*

has given an item that did appear in the TBR material but they gave it in the incorrect serial position. Each TBR item in a trial can give a maximum of 1 point which is given when an item is given in its correct serial position. When an item is given that was part of the TBR array but has been given in the incorrect serial position then up to 0.5 points can be given for this response. The actual amount given varies, it is capped at a half as 1 out of 2 components were correct (item-correct, serial position-incorrect), and is subject to a weight that is determined by the probability of a transposition happening at random which varies from task to task. As a simple example, if a transposition happens purely by chance 20% of the time then the weighting used is $1 - .2 = .8$. Therefore a transposed response would be given .4 points $(.5 * .8)$.

Figure 3.5 shows the density curve for the observed data (black line) and the normal density curve (red line) imposed over the histogram for each scoring method for the OS task. Figures 3.7, 3.8, and 3.6 show the same information for the AS, CC, and WMP tasks respectively. Table 3.8 includes descriptive statistics for each of the complete data scoring methods as well as a traditional span measure for each task.

From Table 3.8 it can be seen that the skew/kurtosis values of all the full-data scoring methods for all tasks do not give too much cause for concern (only the T score for the CC task gives an absolute skew value greater than 1 (-1.02). Perhaps this is better seen through the density plots of the data for each task and scoring method.

3.4 Using the Rasch Model To Assess WM Task Items

By implementing the Rasch model (Rasch, 1960) on results from working memory tasks we can use the properties to quantify the difficulty of certain items. Item Response Theory is a vast area of statistics focused on assessing the quality of measurement tools. The IRT model used here is the basic Rasch model. For more information on this model see Rasch (1960) or for a very accessible account see Bond and Fox (2001). The general features of such models are that each item is probabilistically profiled along with the

Table 3.8

Descriptive Statistics for four scoring methods for each task. The tasks are: os = Operation Span, as = Arrow Span, cc = Colour Corsi, wmp = Working Memory Period. The scoring methods are: span = the maximum list length or level the participant was successful, fta = full trial accuracy score, nc = total number of correct individual memoranda, t = t-score, nc with correction for transposition errors.

	mean	sd	median	min	max	skew	kurtosis
os.span	2.05	1.40	2.00	0.00	4.00	-0.38	-1.18
os.fta	12.18	7.11	11.00	2.00	29.00	0.67	-0.32
os.nc	29.25	8.52	29.00	9.00	48.00	0.02	-0.42
os.trans	31.53	8.41	31.53	11.88	50.80	-0.09	-0.23
as.span	2.02	1.73	3.00	0.00	5.00	-0.07	-1.59
as.fta	17.84	8.51	16.00	3.00	37.00	0.57	-0.48
as.nc	36.16	9.95	36.00	11.00	55.00	-0.12	-0.38
as.trans	41.80	8.96	41.20	19.14	58.54	-0.16	-0.27
cc.span	3.85	1.04	4.00	1.00	6.00	-0.84	0.85
cc.fta	31.90	9.19	31.00	4.00	52.00	-0.24	0.22
cc.nc	42.60	6.99	43.00	17.00	54.00	-0.94	1.61
cc.trans	43.94	6.37	43.94	19.69	54.00	-1.02	2.17
wmp.span	2.66	1.83	2.00	0.00	6.00	0.35	-0.73
wmp.fta	32.77	13.47	32.00	4.00	57.00	-0.22	-0.76
wmp.nc	51.75	7.27	53.00	32.00	62.00	-0.86	0.10
wmp.trans	52.32	6.95	53.91	33.83	62.00	-0.89	0.14

ability of the test-takers. Thus we consider the probability of success on any given item for any given ability level. This becomes useful for identifying items that are required to discriminate between test-takers at different levels of the ability scale. It is important that a sufficient amount of data are used for the parameter estimates from these models to be meaningful. For this analysis I was fortunate enough to gain access to a large dataset collected from the Attention and Working Memory lab ran by Randall Engle and colleagues. The dataset contains data collected on three automated working memory tests. The automated operation span task (Unsworth et al., 2005; Redick et al., 2012), the automated reading span task, and an automated symmetry span task. The tasks were administered to undergraduate students between 2007 and 2008. The complete sample was 1259 for Operation Span, 1271 for Reading Span, and 1277 for Symmetry Span.

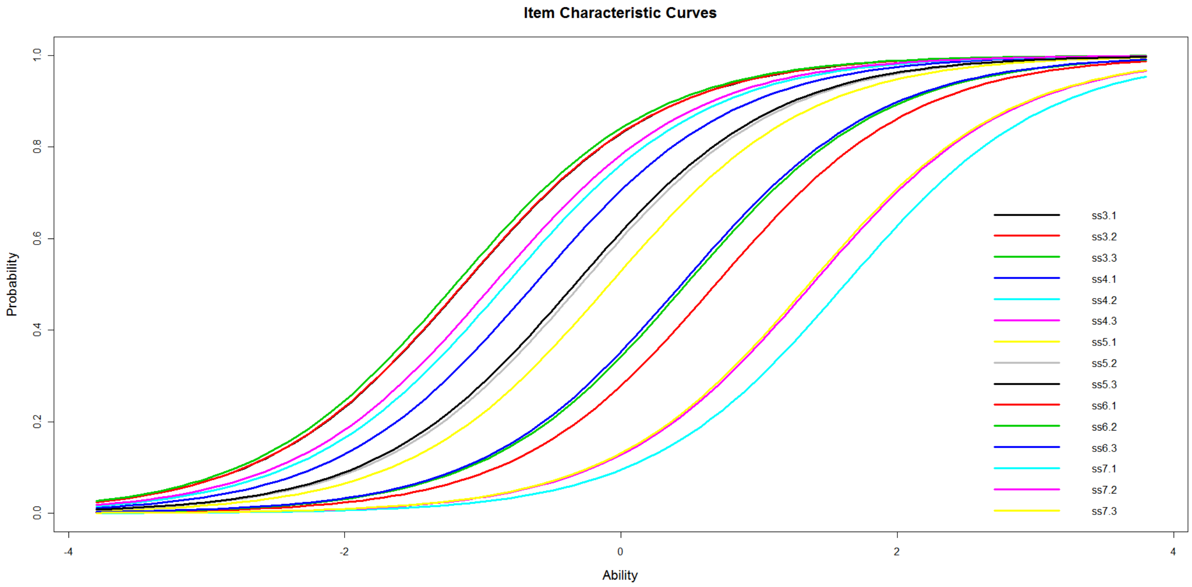


Figure 3.9. Operation Span - Item Characteristic Curves.

The unit of measurement in these analyses is each trial on the span tasks. Therefore for every trial (three at each set size) the participant is either successful or unsuccessful. The models were fit using the ltm package in R (Rizopoulos, 2006). Table 3.9 shows model fit statistics for the Operation Span data and suggests that the standard unconstrained Rasch model was the most parsimonious model. Therefore, that was the model applied in order to gather parameter estimates. Figures 3.9,3.10, and 3.11 show the item characteristic curves (ICC) for each item. From these figures it can be seen that these items function very well. For example, if we look along the x-axis to ability = 0 which represents the average ability of the population, we can see that the probablity of success scales sensibly based on the span length.

Table 3.10 shows the estimated parameters from fitting this model. The $p(x = 1|z = 0)$ column gives the estimated probability of success for each trial for a participant with average ability on the measured construct. The difficulty column is a logit value characterising the difficulty of an item when ability is controlled for. The average logit at each list length

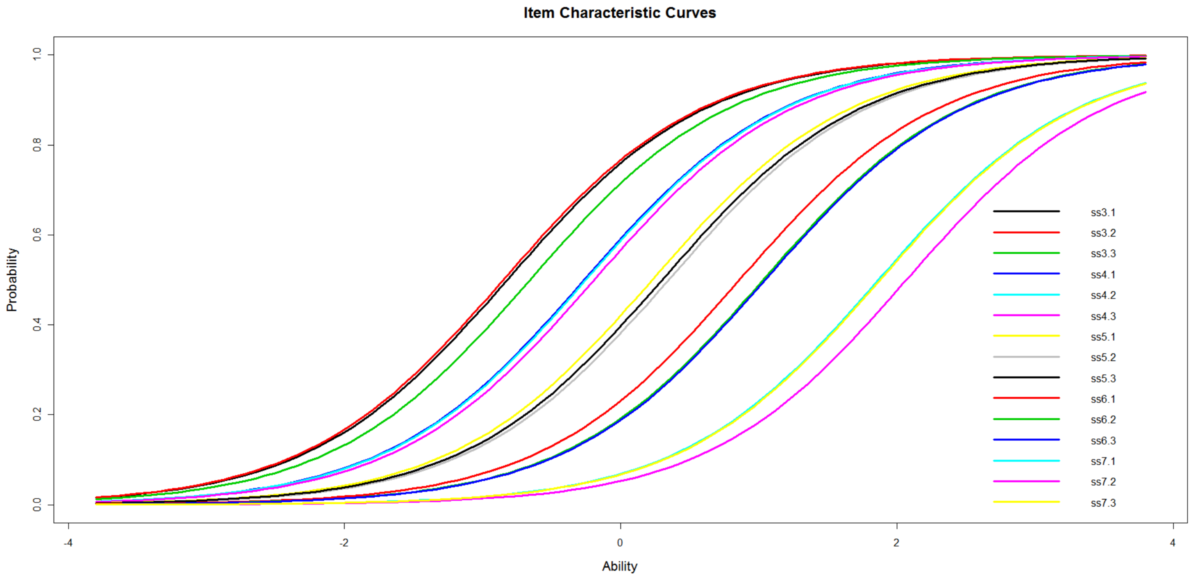


Figure 3.10. Reading Span - Item Characteristic Curves.

Table 3.9

Rasch Model fit statistics for the Operation Span test. AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion.

Model	LogLikelihood	AIC	BIC
constrained rasch	-9920.43	19870.87	19947.94
unconstrained rasch	-9852.4	19736.81	19819.02
2-Parameter logistic model	-9841.73	19743.45	19897.6
unconstrained rasch with guessing parameter	-9849.64	19761.27	19920.55

is -1.158 at ll3, -0.793 at ll4, -0.235 at ll5, 0.532 at ll6, and 1.456 at ll7. The difficulty logit gap between ll3 and ll4 is 0.365, between ll4 and ll5 is 0.558, between ll5 and ll6 is 0.767, and between ll6 and ll7 is 0.924. These results suggests that the increase in difficulty from ’n’ list length to ’n+1’ list length is not a linear progression, the jump in difficulty is a larger jump as list length increases.

The fit of the Rasch variants was also assessed on the remaining complex span tasks and both were fit better by the unconstrained discrimination parameter model. The Reading Span estimates show the average logit at each list length is -0.776 at ll3, -0.235 at ll4, 0.292 at ll5, 0.98 at ll6, and 1.938 at ll7. The difficulty logit gap between ll3 and

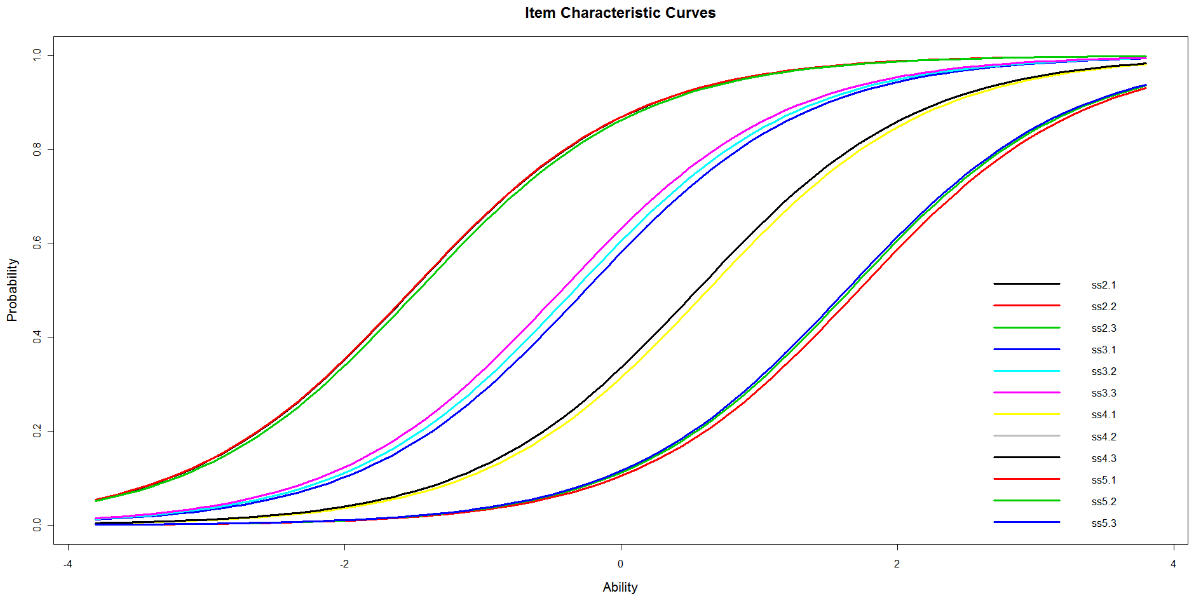


Figure 3.11. Symmetry Span - Item Characteristic Curves.

ll4 is 0.541, between ll4 and ll5 is 0.527, between ll5 and ll6 is 0.688, and between ll6 and ll7 is 0.958. While this pattern is not quite as progressive as that observed in the O-Span analysis due to the first two difficulty jumps being very similar and in fact the second jump is smaller. However, overall the jumps then increase rather than stay the same (linear) or decrease. For the Symmetry Span task the average logit values are; ll2 = -1.497, ll3 = -0.341, ll4 = 0.576, and ll5 = 1.669. This gives jumps of 1.156 between ll2 and ll3, 0.917 between ll3 and ll4, and 1.093 between ll4 and ll5. This is a rather different pattern than seen in the O-Span and R-Span logit values as all three jumps in difficulty are of similar magnitude.

What these analyses show is that when ability is controlled the difficulty gaps between list lengths in serial recall memory span tasks vary from task to task and should not be considered as linear. The implication of this for assessing change in performance on these tasks using measures based on max span or other scoring methods that are all-or-nothing per trial is that an increase of x does not mean the same improvement at each point on the scale.

Table 3.10

Parameter coefficients for Operation Span Items. The $p(x = 1|z = 0)$ column indicates probability of success on that item for a participant of average ability.

Item	Difficulty Estimate (Logit)	std. error	$p(x = 1\|z = 0)$
ll3.3	-1.196	0.069	0.841
ll3.2	-1.141	0.068	0.830
ll3.1	-1.137	0.067	0.829
ll4.3	-0.921	0.063	0.783
ll4.2	-0.831	0.062	0.761
ll4.1	-0.625	0.625	0.705
ll5.3	-0.333	0.057	0.614
ll5.2	-0.288	0.056	0.599
ll5.1	-0.083	0.055	0.529
ll6.3	0.437	0.057	0.353
ll6.2	0.473	0.057	0.341
ll6.1	0.686	0.059	0.278
ll7.3	1.361	0.071	0.131
ll7.2	1.384	0.072	0.127
ll7.1	1.624	0.078	0.095

3.5 Discussion

3.5.1 WMP Properties

These data suggest a number of interesting properties in the WMP task. The FTA score for WMP suggests that the WMP task is useful for discriminating between the ability of individuals. The success rates decrease as level increases except between level five and six. And importantly, the correlation between measures of WMP and OS are significant and of a high magnitude when scoring measures that use all the information are used (.41,.51, and .57).

The additive difficulty of a move up in period level is clearly not as substantial as the analogous increase in list length in a span task. Comparing the success rates of each level of difficulty in each task shows that the pattern for WMP is distinct from the span type tasks. In particular OS and AS show very sharp decreases in success as list length increases to the point that there are hardly any successful responses to the highest list lengths administered. To put this in perspective, the average number of successes for the OS task at list length four is 0.63 out of three(see Table 3.4) which is 21%. At the highest level of the WMP administered (6) the average number of successes was 1.38 out of three (see Table 3.1) which translates to 46%.

An implication of these results may be that WMP is a task that works well as a training task due to having smaller steps in difficulty as 'level' increases. It may be easier for participants to see progression as level increases may seem less insurmountable. Think of a participant who is engaged with an adaptive difficulty version of the OS or AS task as part of a WMT study. They are successful up to span four when they start to make errors therefore they stay at level four for some time. They then manage to get four out of five trials correct so they are given a block of trials at list length five. The jump in difficulty is seemingly huge and the participant is moved down a level to list length four again. If this situation repeats numerous times the participant may lose motivation to get back to span five trials as they are unable to be successful at these trials. If this participant was training on an adaptive difficulty WMP task it is likely that this feeling of reaching one's ceiling may be less likely to occur, or at least less likely to occur too early in training.

3.5.2 Arrow Span Properties

The perceived difficulty of the rotation span task when testing the developed materials for the training studies was the motivation behind assessing arrow span performance. If performance on the arrow span task was especially poor then it would suggest that there would be a serious possibility of floor effects in the rotation span measures. Rather than administer the rotation span task in this study to explore this issue it was decided that there was scope to get more information by administering the short-term memory equivalent, arrow span. This is because if very poor levels of performance were observed then it would be strong evidence that it is the nature of the stimuli that makes the task difficult (as there are no other processes involved, simply storage and retrieval of the arrow stimuli). Also, it will allow a neat comparison between an arrow span dataset and rotation span dataset once rotation span has been administered in a study. The only descriptives available in the Kane et al. (2004) article were for proportion correct of items within trials where their sample scored 65% and 61% for arrow span and rotation span respectively.

It can be seen from Table 3.6 that all participants were able to successfully respond to at least one trial in the AS task as indicated by the minimum value for the FTA score variable being 3 (worst performer was successful at one span three trial and no others). However, the profile of these data reads like a complex span task as opposed to the difficulty of a simple span task. Consider the comparison of the AS task with the CC task and then the OS task. The AS task data are more comparable to the OS data in terms of the difficulty of different list lengths and the range of FTA scores seen. If we then consider that it is only reasonable to assume the addition of processing judgements (rotation span) is going to decrease performance levels the fear of floor effects is realistic.

3.5.3 Scoring Methods

Given the dominant significance tests used in psychological research it is fair to say that when deriving measures of ability it is often best when the resulting variable follows, as close as possible, a Gaussian distribution. In addition, the measurement should reflect each participant's actual ability and therefore be sufficient to discriminate between different levels of ability. When using span measurements one could argue the resultant data reflects a non-continuous variable and is best described as an ordinal variable. As an example, the OS task yielded a max span of 4 and a minimum of 0 (indicating that some participants were incorrect on at least two of the three list length two trials). Therefore, the values in this vector only take one of four values (0,2,3,4). Using span as the DV therefore likely presents violations to assumptions inherent in any of the significance tests these variables are then subjected to.

The reasons for investigating different scoring methods for these tasks are twofold. Firstly, from a psychometric perspective, establishing which methods have the best properties for both reliability and validity considerations. And secondly, to understand how different scoring procedures may be measuring slightly different elements of WM. For example, St Clair-Thompson and Sykes (2010) administered a battery of 5 STM/WM tasks to a group of 7-8 year old children and also obtained scholastic measurements for each child from their school for maths, reading, writing, and science. The authors were primarily interested in the difference in predictive power of absolute scoring (what I have called the FTA (full trial accuracy) method) and a proportion correct (proportion of correctly recalled items within trials averaged over all trials) methods. They found that the proportion correct scores of the STM/WM tasks often explained unique variance in the scholastic attainment scores after controlling for the absolute scores. A further study in 2012 (St Clair-Thompson, 2012) further examined differences between these two scoring methods but also between an administration manipulation, namely whether list lengths of trials were given in ascending order or randomised. St-Clair Thompson found that when administered with LLs randomised scores on the Counting Span task were significantly different but not for the Reading Span task. More important, however, were the results suggesting that only the randomised versions of both tasks were significantly related to RAPM performance, this was the case for both FTA and proportion correct scoring methods. This followed the work of Unsworth and Engle (2007b) who suggested that proportion correct methods of scoring WM tasks produced higher predictive power on criterion tasks of STM/WM as these methods benefit from using data from higher list length trials that are often not recalled in entirety, and these trials tap secondary memory due to primary memory reaching 'capacity'. Therefore the proportion correct scoring method measures primary memory as well as a contribution from secondary memory which is absent in absolute scoring and would also be absent in a 'span' measurement. Thus, while measurements of WM ability are used in a large amount of research programs there is clear evidence that the assessments one makes regarding a person's or group's WM ability can be significantly influenced by decisions regarding the administration of the task (i.e. list length order, termination or not, etc.) and also by the method one chooses to produce scores on the resultant data.

These differences in outcomes based on procedural and scoring variations have been used to discuss the nature of WM functioning. For example Lustig et al. (2001) suggest that the boost in performance for longer list length trials when a descending order of LLs was used is evidence for the important role of suppressing proactive interference in tasks of STM/WM. This is due to the longer LLs being conducted when there have been little to no previous trials and thus a much smaller pool of potentially interfering stimuli from these trials whereas generally the longer LL trials are conducted when the interference pool is at its greatest and are therefore greater influenced by the relevant interference mechanisms in such processing.

In this experiment we did not administer any non-WM assessments which would have allowed for comparison of different scoring methods predictive utility for the four span tasks we administered. However, the interrelationships between the four span tasks can be used to probe the utility of each scoring mechanism. In addition, whereas St Clair-Thompson and Sykes (2010) focused on two scoring methods (FTA and proportion correct) this study extends that by including traditional span scores as well as two other more continuous scoring methods; total items recalled (as per Friedman & Miyake, 2005), and an additional method that attributes value to occurrences where the item was recalled but in the wrong serial position (corrected for chance performance).

The correlational pattern (see Figure 3.7) suggests that when 'Span' is used as the outcome of such tasks the relationship between them is either non-existent or at best unintuitive. However, the more continuous measures provide interrelationships that can be interpreted in the context of the existing WM literature. When using the FTA, total correct, or 'trans' scoring methods the tasks all generally share a positive correlation but also where overlapping domains (verbal/visuospatial) produce correlations of a greater magnitude. The distributional properties of the continuous scoring methods do not give great support for one over the others as they all appear reasonably normally distributed.

The Rasch parameters obtained in these analyses show that the difficulty increase in list length jobs is not linear for all tasks. Therefore if conducting an experiment designed to detect improvements in working memory it is important to include scoring methods that aren't all-or-nothing based. By including methods that give credit for all successfully recalled items such as proportion correct or total number of items will mitigate this affect and maybe not mask small changes at the higher levels of ability.

Chapter 4

Working Memory Training Developmental Study Two

4.1 Introduction

In the first experiment we failed to find any significant and lasting effects as a result of the multi-task adaptive working memory training program. Transfer to tasks commonly used to assess working memory but different in concept to the trained tasks was not seen but more surprisingly there was little evidence to suggest that the children benefitted from repeated exposure to the same task as evidenced by the flat 'growth curves' seen when assessing the practice effects. A second experiment was designed to further investigate the effects of practice on the working memory construct that responded to the weaknesses of the previous experiment.

The primary method shift in the second experiment compared with the first experiment is the decision to use single task training paradigms where participants will be allocated into a specific training group and all their training time will be spent on a single task. This will significantly increase the time spent training on that specific task. Additionally, while this issue was not present in the previous experiment due to a lack of evidence for transfer effects, but if they were to be found it would be more clear what aspect of the training caused the improvement when training was focused on a particular type of task. With a training group that has focused on only one task then there is a greater weight of evidence that a significant change in pre-post measurements could be attributed to the stimulation of the processes involved in that training task. An important question is whether the improvement on the trained task may be the driving force of any potential benefit of the training program or if the overall time spent on the training task is more important than progressing to even higher levels of difficulty. By increasing the time spent on each task it may allow the influence of these tasks to reach the necessary threshold to produce transfer.

A final point to note regarding the difference in training program methods is that the singular task method gives a greater intensity of training as there is less time spent learning what the task requires, less time switching between tasks (every 5 minutes), and while the overall time spent training may not change the amount of training on the one task will be approximately five times that of the training on any one of the tasks in the training battery method. The critical problem with switching to a singular training task method is the greater risk of fatigue on behalf of the participants leading to motivational issues and a lack of focus on the training after a critical threshold. It is hoped that the adaptive nature ensures that participants are constantly facing a challenge and that this aspect of the design mediates the fatigue issue. However, it is clearly a concern for all work where participants are required to repeat tasks.

The training tasks we chose to investigate in this experiment were the Working Memory Period, Colour Corsi, and N-Back tasks. These tasks were selected as they each represent a focus on different processes in working memory. Working Memory Period represents a task that focuses on the ability to deal with increasing processing demands while maintaining the same verbal memory load. The Colour Corsi task involves visuospatial material in addition to requiring storage and recall of multi-feature items (colour and location). Both these tasks are carried over from the previous experiment. An addition for this experiment is the N-Back task. There is a growing literature supporting the effectiveness of working memory training based on tasks categorised as updating tasks. As discussed in the literature review the most prominent updating task used in the training literature is the N-Back task. The meta-analysis on transfer to fluid intelligence (Gf) after n-back training by Au et al. (2015) included both standard and dual versions of the n-back paradigm and showed an overall small but significant effect of transfer to Gf. The 20 studies included in their meta-analysis were studies including healthy adults between the ages of 18 and 50. However, Jaeggi et al. (2011) have investigated the effects of a standard n-back adaptive training paradigm on typically developing elementary and middle school children. They found that the general effect of training to the transfer measure of Gf, a composite score based on performance on the Test of Nonverbal Intelligence(TONI) and Raven's Standard Progressive Matrices (RSPM), was not significant. However, they then formed groups based on low or high gains in the training task over the repeated sessions and found that those in the 'high-gain' group showed significant transfer to Gf. The authors included no near-transfer measures. Zhao et al. (2011) also conducted a study involving typically developing 9-11 year-old children completing an updating based WM training intervention. The intervention tasks used by Zhao et al. were two slight variants on the running memory task whereby participants are given a stream of sequential stimuli and asked to keep track of the most recent four items shown. The trials given to participants were either 5, 7, 9, or 11 in list length. They report significant transfer to Gf (RSPM) after 15 training sessions (each consisting of 20 running-memory trials, 5 at each set size). These authors also did not include any near-transfer measures to assess improvement in WM performance after training. Therefore, while there is reason to be optimistic about the potential effects of updating based working memory training the amount of work studying typically developing children is very small and there has been no demonstration of how these paradigms improve WM function which is the proposed mechanism for the transfer to Gf examined. Therefore it was decided to include an n-back training group in this follow up experiment with an expanded near-transfer battery of pre-post tests.

As a further methodological improvement for this experiment, a larger number of tasks have been included in the pre-post battery. Shipstead et al. (2010) argued that training studies should use dependent measures that are made up of scores from a number of tasks (latent scores) as a way of tackling the issues of test-retest and random error in the test scores. A number of verbal WM (3), spatial WM (2), and processing speed measures (2) were included. These will first be tested with a principal components analysis to ensure that the scores can be combined and if so the latent scores on these factors will be the primary dependent variables of interest. The pre-post battery of tasks includes simple span, complex span, processing speed, and mental arithmetic measures. These will provide a variety of measures that will allow any differential impact of the training regimens on overall WM functioning to be observed.

The hypotheses in this chapter follow from chapter two. It is expected that the WM training groups will show improved scores on measures of near-transfer compared with the active control group. The degree of transfer may vary between training groups and between near-transfer measure due to different task demands of the training as well as the degree of overlap between WM sub-systems responsible for performance on the measures. The inclusion of processing speed and mental arithmetic measures represent 'far-transfer' and as with the previous chapter we remain agnostic with regards the expected outcome of WM training on these transfer measures.

4.2 Method

4.2.1 Participants

The participants recruited for this experiment were 115 children who were students in four classroom groups across two schools. Two classes were year-5 (age 9-10 years) and two were year-6 (age 10-11 years). The participating schools responded to a letter sent by the researcher outlining the goals and methods of the proposed research. Again there

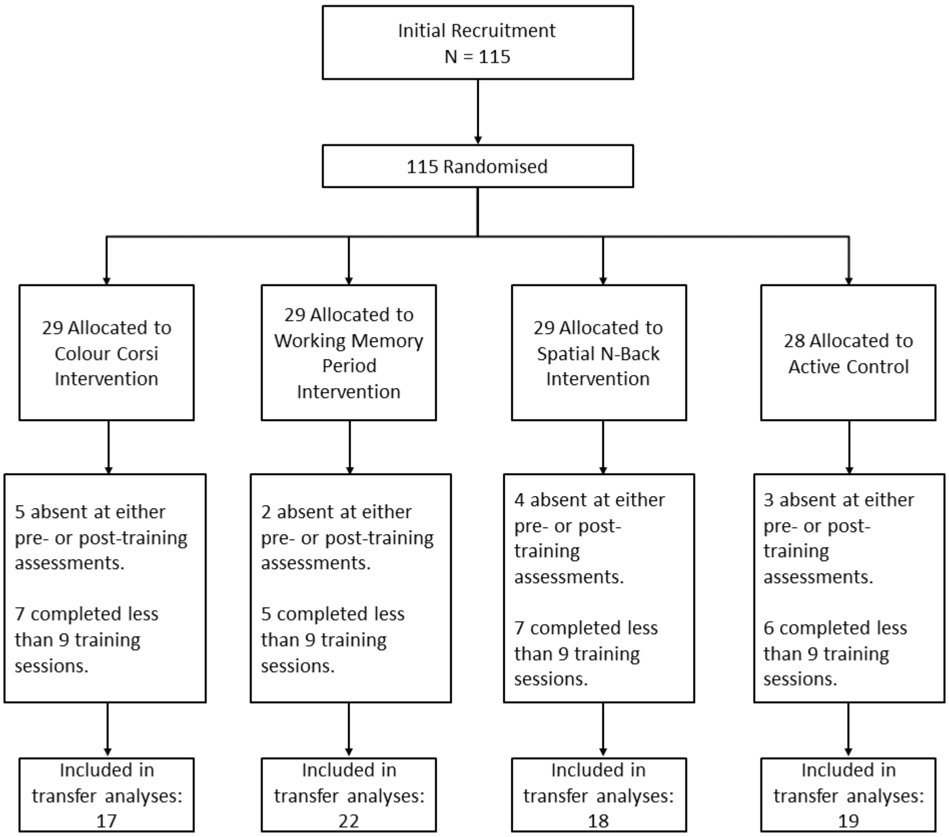

Initial Recruitment
N = 115
115 Randomised
29 Allocated to Colour Corsi Intervention
29 Allocated to Working Memory Period Intervention
29 Allocated to Spatial N-Back Intervention
28 Allocated to Active Control
5 absent at either pre- or post-training assessments.
7 completed less than 9 training sessions.
2 absent at either pre- or post-training assessments.
5 completed less than 9 training sessions.
4 absent at either pre- or post-training assessments.
7 completed less than 9 training sessions.
3 absent at either pre- or post-training assessments.
6 completed less than 9 training sessions.
Included in transfer analyses: 17
Included in transfer analyses: 22
Included in transfer analyses: 18
Included in transfer analyses: 19

Figure 4.1. Flow diagram highlighting participant involvement throughout WMT study 2.

are phases where some participants may not have provided data for various reasons and this will be documented in the results section. Figure 4.1 shows the participant progress throughout the study.

4.2.2 Design and Materials

The experiment is a randomised controlled trial that consists of three phases; baseline phase, training phase, and post-training phase. The training phase was administered over a five-week period and during this phase participants either completed adaptive WM training (three different groups) or were part of an active control group engaging in standard puzzle tasks. Participants were randomly assigned to one of the four groups prior to commencement of the experiment. To measure any transfer of training a battery of tasks was completed before and after the training phase.

All computerised tasks used in this experiment were programmed using the Java programming language using the framework provided by the Tatool package (von Bastian, Locher, & Ruflin, 2013). Due to the popularity of the Java language a JRE is installed on most machines as standard due to it being a requirement for so many programs. This makes it much easier to take the application and use it on a wide range of computers with relative ease such as those maintained by schools as well as individual home machines.

Training Tasks

Working Memory Training (WMT) Groups

Spatial N Back The Spatial N-Back task developed based on the description of the task used by (Jaeggi et al., 2011). Participants were presented with 10 red circles dispersed on the screen that represented 'locations'. A stream of given locations was presented to participants by changing the colour of the locations to green one at a time. Each location would be highlighted for 500ms followed by a 2500ms inter-stimulus interval. The participant was required to indicate when a match occurred between the currently presented location and the location highlighted n items ago, where n was denoted by the current level of difficulty. Each location can be considered a trial and these were distributed in blocks where each block contained *15+n* trials. At the end of each block the level adapted based on the performance during the previous block. If less than three errors were made the level would increase. If more than four errors were made the level would decrease. The level remained the same if three or four errors were made.

Colour Corsi The Colour Corsi task is taken from the first empirical study described in this thesis and therefore I won't repeat all the details here (Figure 4.2 is reproduced here as a reminder).

Performance was assessed every five trials and if the participant had correctly re-

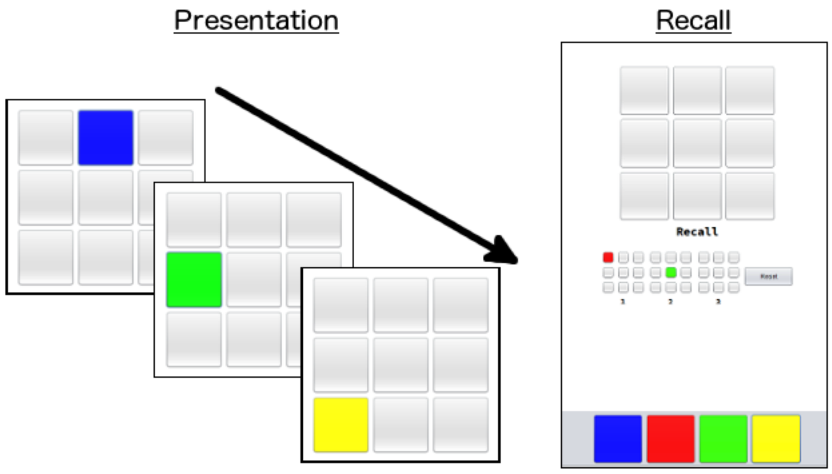


Figure 4.2. Illustration of the Colour Corsi task

sponded to 4 or more trials then the level would increase by one. If however, two or fewer were correctly recalled then the level was dropped by one. The current level dictated how many grid-colour combinations each trial consisted of.

Working Memory Period The working memory period task used in this study is also taken from the previous study (Figure 4.3 is reproduced here as a reminder). One difference was that the stimuli were not pre-determined and taken from previous research. Instead the mathematical operations were constructed by the program when required. In the first study there were only three potential levels of difficulty and some items would end up repeating a number of times over the course of the training phase. This time there are no restrictions on what level participants can get to as the program will generate operations that match any level.

The adaptive difficulty works identical to the procedure for Colour Corsi task described above.

Active-Control (AC) Group

As with the previous study the active control group needed to be involved in a procedure that as closely matched up to the training groups as possible without engaging them in

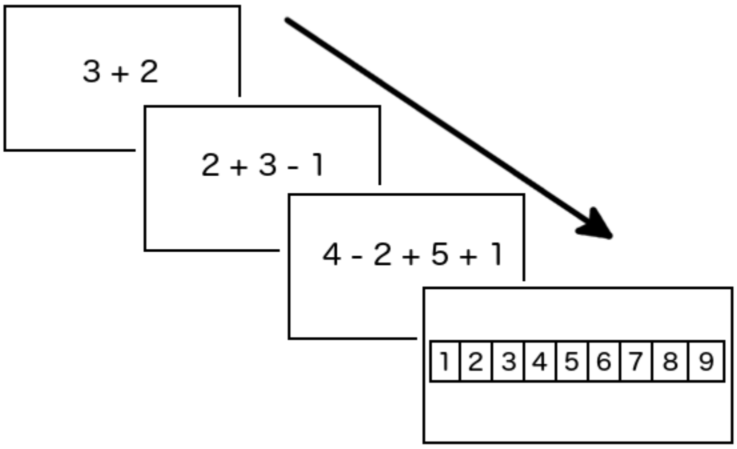


Figure 4.3. Illustration of the Working Memory Period (WMP) Task

any overt working memory training. For this reason, three puzzle tasks were developed in the same framework as the training tasks (Java/Tatool). The overall look of the software was the same and there was a level system in place just like the training groups. The puzzle tasks were; Wordsearch, Jigsaw, and Sudoku.

Wordsearch The Wordsearch task presented participants with a standard Wordsearch puzzle in the centre of their screen. The size of the grid was dependent on the current level the participant was at. The grids started at 8x8 letters and as the level increased the grid got larger (making words harder to spot). There were twenty themed word lists (i.e. Countries, Animals etc.), each trial was one Wordsearch grid that was generated using one of these lists (selected at random). The letter placements were generated by the program so they were different each time even if it was re-using the same word list.

Alongside the Wordsearch grid was the list of words to find. Participants selected words by pressing letters one at a time to spell out a word from the list (the program would only allow valid selections, so once a letter was selected you could only pick an adjacent/diagonal letter and then after that they could only continue in their specified direction). Participants would then press a submit button, if the word was valid then it was highlighted for the rest of the game and removed from the list of words to be found.

Jigsaw The Jigsaw task presented participants with an image that was broken up into Jigsaw pieces. The number of pieces the image was broken into was determined by the level the participant had reached. The top left corner showed what the image looked like when completed and the rest of the screen was a 'canvas' where participants could move the pieces around. To join two pieces together they simply needed to move one piece close to the connecting part of the other piece. If they were supposed to go together they would 'snap' into place and would be joined together. Upon completing the Jigsaw the participant was shown a congratulatory message and the amount of time it had taken them to complete it.

Sudoku The Sudoku task generated a traditional Sudoku puzzle for the participant to complete on each 'trial'. The difficulty adapted to the current level by means of how many empty spaces were present at the start of the trial, as level increased there were more spaces to fill making it more difficult. As this can be the type of puzzle that some people are not comfortable with a 'help' mechanism was included that could be used 5 times. By selecting the help button and pressing on a number (from 1-9) the program told the user which spaces this number could go in which provides a significant help.

Pre- and Post-Training Assessement

In this experiment we used a larger battery of tasks for the pre/post training assessments. Having more performance measures at this stage gives a greater chance of detecting the possible effects of the training interventions. Additionally, as argued by Shipstead et al. (2010), measuring constructs on any one single task is not an optimum method of measuring the constructs and therefore where possible a combination of tasks should be administered per construct of interest. Using latent factor scores as opposed to raw scores on individual tasks would further support the conclusion that any change has occurred due to a change in ability as opposed to being an artefact of repeated measurement and the random variation one sees as a product of external influences (measurement error).

Task selection was informed by previous research, specifically Kane et al. (2004) and Nettelbeck and Burns (2010). The latent variable analyses conducted by these researchers were used to select tasks that had already been shown to relate to the suggested constructs. The abilities we wanted to measure were visuospatial working memory, verbal working memory, short-term memory, and processing speed (PS). Three tasks were selected to test verbal WM, while two were selected for visuospatial WM and PS abilities. In addition to the memory constructs we included a mental arithmetic task as a far-transfer measure.

Visuospatial Working Memory

Symmetry Span In the symmetry span task participants are required to remember grid (4x4) locations presented to them in the correct serial order. Figure 4.4 shows a schematic representation of this task. As is shown, participants are given a processing operation to complete after each TBR grid is presented. This processing element requires them to make a judgement of whether the presented pattern is symmetrical along the vertical axis or not using the left/right arrow keys (8x8 grid used for giving patterns). When the grid locations were highlighted in the storage phase a blue colour was used while the pattern presented during the processing element was produced using a black fill.

After the presentation phase had been completed (all storage-processing pairs) the recall phase began. Responses were recorded by presenting participants with the 4x4 grid and allowing them to click the boxes in the order they recall seeing them. When a box was selected it turned blue so participants could keep track of their responses.

Test trials consisted of three at each list length between two and six.

Matrix Span The matrix span task was the STM equivalent of the symmetry span task. The procedure is the same as described for symmetry span except for the removal of the processing element. The grids were highlighted for 1000ms with an ISI of 1000ms before the next grid was shown.

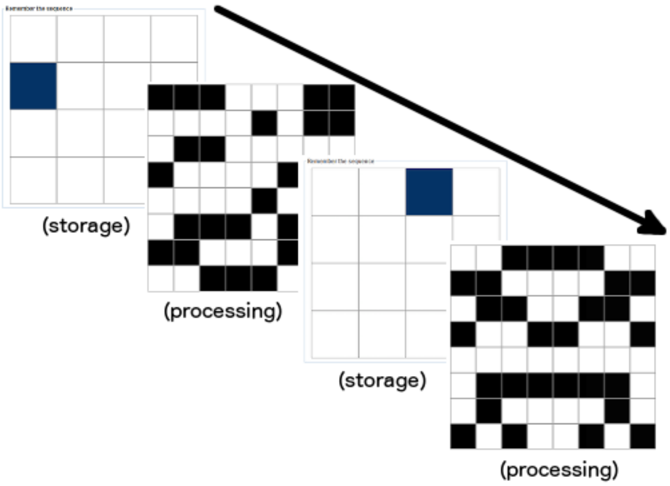


Figure 4.4. Illustration of the Symmetry Span task

Test trials consisted of three at each list length between two and seven.

Verbal Working Memory

Operation Span Figure 4.5 presents a diagram which clearly describes the operation span task and how it was employed in this particular study. The TBR items in the storage phase were digits (between 10 and 99). A difference in the procedure here with how the operation span task is often reported (e.g. Unsworth et al., 2005) is that the processing phase occurs after the storage item is presented as opposed to before. In our previous study we followed the traditional protocol. But, it seems more appropriate to have the processing element after the storage element, particularly in the case of span size 2 trials. In this instance there is a clear argument to suggest that the first processing element wouldn't interfere with storage in the original procedure and therefore half of the processing is to some extent irrelevant.

The processing element was once again a mathematical operation presented in the centre of the screen with an answer. The participant needed to indicate (using the left/right keys) whether this answer was the correct answer to the question. The recall process asked participants to input the numbers they remember in order by presenting "Number

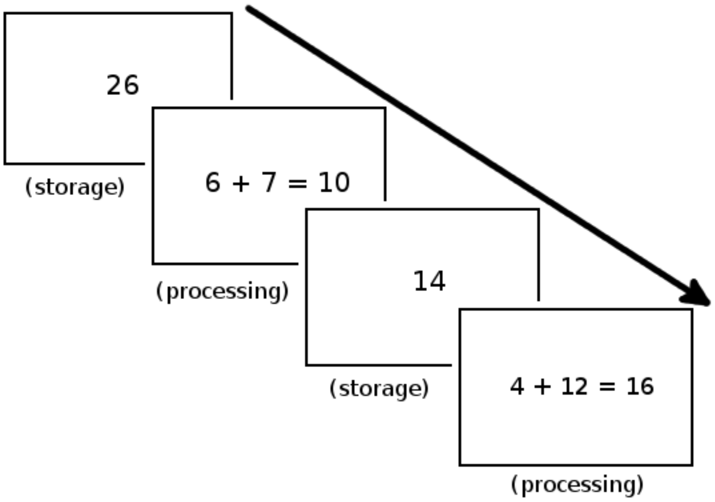


Figure 4.5. Illustration of the Operation Span task

1: " in the centre of the screen with a text box for inputting the answer, upon pressing enter the text changed to "Number 2: " and the text box emptied ready for the next response.

Test trials consisted of three at each list length between two and five.

Reading Span The reading span task (Figure 4.6) differed from the operation span task only in the processing element. Rather than having to verify a mathematical operation, the participants were presented with a sentence that they had to decide if it made sense or not.

Test trials consisted of three at each list length between two and five.

Digit Span The digit span task is the STM equivalent of the reading/operation span task. The procedure is the same as described for those except for the removal of the processing element. Participants were presented with the digits on screen for 1000ms and the inter-stimulus interval (ISI) was also 1000ms.

Test trials consisted of three at each list length between two and six.

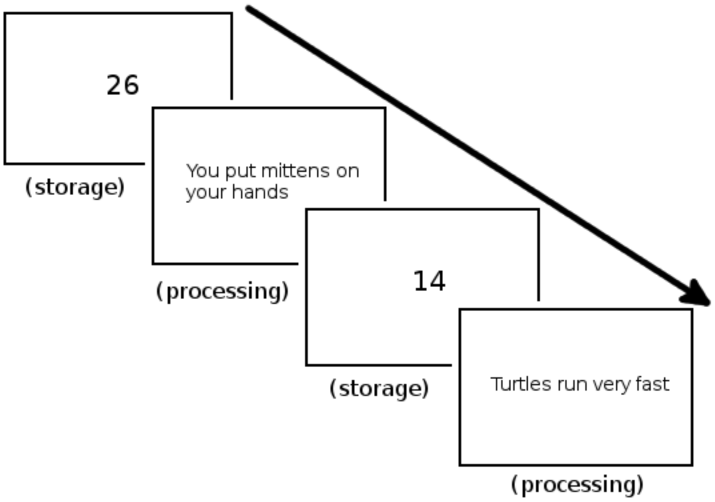


Figure 4.6. Illustration of the Reading Span task

Processing Speed

Odd One Out The Odd One Out task presents participants with a simple task of selecting the red light (out of three presented) that is furthest away from the others. Figure 4.7 shows the paradigm. In this figure the left screen is the resting state. The participant must press (using the left mouse key) the home button (yellow button at the bottom of the screen) and keep the home button pressed. After a short random delay three of the blue lights will turn red. The participant can then make their decision, release the home button, and press the odd one out. The time between the red lights being presented and the participant releasing the home button is their decision time (OOO-DT). One can also calculate a movement time (OOO-MT) by taking the time between the release of the home button to the response being given.

Participants completed a total of 50 trials on the OOO task.

Inspection Time The Inspection Time task presents participants with a target figure of two vertical lines where one of the lines is always shorter than the other. After the SOA (stimulus-onset-asynchrony) passed the target figure was masked. Figure 4.8 shows the two states of the Inspection Time task. Participants need to respond with the line

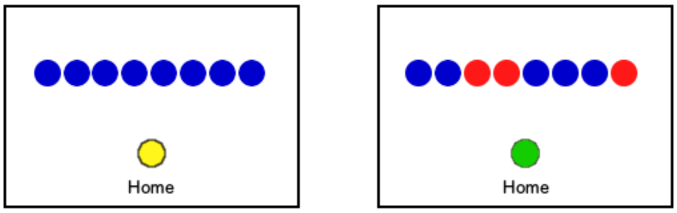


Figure 4.7. Illustration of the Odd One Out task

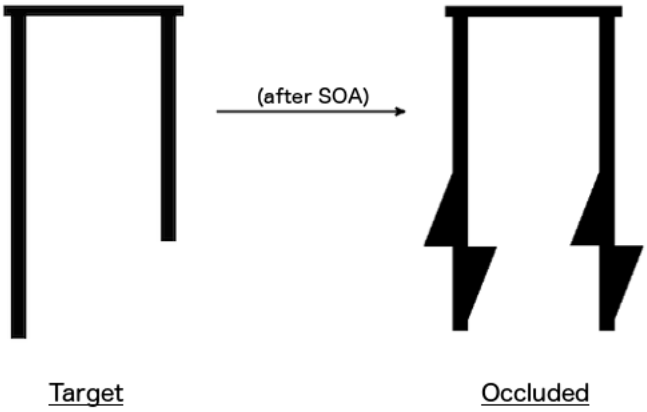


Figure 4.8. Illustration of the Inspection Time task

they believe was the shortest using the left and right arrow keys. Following the procedure of Nettelbeck and Burns (2010) the SOA is set to 250ms initially and varies according to performance. After three consecutive correct responses the SOA is reduced by 17ms while any incorrect response increases the SOA by 17ms. Trials are presented to the participant until they experience eight reversals of direction on the SOA. The measure of performance is the mean SOA at the end of the task.

Further Transfer

Mental Arithmetic The mental arithmetic task used here was identical in procedure to the one used in our first developmental study. The only difference here was after that initial study we decided to alter a small number of questions just to ease the transition of difficulty. The first study used questions that had a strict stepwise difficulty jump. For example, the first ten addition questions were all very easy e.g. ”5 + 4”, the next ten were all similar but used double digit numbers e.g. ”16 + 11” and the final ten were all triple figure numbers such as ”145 + 394”. The magnitude of the numbers in the questions changed in a more linear fashion this time around. The pool of questions used in each study is provided in the appendix.

As before the test consisted of six blocks of questions where the participant was given one minute per blocks to answer as many questions as possible. The blocks included specific ’types’ of operations; addition without carry, addition with carry, subtraction without carry, subtraction with carry, multiplication, and division.

4.2.3 Procedure

Ethical approval was obtained for the current study, details of which are available in section A.1 of the appendices. The pre- and post-training measurements were spread across two testing sessions that occurred on consecutive days and lasted between 30-45 minutes. These sessions were conducted in classroom groups where each task was explained and demonstrated by the researcher in advance of them beginning. Once a participant reached the end of a task they were asked to wait the short amount of time it would take for others to catch up, when the next task could be explained and demonstrated. The primary researcher was accompanied by the class teacher at each of these sessions. As with the previous WM training study it was not possible to implement a full blinding procedure. At the time pre- and post-training assessments were completed the primary researcher (who controlled these sessions) was blind to the grouping allocation. However, the school teachers were responsible for scheduling the training within their classroom and thus they and other children are likely to have observed that some children were completing different training tasks.

The training phase began two weeks after the initial baseline assessment sessions occurred. The training phase lasted 5 weeks during which the children were asked to complete 15-minute training sessions on the task that corresponded to the training group they had been assigned. The class teachers were given the flexibility of being able to set the training schedule on a basis of 3-4 children completing training sessions at any one time in the classroom with the guidelines that each child should try to complete 3-4 training sessions per week. Upon completion of a training session the data was uploaded to a secure server accessible only by the primary researcher.

The post-training assessments were made the week following the end of the training phase period.

4.3 Results

4.3.1 Comparison of excluded participants and retained participants at T1

To ensure there was no evidence that the excluded group of participants formed a systematic grouping when compared to the retained participants a series of one-way ANOVA tests were conducted on the primary dependent variable for each of the 8 individual tasks conducted at T1 with group (excluded vs retained) as the IV. Each result provided $p > .05(.19 - .83)$ suggesting the two groups did not differ in baseline ability.

4.3.2 Correlation/PC Analysis of T1 data

As stated in the introduction section to this study we wanted to have a larger battery of pre-post tasks in this study. To ensure that the individual task scores can be combined and extract the coefficients to use to combine them into latent factors we ran a principal components analysis on the pre-training battery of tasks.

Correlations The following analysis was conducted using the full dataset from the pre-training phase which included 115 participants. A 90% winsorisation procedure was performed on the processing speed measures before computing the correlation matrix (Table 4.1). There is a large degree of collinearity within the data as one would expect. The processing speed measures appear to share little to no linear relationship which is a surprising result and casts doubt on our ability to combine these meaningfully into one processing speed composite measure. They do at least correlate with each other $(.19, p < .05)$ but they have stronger correlations with other measures than each other.

Table 4.1

Correlation Matrix of each primary dependent variable; Digit = Digit Span t-score, Op = Operation Span t-score, Reading = Reading Span t-score, Matrix = Matrix Span t-score, Symm = Symmetry Span t-score, IT = Inspection Time mean SOA, OOO = Odd One Out decision time, M-Arith = Total correct operations across all blocks

	Digit	Op	Read	Matrix	Symm	IT	OOO
Digit							
Op	0.63***						
Read	0.52***	0.71***					
Matrix	0.44***	0.44***	0.43***				
Symm	0.40***	0.54***	0.52***	0.63***			
IT	-0.36***	-0.44***	-0.35***	-0.28**	-0.46***		
OMO	-0.23*	-0.11	-0.12	-0.17	-0.25**	0.19*	
M-Arith	0.42***	0.52***	0.53***	0.38***	0.49***	-0.38***	-0.13

The results of Bartlett's test was $\chi^2 = 292.01, p < .0001$ and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy = .8 which is a 'good' score (Kaiser, 1974, Hutcheson & Sofroniou, 1999) thus suggesting the data is suitable for a PCA.

Initial Solution Note that the mental arithmetic variable was included in the correlation matrix for descriptive purposes and for some discussion later but was not included in this analysis.

Figure 4.9 shows the scree plot produced by the initial PCA while Table 4.2 shows

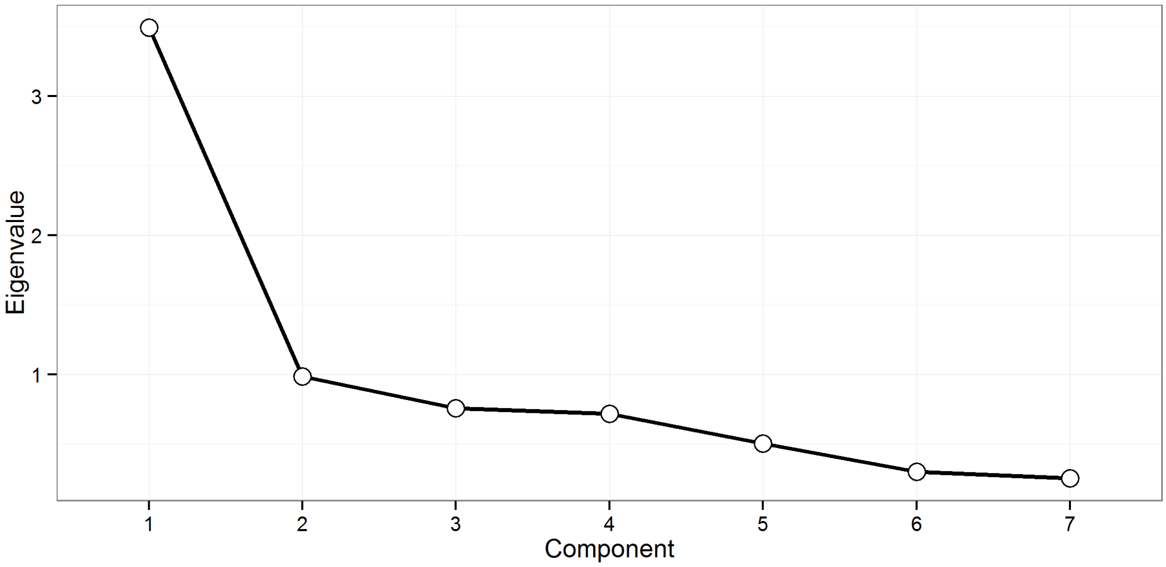


Figure 4.9. Scree plot for initial PCA solution

Table 4.2

Variance explained by the extracted principal components (initial unrotated solution)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SS loadings	3.493	0.983	0.756	0.716	0.502	0.299	0.251
Proportion Var	0.499	0.140	0.108	0.102	0.072	0.043	0.036
Cumulative Var	0.499	0.639	0.747	0.850	0.921	0.964	1.000

the variance explained properties of the outcome. We can see that only one PC has a high eigenvalue (> 1). However, this one PC explains only half the variance in the individual variables. From a cumulative variance accounted-for perspective a 3 PC would be required to explain 75% and would also give the possibility of the three PCs we want which when we investigate rotation strategies will likely spread the variance across those PCs differently.

3-Factor solution The PCA analysis was conducted again extracting only three components and using a promax rotation strategy as we know that the extracted components should have some shared variance (e.g. Kane et al., 2004). Table 4.3 shows the loadings matrix for the this solution. From this table of loadings it is clear that a processing speed PC does not emerge. The Inspection Time measure is combined with the verbal memory PC while the Odd One Out task is catered for by PC2 exclusively.

Table 4.3

Loadings matrix for the 3-PC promax solution

	PC1	PC3	PC2
Digit T-Score	0.78		
Op T-Score	0.91		
Read T-Score	0.80		
Matrix T-Score		0.95	
Symm T-Score		0.71	
IT SOA	-0.67		
OMO DT			0.98

Given the confusing pattern of loadings seen in this three factor extraction and given the processing speed measures had already been flagged as a potential problem it seems sensible to conclude that there will be no solution where a processing speed exclusive PC emerges. Thus in order to generate loadings that could be used to generate two memory components the PCA will be recalculated using only the memory span tasks. The processing speed measures will be considered individually for transfer.

Memory Span PCA Table 4.4 shows the variance summary information for the extracted PCs in the initial solution while Table 4.5 shows the rotated loadings matrix for this solution. The solution provides a clear distinction between the spatial and verbal memory measures with both PCs correlating at .57. The impact of this is that it is reasonable to create a composite score for verbal and spatial factors and not a processing speed factor with the loadings from this solution providing suitable weights for the composite scores.

Table 4.4

Variance explained by the PCs (memory span only - initial solution)

	PC1	PC2	PC3	PC4	PC5
SS loadings	3.120	0.767	0.521	0.335	0.258
Proportion Var	0.624	0.153	0.104	0.067	0.052
Cumulative Var	0.624	0.777	0.881	0.948	1.000

Table 4.5

Loadings matrix for the 2-PC promax solution (memory span only)

	PC1	PC2
digit.trans.score	0.86	
op.trans.score	0.89	
read.trans.score	0.82	
matrix.trans.score		0.94
symm.trans.score		0.84

discussion point: lack of processing speed factor The failure to produce a processing speed principal component was surprising and perhaps deserves further investigation at a later point. It may be the case that my implementation of one of the tasks was not up to the standard of the versions used in other studies. The correlations between the OOO decision time and IT mean SOA were .399 for adults and .396 for children in Nettelbeck and Burns (2010), while a correlation of .354 emerges in O'Connor and Burns (2003) with an adult sample. This compares with a .19 correlation obtained from this data. It may be a result of a lack of sensitivity in the equipment used given that we were restricted to conducting these tests on the computers provided by the schools that volunteered to participate. Alternatively it may be that the values found by Nettelbeck and colleagues are on the higher end of the error range while the data here place in the lower end of that range.

4.3.3 Practice Effects

In this section the degree of improvement in performance (or lack of) over the course of the 5-week training phase will be assessed. While the pre-post analyses assessing transfer effects focus on a subset of the overall sample due to various reasons (outlined above) no such exclusions will apply here. If a participant has only conducted three sessions then that is still some worthwhile information to include in the assessment of any changes as sessions increase. The hierarchical approach to analysing the data means we can handle the unbalanced structure of the dataset.

sessions/number of trials A potential issue in using amount of sessions completed as an indicator of the amount of training a participant has completed is that due to the training sessions being capped by a time limit as opposed to a number of trials, the number of trials in any session can vary. This coupled with there being some scope for children to be distracted if the teacher was not able to ensure a completely favourable environment for any given session leads to some sessions having a small number of trials compared with the average. Figure 4.10 graphically illustrates this issue by plotting the total number of sessions a participant has completed with the total number of trials they completed, separated for each training group. As one would expect there is a clear increasing linear relationship where the number of trials completed increases for each additional session a person has completed. However, one can also see the variability 'within' each total session count. For example, if you look at the WMP facet of the plot, specifically at (for example) those who have completed 9 sessions. There is a cluster of six participants who completed nine sessions and totalled between 125-250 trials. The difference between 125 and 250 trials would already suggest a large difference in the amount of training completed, but the issue becomes even more apparent when you consider that there are other participants at nine sessions completed but they are up between 375 and 500 total trials.

For this reason the training data are presented in two forms, one where performance is measured based on aggregate measures from whole 'sessions' which is a 15 minute block

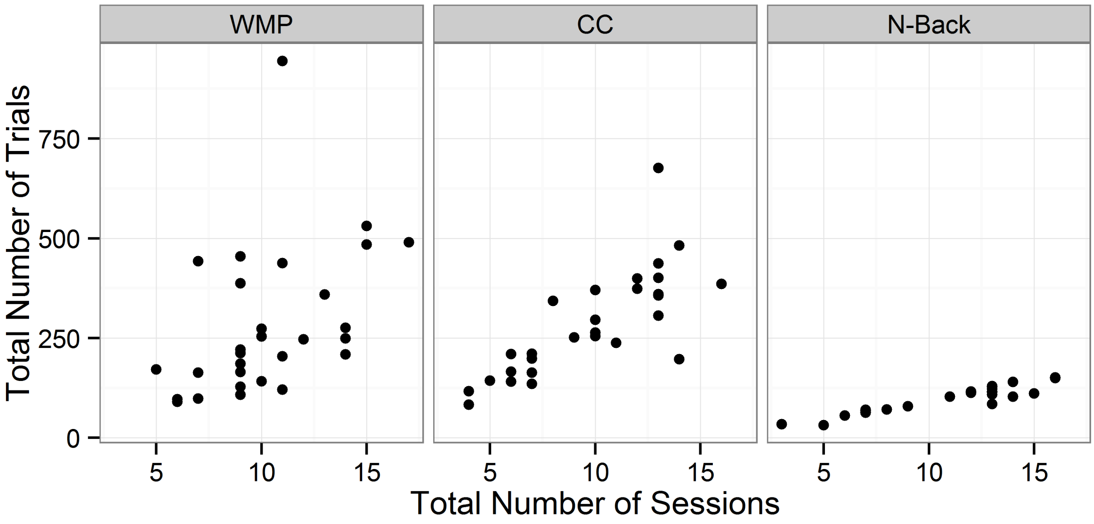


Figure 4.10. Scatterplot showing the relationship between number of 'sessions' completed and the number of trials this resulted in for each person

regardless of the number of trials completed. In addition, the training data were also split into trial 'blocks'. The number of trials that made up a block was decided based on the average number of trials completed by the participants. The mean values were 284 for CC, 280 for WMP and 96 for n-back (a 'trial' on n-back is one stream of $15+n$ items) and therefore the splits were set at 20, 20, and 8 respectively. This method ensures that the comparison of blocks between participants is valid with respect to the amount of trials completed in the whole training phase up to that point, i.e. all block 3 measurements refer to trials 41-60 for WMP/CC participants and trials 17-24 for n-back participants.

Figure 4.11 shows the participant attrition for the training data by plotting the number of participants who completed each session number. The profile is very similar for each group showing the majority of participants completed 8-10 sessions while few completed a higher or lower number of sessions. This figure is useful when looking at the results highlighting performance level at each session for evaluating the width of the confidence intervals. The participants who completed less than 8 sessions were excluded from the pre-post analyses. This shows how much data was lost in the pre-post analyses from the original recruitment; 28 participants started in the CC group and 11 of those completed

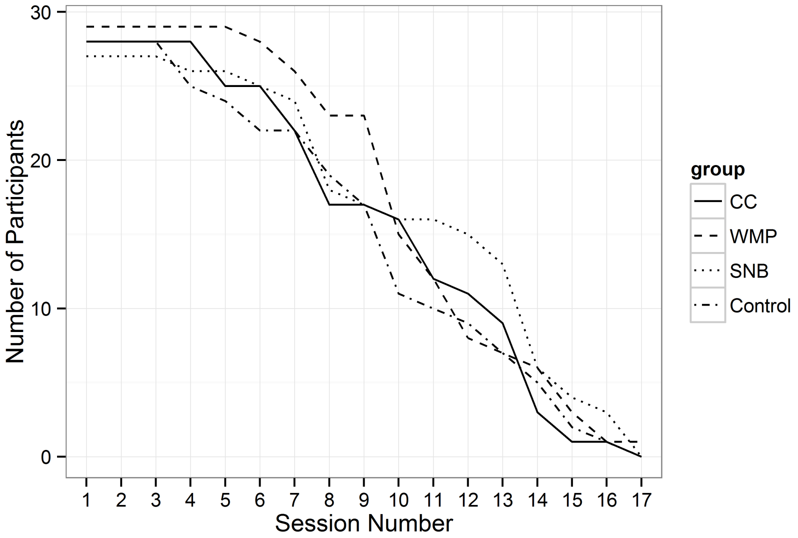


Figure 4.11. This figure highlights the number of participants who reached each stage of the training for each of the groups.

less than eight training sessions.

As detailed in the method section the training tasks implemented an adaptive difficulty algorithm which adjusted the level (where higher means more difficult) of the task in accordance with current performance. Therefore an aggregate of the level at which a participant was working at for a session is a reasonable indicator of overall performance. It is important to note that performance carries over so that if a participant was working at level four when the previous session ended they will start at level four for the current session. To analyse if the mean session level varied as a function of session number a generalised linear model was fitted, again using a random effect at the level-two unit of subject. This is in place of general descriptives due to the mean values at each session being uninterpretable without controlling for the participant attrition as session number increases. A suitable quantification of the group performance at each session number can be attained by fitting the hierarchical linear model with the random intercept at the subject unit. By taking into account the individual variability in starting scores the parameter estimates at each session number are not biased as a result of drop-out.

The fixed effect element of the model is simply mean session level as the dependent

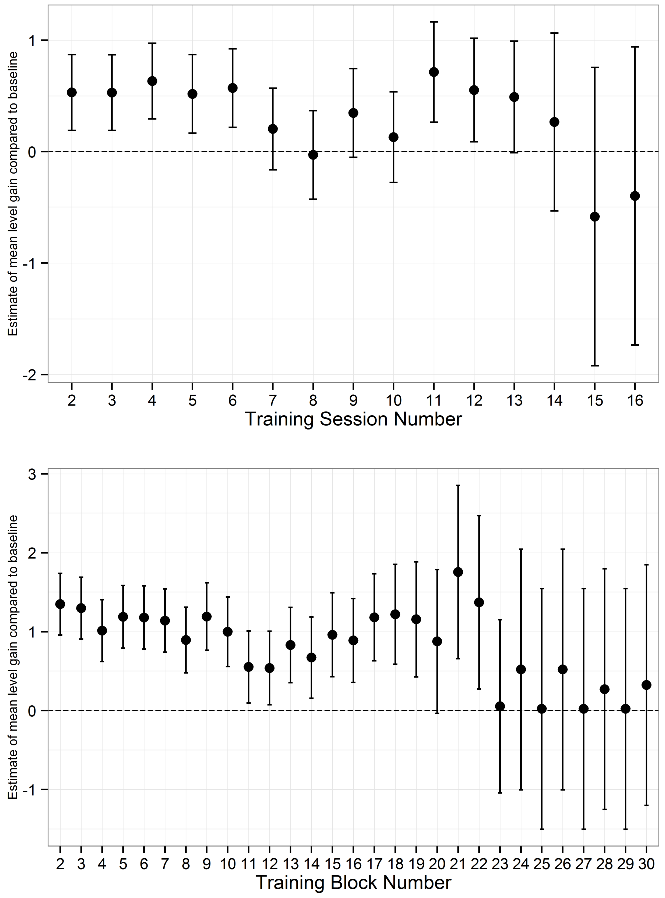
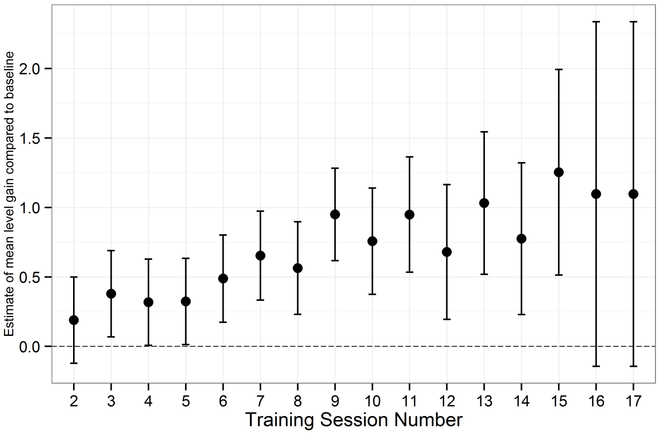


Figure 4.12. CC; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks

variable with a fixed intercept and session ID/trial block as predictors. Then the random effects part of the model is a subject-specific random intercept with a mean of the fixed effect intercept and a calculated variance. By treating session ID/trial block as a categorical variable we can obtain a beta estimate of the amount that the mean session level varies compared to the 'reference' which will be session one (baseline) scores. Figure 4.12 shows graphically what these beta estimates are with their 95% confidence intervals.

Figure 4.11 shows that of the 29 participants that began the training phase, six of them will be removed for the pre-post analysis (< 8 sessions). Figure 4.13 shows the profile of mean level change as WMP sessions progressed. The values for mean level change in the level of n for the SNB task is shown in Figure 4.14



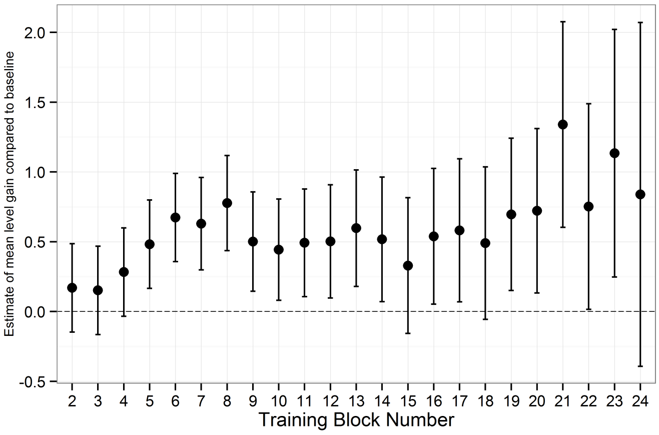


Figure 4.13. WMP; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks

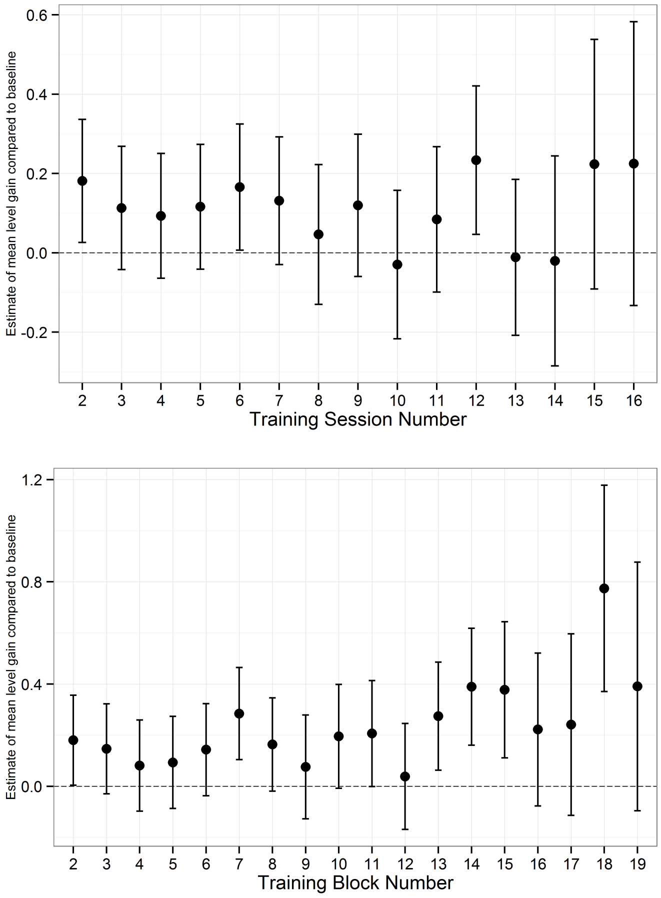


Figure 4.14. N-Back; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks

Table 4.6

Pre- and post-training performance on each of the transfer tasks for the three training groups and active control. Values represent means with standard deviation in parenthesis. Tasks; CC = Colour Corsi, WMP = Working Memory Period, SNB = Spatial N-Back

Task/Measure	CC		WMP		SNB		Control	
	pre	post	pre	post	pre	post	pre	post
Digit Span								
FTA Score	13.71(5.6)	16.06(5.5)	13.32(4.1)	12.7(4.9)	14.06(5.6)	16.11(6.8)	11.22(4.6)	13.53(4.1)
T-Score	25.45(6.6)	28.44(6.6)	26.47(6)	24.64(8.1)	27.34(7.2)	28.8(7.6)	24.24(6.7)	25.62(5.6)
Operation Span								
FTA Score	7.59(5.4)	8.59(7.4)	6.82(6.9)	6.2(6)	7(6.7)	8.56(5.8)	7.06(5.9)	7.4(7.5)
T-Score	16.44(7)	15.04(9.2)	13.88(7.7)	12.58(7.6)	15.23(8.1)	17.19(8.5)	14.19(7.9)	14.3(9.3)
Reading Span								
FTA Score	7.76(6.6)	6.47(6)	6.86(4.8)	7.1(4.7)	9.28(5.5)	7.72(5.8)	7.89(6.4)	4.83(6.5)
T-Score	15.03(8.5)	9.9(7.8)	13.81(6.9)	12.43(7.3)	16.59(7.1)	14.91(9)	13.89(7.7)	9.44(9.1)
Matrix Span								
FTA Score	34.71(11.9)	36.18(8.3)	27.68(9.9)	33.14(9.1)	31.89(9.8)	38.06(13.2)	28.68(11.6)	30.33(10.2)
T-Score	50.67(8.8)	49.04(7.6)	45.18(8.7)	47.92(8.1)	49.36(7.4)	51.81(7.9)	45.05(9.6)	46.3(8.5)
Symmetry Span								
FTA Score	15.94(10.2)	16.65(10.7)	11.41(8.9)	14.29(10.9)	19.61(12.4)	17.78(11.3)	13.42(10.9)	14.06(11.3)
T-Score	29.24(9)	26.06(10.7)	23.17(9.6)	24.67(11.9)	30.87(9.8)	29.02(11.3)	24.74(10.9)	24.17(11.2)
Inspection Time								
Mean SOA	193.65(75.2)	157.49(37.9)	208.38(96.9)	186.99(55.6)	161.24(69.2)	161.31(51.5)	211.44(83.6)	186.22(73.8)
Odd One Out								
Decision Time	532.71(146.8)	507.62(134.6)	543.59(170.6)	461.66(105.2)	507.72(148)	503.87(136.4)	606.13(199.4)	531.24(151.6)
M-Arithmetic								
Total Correct	58.35(23.3)	55.29(17.62)	52.86(22.4)	43.35(25.6)	65.33(26)	64.17(27.7)	63.33(16.3)	46.18(19.13)

4.3.4 Transfer Effects

Table 4.6 shows descriptive statistics for the main dependent variables pre-training and post-training for the CC, WMP, N-Back, and Active-Control groups respectively.

Pre-Post Analysis of Verbal/Spatial Latent Factors Composite scores were calculated combining the three verbal WM tasks (Digit Span, Operation Span, and Reading Span) and the two visuospatial WM tasks (Matrix Span and Symmetry Span) using the weights derived from the principal components analysis (see Table 4.5). Figure 4.15 shows the pre- and post-training mean scores for these measures. These plots suggest that overall there is not much change from pre-training to post-training for any of the groups. This was formally tested by means of an ANCOVA modelling post-training scores with training group as the IV and scores at pre-training as a covariate. The top two rows in Table 4.8 show the F- and p-values from these analyses along with effect sizes that are Hedges' g effect sizes using the adjusted means (adjusted for covariate) of each pairwise comparison comparing the WM training groups to the active control group. This analysis confirms what Figure 4.15 suggests that there is no substantial evidence from this dataset that any of the interventions improved overall performance on the verbal or visuospatial composite scores.

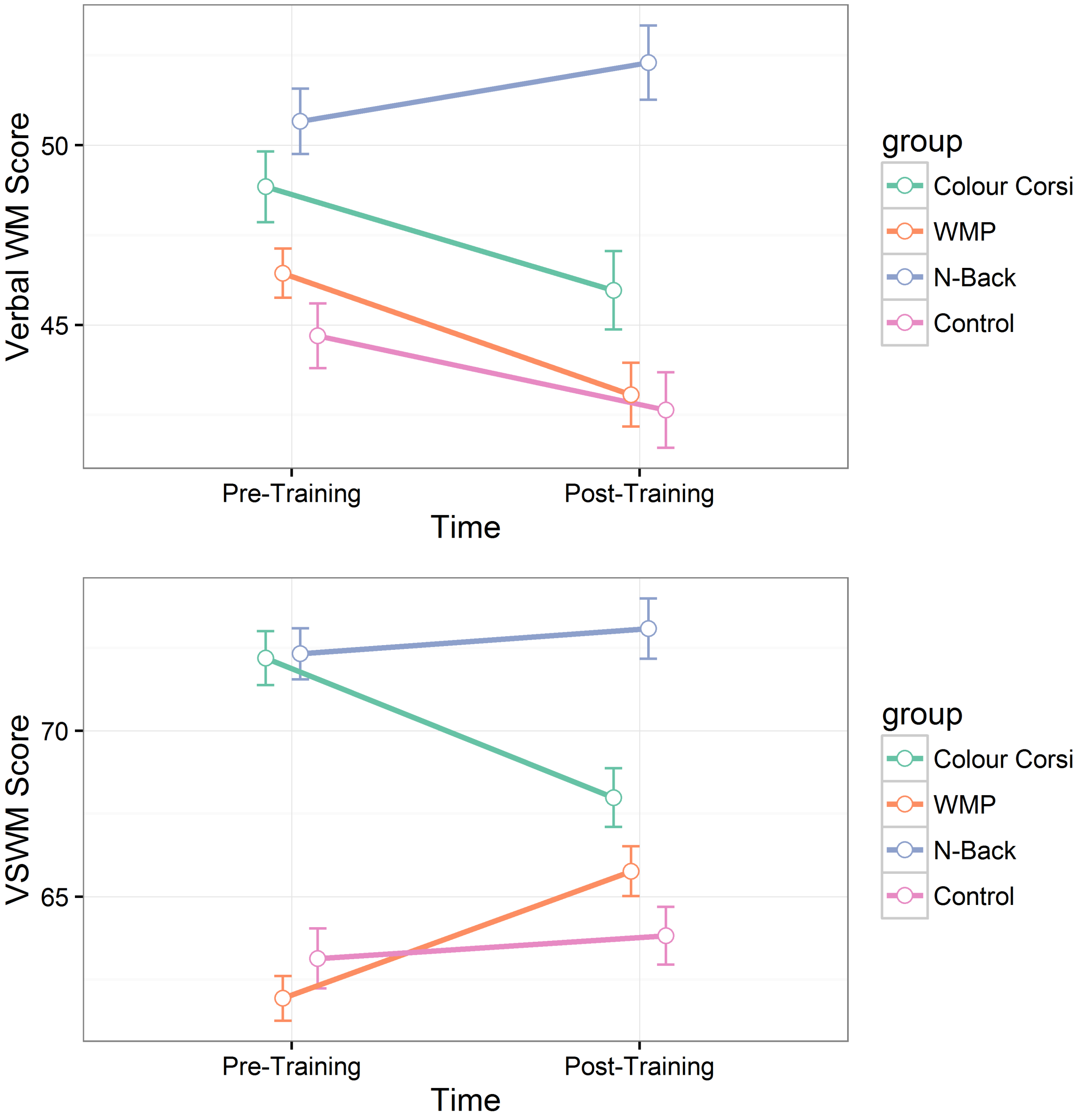


Figure 4.15. Pre- and Post-Training scores for the verbal and visuo-spatial WM PC scores as determined by the coefficients derived from pre-training PCA analysis with standard errors.

The effects of the different interventions compared to the active control group were also assessed by means of a generalised linear model with random effect structure to account for the within-subject variability. By taking into account the two level structure of the data and allowing the intercepts to vary between participants (modelled as a normally distributed random variable) we resolve the issue of the vector of scores that makes up the dependent variable not being independent without this parameter. A group by time interaction in this design would indicate differing slopes between pre- and post-training time points. This method has the advantage of relaxing the homogeneity of slopes assumption that is important for the ANCOVA results. By providing results from both these types of analysis a comprehensive view of what the data is saying can be seen. Note the processing response times were analysed using the mixed model only. Table 4.9 indicates whether the fixed effect interaction of time by group is significant (as the models are nested a likelihood ratio test can be conducted to assess if the addition of the interaction term is significant) and provides the beta values produced for the interaction effect which indicate in raw units the difference at post-training compared to the control group for further descriptive purposes (due to general non-significance). The results for the composite measures confirm the ANCOVA results.

Pre-Post Analysis of Individual Tasks One of the primary goals of this follow-up WM training study was to focus assessment of transfer on composite scores made up of multiple tests. But as already noted the processing speed measures seem to share too little variance (see PCA section) to have confidence in the validity of a composite as an enhanced indicator of processing speed. In addition the mental arithmetic task included as a far-transfer measure needs to be assessed individually. Additionally, while the composite measures for the memory span tasks yielded non-significance it was decided that it would be useful to also conduct the same analyses on those tasks individually. Therefore Tables 4.8 and 4.9 also show the results when individual task dependent variables were assessed. The results suggest no effect of group on any of these measures for each task except for mental arithmetic. This significant effect is due to the N-Back and Colour Corsi groups not decreasing as much as the control group (but still decreasing overall) (see Table 4.6 for the mean values for each group at each phase).

Table 4.7

Summary of performance on the processing phase of complex span tasks at pre- and post-training. Accuracy = proportion correct, RT = response time in milliseconds

	Colour Corsi		WMP		N-Back		Control	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Operation Span								
Accuracy	.71	.72	.6	.63	.72	.74	.72	.7
RT (ms)	2841	2821	2362	2634	3146	2878	3141	2656
Reading Span								
Accuracy	.73	.74	.7	.7	.77	.75	.73	.66
RT (ms)	2773	2204	2492	2336	2791	2136	2588	2061
Symmetry Span								
Accuracy	.83	.83	.76	.77	.84	.78	.83	.76
RT (ms)	2152	1657	2064	1543	2105	1458	2179	1511

Table 4.7 includes each group mean value of the median RT for correct operations for each of the complex span tasks at pre- and post-training. ANCOVA analyses suggest no difference at post-training controlled for pre-training values (Hedges' g with 95% CI for each pairwise comparison against the control group for post-training adjusted means in the following order; CC, WMP, SNB); Operation Span - F(3,66) = 0.169, p = .917, g = 0.2[-0.46,0.87], 0.19[-0.43,0.81], and 0.16[-0.49,0.82], Reading Span - F(3,69) = 0.382, p = .766, g = 0.08[-0.59,0.74], 0.29[-0.34,0.91], and 0.01[-0.64,0.67], Symmetry Span - F(3.69) = 0.543, p = 0.654, g = 0.33[-0.34,1], 0.23[-0.39,0.85], -0.01[-0.66,0.65]. The processing accuracy values were also tested and yielded; Operation Span - F(3,66) = 0.737, p = 0.534, g = 0.24[-0.42,0.91], -0.03[-0.65,0.6], and 0.37[-0.29,1.03], Reading Span - F(3,69) = 1.246, p = 0.299, g = 0.54[-0.13,1.25], 0.32[-0.3,0.95], and 0.53[-0.13,1.2], Symmetry Span - F(3,69) = 1.256, p = .296, g = 0.53[-0.14,1.21], 0.43[-0.2,1.06], and 0.11[-0.55,0.76].

Table 4.8

Summary of Ancova results plus the effect size (Hedges' g) for pairwise comparisons comparing the adjusted means for each training group to the active control group.

	F	*p*	CC	WMP	N-Back
PC Scores					
Verbal PC Score	0.625	.6	-0.01[-0.36,0.35]	0.02[-0.6,0.64]	0.35[-0.31,1.01]
Spatial PC Score	1.07	.368	-0.2[-0.86,0.47]	0.29[-0.34,0.91]	0.27[-0.39,0.93]
Digit Span					
fta.score	0.124	0.945	0.2[-0.46,0.87]	-0.39[-1.02,0.23]	0.17[-0.48,0.83]
trans.score	1.592	0.2	0.27[-0.39,0.94]	-0.34[-0.96,0.29]	0.17[-0.49,0.82]
Operation Span					
fta.score	0.377	0.77	0.06[-0.42,0.53]	-0.2[-0.62,0.22]	0.11[-0.35,0.57]
trans.score	0.706	0.552	-0.06[-0.72,0.61]	-0.07[-0.69,0.55]	0.32[-0.34,0.98]
Reading Span					
fta.score	1.119	0.348	0.21[-0.45,0.88]	0.32[-0.31,0.95]	0.27[-0.39,0.92]
trans.score	1.606	0.196	-0.01[-0.67,0.66]	0.41[-0.21,1.04]	0.55[-0.11,1.22]
Matrix Span					
fta.score	1.653	0.185	0.34[-0.33,1.01]	0.36[-0.26,0.99]	0.72[0.04,1.39]
trans.score	1.465	0.232	-0.06[-0.72,0.6]	0.28[-0.35,0.9]	0.54[-0.13,1.2]
Symmetry Span					
fta.score	0.124	0.945	0.16[-0.51,0.82]	0.16[-0.47,0.78]	0.04[-0.78,0.82]
trans.score	0.335	0.8	-0.12[-0.78,0.55]	0.19[-0.43,0.81]	0.07[-0.58,0.73]
Inspection Time					
mean.soa	0.876	0.458	-0.49[-1.17,0.18]	-0.23[-0.85,0.4]	-0.1[-0.75,0.55]
Odd One Out					
median.dt	1.079	0.364	-0.31[-0.97,0.36]	-0.55[-1.18,0.09]	-0.23[-0.89,0.42]
Mental Arithmetic					
total.corr	4.349	0.007*	0.88[0.18,1.58]	0.52[-0.12,1.15]	1.08[0.38,1.78]

Table 4.9

Summary of the time by group interaction fixed effect for each depdendent variable of interest. The likelihood ratio test compares the fit of the model with and without the interaction effect where a significant result indicates improved model fit

	L-Ratio	p-value	T2:CC	T2:WMP	T2:N-Back
PC Scores					
Verbal PC Score	1.48	.69	-0.82	-0.08	3.69
Spatial PC Score	4.47	.22	-3.65	3.28	1.31
Digit Span					
fta.score	4.21	.24	-0.02	-2.8	-0.32
trans.score	4.75	.19	1.4	-3.1	-0.11
Operation Span					
fta.score	0.9	.82	0.05	-1.17	0.61
trans.score	1.98	.58	-1.33	-0.65	2.02
processing.rt	3.04	.39	470.78	762.21	223.25
Reading Span					
fta.score	3.42	.33	1.82	3.37	1.56
trans.score	3.5	.32	-0.54	3.11	2.92
processing.rt	2.16	.54	-31.63	379.48	-117.38
Matrix Span					
fta.score	3.54	.32	0.31	3.78	5.01
trans.score	4.66	.2	-2.19	1.67	1.9
Symmetry Span					
fta.score	1.99	.57	0.43	2.35	-2.11
trans.score	2.33	.51	-2.06	2.03	-0.72
processing.rt	1.47	.69	186.32	176.7	34.66
Inspection Time					
mean.soa (ms)	4.41	.22	-6.84	2.38	29.39
Odd One Out					
median.dt (ms)	2.75	.43	52.84	-3.3	74.07
Mental Arithmetic					
total.corr	11.62	.01*	14.69 (p=.01)	10.1 (p=.05)	16.58 (p<.01)

4.4 Discussion

Transfer Effects The general pattern of results seen in this experiment once again point to a lack of generalisable improvement in the working memory construct. By using composite scores calculated based on performance on multiple individual tasks more reliable indicators of verbal and visuospatial working memory were assessed. None of the WM training groups improved on these composite measures above that of an active control group. The training paradigms tested in this experiment each focused on different aspects of working memory. The Working Memory Period training task adapted difficulty by increasing the amount of processing required while maintaining the same amount of verbal domain to-be-remembered material. The Colour Corsi task is a more typical simple span paradigm with sequences of TBR memoranda presented and required to be recalled in correct serial order. The difficulty is adapted via increasing the list length of TBR memoranda which are multi-feature (colour-location) visuospatial type material. And finally, the spatial n-back (SNB) task requires the constant maintenance and updating of spatial material where the number of constantly held items matches the current level of n. Therefore, the data presented in this study is a broad assessment of WM training including training paradigms that 'hit' different elements of working memory.

The results outlined in Tables 4.6, 4.8, and 4.9 offer strong support for no generalisable change in the tested constructs as a result of either intervention. With particular interest in the results for the composite scores for verbal and visuospatial WM it can be seen there is no effect of training intervention (p-values of .6 and .368 respectively). The contributors for the drop in performance are not ubiquitous. Both the CC and Control training group showed much weaker scores on the reading span task at post-training which accounts for the overall drop in the verbal PC score. However, the WMP training group did not exhibit this drop-off in the reading span task but overall showed slightly worse scores at post-training. The N-Back training group also showed weaker scores at post-training for reading span but not to the same degree, this coupled with improved scores on the other verbal measures ensured the overall score change was in the positive direction. Rather surprisingly the only training group to perform worse on the spatial tasks at post-training was the CC training group while the WMP training group improved the most.

However, interpreting these changes as anything other than standard fluctuations over multiple testing is unwarranted given the magnitude of changes compared with the observed variability. In some instances the interpretation of these results may tend to focus on the comparisons where a positive effect size emerges and discuss this as a trend rather than a significant result. For example, the SNB training group did yield results where the effect size was in the positive direction indicating improved performance over the active control group (the negative effect sizes for the processing speed tasks indicate faster times and hence a decrease in the measure). However, only the FTA measure for the Matrix Span task individually yields an effect size where zero is not in the 95% interval range. this is not including the Mental Arithmetic measure due to the effect being mediated by a large decrease seen in the active control group scores at post-training rather than increased improvement as a function of the intervention. Spurious results such as that seen with the Mental Arithmetic task are possible due to noise and regression to the mean effects amongst other potential explanations. Such issues are magnified when n is small. I would be hesitant to describe these data as suggesting anything other than a lack of generalisable near-transfer to working memory and a lack of far-transfer to measures of processing speed and mental arithmetic. To focus on the trend of positive effect sizes for the N-Back training group would be overstating the presented evidence.

One such way in which participants may improve their WM performance is by becoming more efficient at the processing elements of the complex span tasks or improving the resilience of the TBR material in the face of said processing. The WMP task provides a concurrent processing focus to the WM training. The results show that with regards the response time, and also accuracy, for the processing elements of the complex span measures there was no beneficial impact of the WMP intervention nor the alternative training paradigms. The effect sizes regarding the concurrent processing aspects of the complex span measures also yield trends towards effects with a number of g values above 0.4 but again the confidence range for these estimates includes 0 for each due to the large variance observed and relatively low n. Additionally, the trends are generally the result of a decrease in performance for the control group and when this occurs the improvement required in the training groups to generate a meaningful effect size decreases.

Practice Effects The general lack of transfer once again observed suggests that for these training programs, at these 'doses' there is no genuine improvement in working memory performance as a consequence of the interventions. However, there was also a lack of practice effects observed in the data across the different training regimes. That no transfer effects were observed is not a result which significantly conflicts with the literature if it is judged through a critical lens as covered in the literature review for this thesis and in the published literature (Shipstead et al., 2010, 2012). However, the observation that the children in this study showed very little improvement on the training tasks over repeated sessions is more surprising and has ramifications for assessing the results on the transfer analyses. As the Working Memory Period and Colour Corsi tasks represent somewhat unique additions to the training literature there is little literature to compare these results with to attempt to explain these results with regards to whether these tasks for this development group are generally able to improve over repeated sessions, or if the environment variables due to the group/classroom setting are the significant cause. There are comparable datasets discussed in the literature to compare for the spatial N-Back training task due to the growing literature around that paradigm. Jaeggi et al. (2011) found significant practice effects by comparing mean level performance during the first two training sessions to that of the final two training sessions. At a group level their training group was able to improve 0.76 levels (from 2.17 to 2.93). The data we observed showed much smaller overall gains and a generally lower level of performance as well. Performance at some sessions was above baseline performance (session one) but no subsequent session produced significantly better performance than that observed in the second session. The mean level of performance across all spatial n-back training was 1.67 (full spaghetti plot provided in appendix, see Figure B.5) which shows that the general profile of performance was a move up to level two where much of the time the participants were unable to be successful enough to maintain that level and thus would have been moved back down to 1-back. It is surprising how low performance was in the task as the children sampled in this study were slightly older than those in the Jaeggi and colleagues study (mean age 9.12 but sd of 1.52 indicating a proportion of the sample would have been equivalent ages to those tested here). It is not specified how the training sessions were administered in Jaeggi and colleagues paper and thus it is not clear if factors attributed to that aspect of the experimental design can be considered as prime candidates to explain the differences observed. Zhao et al. (2011) provide a further example with more pronounced practice effects for a WM updating but this is based on the Running Memory task (both versions used; visual/spatial stimuli although visual stimuli like conflated with verbal label attachments by participants). The literature on typically developing children is thin, as discussed in the introduction. There are numerous published studies using dual/single n-back training paradigms and the general pattern of practice effects observed is of improved performance session on session (see for exmaple; Li et al., 2008; Jaeggi et al., 2008; Redick et al., 2013; Thompson et al., 2013; Colom et al., 2013).

Limitations The large variability observed in the majority of variables coupled with some spurious results such as large decrease in performance between pre- and post-training for some measures points towards the environmental factors impacting the results. While this clearly needs to be taken into account when assessing the data provided here and what it means for whether WM training can possibly produce generalisable improvements in the construct, there remain useful conclusions that can be drawn with regards the evidence for how and when such transfer can occur and whether this type of administration of training can be effective. The commercialisation of 'brain-training' products that claim to be scientifically backed are generally some variant on the types of WM training assessed in this thesis and the literature discussed. Not all of these commercial products are sold with a manual of how the training needs to be conducted to be successful. A product targeting the education sector might suggest the training can be conducted in groups as part of the normal curriculum but data obtained here either suggests that these adaptive WM training interventions may not produce robust generalisable improvements in the WM construct and if they can produce such effects, the consequences of conducting the training in the classroom environment provide a significant barrier for the progress of such effects. A limitation of much psychological research is that of running studies with low power for detecting effects (Cumming, 2014). Post-Hoc power analyses (see section 2.3.5) showed that for 80% power the number of participants in each group would have needed to be 27 for verbal near-transfer and a much larger 60 for visuospatial near-transfer. As with the first WM training study outlined in this thesis we were unable to enrol as many participants in each group as we would initially have liked, and then due to the real-world setting of the research that reduces the control we have as experimenters, the amount of data from each participant (number of training sessions) was compromised. The combination of these two factors leads to analyses with much less power than originally intended and required according to the power analysis. This is particularly evidenced when assessing the observed effect sizes. There are numerous effect sizes reported in the 0.3-0.6 range (improvements of 0.3-0.6 standard deviations above that seen in the control group) but due to the low power and high variability the uncertainty in these estimates gives wide confidence intervals. Effect sizes of this magnitude compare well with the literature when multiple analyses are compared such as from meta-analyses (Melby-Lervåg & Hulme, 2013; Au et al., 2015).

Chapter 5

Working Memory Training Adult Study

5.1 Introduction

The results observed in this thesis so far with regards to working memory training relate to classroom based group training in typically developing children. There are only a small number of published studies using typically developing samples and often these use pre-school aged children (e.g. Thorell et al., 2009; Bergman Nutley et al., 2011). Therefore a potential explanation for the results described showing a lack of improvement in working memory after training could be that either, a) the interventions are ineffective for this group, or b) factors relating to the environment in which the interventions took place impede the mechanisms of improvement. Explanation (a) may reflect characteristics of the group beyond their age and typical development status. Instead one might point to motivational factors as a candidate for explaining these findings. The children who participated did not volunteer themselves rather their schools signed up to the research program. Jaeggi, Buschkuehl, Shah, and Jonides (2014) provided data that suggest a modest relationship between intrinsic motivation as measured by self-reported engagement levels and the magnitude of gain observed in the trained tasks. Explanation (b) offers a number of extrinsic factors that could affect ones attention during the training phases despite the best efforts of researchers and teachers to ensure a suitable distraction free environment.

We therefore set out to carry out a comparable working memory training study that sampled healthy adult participants. This sample will allow comparison to a greater amount of published work in the field. If the effects of practice on these working memory tasks are notably different in the adult sample then it may suggest that the factors noted above are impeding the mechanisms improving working memory function in those children. However, if the results are consistent with the previous studies, and no generalised improvement to working memory performance is observed, then the generalisability of these findings is improved.

The updating training paradigm of n-back (in particular dual n-back) has gained significant traction based on publication of a number of studies showing significant transfer to Gf (Jaeggi et al., 2008; Rudebeck, Bor, Ormond, O'Reilly, & Lee, 2012). Au et al. (2015) conducted a meta-analysis including only studies that utilised an n-back training procedure and at least one transfer task measuring fluid intelligence. Their meta-analysis included 20 studies with 24 comparisons and found a significant positive overall effect size of 0.241 (se = 0.07). Unfortunately this meta-analysis may not be as conclusive as one would hope form a meta-analysis for a number of reasons. Firstly, as noted by Moody (2009) the administration of popular measures of Gf often strays from the prescribed protocol and in doing so invalidates the measure. The particular problem Moody noted was that in Jaeggi and colleague's (2008; 2010) methodology they administered the BOMAT with a 10-minute time limit. The BOMAT is made up 29 test items of increasing difficulty and recipients are supposed to be allowed 45 minutes to complete these items. Moody argues that the reduction in allotted time significantly alters the meaning of the dependent variable (the number of successfully answered visual analogies) as it is now a test of speed on the easier analogies as opposed to a measure indicating the highest difficulty the participant is able to successfully complete. This methodology was applied in five of the

24 comparisons included in the meta-analysis. Moreover the method of using effect sizes assessing the difference in post-tests scores for the treatment and control in the Au et al. meta-analysis leads to spurious effect sizes. As an example, one of the studies included in the analysis was conducted by Salminen, Strobach, and Schubert (2012) and was reported as providing a Hedge's g of 0.816 with regards n-back training transfer to Gf. However, Salminen et al. (2012) find no transfer to Gf as a function of the n-back training paradigm. In fact they find a significant interaction that is explained by an increase from pre- to post-training on Raven's Advanced Progressive Matrices for the control group and a lack of change for the training group. Thus, even though at the post-training phase the trained group provided higher scores than the control group this could not be attributed to the training intervention. An effect size using pre-training adjusted values would therefore provide a negative effect size for this comparison.

The drawbacks of the updating training evidence noted above illuminate the issue that the currently perceived most successful training paradigm is still built on an evidence base that is unclear with regards to the generalisable benefits to higher order cognitive functions such as fluid intelligence. Therefore, the same arguments presented in this thesis regarding a need to understand the mechanisms by which training programs affect the working memory system in order to produce and explain robust transfer effects also applies to the updating training literature. This issue in regards n-back training is further complicated by the lack of a clear understanding of the relationship between n-back and commonly used complex span tasks (Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Redick & Lindsey, 2013).

The core research question in this study extends the work described so far - does practice on adaptive-difficulty working memory based tasks produce robust improvement to working memory performance in a healthy adult sample? The three training groups will each train on a single task providing three different training groups that will be compared to an active control group. The design of the experiment is such that the experiment is as close to the previous design using a typically developing sample. The only significant difference is the use of a dual n-back procedure replacing the spatial n-back task. Generalised working memory improvements will be assessed using a battery of tasks that provide sufficient procedural differentiation from the training tasks that will also be merged into a number of composite measures of working memory. The training programs provide sufficient differences in their working memory demands to yield interesting comparisons. Working Memory Period is an exclusively verbal domain task that stresses increased resilience of memory representations as processing demands increase in an adaptive manner. The Colour Corsi task provides a visuospatial domain training program that may also recruit additional executive processes due to the required feature binding. And finally, the dual n-back procedure incorporates verbal and visuospatial streams of stimuli and requires the participant to monitor these independently.

5.2 Method

As this experiment was an extension to the second developmental study described in the previous chapter, this method section will focus on the additions in this study and refer to the previous method section where information overlaps.

5.2.1 Participants

Participants consisted of adults recruited from the participant recruitment system within the Psychology department at the University of Lancaster. All participants were students at the University. Participants were paid for their time up to a maximum of 45 pounds each which covered time spent in lab sessions pre-and-post-training as well as the training sessions they completed. Sixty participants were recruited initially. Of these 60, 15 withdrew participation during the training phase (all due to falling behind in the training schedule) therefore a further 15 were recruited to replace these. The final sample had a mean age of 21.27 (sd = 2.28) years. Figure 5.1 shows the participant progress throughout the study.

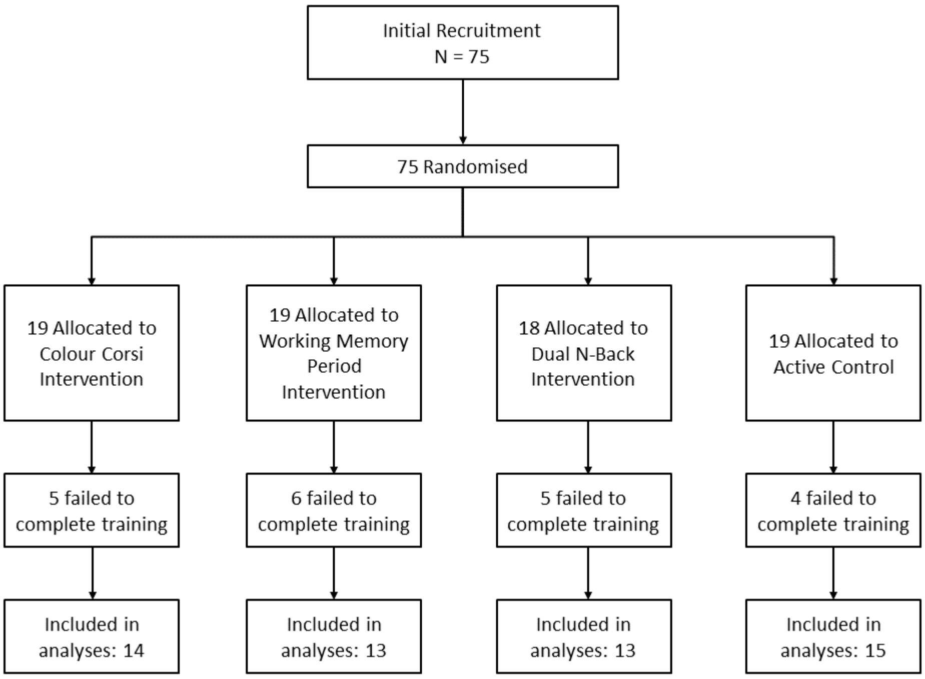

Initial Recruitment
N = 75
75 Randomised
19 Allocated to Colour Corsi Intervention
19 Allocated to Working Memory Period Intervention
18 Allocated to Dual N-Back Intervention
19 Allocated to Active Control
5 failed to complete training
6 failed to complete training
5 failed to complete training
4 failed to complete training
Included in analyses: 14
Included in analyses: 13
Included in analyses: 13
Included in analyses: 15

Figure 5.1. Flow diagram highlighting participant involvement throughout the WMT adult study

5.2.2 Design and materials

The experiment is a randomised controlled trial that consists of three phases; baseline phase, training phase, and post-training phase. The training phase was administered over a five-week period and during this phase participants either completed adaptive WM training (three different groups) or were part of an active control group engaging in standard puzzle tasks. Participants were randomly assigned to one of the four groups prior to commencement of the experiment. To measure any transfer of training a battery of tasks was completed before and after the training phase.

All computerised tasks used in this experiment were programmed using the Java programming language using the framework provided by the Tatool framework (von Bastian, Locher, & Ruflin, 2013). Stone and Towse (2015) provides a more detailed description of the software.

Pre- and Post-Training Assessment

As noted in the same section in the method for the previous study, task selection was informed by previous research, specifically Kane et al. (2004) and Nettelbeck and Burns (2010). The tasks that form the pre-post battery in this study were broadly (slight variation in list length trials for span tasks) the same 8 tasks as used in the previous study with the addition of Rotation Span and Free Recall.

Visuospatial Working Memory

Rotation Span The rotation span task used here is an adapted version of the task used by Shah and Miyake (1996). Figure 5.2 shows a schematic representation of a rotation span trial showing the storage and processing parts of the task. The to-be-remembered (TBR) stimuli in the rotation span task are images of arrows that are differentiated in two characteristics. Any one arrow can differ in its length; long (300 pixels) or short (100 pixels), or it can differ in its angle of rotation (0°, 45°, 90°, 135°, 180°, 225°, 270°, or 315°). Therefore the storage phase of this task is to remember the arrows presented in their correct serial position.

The processing operation in this complex span task presents participants with a letter (F, G, or R) that may be normal or a mirror image. It may also be rotated at one of the 45 degree rotations. The participant must mentally rotate the image back to normal orientation (05°) so that they can make a judgement on whether the letter is a normal or mirror representation using the left/right keys.

After all the storage-processing elements of a trial were completed the recall phase began. The recall screen presented the 16 possible arrows in a 2 by 8 grid where the top row of arrows were all the short arrows and the bottom row were all the long arrows. Participants used the mouse to select the arrows they remembered seeing in the correct order.

Test trials consisted of three trials at each list length between two and six and were

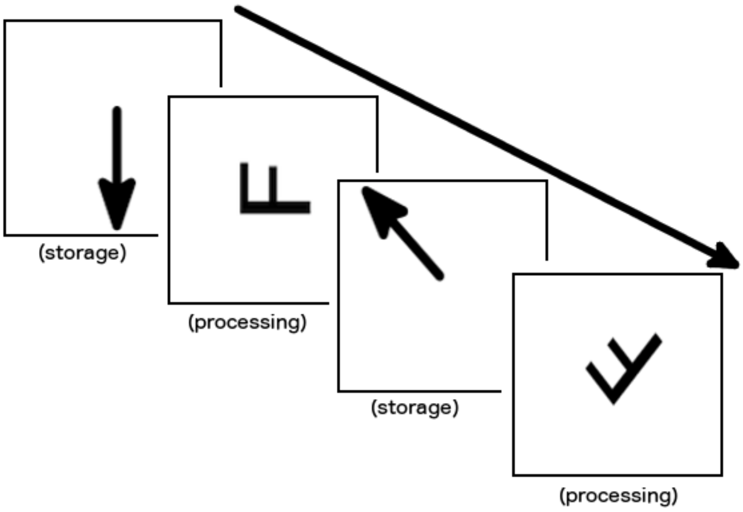


Figure 5.2. Illustration of the Rotation Span task

administered in a randomised order.

Symmetry Span Please refer to Figure 4.4 in the previous chapter for a schematic diagram of the Symmetry Span task. Test trials consisted of three trials at each list length between two and seven and were administered in a randomised order.

Verbal Working Memory

Operation Span Please refer to Figure 4.5 in the previous chapter for a schematic diagram of the Operation Span task. Test trials consisted of three trials at each list length between two and six and were administered in a randomised order.

Reading Span Please refer to Figure 4.6 in the previous chapter for a schematic diagram of the Reading Span task. Test trials consisted of three trials at each list length between two and six and were administered in a randomised order.

Short-Term Memory

Digit Span Test trials consisted of three trials at each list length between two and seven and were administered in a randomised order.

Matrix Span Test trials consisted of three trials at each list length between two and seven and were administered in a randomised order.

Free Recall The free recall presented participants with 15 words to remember. The words were presented on screen for 1000ms with an ISI of 1000ms (*double check the timings*). After all words were presented in a trial they were given 60 seconds to type in as many words as they could remember. As they submitted words the list of recalled words grew on screen showing them what responses had been given up to that point. When scoring this task typos were allowed but variants of given items were not allowed. For example if a given word was 'Dancer' and a participant responded 'Dancing' then this was marked as incorrect.

The pool of words that the task randomly selects from was generated using the English Lexicon Project (http://elexicon.wustl.edu/) tools. Using this tool, 156 words were selected to include in the pool of potential to-be-remembered words. All selected words were of length 5-6 characters and were disyllabic and had a HAL frequency of at least 7,000 (Mean: 5,636).

Processing Speed

Odd One Out Please refer to Figure 4.7 in the previous chapter for an illustration of the Odd One Out task. Participants completed a total of 50 trials on the OOO task.

Inspection Time Please refer to Figure 4.8 in the previous chapter for an illustration of the Inspection Time task. As before, participants complete trials until they experience eight reversals of direction on the adapting SOA.

Further Transfer

Mental Arithmetic As before the test consisted of six blocks of questions where the participant was given one minute per blocks to answer as many questions as possible. The blocks included specific 'types' of operations; addition without carry, addition with carry, subtraction without carry, subtraction with carry, multiplication, and division. The pool of questions used in each study is produced in the appendix *(include reference)*.

Training Phase

The training phase was a five week period that began as soon as the participant had downloaded and setup the training software on their own computer. The software was distributed via an executable JAR (java archive file). As long as a java runtime environment was installed on the machine the .jar file could be executed and the training software began.

Working Memory Training (WMT) Groups

Dual N Back This training group completed the dual n-back task as described by Jaeggi et al. (2008). Figure 5.3 shows a schematic representation of the dual n-back task. Participants were required to monitor two stimuli streams, audio and visuospatial. You can think of one item on the dual n-back task as a presented letter given auditorily AND a grid location which would be presented visually (by the grid filling black for 500ms). For each item the participant could press the '1' key when there was a letter match and/or the '0' key if there was a grid match. A letter match had occurred if the current letter was the letter presented n letters ago, where n is the current level of difficulty. Similarly, a grid match had occurred if the highlighted grid was the same grid as the one presented n grids ago.

Taking the five items in Figure 5.3 as an example. If the current level of difficulty is one then in the visual stream there is only one match (item 4) because n grids ago the same one was presented. Similarly there is only one match in the auditory stream and this occurs at item 2. If the current level of difficulty was three then there would also only be one match per stream and it occurs at item 5 for both streams.

One block of items consisted of $20 + n$ items. Stimuli were randomly generated by the

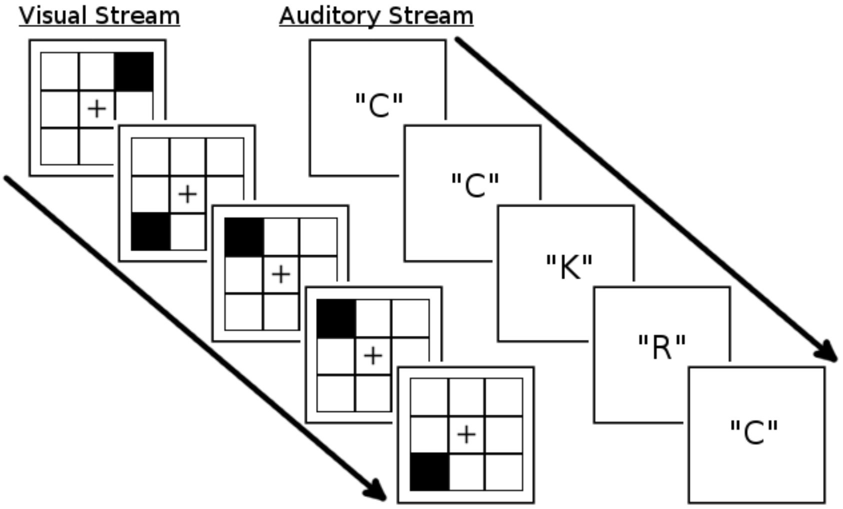


Figure 5.3. Illustration of the Dual N-Back task

program but it was constrained to always have four independent matches per stream (i.e. a letter match when there wasn't a grid match and vice versa) as well as two concurrent matches (grid/letter match occurs at the same item) as described in Jaeggi et al. (2008).

If participants made less than 4 errors in any block then they were upgraded a level. If more than 10 errors were made then the level was downgraded by one. The level remained the same for errors in the range of 4-10.

Colour Corsi This group trained solely on the Colour Corsi task which was identical to the implementation in the previous study/chapter. Refer to Figure 4.2 for a reminder on the demands of this task.

Working Memory Period This group trained solely on the Working Memory Period task which was identical to the implementation in the previous study/chapter. Refer to Figure 4.3 for a reminder on the demands of this task.

Active-Control (AC) Group

The active control group completed the same rotation of tasks as described in the last chapter; Sudoku, Wordsearch, and Jigsaw. Each session consisted of just one of the three tasks.

5.2.3 Procedure

Ethical approval was obtained for the current study, details of which are available in section A.1 of the appendices. The pre- and post- training sessions were conducted in small groups of 3-4 participants in a research computer laboratory on 21.5" iMac computers. Participants used their own computers at home for the training phase. Participants attended an initial lab session where they completed all of the tasks outlined above in the pre/post training section. When each task was completed they could take a short break before launching the next task. After participating in the initial lab session participants were sent an e-mail to an experiment website which detailed all the instructions for downloading and running the training software. Blinding was not possible due to the primary researcher being the only person involved in any aspect of the data collection.

As soon as a participant conducted their first training session the training phase begun. During the next 5 weeks they were asked to complete four training sessions per week for a total of 20 sessions. Session information was uploaded to a secure server accessible only to the primary researcher researcher automatically each time a participant completed a training session. If a participant fell behind schedule they were sent a gentle reminder e-mail which stated how many sessions they had completed to that point and how much time was remaining in their training period.

As the training phase drew to a close the participants were invited back to the lab for the post-training phase. At this point the participants completed the pre/post measures once again. No more than a week passed between the end of a participants training phase and the post-training lab session.

5.3 Results

5.3.1 Correlation/PC Analysis of T1 data

Principal Components - Full Data As a precaution, and as a way of deriving weights for calculating composite scores, these data were subject to principal components analysis (PCA) in order to verify the assumed components would be extracted based on previous work. Namely, that it is appropriate for this data to combine verbal memory tasks, visuospatial tasks, and processing speed tasks into separate composite scores. The same procedure was followed as in the previous study using a developmental sample. As was shown in those analyses it was not appropriate to combine the processing speed measures into a single construct, the results here may illuminate the cause of this with regards to whether it may be related to our particular execution of these tasks or if was a product of the developing population.

Full data for pre-training baseline tasks was available for 73 participants and these data were used in the PCA. Each variable included in the analysis was checked for extreme values and where absolute z-scores above 3 were found that variable was subject to a 90% winsorisation transformation. This was only necessary on the odd one out median decision time (3 scores) and inspection time mean SOA (2 scores). Table 5.1 shows the overall Pearson correlation matrix for the 9 measures that will be subjected to PCA plus the mental arithmetic measure is included here so it can be discussed later. The absolute scoring method (total number of individually recalled items in correct serial position) with an adjustment for transpositions was used for the span tasks. The first observation to note is the lack of correlation between the two processing speed measures (odd one out and inspection time, -0.11). This points to a conclusion of no linear relationship between the two tasks for this dataset which is clearly an even more significant issue than what was observed in the previous study. It is already clear that there is not going to be a factor solution that involves a processing speed factor that these measures load heavily onto as was expected. Therefore, it is likely at this point that the processing speed measures will have to be analysed as individual task indicators of change rather than as a composite as was hoped. Further analysis needs to be undertaken to assess why these variables are showing no linear relationship given the established relationship within the literature (O'Connor & Burns, 2003; Nettelbeck & Burns, 2010).

Table 5.1

Correlation Matrix of each primary dependent variable; OOO = Odd One Out decision time, IT = Inspection Time mean SOA, M-Arith = Total correct operations across all blocks, F-Recall = Total correctly recalled items across all FR trials, Digit = Digit Span t-score, Op = Operation Span t-score, Matrix = Matrix Span t-score, Symm = Symmetry Span t-score, Rotation = Rotation Span t-score, Reading = Reading Span t-score.

	OOO	IT	M-Arith	F-Recall	Digit	Op	Matrix	Symm	Rotation
OOO									
IT	-0.05								
M-Arith	-0.34*	-0.13							
F-Recall	-0.15	-0.27*	0.15						
Digit	-0.28*	-0.32*	0.49*	0.50*					
Op	0.00	-0.35*	0.41*	0.26*	0.63*				
Matrix	-0.04	-0.38*	0.09	0.23	0.27*	0.21			
Symmetry	-0.04	-0.54*	0.13	0.38*	0.33*	0.31*	0.69*		
Rotation	-0.02	-0.36*	0.17	0.29*	0.40*	0.43*	0.50*	0.69*	
Reading	-0.13	-0.30*	0.28*	0.35**	0.55*	0.76*	0.32*	0.40*	0.56*

The results of some diagnostic checks suggest these data are suitable for PCA - Bartlett's test: $\chi^2 = 280.71, p < .0001$ and Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy = .73 which is reasonable (Kaiser, 1974, Hutcheson & Sofroniou, 1999). However, the individual value of KMO for the odd one out measure is only .32 which further adds to the concerns over the suitability of the processing speed measures. Both of these are good indicators that PCA is appropriate for these data but with a warning tag on the odd one out DT variable.

Initial Solution Figure 5.4 shows the scree plot obtained from the PCA while Table 5.2 shows the breakdown of the variance components for each extracted PC. It shows that 3 PCs are required for 71.1% to be accounted for while four PCs gives 80.4%. Both the three and four factor solutions could prove to be sensible depending on the behaviour of the extracted factors, and while the eigenvalue of PC4 is < 1 it is still above $> .7$ which has been suggested to be appropriate (Joliffe, 1972, 1986) but this cut-off might be considered too low. However, the fourth PC must be integral to the interpretation of the extracted PCs for it to be suitable to include a PC with eigenvalue < 1. Rotation may also enhance the contribution of the fourth factor.

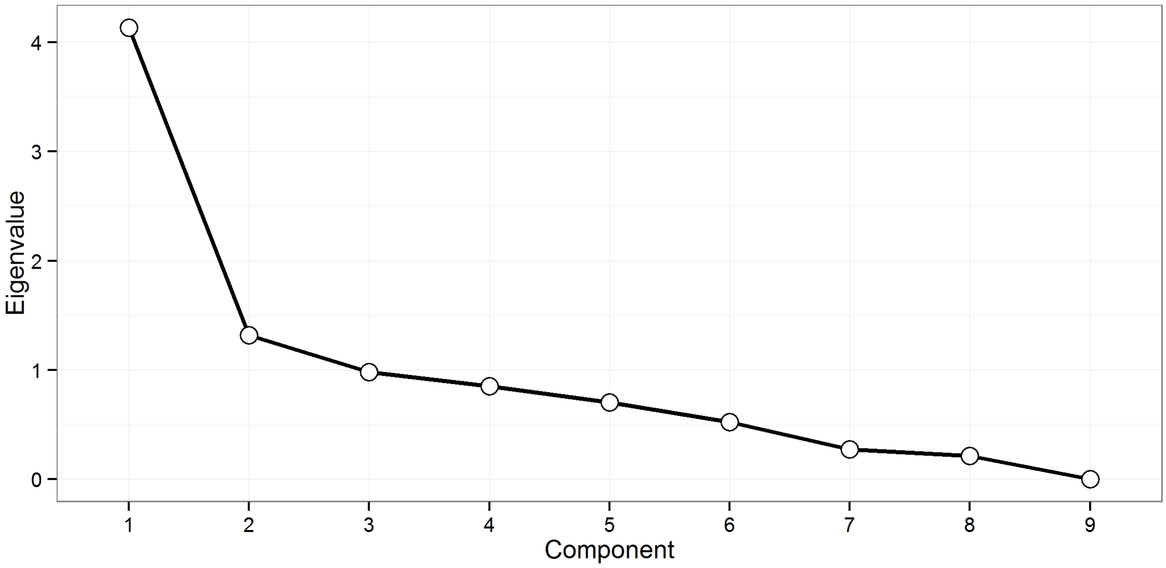


Figure 5.4. Scree plot highlighting the eigenvalues for extracted PCs in the initial solution (Adult - WMT3)

Four factors - promax rotation A PCA was carried out again extracting only 4 factors and using a promax rotation method as it would be unreasonable to expect the extracted factors to not correlate with each other. Table 5.3 shows the loading matrix for the 4 factor solution with promax rotation. It is difficult to have a clean interpretation of this solution. PC2 appears to reflect spatial memory abilities with a contribution from inspection time. PC1 is primarily made up of the verbal memory tasks but with overlap

Table 5.2

Variance explained by the Principal Components from the initial unrotated solution with all variables included

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
SS loadings	3.892	1.409	1.100	0.833	0.615	0.411	0.355	0.227	0.157
Proportion Var	0.432	0.157	0.122	0.093	0.068	0.046	0.039	0.025	0.017
Cumulative Var	0.432	0.589	0.711	0.804	0.872	0.918	0.957	0.983	1.000

from the rotation span task. PC4 and PC3 appear to be there to mop up the remaining tasks, inspection time loads onto these as well as PC2. PC3 effectively only caters to the odd one out task (inspection time loading of 0.32 isn't too high, .3 cut-off has been applied for presentation purposes in this table).

Table 5.3

Loadings matrix for the 4-PC promax solution (loadings below 0.3 suppressed)

	PC2	PC1	PC4	PC3
OOO				0.95
IT	-0.39		-0.45	-0.32
F-Recall			0.94	
Digit		0.58	0.43	
Op		1.01		
Read		0.90		
Matrix	0.96			
Symm	0.91			
Rotation	0.70	0.36		

Three factors - promax rotation Table 5.4 shows the loading matrix for the 3-PC solution.

The 3-PC solution provides a 'simple' solution (not unreasonable to exclude the -.35 value for digit span on PC3). However, the interpretation of the solution is not what would have been hoped and is difficult to properly explain. In this dataset there is clearly a stronger relationship between inspection time performance and the spatial memory measures than with the alternate processing speed measure (odd one out decision time) or with the verbal/free-recall measures. PC3 appears to be a 'mopping up' PC rather than a clearly useful latent measure. At each stage it appears as though free recall and odd one out have struggled to find a place in the solutions. I would tentatively conclude at this stage that rather than trying to force a solution upon the full dataset it may be more appropriate to move forward extracting two factors from the simple/complex span measures and leaving the other tasks (odd one out, inspection time, and free-recall) as individual task measures.

Table 5.4

Loadings matrix for the 3-PC promax solution (loadings below 0.3 suppressed)

	PC1	PC2	PC3
OOO			0.95
IT	-0.61		
F-Recall			-0.43
Digit		0.71	-0.35
Op		1.06	
Read		0.89	
Matrix	0.95		
Symm	0.98		
Rotation	0.64		

Principal Components - simple/complex span data A sensible structure should a formality from what we have seen, but the PCA from scratch using just the six simple/complex span memory tasks was run to confirm it was suitable and to derive weightings for creating composite scores. Re-running the diagnostic checks on the span subset data yielded no cause for concern; Bartlett's test gives $\chi^2 = 212.24, p < .0001$, and the KMO measure of sampling adequacy = .76 which is reasonable (Kaiser, 1974, Hutcheson

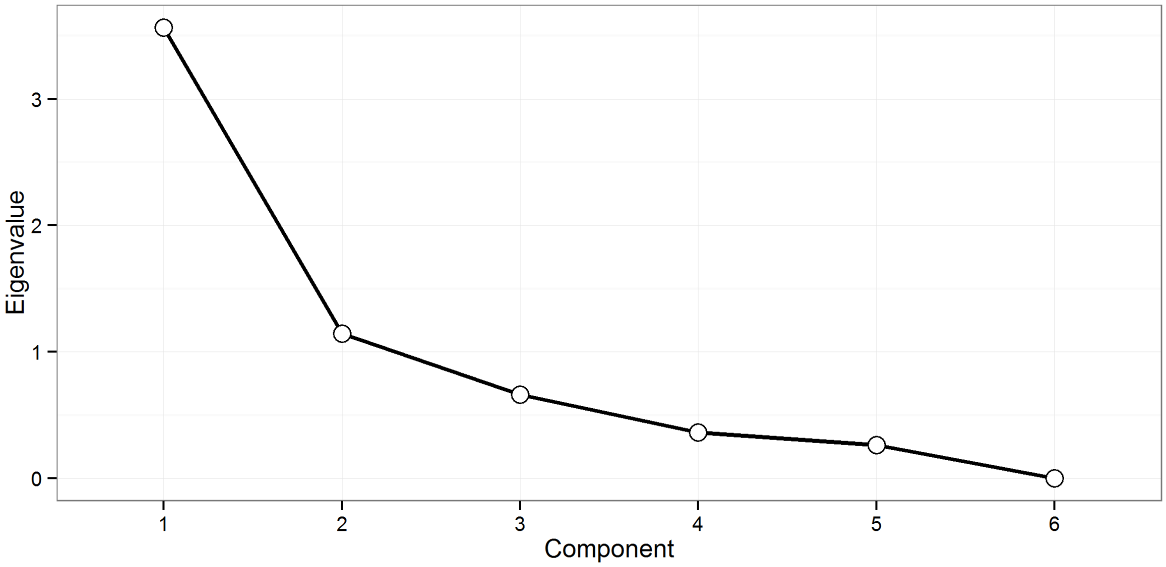


Figure 5.5. Scree plot for initial PCA solution using the span subset data

& Sofroniou, 1999).

Figure 5.5 shows the scree plot for these data and confirms a two-factor solution being the most appropriate. Table 5.5 shows the loading matrix for this PCA. The loadings provided by this analysis will be used to calculate composite scores for verbal and visuospatial memory performance at pre- and post-training phases.

Table 5.5

Loadings matrix for the 2-PC promax solution with only memory span measures included (loadings below 0.3 suppressed)

	PC1	PC2
Digit	0.82	
Op	0.97	
Read	0.84	
Matrix		0.93
Symm		0.93
Rotation		0.69

Table 5.6

Correlation matrix of processing speed measures from complex span tasks and primary dependent variables from processing speed tasks

	OS	Read	Symm	Rotation	IT
OS					
Read	0.53***				
Symm	0.35**	0.49***			
Rotation	0.53***	0.50***	0.60***		
IT	0.06	0.09	0.05	0.04	
OOO	0.40***	0.23*	0.12	0.13	-0.04

Discussion point - Note on processing speed measures

To investigate the confusing results relating to the processing speed measures an additional correlation matrix was produced which included the two specific processing speed measures along with a measure of speed of processing from each of the four complex span tasks administered at this time point in the WMT adult study (Table 5.6). The inspection time measure is somewhat remarkable in that it does not correlate with any of the other measures. Decision time on the Odd One Out task correlates with the response time on verbal complex span processing tasks but not the visuospatial processing tasks which is a curious result.

5.3.2 Practice Effects

As with previous analyses of performance on the training tasks over the repeated sessions all the information collected throughout the study was utilised. Therefore where a participant may have dropped out after completing less than the required 15 training sessions they were included in the practice effect analyses but not transfer effect analyses. Figure 5.6 shows the profile of the sample size as session number progresses. The analyses are presented split by session and by trial blocks as discussed in the previous chapter. There is less attrition in this study when compared with the developmental studies. This is to be expected based on the manner in which recruitment occurs with adult participants. The participants selected themselves by responding to the study information page and are financially compensated for their participation based on how much of the study they complete.

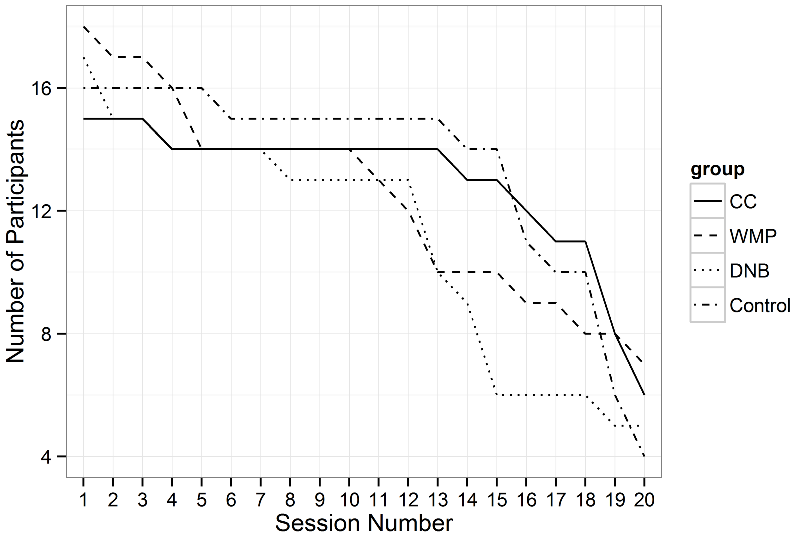


Figure 5.6. Number of participants who reached each stage of the training for each of the groups.

Performance trends over repeated training sessions were analysed using the same mixed model method outlined in previous chapters. As a brief recap we fit a random intercept effect for subject and include the session number as a categorical fixed effect predictor. Then by plotting these parameter estimates with the respective confidence intervals we

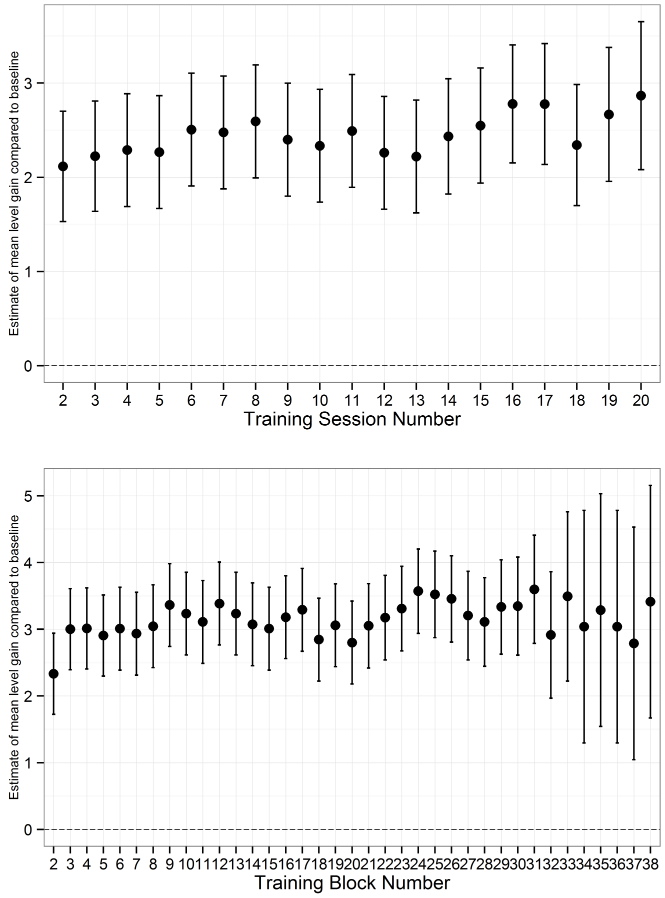


Figure 5.7. CC; Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks

can assess the trend in variables over repeated sessions utilising all the information and removing the bias if we used such a method without the random effect. Figure 5.7 shows the trend in performance over training sessions / trial blocks for the Colour Corsi task showing the derived parameter estimates of the overall effect of session/block number. In terms of the estimates of the overall effect of session/block the practice 'curve' is in the same vein to the developmental training analyses in that there is a clear increase between the first and second units but then no reliable increases beyond that. The average level worked at for the first session was 3.17. As we can see from Figure 5.7 the performance after session one is somewhat invariant around two levels above baseline. This increase between session one and two is more of an artefact of the administration procedure based on participants starting at level one and the level adjusting every five trials based on performance. Therefore the initial mean value around level three more often than not represents success over the first 20-25 trials where the program was working through the lower levels. As future sessions continue from where the previous session left off this means that by session two participants are now starting at their current ability level. Figure C.6 in the appendix shows this pattern per participant and the relative lack of variability in that aside from two strange trajectories there is a level of homogeneity with the performance profiles.

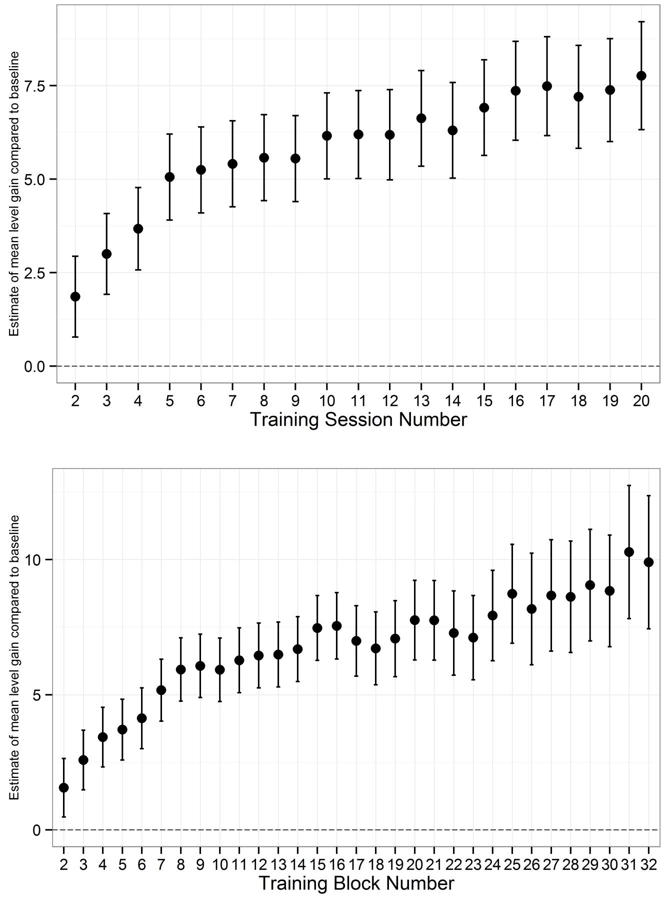


Figure 5.8. WMP; - Top - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number, Bottom - Split into trial blocks

The performance profile for the Working Memory Period task is shown in Figure 5.8 There appears to be two parts to this growth 'curve'; a sharp gradient over the first

5 sessions / 8 trial blocks, followed by a steadier but consistent positive gradient in the remaining sessions/blocks. This differs from the corresponding plot in the second developmental training study where a slight but consistent improvement was seen as the training phase progressed. The average level reached in the first training session was 2.48. Figure 5.9 shows the corresponding figure for the Dual N-Back training performance. Again it can be seen that a pattern emerges that is different from what was seen in the developmental sample albeit a single N-Back task was used previously. The average level of n in the first training session was 1.75. Figures C.5 and C.7 provided in the appendix show the overall performance over the repeated sessions for each participant for the WMP and DNB tasks. These tasks show a more variable pattern than what was observed with the CC task.

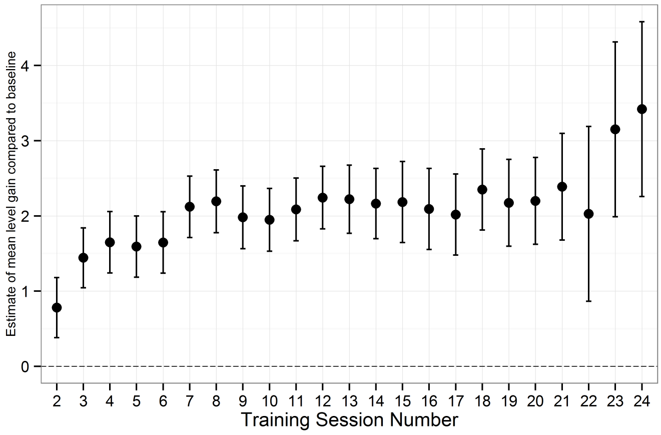


Figure 5.9. Dual N-Back - The parameter estimates (with 95% confidence intervals) obtained from the mixed model showing the extent to which performance varied as a function of session number

5.3.3 Transfer Effects

Verbal/Visuospatial composite scores Figure 5.10 shows the mean pre- and post-training scores for each training group on the verbal and spatial composite factors (using the coefficients extracted from the PCA, see Table 5.5). The top two rows in Table 5.7 show that ANCOVA analyses controlling for pre-training scores suggest no effect of training group for either composite measure. This table shows the value of F and p for the group effect as well as the Hedges' g effect size for each pairwise comparison of the adjusted post-training means for the WM training groups compared to the active control group. As before (see transfer effects section of previous chapter for more details) these effects were also assessed by means of testing the interaction effect of Time x Group in a hierarchical linear model. These results are shown in Table 5.8.

The composite measures represent the most robust measures of generalised change in the WM construct and these results suggest that the interventions have not had an impact on these measures. Performance on the simple/complex span tasks was also assessed individually in addition to the processing speed tasks (Odd One Out and Inspection Time), Free Recall, and Mental Arithmetic.

Pre-Post Analysis of Individual Tasks Mean values for full trial accuracy (FTA) and absolute with transposition adjustments (T-Score; see task validation chapter for more information on these scoring methods) for pre- and post-training performance for each training group on the memory tasks can be seen in Table 5.9 while Table 5.11 shows the basic information for the remaining transfer tasks. Tables 5.7 and 5.8 introduced previously also include the formal analysis of individual tasks as well as the composite measures and show that none of these individual analyses yields a 'significant' result.

Complex Span Processing Measures The processing elements of the complex span tasks are a significant part of the WM measurements and were also analysed for pre-post changes as a function of the interventions. Table 5.10 gives a summary of these measures with regards the accuracy (as a proportion of processing elements successfully completed) and the mean of the median response times for the processing operations. For the accuracy measures ANCOVA analyses suggest no difference at post-training controlled for pre-training values (Hedges' g with 95% CI for each pairwise comparison against the control group for post-training adjusted means in the following order, CC, DNB, WMP);

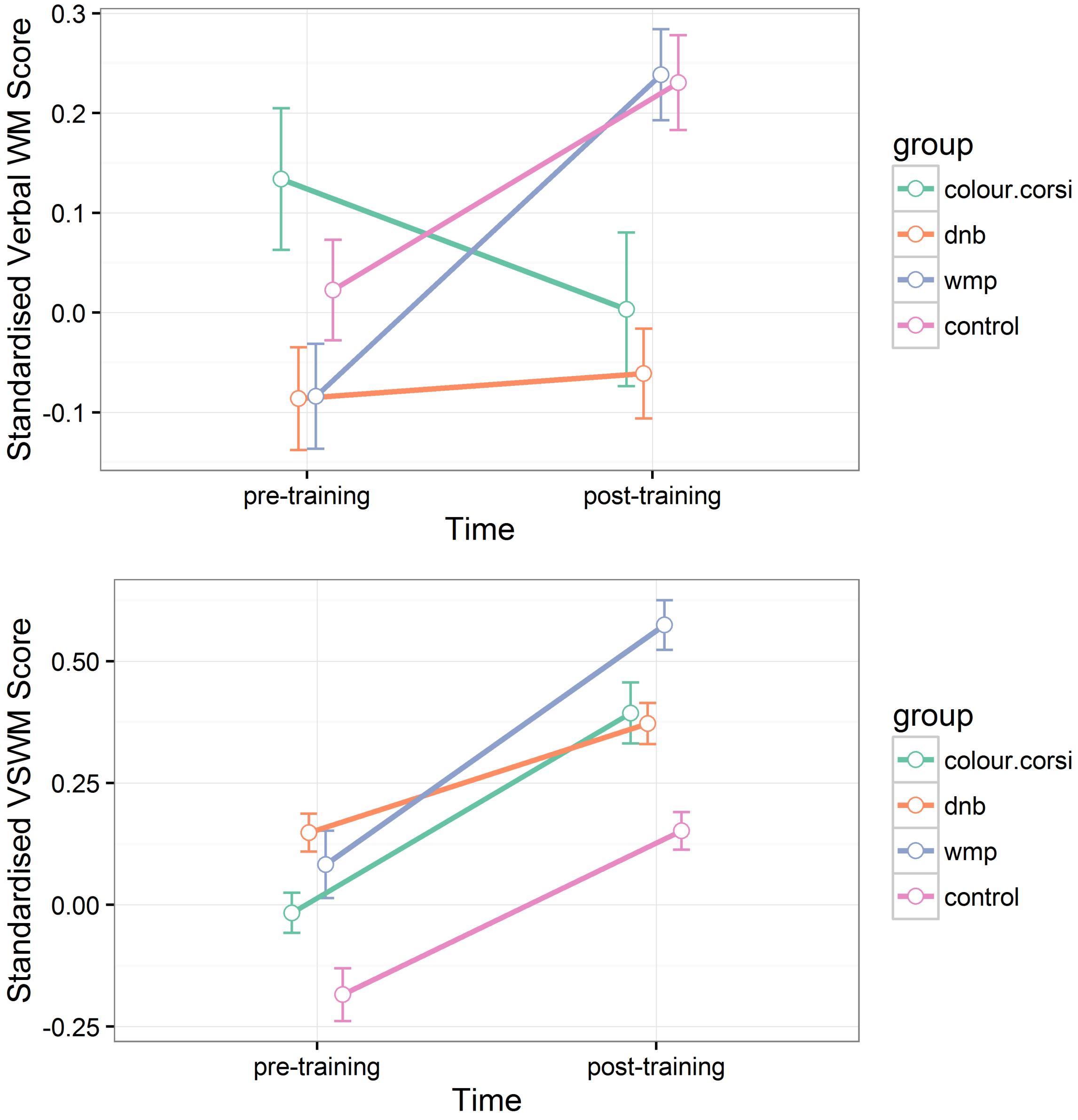


Figure 5.10. Standardised Pre- and post-training scores for the verbal (top) and visuospatial (bottom) composite measures as determined by the coefficients derived from the PC analysis with standard errors

Table 5.7

Summary of ANCOVA results plus the effect size (Hedges' g) for pairwise comparisons comparing the adjusted means for each training group to the active control group. Tasks; CC = Colour Corsi, DNB = Dual N-Back, WMP = Working Memory Period. Scores; FTA = full-trial accuracy, T = transposition score (number of indivdidual items correct plus partial credit for transpositions)

	F	p	CC	DNB	WMP
PC Scores					
Verbal PC Score	2.143	.11	-0.65[-1.41,0.11]	-0.41[-1.19,0.34]	0.19[-0.57,0.95]
Spatial PC Score	1.264	.29	0.27[-0.47,1.02]	-0.05[-0.81,0.7]	0.6[-0.18,1.37]
Digit Span					
FTA Score	0.557	.65	0.09[-0.43,0.62]	-0.19[-0.74,0.35]	0.3[-0.24,0.85]
T-Score	0.536	.66	-0.03[-0.77,0.71]	-0.31[-1.07,0.45]	-0.38[-1.14,0.39]
Operation Span					
FTA Score	0.995	.4	-0.52[-1.27,0.23]	-0.56[-1.33,0.21]	-0.4[-1.17,0.36]
T-Score	2.507	.07	-0.56[-1.32,0.19]	-0.3[-1.06,0.46]	0.43[-0.36,1.22]
Reading Span					
FTA Score	1.174	.33	-0.38[-1.13,0.36]	-0.6[-1.37,0.17]	-0.05[-0.8,0.71]
T-Score	1.922	.14	-0.7[-1.47,0.06]	-0.41[-1.17,0.35]	0.07[-0.69,0.83]
Matrix Span					
FTA Score	1.197	.32	0.51[-0.24,1.31]	-0.04[-0.82,0.74]	0.41[-0.35,1.18]
T-Score	0.315	.81	0.25[-0.49,0.99]	-0.02[-0.77,0.74]	0.24[-0.52,1]
Symmetry Span					
FTA Score	2.361	.08	0.63[-0.13,1.43]	0.41[-0.36,1.17]	0.95[0.15,1.74]
T-Score	1.978	.13	0.29[-0.23,0.81]	0.17[-0.38,0.71]	0.88[0.34,1.42]
Rotation Span					
FTA Score	0.294	.83	-0.2[-0.94,0.54]	-0.21[-0.97,0.55]	0.07[-0.68,0.83]
T-Score	0.188	.9	0.04[-0.7,0.78]	-0.2[-0.96,0.56]	0.05[-0.71,0.81]
Inspection Time					
mean.soa	0.96	.42	0.15[-0.59,0.9]	-0.08[-0.84,0.67]	0.46[-0.31,1.22]
Odd One Out					
median.dt	1.44	.24	-0.12[-0.87,0.62]	0.55[-0.22,1.32]	0.38[-0.39,1.14]
Mental Arithmetic					
total.corr	1.089	.36	0.09[-0.66,0.83]	-0.54[-1.31,0.23]	-0.11[-0.86,0.65]
Free Recall					
total.recalled	2.081	.12	0.06[-0.68,0.8]	-0.22[-0.98,0.53]	-0.79[-1.57,0]

Table 5.8 Summary of the time by group interaction fixed effect for each dependent variable. The likelihood ratio tests the significance of the effect. FTA = full-trial accuracy, T = transposition score (number of indivdidual items correct plus partial credit for transpositions)

	L-Ratio	p-value	T2:CC	T2:WMP	T2:Dual N-Back
PC Scores					
Verbal PC Score	6.79	.08	-8.5	3	-4.47
Spatial PC Score	3.42	.33	2.9	6.67	-3.15
Digit Span					
FTA Score	2.48	.48	0.19	2.71	-2.06
T Score	1.87	.6	-0.5	-2.57	-1.86
Operation Span					
FTA Score	4.18	.24	-4.95	-3.35	-4.73
T Score	8.32	.04*	-4.22	3.38	-1.63
processing.rt	3.62	.31	-171.06	82.56	-321.09
Reading Span					
FTA Score	2.95	.4	-1.57	1.01	-3.14
T Score	5.82	.12	-4.76	2.18	-1.62
processing.rt	2.41	.49	-245.7	84.46	-242.08
Matrix Span					
FTA Score	2.94	.4	3.2	2.42	-4.19
T Score	0.61	.9	-0.1	-0.1	-1.47
Symmetry Span					
FTA Score	6.6	.09	6.75	11.62	3.93
T Score	5.71	.13	2.17	8.01	0.05
processing.rt	1.88	.6	-48.75	17.62	-100.69
Rotation Span					
FTA Score	0.97	.81	-0.8	-1.18	-2.8
T Score	1.48	.69	1.41	-1	-2.64
processing.rt	4.93	.18	-394.3	-183.75	-326.1
Inspection Time					
Mean SOA (ms)	2.38	.5	13.84	49.38	-10.72
Odd One Out					
Decision Time (ms)	1.29	.73	32.16	45.69	37.93
Mental Arithmetic					
Total Correct	3.23	.36	0.58	-1.22	-3.61
Free Recall					
Total Recalled	7.37	.06	0.7	-4.12	-1.66

Table 5.9

Pre- and post-training performance on the memory span transfer tasks for the three training groups and active control. Values represent means with standard deviation in parenthesis.

Task/Measure	CC		WMP		DNB		Control	
	pre	post	pre	post	pre	post	pre	post
Digit Span								
FTA Score	16.79(6.2)	18.57(11.4)	14.23(4.5)	18.54(6.8)	16.77(4.9)	16.31(7.9)	15.2(7.2)	16.8(7.5)
T-Score	45.21(6.9)	48.13(9.6)	45.69(7.2)	46.54(5.7)	44.06(7.5)	45.62(6.8)	43.73(9.6)	47.16(9.7)
Operation Span								
FTA Score	15.86(10.9)	14.64(12.9)	13.77(7.7)	14.15(5.8)	14.46(6.5)	13.46(6.7)	12.53(6.7)	16.27(7.2)
T-Score	36.08(11.8)	33.09(12.4)	33.11(7.3)	37.74(6.4)	33.48(6.1)	33.09(6.8)	34.88(5.9)	36.12(7.7)
Reading Span								
FTA Score	11.07(8.4)	9.57(7.11)	9.92(6.3)	11(5.4)	11.69(5.5)	8.62(4.9)	11.8(6.2)	11.87(4.7)
T-Score	31.62(11.1)	28.03(10.4)	27.98(8)	21.32(6.8)	29.26(7)	28.81(6.8)	31.28(9.1)	32.45(5.5)
Matrix Span								
FTA Score	49.21(12.8)	59.14(16.6)	47.92(15.2)	57.08(16.7)	51.08(13.29)	53.62(10.92)	40.87(15.2)	47.6(13.3)
T-Score	68.87(6.8)	73.28(7.1)	68.75(8)	73.17(10.1)	68.44(6.9)	71.48(5.2)	64.09(8.9)	68.61(6.9)
Symmetry Span								
FTA Score	30.29(15.9)	38.57(24.2)	26.31(16.5)	39.46(12.9)	30.77(13.4)	36.23(13.4)	27.13(19.2)	28.67(13.5)
T-Score	53.65(14.52)	59.13(20.9)	52.3(16.8)	63.6(7.8)	68.44(6.9)	71.48(5.2)	52.07(14.4)	55.38(8.8)
Rotation Span								
FTA Score	7.14(7.1)	11.14(8.81)	12.77(13)	16.38(10.33)	11.69(9.8)	13.69(7.9)	8.67(6.4)	13.47(7.9)
T-Score	23.95(9.9)	29.29(12.1)	30.41(14.2)	33.35(11.3)	29.22(9)	30.51(9.2)	26.73(10.9)	30.67(9.8)

Table 5.10

Summary of processing elements of complex span tasks at pre- and post-training

	Colour Corsi		WMP		Dual N-Back		Control	
	Pre	Post	Pre	Post	Pre	Post	Pre	Post
Operation Span								
Accuracy	.95(.05)	.93(.05)	.93(.04)	.92(.07)	.95(.04)	.94(.03)	.94(.06)	.96(.05)
RT (ms)	2418(924)	1908(686)	2101(747)	1844(873)	2642(1244)	1982(1021)	2403(950)	2065(665)
Reading Span								
Accuracy	.89(.08)	.9(.05)	.91(.05)	.92(.06)	.93(.04)	.94(.03)	.91(.05)	.93(.04)
RT (ms)	2844(1293)	2357(898)	2014(518)	1857(704)	2102(812)	1618(664)	2320(850)	2079(668)
Symmetry Span								
Accuracy	.96(.05)	.97(.03)	.95(.06)	.97(.03)	.96(.03)	.97(.02)	.89(.16)	.94(.14)
RT (ms)	1150(539)	932(342)	935(250)	784(141)	1146(299)	876(173)	1221(520)	1052(318)
Rotation Span								
Accuracy	.83(.18)	.92(.08)	.9(.09)	.87(.13)	.84(.18)	.91(.11)	.85(.17)	.89(.15)
RT (ms)	1888(633)	1505(443)	1612(576)	1440(598)	1924(827)	1610(737)	1726(428)	1738(489)

Table 5.11

Pre- and post-training performance on the non-span transfer tasks for the three training groups and active control. Values represent means with standard deviation in parenthesis. IT = Inspection Time Mean SOA (ms), OOO = Odd One Out Decision Time (ms), F-Recall = Free Recall Total Items Recalled, M-Arith = Mental Arithmetic Total Correct Operations.

Task/Measure	CC		WMP		DNB		Control	
	pre	post	pre	post	pre	post	pre	post
IT	145.9(106)	162.9(105)	144.3(24.8)	182.2(63.8)	143.4(47.8)	125(14.2)	120.7(20.8)	129.6(13.2)
OOO	457.7(132)	381.2(76)	480.9(217)	430.2(104)	439.6(76.8)	410.1(112.8)	427.6(96)	411.6(75.5)
F-Recall	30.14(7.58)	31(5.79)	29.62(6.09)	32.69(6.3)	28.54(5.22)	31.92(6.42)	30.93(7.58)	33.87(7.21)
M-Arith	106.86(24.2)	110.79(19.87)	100.54(21.16)	109.38(19.88)	105.92(20.64)	109.15(22.12)	105.53(15.27)	113.67(15.58)

Operation Span - F(3,50) = 2.112, p = .11, g = -0.78[-1.55,-0.01], -0.57[-1.35,0.2], and -0.79[-1.58,0], Reading Span - F(3,50) = 0.933, p = .432, g = -0.43[-1.18,0.32], 0.11[-0.65,0.87], and -0.33[-1.09,0.43], Symmetry Span - F(3,50) = 0.026, p = .994, g = 0.02[-0.72,0.77], -0.05[-0.81,0.7], and 0.05[-0.71,0.8], Rotation Span - F(3,50) = 1.232, p = .308, g = 0.35[-0.39,1.1], 0.24[-0.52,1], and -0.32[-1.08,0.44].

The corresponding analyses for response times also yields no 'significant' results but some interesting pairwise comparison effect sizes; Operation Span - F(3,50) = 0.671, p = .574, g = -0.32[-1.06,0.43], -0.45[-1.22,0.32], and -0.04[-0.8,0.71], Reading Span - F(3,50) = 1.103, p = .357, g = 0[-0.74,0.75], -0.59[-1.36,0.18], and -0.1[-0.86,0.65], Symmetry Span - F(3,50) = 2.507, p = .0696, g = -0.56[-1.32,0,2], -0.92[-1.72,-0.12], and -0.8[-1.59,-0.01], Rotation Span - F(3,50) = 1.369, p = .263, g = -0.69[-1.46,0.07], -0.51[-1.28,0.26], and -0.5[-1.27,0.27].

Training Gain as Explanatory Variable Figure 5.11 shows two rows of figures where the mean training change (aggregate of each session performance divided by baseline performance) is on the x-axis and the top row displays verbal PC on the y-axis while the bottom row displays the visuospatial PC performance. These plots show a scatter plot of the raw data for each training group between these two variables. In addition to the raw data is the fitted regression lines and text information summarising key output from regressing training change onto the difference scores.

The figures show that for the colour corsi group there is some evidence to suggest that the level of improvement on the colour corsi training is a predictor for the pre-post improvement on the verbal/visuospatial PCs with adjusted R^2 values of .16 and .48 respectively. The DNB training group show some relationship between these variables for the visuospatial PC (Adjusted R^2 of 0.18) but not for the verbal PC. WMP shows no evidence for the rate of improvement predicting pre-post differences.

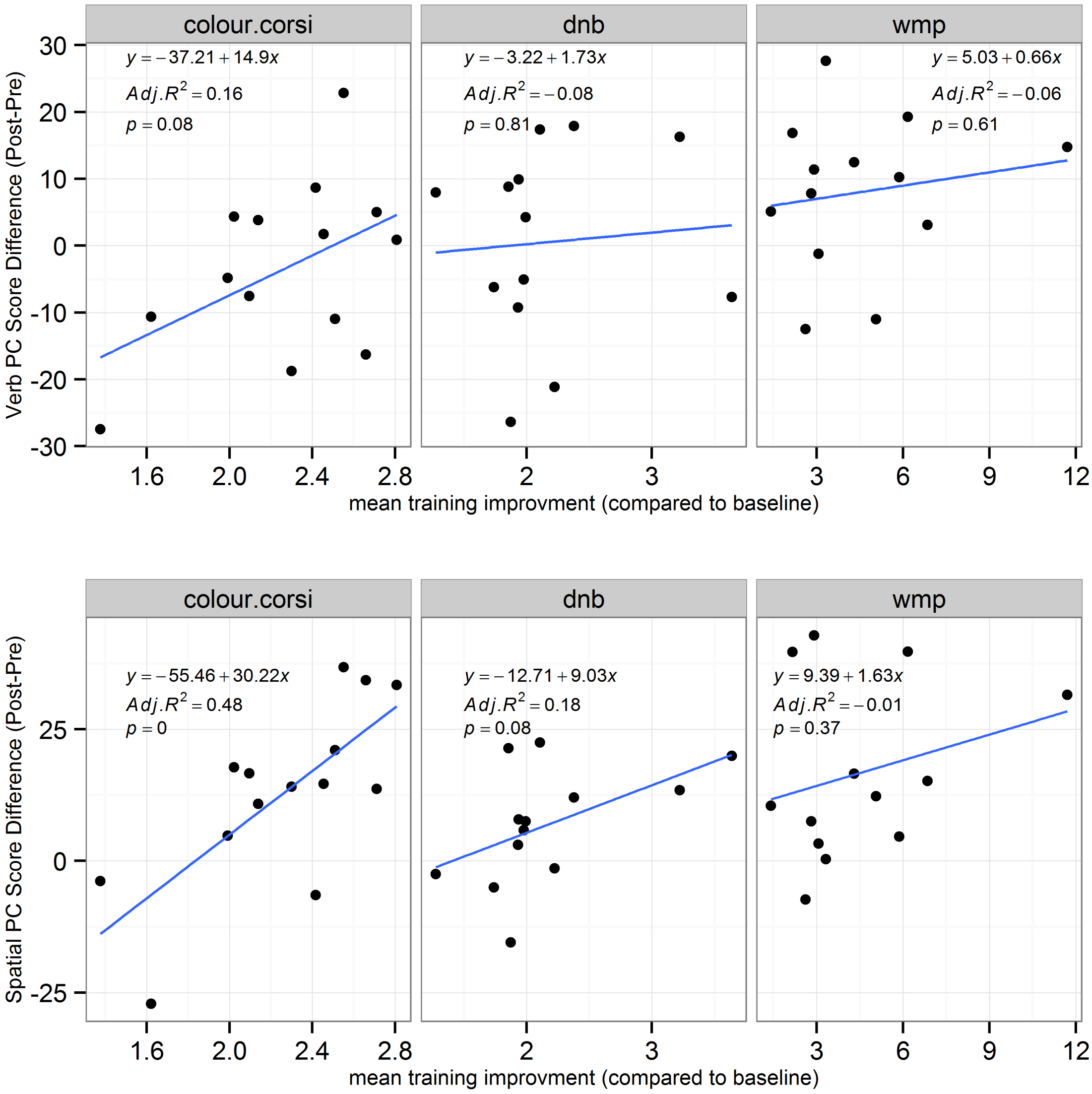


Figure 5.11. Scatterplots showing the observed mean training improvement against difference scores on the PCs (top - verbal; bottom - spatial)

5.4 Discussion

The general pattern of results observed in this study suggest almost no evidence to support any generalisable benefits of the assessed Working Memory based adaptive difficulty interventions on the composite measures of WM (verbal/visuospatial), or when assessing the weaker measures of individual tasks measuring simple span, complex span, processing speed, free recall, or mental arithmetic speed. With regards general transfer effects, the results obtained match that previously observed in the prior studies presented in this thesis.

Assessment of Dual N-Back intervention A particular goal of this study was to replicate the training regime used by (Jaeggi et al., 2008) which produced impressive far transfer effects to measures of fluid intelligence and formed the starting point for a thread of training research which has gathered the most support - Working Memory Updating paradigms (Jaeggi, Studer-Luethi, et al., 2010; Jaeggi et al., 2011; Salminen et al., 2012; Redick et al., 2013; Colom et al., 2013; Jaeggi et al., 2014). Jaeggi et al. (2008) methodology consisted of 69 participants who can be divided into training (n=34) and control (n=35). These groups were then sub-divided into a 'dose' group. Jaeggi and colleagues do not explicitly state the number of participants within each subgroup in their method section but they do state that for each overall dose condition the Ns were 16, 22, 16, and 15 for the ascending levels of dose (days; 8, 12, 17, and 19). Assuming an equal split between control and training in these conditions that gives training Ns of 8, 11, 8, and 7. Figure 2 in their paper shows the mean n-back level achieved through each training session per group. This can be compared with Figure 5.9 which shows a plot of the parameter estimates obtained from the mixed model with mean n-back level per session as dependent variable and session number as predictor. The practice effect seen between the datasets shares similarity in a generally upwards trend. The linearity of this effect seems to be apparent in the Jaeggi et al. data from the first through to the last session with mean levels of n ranging from three to over five. The trend observed in the data provided here suggests a linear trend towards an overall two level gain over 7-8 sessions but then a plateau is seemingly reached. This represents a clear difference in the datasets in terms of trend and maximum levels of n given that the maximum two level increase observed in this study represents an improvement to level 4 (mean initial level was almost 2). Furthermore, it is likely at least some of the initial improvements in the early sessions can be attributed to an increased familiarity with the demands of the task. This applies to all repeated tasks but particularly a task such as the Dual N-Back given the relative complexity with adjusting to monitoring two streams of information. Therefore the estimated degree of improvement that can be considered an improvement above an initial 'cap' is an unknown amount, but must be less than the observed improvement. The analyses carried out by Jaeggi et al. are reported to support the notion that the dual n-back training has a significant impact on fluid intelligence (Gf), but the authors show that the training in their sample did not significantly improve performance on a complex span task (however, an effect was seen for a simple span task - Digit Span). This leads them to the suggestion that DNB training improve fluid intelligence but with little evidence that this is mediated by an improvement in Working Memory. In this study we have added to the literature by focussing on the effects of such training methods on the WM construct in order to provide a) evidence that WM can actually be improved via adaptive difficulty training, and b) begin to understand what aspects of WM may be trainable to create the overall improvements in the measures we assess. While the group sample sizes are low in this study there is strength in the range of tasks used to assess near-transfer and in the analysis of composite scores made up of multiple tasks. Given that no effects of training occurred from any of the training groups assessed here the issue of explaining far-transfer results showing improved fluid intelligence as a consequence of improved WM after training is now more compromised than already discussed.

A further high impact finding from Jaeggi et al. (2008) was that the dose of training mediated the amount of improvement made on the fluid intelligence measures. This result matched well with their findings that participants continued to improve on the DNB task at every session up to the maximum number administered (19). However, given a training profile observed in the data provided here it may not be reasonable to suggest an additional benefit of training dose beyond 8 sessions which represents a very small amount of time spent engaging with the activity. If participants do not continue to improve after a certain number of sessions but did in fact continue to see positive benefits on the transfer to Gf then the interpretation of what is occurring cognitively strays into the same difficult to explain territory as discussed with regards far-transfer in the absence of near-transfer. This relates to an alternative issue in the literature which focuses on what does the n-back test measure and therefore what exactly is it training when deployed as an adaptive difficulty intervention. While we do not include a measure of Gf (The measures of Gf that Jaeggi et al. (2008) used have been criticised (Moody, 2009) due to the way the significant shortening of the item pool) in the vein of a Raven's or BOMAT type test, given no near-transfer to STM or WM tasks, no transfer to processing speed measures, and no transfer to mental arithmetic, it is difficult to be enthusiastic with regards the possibility of DNB training producing a genuine cognitive improvement.

Colour Corsi and Working Memory Period Interventions Despite no clear effects of these alternate interventions on WM there are some interesting patterns to note with regards the practice effects in these tasks. WMP performance shows clear and consistent growth in this sample (refer to Figure 5.8) while CC performance was generally static (Figure 5.7). As already mentioned an improvement between the first two training sessions is much more likely to represent an increased familiarity with the demands of the task as well as the adaptive difficulty algorithm working through lower levels that are below the participant's ability cap. Therefore the data presented here suggests that performance on the CC task does not improve at all over 15-20 sessions. This matches the findings for this task from both developmental datasets explored previously. However, the profile of WMP performance over this number of training sessions is very different from that seen in the developmental datasets.

Individual Differences in Training Gain and Transfer Jaeggi et al. (2011) found that in their initial analyses no significant transfer was observed from N-Back training to a composite measure of fluid intelligence. Additionally, the same group have published research which highlights potentially important individual difference factors that affect whether transfer occurs such as intrinsic motivation and belief about the malleability of intelligence (Jaeggi et al., 2014). As there was a varied level of observed training gain in these data the opportunity was taken to assess the relationship between training gain and transfer. Given the low sample size and the loss of power attributed to using median split techniques to group participants a regression approach was taken to assess these effects. The CC group showed some evidence for a relationship on both PCs but of a higher magnitude for visuospatial WM. This is interesting due to the overall lack of practice effects at the group level. The DNB gain was very modestly related to visuospatial WM but not verbal WM. WMP gains shared no variance with either component of WM. This is of particular interest due to WMP being the task that provides the most potential for improvement over baseline. The mean value of training change for the WMP group was 4.48 compared to 2.26 and 2.16 for the CC and DNB groups respectively. A possible interpretation for these results pertains to what practice effects represent for these tasks. As the WMP task is based on phonological to-be-remembered memoranda the use of rehearsal strategies are likely to be more readily identified (Naus, Ornstein, & Aivano, 1977; Cowan, Cartwright, Winterowd, & Sherk, 1987). Thus gains on the alternative tasks while of a smaller magnitude may reflect factors other than just strategy development. A problem with this account is that useful strategies to boost performance on the WMP task would likely be applicable to the verbal WM tasks that measured transfer. Therefore, it would be expected that even strategy related gains on the WMP task would show some transfer but only to verbal WM. This issue is further discussed as it relates to practice effects across all three WMT experiments in the general discussion.

Chapter 6

General Discussion

6.1 Brief Overview

Can working memory be improved through repeated practice, and if so, to what extent? This is the primary question this thesis has tackled. The recent emergence of the working memory training (WMT) literature suggesting that relatively short interventions produce generalisable cognitive benefits with the most striking findings suggesting improvements in fluid intelligence (Klingberg et al., 2002; Jaeggi et al., 2008; Zhao et al., 2011). In this thesis a number of experiments were conducted to attempt to illuminate the effects of WMT on the trained construct itself which may then be used to explain how such far-transfer to Gf occurs.

In experiment one a typically developing sample of children aged 9-11 years completed a WMT intervention that was a battery of WM tasks, an approach used in commercial products that have been tested in the academic literature such as CogMed (e.g. Klingberg et al., 2005; Holmes et al., 2009, 2010; Prins, Dovis, Ponsioen, ten Brink, & van der Oord, 2011; Kronenberger, Pisoni, Henning, Colson, & Hazzard, 2011; Brehmer et al., 2012) as well as other combinations of tasks that are non-commercial (Ball et al., 2002; Dahlin, Nyberg, et al., 2008; Thorell et al., 2009; von Bastian, Langer, et al., 2013). Another group formed the Active Control (AC) group and completed a selection of tasks that were not dependent on STM/WM and were designed to provide an appropriate alternative to the training that their peers in the WMT group were completing. The training intervention did not produce significant increases in single task measures of verbal working memory, visuospatial working memory, working memory and long term memory interaction, or mental arithmetic.

Experiment two also consisted of 9-11 year old typically developing children and investigated multiple WMT interventions where each one was comprised of a single task. The N-Back training task was selected based on the growing literature suggesting the updating paradigm may be the most effective training intervention (e.g. Jaeggi et al., 2011; Zhao et al., 2011). The alternate training groups completed tasks taken from the battery used in experiment one to provide a diverse set of training groups with regards the type of WM processes being trained. The Colour Corsi task while also in the visuospatial domain involves the binding of separate features while the Working Memory Period task involves storing verbal information for recall while engaging in concurrent numerical processing. The WMP task differs from the N-Back and Colour Corsi tasks in that difficulty is increased by increasing processing demands exclusively rather than adding to the storage requirements. In addition to the three WMT interventions again an AC group was included who completed puzzle games within the same software framework as the WMT interventions. The results provided no support for any of the interventions leading to an improvement in composite measures of verbal WM or visuospatial WM, or for individual tasks measuring processing speed and mental arithmetic.

The final WMT experiment presented in this thesis was a direct extension of experiment two with a healthy adult sample. In this extension the N-Back task was used was the Dual N-Back variant. Again, the results did not support any notion that the interventions improved aspects of WM, processing speed, free recall, or mental arithmetic.

The lack of near-transfer across these studies was unexpected. As discussed in the literature review (Chapter 1) the pattern of near-transfer results is somewhat mixed based in part on often being overlooked in favour of seeking far-transfer. The near-transfer results are not consistent but they occur often enough in the literature (e.g. Klingberg et al., 2002, 2005; Borella et al., 2010; Brehmer et al., 2012) that it was expected that with a variety of interventions and a variety of untrained WM tasks as near-transfer measures that some significant results would be observed. These effects could then be analysed in more detail to profile what specific aspects of WM function is malleable and leads to the improvement observed in simple/complex span performance. As no near-transfer was observed it was impossible to build on significant findings to attempt to demonstrate robustness.

The improvement in one 'unit' in span length on a Memory Span or Dual N-Back paradigm would seemingly be large and therefore the empirical studies were designed so as to permit multiple methods for scoring the working memory tasks. Examples of more continuous scoring methods discussed in previous chapters include proportion correct, absolute score (total number of individually correct items in correct serial position), and the t-score introduced here building on the absolute score by attributing some additional points when transposition errors occur (adjusted for chance). Improvements in these measures can be obtained by being more cognitively efficient. It is probable that sometimes a participant responds incorrectly to trials that are well within their cognitive capabilities. Such errors can occur because of a variety of factors such as a dip in concentration, external distraction (environmental), internal distraction (e.g. mind wandering), or a number of other potential reasons. The observation that none of the interventions led to improved scores on these measures can be used to suggest that neither the capacity nor the efficiency of the cognitive engine being trained was successfully trained. The analysis of the properties of these scoring methods in chapter three show that generally these continuous measures have more favourable psychometric properties which supports previous results (Friedman & Miyake, 2005; St Clair-Thompson & Sykes, 2010). However, while these continuous scoring measures can be more sensitive, reliable, and thus pick up on finer changes in performance but unless these were of a certain magnitude it is difficult to interpret with regards to what that improvement represents if it is not coupled with an increase in maximum span. As no improvement was seen in these studies it becomes a less critical issue to assign meaning to the (lack of) change irrespective of method. Nonetheless it remains a potentially relevant issue for future work to consider what any differences may actually represent.

The remainder of this chapter will address these findings in more detail, consider them in the context of the literature, assess the strengths and limitations of this work, and consider what future WMT studies should address.

6.2 Does Practice Lead To Enhanced Working Memory?

Improved WM ability/capacity is often claimed in WMT studies based on significant practice effects which is often assessed by comparing early vs. late performance on the trained task. However, these practice effects may reflect processes that are not reflective of a generalised improvement of cognition but instead may be brought about by the usage of strategies that lead to better performance by improved utilisation of cognitive faculties. These strategies may or may not be useful in tasks of near-transfer depending on the degree of overlap between the demands and mechanics of the training and transfer tasks. The experiments in this thesis used a range of simple/complex span tasks in order to assess potential effects of practice on the various interventions used. The practice effects in the experiments presented in this thesis show mixed patterns based on both sample characteristics (children/adults) and the specific task.

A potential criticism of the validity of conclusions in this thesis - that these interventions did not improve WM based on the lack of near-transfer to untrained WM tasks would be that the near-transfer would not be expected due to the lack of practice effects observed. Therefore it is important to understand why a number of the observed training groups showed very little practice effects and how this relates to the wider literature with regards to the generally observed magnitude of observed practice effects.

In experiment one a battery of tasks were used in the training intervention. The WMP task improvement peaked at a 0.3 level increase (lower confidence limit 0.184) by session six, CC level improvement peaked at 0.35 (lower confidence limit 0.06) at session five, Stroop at 0.39 (lower confidence limit 0.12) by session three, and no improvement on the variant of WMP or Memory Updating.

The absence of substantial practice effects with the CC paradigm was replicated with a typically developing sample in WMT experiment 2 (improvement peaked at session two approximately 0.54 levels above baseline) but also with healthy adults in WMT experiment 3 where the performance increase from session one to two was much larger (approximately 2 levels) but also represented the only significant improvement as performance never significantly improved over this mark. In WMT experiment 2 the children were able to improve on the WMP task by a larger margin than their counterparts in WMT experiment 1 as the peak improvement was by approximately one level by session nine. Since the children in this experiment spent more time specifically on the WMP task by virtue of it being their only training task, it is unsurprising that they improved to a greater degree than the previous sample. However, it is not the case that a linear trend emerged in WMT experiment 1 that is simply extended in WMT experiment 2 due to more and extended sessions. The pattern of WMP improvement is qualitatively different to the adult trainees used in WMT experiment 3 where improvement was seen session on session until the end of the training regime (20 sessions) and culminated in an approximate 7.5 level improvement by session 20. A spatial N-Back task was deployed in expriment 2 and the results suggested no substantial practice effects for this paradigm. However, a Dual N-Back training task given to the adults in experiment 3 showed significant practice effects peaking at an improvement of just over two levels by session eight.

To decipher whether where these differences relate to task, sample characteristics, or both, it helps to examine the literature. Thorell et al. (2009) gave a WMT regime (WM training group $n = 17$) to preschool children (aged 4-5) where the training paradigm was a visuospatial simple span task (from CogMed). In the data provided by Thorell and colleagues there is some improvement between sessions 4 and 5 which seems analogous to the improvement jump observed in the Colour Corsi groups between first and second sessions in the experiments presented here. The practice effects reported in Thorell et al. (2009) correspond to the WM training group improving significantly on the training task but this result is derived by a repeated measures t-test on the average performance of sessions two - four versus the average of the final three sessions. No effect size is given but from the provided information (Figure 1 in their paper) it is clear this would be of a very small magnitude and their finding is likely influenced by *not* including session five in the early measurement. This is a problem that arises when researchers are free to select arbitrary sessions to compare as opposed to assessing the overall pattern. While the sample used by Thorell and colleagues were typically developing they were a much younger sample than those tested in the studies presented here. It also seems important to note that there was an alternative training group who received an Inhibitory Control training paradigm consisting of the Go/No-Go and Flanker tasks. Both of these showed more substantial practice effects (Go/No-Go profile showed an initial improvement then stabilised while the Flanker performance showed improvement over the first half of the sessions before stabilising). Thus the age and environment aspects are not satisfactory explanations for a lack of practice effects for the WMT group in Thorell and colleagues study which has implications for interpreting the data provided in this thesis. Other WMT studies with a typically developing sample include Bergman Nutley et al. (2011) who administered a WM training condition using CogMed but failed to report any results regarding performance on the trained tasks. Zhao et al. (2011) trained similar age (9-11 years) children using variants of the Running Memory task and scores improved in a linear fashion for 6-7 sessions before slowing for each variant. Loosli, Buschkuehl, Perrig, and Jaeggi (2012) also trained 9-11 year olds but using a visual complex span task and observed very little improvement on the trained task (although again picking certain sessions to compare via a t-test enabled the authors to conclude some significant improvement). And

Jaeggi et al. (2011) used a spatial N-Back training paradigm and found children (mean age 9 years) improved from a baseline level of 2 to 3 across 19 training sessions. The growth seems nonlinear in that there is a swift increase to 2.5 then stable and a resurgence in improvement in the latter training sessions up to n of 3.

Thus the pattern of practice effects in the literature is not entirely consistent. It is clear that it is no guarantee that practising on a WM task will bring significant gains on that specific task as shown by the CC data across the experiments in this thesis, by other published WMT assessing typically developing children (e.g. Thorell et al., 2009; Loosli et al., 2012), WMT studies assessing atypically developing children (Gibson et al., 2011), and WMT studies assessing older adults (e.g. Buschkuehl et al., 2008). Often practice effects on the trained tasks are ignored in the published literature (e.g. Klingberg et al., 2005; Green et al., 2012; Dahlin, 2011; Söderqvist, Nutley, Ottersen, Grill, & Klingberg, 2012). Often this is the case when the study has been setup to assess the effect of practice on WM by means of a criterion task that has significant overlap with the training task e.g. visuospatial simple span training and Span Board administered pre/post.

It is important to ask several questions relating to practice effects. What factors affect whether practice effects emerge? Are practice effects only a reflection of the development and successful utilisation of better strategies to cope with task demands? And also, as the focus of this thesis relates to the WMT literature claims regarding far-transfer to wider cognition in the absence of robust evidence for nearer-transfer to the WM construct itself, it is important to understand if there can be any claims of WM improvement when near-transfer (to untrained WM tasks) is found without corresponding practice effects?

Given that WMT studies generally use an adaptive difficulty paradigm in order to ensure the trainee is working at, or near to, their maximum ability throughout each training session all trainees are going to reach a point at which the task demands become too much for continued success. At this point what does the participant do? The mechanism by which improvement is supposed to occur (e.g. Klingberg et al., 2002; Olesen et al., 2004; Jaeggi et al., 2008) was termed a naive physical-energetic model by Melby-Lervåg and Hulme (2013). Under this proposition by continuing to work on the trials that are currently too difficult the trainees are strengthening the neural system responsible and this leads to improvement in the same way an athlete sees cardiovascular improvement through repetition (Jaeggi et al., 2011). At the same time however, trainees may also begin to think about the way in which they are approaching the task after a particular threshold is reached with regards to being unsuccessful. When a person begins the task they likely approach it in a somewhat straight forward manner as they familiarise themselves with the demands of the task and are not being given much of a challenge due to the low initial demands. As the demands increase and successful responses decrease the participant may begin to approach the task differently in an attempt to overcome the plateau they have reached (Salamé & Baddeley, 1986). Therefore any consideration of practice effects both in terms of behavioural and physiological measurement (e.g. Olesen et al., 2004; Dahlin, Neely, et al., 2008; Langer et al., 2013) needs to be made with both explanations in mind. If WMT does not lead to generalisable improvements in WM by way of increasing the 'capacity' then practice effects and near-transfer effects to untrained WM tasks that have been observed both in this thesis (practice effects) and the wider literature (both) would need to be accounted for. An account based on being able to identify strategies for certain tasks and implement them when a plateau is reached would be a very viable candidate (Dunning & Holmes, 2014; Peng & Fuchs, 2015).

The pattern of practice effects observed across the studies presented in this thesis may be explained by a strategy-based account. With regards the Working Memory Period patterns observed there was a very clear difference between the practice effects in both studies assessing 9-11 year old children and adults. The pattern for adults showed large improvements with regards the average level performance over the training sessions which is in contrast to the relatively flat performance curves seen for children. It would seem unlikely that this difference is explained by differences in the effect of the repeated sessions on the neural processes involved as one might expect the pattern to be reversed and the younger brains more amenable to these effects (Bates et al., 2001; Bryck & Fisher, 2012). Thus strategy identification, experimentation, and successful utilisation is likely the set of processes that separate the groups. Recent results suggest that some individual differences may have an influential impact on the success of WM training such as motivation to complete the training and belief in 'brain training' (Jaeggi et al., 2014). It may be that these factors differ significantly between children and adults and therefore could moderate the effects observed in this thesis. Future research studies should include measures of such individual differences when studying different populations to assess this possibility. Importantly, it may be that a factor such as motivation to improve drives strategy development (see Dunning & Holmes, 2014), therefore there is still a need to design studies that minimise the possibility of strategy overlap to argue for individual differences affecting the success of WM training.

The WMP paradigm is one which is likely heavily influenced by the ability of individuals to construct and implement strategies with regards to the processing phase as well as the storage/recall of the to-be-remembered memoranda. The processing phase of the WMP task involves solving relatively simple but increasingly lengthy arithmetic operations. There are a number of ways in which these operations can be solved (Geary, 1990). A trainee may solve the operation by retrieving the answer from long term memory if it is recognised. The degree to which this will occur is mediated by the amount of previous exposure the trainee has had with that operation or very similar operations. When the answer is not immediately available from long term memory a set of processes needs to be carried out to arrive at the solution and through this method various strategies are available and vary between individuals and between trials (Siegler, 1994). As well as general arithmetic strategies there is potential for task specific strategies when carrying out the WMP based on the lengthening operations as level increases. Take for example this level 3 operation, *3 + 2 + 3 - 2*. This operation is considered a level above the operation *3 + 2 + 3* because of the fourth digit and the third individual +/- element. However, if a participant is able to quickly recognise answers to smaller parts of the operation and retrieve those answers from long term memory then the solution can be derived more quickly. For example, hypothetical trainee A may be able to instantly recognise that *3 - 2 = 1* and therefore the actual operation they process is *3 + 2 + 1*. Trainee B may recognise that they can cancel the *+2* and *-2* as they offset and therefore they are left with the operation *3 + 3*.

The CC task led to consistent phenomena or performance on each of the training experiments. That consistency was the result of an absence of evidence for practice effects in this task for either group. This suggests an inability in either group to improve as a function of neural plasticity or utilisation of helpful strategies. Without using some phonological code to represent spatial memoranda the most likely candidate for rehearsal strategies is eye movements (Baddeley, 1986; Pearson & Sahraie, 2003). If trainees were to attempt to encode the spatial memoranda using a phonological representation so as to use verbal rehearsal this would present a conflict with the likely phonological encoding of the to-be-remembered colour sequence also. The lack of practice effects may indicate that none of these strategies may be useful in this task. Whether the location-colour features are stored independently or as bound units rehearsal strategies would still depend on simultaneously rehearsing either two phonological sets of features, two visuospatial sets of features, or one set of each, as it is unlikely any rehearsal could be effective acting on bound items.

The Spatial N-Back task given to the developmental sample is undoubtedly an easier task than the Dual N-Back given to the adults and yet the adults were able to show performance gains for at least a subset of the training sessions. The children were generally not able to master the 2-back level of the SNB task while the adult training group were able to improve to the 4-back level of the DNB. Practice effects on N-Back paradigms appear to be fairly robust (Li et al., 2008; Jaeggi et al., 2008; Redick et al., 2013; Thompson et al., 2013; Colom et al., 2013). The SNB task was identical in mechanics to that used by Jaeggi et al. (2011) who tested a similar age group (though a little wider range). They obtained larger practice effects generally as the performance curve improved from an average level of 2 to very near the level 3 mark after 15 sessions. Zhao et al. (2011) also saw significant practice effects in the training performance of aged 9-11 typically developing children on an updating-WM paradigm, albeit using the Running Memory task and not an N-Back variant. Therefore the lack of practice effects observed for the SNB task may be specific to the tested sample rather than a property of the training task for this population. This analysis of practice effects somewhat supports the developing narrative in the literature that WM updating paradigms potentially show the most promise as an adaptive difficulty training task. However, as discussed previously, practice effects are only the first chain in a sequence of effects that would need to be shown to be robust to provide evidence for generalised cognitive improvement as a result of a training intervention. In WMT studies two and three the results indicate that even if the N-Back tasks pass the practice effect test they stutter when assessed on near-transfer measures.

6.3 Practical and Theoretical Implications

Throughout three studies involving numerous different types of WM training task there was no evidence for transfer to a variety of untrained tasks in both typically developing children aged 9 to 11 years or healthy adults. In the WMT literature far-transfer claims are sometimes made suggesting that a WMT intervention resulted in improved higher order cognitive abilities such as fluid intelligence but with unsatisfactory or no evidence to suggest even a generalised WM improvement (e.g. Jaeggi et al., 2008; Thorell et al., 2009; Zhao et al., 2011). The results presented in this thesis show that when an active control group completing a suitably engaging control paradigm is included and WM improvement is measured by a suite of untrained tasks that no evidence emerges of any improvement in WM. Somewhat surprisingly this holds for all tested individual transfer tasks for all training groups, even those where the modality of the training task memoranda match that of the transfer task. For example, the nearest transfer assessments made in the WMT studies would likely be the CC task to Matrix Span based on the overlapping mechanics of encoding and recall of spatial sequences in serial order. The WMP task to Operation

Span is also a very near-transfer assessment due to the recall of digits in serial order and the need to complete a processing phase concurrently. And yet despite the high degree of overlap there was no evidence of transfer for these assessments.

The problem for advocates of WMT interventions is that based on the rationale they provide the existence of genuine far-transfer should not occur without corresponding near-transfer to working memory. It may be said that the measures used here to assess near-transfer do not tap into the processes that are hypothesised to improve. This may be a persuasive argument but the onus should be on the advocates of WMT interventions to provide a more nuanced theoretical framework for their studies that outlines which processes are being trained, how these are going to be assessed pre- and post-training, and how these processes can improve other constructs through their advancement.

Some published studies do meet this criterion. Gibson et al. (2013) used the dual-component model of working memory (Unsworth & Engle, 2007a, 2007b) as a framework for their WMT intervention study and provided evidence that interventions could be set up to target the specific components of WM. While Gibson and colleagues focused only on near-transfer measures, von Bastian and Oberauer (2013) based their interventions on targeting each element of the facet model (Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000) and were able to observe evidence of near- and far-transfer to tasks with overlapping sub-processes in the Storage-Processing and Supervision training groups. These studies are currently the exception to the general rule however. Namely, that the evidence required to understand the causal links between training and transfer is often not sought or relegated to an afterthought by means of only providing weak evidence of construct improvement.

An important contrast between models of working memory rests first on whether or not they include a single system or multiple components and second where the capacity (or resource limitations) is limited. Multi-Component models of working memory such as the seminal framework proposed by Baddeley and Hitch (Baddeley & Hitch, 1974; Baddeley, 1986, 2000) include subcomponents that store information and therefore there are various limiting factors on the amount of material one can hold as it may vary depending on the subcomponent required. Models that take a more domain-general approach without dividing storage into multiple components such as the position of Cowan and colleagues (Cowan, 1995, 1999; Cowan et al., 2005) where WM is conceptualised as activated representations of items from long-term memory and performing tasks using these representations requires the navigation of these items within the focus of attention i.e. 'zooming in' on representations particularly relevant to the current goal. An alternative conceptualisation of such a model is provided by Oberauer (Oberauer, 2002, 2009) whereby working memory performance involves the activated items form long-term memory, a capacity limited selection of items for direct access, and the focus of attention. If generalised WM improvements are able to be produced by adaptive WMT interventions then under the assumptions of models with one domain-general capacity driver a wider array of tasks would be able to provide a wider array of transfer as they should all be utilising the domain-general component. Models whereby different components are described for storage of different types of material such as the phonological loop, visuospatial sketchpad, and episodic buffer described by Baddeley and Hitch would be more selective in what training tasks are able to produce transfer to specific domains of task. The training tasks in this study covered various aspects of WM as well as the spread of transfer measures used (particularly in WMT experiments two and three).

While the work in this thesis was focused on the applied domain with regards to the efficacy of WMT interventions there are some observed results that can speak to the theoretical considerations that are important. The general correlational pattern observed between the span tasks show that while it was suitable to extract verbal and visuospatial principal components there was a significant relationship between these components and this held for both the developing and adult samples. These results concur with previous work where SEM analysis only extracts different components when a significant relationship between the two is observed (Kane et al., 2004). The range of training assessed includes focus on the verbal domain (WMP), visuospatial domain (CC,SNB), a concurrent maintenance of both (DNB), emphasis on increased processing cost (WMP), and a requirement to encode and maintain a series of items with multiple features (CC). The absence of any increased performance on highly related near-transfer measures - such as CC training on Matrix/Symmetry Span measures - could be interpreted as evidence against domain-specific stores. If different systems were responsible for different materials then there are more opportunities for components to increase their capabilities. It must be said however that given there were no evidence of any transfer effects then it could be concluded that whatever model of WM you assess in light of these results, whatever component is deemed to be the capacity restricting factor is immutable with regards that capacity.

Few studies in the WMT literature have tested a variety of untrained WM tasks to assess near-transfer and those that have present mixed evidence for different models of working memory. Some results would support domain-general accounts. For example Jaeggi et al. (2014) found transfer to a composite measure of visuospatial WM but not verbal WM (all simple span tasks) after verbal N-Back training. However some patterns of transfer meet the assumptions of multi-component domain-specific models. Bergman Nutley et al. (2011) found transfer restricted to the trained domain as visuospatial WM was shown to increase but verbal WM did not after visuospatial training. Working memory training studies present an opportunity to address some of the important theoretical issues in the field but are currently rarely set up in such a way to offer this insight.

6.4 Strengths and Limitations

It is important to frame the results discussed so far within the scope that the evidence acquired can provide. The difficulties of carrying out research as resource intensive as intervention studies combined with other factors limited various aspects of the studies presented in this thesis. These difficulties manifest in two specific issues. Firstly, the resultant training doses ended up being shorter than planned and having to settle for approximately 10 sessions for many in the developmental studies as opposed to the planned 15-20 for all trainees. Secondly, a number of small properties of the procedures that fall under the rubric of 'training fidelity'. The term was coined by Jaeggi et al. (2014) in describing the process of completing computerised training at home as opposed to supervised by an experimenter in the lab. Thus factors that impact on training fidelity are primarily environmental factors that may reduce the effectiveness of training.

The reduced number of training sessions that particularly affected WMT studies one and two was still equal to or larger than the length of some published WMT work claiming transfer effects (Rueda et al., 2005; Borella et al., 2010; Van Der Molen, Van Luit, Van Der Molen, Klugkist, & Jongmans, 2010; Colom et al., 2010; Prins et al., 2011; Loosli et al., 2012; Borella, Carretti, Zanoni, Zavagnin, & De Beni, 2013). So while I acknowledge this is a significant limitation of these studies as the obvious retort is that the participants may simply have not spent enough time on the training task to elicit the positive effects. It must also be said that the true effect sizes (if indeed there are positive effects to be found) are almost certainly going to be small as evidenced by the results of meta-analyses (Melby-Lervåg & Hulme, 2013; Au et al., 2015) showing mean effect sizes in the .2 to .3 range. As discussed in the literature review it is likely both of these meta-analyses yield slightly positively biased results. The near-transfer measures of verbal and visuospatial WM in Melby-Lervåg and Hulme (2013) included multiple effects from studies where the near-transfer was assessed using the trained task.

Despite these limitations there are several counterpoints. As already stated the amount of training is comparable to a number of studies that make claims regarding transfer and some also involve training session in the classroom environment (e.g. Loosli et al., 2012). Additionally, only the WMP adult training group in WMT study three showed a training curve that suggested further training sessions may lead to further improvements and as discussed it is much more likely that the array of possible strategies to aid performance is more likely to explain that curve. The issues relating to training fidelity are almost entirely a result of the group training procedure we employed in the classroom for the developmental studies. While these sessions were attended by researchers and teachers there is always going to be a degree of distraction and disruption in a classroom of over 20 children. With the influx of commercialised WM training programs in the mass market and entering the education sector (notable examples include CogMed and Meemo which is specifically designed to be conducted in the classroom) studies using a grouped training procedure provide valuable insight into the efficacy of such interventions with this administration method. Therefore while a harsh critic might suggest this procedure does not lead to generalisable conclusions on WMT interventions as a whole, there is still value in the obtained results with regards to grouped training in a developmental setting.

A further strength of the work presented here is the breadth of types of working memory training interventions combined with a focus on assessing the generalisable effects of these interventions on the trained construct itself. It is important that the field of working memory training can explain what is happening to the working memory system itself as a result of the adaptive training regimes to fully explain and validate any claims of far-transfer. While the field of working memory training has seen a recent surge in scientific publications, very few of these assess the interventions at stages of typical development. This is understandable given the identified role of deficits in working memory plays in various intellectual development disorders (Barkley, 1997; Kerns et al., 1999; Kuntsi, Oosterlaan, & Stevenson, 2001; Rapport et al., 2008). The studies presented in this thesis therefore add to a very small segment of the WMT literature that is currently underdeveloped.

This inevitably leads to the fact that the studies presented in this thesis can only speak strongly about the effects of working memory training programs on typically developing children and healthy adults. Much of the updating based training studies focus on healthy young- and old-adults (Dahlin, Nyberg, et al., 2008; Li et al., 2008; Jaeggi et al., 2008; Jaeggi, Studer-Luethi, et al., 2010; Salminen et al., 2012; Redick et al., 2013; Thompson et al., 2013) while the other dominant training paradigm CogMed (reminder that this is essentially just visuospatial STM training) dominates the atypically developing segment of the WMT literature (Klingberg et al., 2002, 2005; Holmes et al., 2009, 2010; Van Der Molen et al., 2010). Therefore while the myriad criticisms regarding many of the methodological shortcomings of many of these papers the evidence provided in this thesis cannot necessarily be used to refute claims that interventions such as CogMed are beneficial to groups with diagnosed intellectual development disorders where low working memory is a known symptom/contributor. However, one hopes that the discussions of the overall shortcomings in methodological rigour discussed here and by others (Shipstead et al., 2010; Melby-Lervåg & Hulme, 2013) combined with the currently vague causal models put forward to explain transfer - that rarely is supported due to a lack of robust evidence for working memory improvement - is enough to raise awareness regarding the efficacy of these interventions.

A considerable limitation of the WMT studies discussed in this thesis is a lack of statistical power. In each of the WMT experiments presented the number of participants completing the study was smaller than the post-hoc power calculations (presented in section 2.3.5). While some studies have reported extremely large effect sizes such as Holmes et al. (2009) who reported a Cohen's d of 2.38. However, effect sizes of this magnitude are considerably distant from the average effect size reported in meta-analyses (Melby-Lervåg & Hulme, 2013). It is essential that future studies in this area use sensible effect sizes to calculate sample size requirements and not those that show irregularly large effects.

In the studies presented in this thesis there were no intent-to-treat (ITT) analyses conducted. If more demographic information was available then ITT analysis would have been potentially important to pinpoint what makes a participant more likely to complete the training up to certain thresholds. For example, studies such as that by (Jaeggi et al., 2014) which focus on individual differences in WM training results may be able to use ITT with information such as motivation, belief in 'brain-training' and perceived difficulty to assess which factors affect compliance.

6.5 Methodological Implications

The methodological rigour of the field of working memory training has been criticised (Shipstead et al., 2010, 2012; Redick et al., 2013) and the results presented in this thesis can further support that assessment. Firstly, from an analysis perspective, a repeated observation from the studies presented here is that the results of inferential tests need to be considered in conjunction with the profile of change in an active control group. There are a number of results discussed in their respective sessions that could be used to support the notion of transfer but when more closely inspected are a consequence of lower post-training scores in the control group. A related issue stems from the analysis of post-training differences only without controlling for pre-training scores. For example in the Au et al. (2015) meta-analysis each effect size input into the analysis was computed using the post-training differences between groups only. This resulted in some effects being misrepresented in situations where pre-training differences might have existed. Using this methodology it is possible to obtain positive effect sizes for a group factor with a reduced score at post-training for the intervention group. This is clearly an inappropriate interpretation of such results. That method is based on the assumptions one can make when true randomisation has been used to assign groups to eliminate pre-training differences. However, when smaller samples are used there is still a possibility that differences between groups are seen. Often analyses are presented to show that these differences are not significant but this does not protect the researcher completely from those differences in their post-training analyses. Therefore it is suggested that all analyses use adjusted means based on pre-training scores to prevent this issue.

The quality of control groups is a point of emphasis for the methodological quality of intervention studies. The control group only acts as a baseline measure if the prescribed activities for the group can reasonably be assumed to match the experimental intervention in all facets except what the intervention is targeting. Many studies in the field do not use a control group (e.g. Holmes et al., 2010; Kronenberger et al., 2011; Loosli et al.,

2012) or use a passive control group who only participate by providing pre- and post-training information (e.g Westerberg et al., 2007; Jaeggi et al., 2008; Chein & Morrison, 2010). Throughout this thesis there have been examples of analyses that would have pointed to significant effects of training without the presence of the data provided by the control group. These have been discussed in their respective sessions but as a reminder include transfer to the silly sentence task in WMT experiment one, transfer to mental arithmetic, and transfer to WM across multiple tasks with regards speed of processing in WMT experiment three.

Additionally, it may be the case that design choices in the administration of a training intervention have implications for detecting important effects. Group testing clearly presents a greater degree of potential distraction particularly when testing children in the classroom. As already discussed this is a potential source of the difference in practice effects observed between the developing sample used in WMT studies one/two in this thesis and the adult sample used in study three. This may to some degree explain the differences between the results presented here and that of the numerous studies showing significant transfer as individual testing is most common in the literature.

Shipstead et al. (2010, 2012) highlight the importance of task selection in WM training studies. Chapter three of this thesis expands on this issue by assessing different methods of scoring WM tasks and what differences in measures of 'memory span' actually represent. The results of the analyses presented in chapter three suggest that it is not appropriate to consider span as an interval scale as the difficulty of increasing span items does not increase at a constant rate. WM training studies need to consider this issue as smaller value increases at higher span values may actually be representing a larger cognitive improvement than a seemingly equivalent improvement at the lower range of span scores.

6.6 Future Work

Despite the results provided in this thesis and the methodological concerns over much of the working memory training literature, there are undoubtedly many more intervention studies to come. N-Back appears to be the paradigm of choice for those who make claims of far-transfer benefits (Au et al., 2015). An important point when discussing the N-Back training task which relates to the wider literature regarding WM measurement via N-Back paradigms is that of construct validity (Kane et al., 2007; Jaeggi et al., 2010). Questions regarding construct validity stem from repeated results suggesting a non-significant or low correlation between n-back and complex span tasks of WM. Redick and Lindsey (2013) conducted a meta-analysis focused on the relationship between n-back and complex-span performance. They found that the weighted average correlation between the two tasks was $r = .20$ (95% $CI = .16 - .24$). The authors also collected information on the relationship between simple span measures and n-back performance where available. They found that the weighted average correlation between these measures was $r = .25$ (95% $CI = .21 - .3$). These results indicate that while the shared variance between complex span and n-back is significant, it is of a magnitude lower than would be expected if they measured the same construct (reaffirming the conclusion drawn by Kane et al., 2007). Additionally, the strength of relationship between simple span and n-back is not 'statistically significant' from that of complex span and n-back. The results presented here showing no transfer of N-Back training to WM span tasks replicates other failed near-transfer to WM assessments (e.g. Jaeggi et al., 2008; Redick et al., 2013). This suggests that N-Back to standard measures of WM may be further transfer than generally considered in the WMT literature and further than I have treated it in pursuit of this research program.

In defence of the use of N-Back tasks as a working memory measure, much of the work showing only a small relationship between N-Back and complex span measures of working memory use varying dependent variables as indicators of N-Back performance. There appears to be no definitive outcome measure of n-back performance that is consistently used in much of the behavioural work where it is deployed. Redick and Lindsey (2013) used an overall accuracy dependent variable for n-back performance as this was the measure that was available most often. However, often overall accuracy may be included in the results (or may have been obtained from the authors via correspondence) but are not the dependent variable focused on by the authors in their analyses. Kane et al. (2007) is an example of such a situation where the focus was on lure performance and signal detection estimates as opposed to an overall accuracy measure. Alternatively, Jacola et al. (2014) assessed overall accuracy and response times as primary DVs and found considerable ceiling effects in their accuracy outcomes. The common administration of n-back in studies comparing performance to complex span appears to be a pre-defined selection of blocks at specific levels of n. For example, Kane et al. (2007) gave participants 8 experimental blocks alternating between 2-back and 3-back. Jacola et al. (2014) asked participants to complete one experimental block at each 0-,1-, and 2-back for both a verbal and an object stimulus type. It would seem likely that n-back performance is more likely to match complex span performance when the dependent variables selected are comparable.

As discussed in this thesis already there are numerous methods to score complex span tasks (Conway et al., 2005; Friedman & Miyake, 2005). Generally a participant is going to receive a score which is based on the maximum span they were able to reach while still answering questions successfully or a total number of individually recalled items in correct serial order which is going to be mediated by how far they were able to go in the task. An analogous measure for n-back performance might be to continue to increase n while participants are maintaining a certain level of success in their responses and terminating the procedure when a critical threshold of errors is reached. The level of n itself clearly affects the processes involved in being successful at the task as shown by the correlational pattern in Jaeggi and colleagues (2010) work where the correlations between 1-, 2-, and 3-back variants of n-back and Raven's advanced progressive matrices (RAPM) showed that the 3-back versions (visuospatial, verbal, and dual) were significantly correlated with RAPM whereas lower levels of n were not. Therefore, when N-Back tasks are used in an adaptive-difficulty paradigm as in training studies participants are working at or above the highest level of n they are able to succeed at. It may be that at these levels of n the task shares more properties with more traditional WM measures. Further work needs to be carried out to clarify this relationship to either justify the use of N-Back as a working memory training task, or to identify alternative functions that may be benefitting from N-Back training in place of working memory.

This thesis has drawn attention to the importance of and relationship between three types of training impact; practice, near-transfer, and far-transfer. The conclusions drawn from the data presented here suggest that the first of these types of effects - practice effects - may be an under-studied area of the field. Practice effects are rarely given any consideration in published working memory training papers (Klingberg et al., 2005; Green et al., 2012; Söderqvist et al., 2012; Dahlin, 2011) but there is potentially a lot to learn about what is actually happening cognitively throughout training by giving due thought to the practice effects. The results presented here show differential patterns of practice effects between tasks and sample characteristics. As discussed in a previous section it is very likely that the role of cognitive strategies are integral to understand practice effects and in turn affect the way in which transfer results would then be interpreted. If the practice effects can be explained by strategy utilisation and the developed strategies are applicable to the transfer tasks then transfer may be observed without a raising of the limits of the 'cognitive engine'. The role of rote rehearsal is well known to be integral to performance on memory span tasks, most commonly with regards phonogological memoranda (Naus et al., 1977; Cowan et al., 1987) but also spatial memoranda (Pearson & Sahraie, 2003). Covert rehearsal is relatively automatic in adults when encountered with the types of tasks used in short-term/working memory research (Guttentag, 1984). Naus et al. (1977) found that sixth grade (ages 12-13) children spontaneously used covert rehearsal strategies to help performance but second/third grade children did not. It has been shown that adult performance on a simple span task (word span) can be reduced to similar levels as young children by blocking their ability to engage in rehearsal (Cowan et al., 1987). Cognitive tips and tricks clearly affect the level of performance on assessments of memory and the current stage of development is a significant predictor in whether participants are likely to engage in these without being prompted. Future WMT studies should focus on understanding if strategies are the driving force behind observed effects and consider interventions based around teaching strategies and methods to ensure these strategies transfer to a wide range of activities.

6.7 Conclusions

Research attempting to identify interventions to improve cognitive abilities has a long and influential history. Earlier interventions were highly intensive and considerable in their duration such as the Abecedarian project. Despite the duration of such interventions being measured in years and applied at the earliest stage of development the overall effects on cognitive faculties while significant were relatively modest (e.g. Ramey & Haskins, 1981). Therefore when Torkel Klingberg and colleagues began publishing work showing that a very short adaptive-difficulty training intervention targeting working memory produced generalisable cognitive benefits via far-transfer to fluid intelligence (Klingberg et al., 2002, 2005) a wave of optimism encompassed the field. A number of studies followed that were reported to replicate these findings, extend them to additional populations, and show that a variety of working memory tasks can produce such effects. The work of Susanne Jaeggi and colleagues (Jaeggi et al., 2008; Buschkuehl et al., 2008; Jaeggi et al., 2011) has been particularly influential by introducing the Dual N-Back paradigm as a successful training task. However, the field was built on a number of studies with very weak methodological rigour (Shipstead et al., 2010, 2012) due to either the absence of control groups or use of a passive control group, weak evidential claims of transfer by using variants of the same task as pre-post assessments, not using the most powerful statistical analysis to detect effects, and a significant amount of potential for a conflict of interests due to the rapid commercialisation of the interventions.

Therefore the goal of this thesis was to carry out independent research assessing a variety of working memory training interventions, both novel and replicating successful paradigms from the literature. Each study presented in this thesis uses a control group that is matched as closely as possible to the training group in terms of the attention they receive from the researchers and the time spent engaging with adaptive computerised tasks. These studies focus on whether repeatedly training on working memory tasks improves the construct itself as this is the most basic causal explanation for the observed far-transfer in the literature - training increases working memory capacity which is intergral to numerous facets of cognition. Additionally, work was conducted to ensure that the tasks used for assessment were appropriate and alternative scoring methods were assessed and used based on their properties.

The pattern of results with regards near-transfer to untrained tasks of working memory is consistent across each study (typically developing children and adults) in that there is none. The range of training tasks covers various facets of working memory in terms of verbal and visuospatial tasks, multi-feature binding tasks, and updating paradigms (spatial N-Back for developing sample and Dual N-Back for adults). The pattern of practice effects varied between tasks and between samples. It is posited that given the lack of near transfer observed, where practice effects were observed it is more likely that it was a result of strategy use than a neural-based improvement in a capacity limited component of working memory. These results cast doubt on the efficacy of adaptive difficulty working memory training paradigms.

References

Abikoff, H. (1991). Cognitive training in adhd children: Less to it than meets the eye. *Journal of learning Disabilities*, *24*(4), 205–209.

Alloway, T. P. (2009). Working memory, but not iq, predicts subsequent learning in children with learning difficulties. *European Journal of Psychological Assessment*, *25*(2), 92–98. doi: 10.1027/1015-5759.25.2.92

Arnett, P., Higginson, C., Voss, W., Wright, B., Bender, W., Wurst, J., & Tippin, J. (1999). Depressed mood in multiple sclerosis: Relationship to capacity-demanding memory and attentional functioning. *Neuropsychology*, *13*(3), 434–446. Retrieved from `http://dx.doi.org/10.1037/0894-4105.13.3.434`

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *Psychology of learning and motivation*, *2*, 89–195.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic bulletin & review*, *22*(2), 366-377. doi: 10.3758/s13423-014-0699-x

Baddeley, A. (1968). A 3 min reasoning test based on grammatical transformation. *Psychonomic Science*, *10*(10), 341-342. Retrieved from `http://dx.doi.org/10.3758/BF03331551` doi: 10.3758/BF03331551

Baddeley, A. (1986). *Working memory.* London: Oxford University Press.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, *4*(11), 417–423.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psy-*

chology of learning and motivation (pp. 47–89). New York: Academic Press. doi: 10.1016/S0079-7421(08)60452-1

Baddeley, A., & Lieberman, K. (1980). Spatial working memory. In R. Nickerson (Ed.), *Attention and performance viii* (pp. 521–539). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Ball, K., Berch, D., Helmers, K., Jobe, J., Leveck, M., Marsiske, M., … others (2002). Effects of cognitive training interventions with older adults. *The Journal of the American Medical Association*, *288*(18), 2271–2281.

Barkley, R. A. (1997). Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of adhd. *Psychological bulletin*, *121*(1), 65–94.

Bates, E., Reilly, J., Wulfeck, B., Dronkers, N., Opie, M., Fenson, J., … Herbst, K. (2001). Differential effects of unilateral lesions on language production in children and adults. *Brain and language*, *79*(2), 223–265. doi: 10.1006/brln.2001.2482

Beck, S., Hanson, C., Puffenberger, S., Benninger, K., & Benninger, W. (2010). A controlled trial of working memory training for children and adolescents with adhd. *Journal of Clinical Child & Adolescent Psychology*, *39*(6), 825–836.

Bergman Nutley, S., Söderqvist, S., Bryde, S., Thorell, L., Humphreys, K., & Klingberg, T. (2011). Gains in fluid intelligence after training non-verbal reasoning in 4-year-old children: a controlled, randomized study. *Developmental science*, *14*(3), 591–601.

Binet, A., & Simon, T. (1905). New methods for the diagnosis of the intellectual level of subnormals. *L'annee Psychologique*, *12*, 191–244.

Boake, C. (2002). From the binet–simon to the wechsler–bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 383–405.

Bond, T. G., & Fox, C. M. (2001). *Applying the rasch model.* New Jersey: Lawrence Erlbaum Associates London.

Borella, E., Carretti, B., & De Beni, R. (2008). Working memory and inhibition across the adult life-span. *Acta psychologica*, *128*(1), 33–44. doi: 10.1016/j.actpsy.2007.09.008

Borella, E., Carretti, B., Riboldi, F., & De Beni, R. (2010). Working memory training in older adults: evidence of transfer and maintenance effects. *Psychology and aging*, *25*(4), 767–778. doi: 10.1037/a0020683

Borella, E., Carretti, B., Zanoni, G., Zavagnin, M., & De Beni, R. (2013). Working memory training in old age: an examination of transfer and maintenance effects. *Archives of clinical neuropsychology*, *28*(4), 331–347. doi: 10.1093/arclin/act020

Bower, G. H. (1970). Analysis of a mnemonic device: Modern psychology uncovers the powerful components of an ancient system for improving memory. *American Scientist*, 496–510.

Bower, G. H. (2000). A brief history of memory research. *The Oxford handbook of memory*, 3–32.

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working memory training in younger and older adults: training gains, transfer, and maintenance. *Frontiers in human neuroscience*, *6*, 1–7.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, *64*(3), 205–215.

Brooks, L. R. (1968). Spatial and verbal components of the act of recall. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *22*(5), 349–368.

Brown, A. L., & Barclay, C. R. (1976). The effects of training specific mnemonics on the metamnemonic efficiency of retarded children. *Child Development*, 71–80.

Brown, A. L., Campione, J. C., Bray, N. W., & Wilcox, B. L. (1973). Keeping track of changing variables: Effects of rehearsal training and rehearsal prevention in normal and retarded adolescents. *Journal of Experimental Psychology*, *101*(1), 123–131.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, *10*(1), 12–21.

Bryck, R. L., & Fisher, P. A. (2012). Training the brain: practical applications of neural plasticity from the intersection of cognitive neuroscience, developmental psychology, and prevention science. *American Psychologist*, *67*(2), 87-100. doi:

10.1037/a0024657

Bühner, M., König, C., Pick, M., & Krumm, S. (2006). Working memory dimensions as differential predictors of the speed and error aspect of multitasking performance. *Human Performance*, *19*(3), 253–275. doi: 10.1207/s15327043hup1903_4

Bunting, M. (2006). Proactive interference and item similarity in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(2), 183–196. doi: 10.1037/0278-7393.32.2.183

Buschkuehl, M., Jaeggi, S. M., Hutchison, S., Perrig-Chiello, P., Däpp, C., Müller, M., . . . Perrig, W. J. (2008). Impact of working memory training on memory performance in old-old adults. *Psychology and aging*, *23*(4), 743. doi: 10.1037/a0014342

Butterfield, E. C., Wambold, C., & Belmont, J. M. (1973). On the theory and practice of improving short-term memory. *American journal of mental deficiency*, 654–669.

Calkins, M. W. (1894). Association. *Psychological Review*, *1*, 476–483.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, *97*(3), 404–431.

Carretti, B., Borella, E., & De Beni, R. (2007). Does strategic memory training improve the working memory performance of younger and older adults? *Experimental Psychology*, *54*(4), 311–320. doi: 10.1027/1618-3169.54.4.311

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of experimental child psychology*, *33*(3), 386–404. doi: 10.1016/0022-0965(82)90054-6

Chein, J., & Morrison, A. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic bulletin & review*, *17*(2), 193–199.

Clark, R. E., & Sugrue, B. M. (1991). Research on instructional media, 1978-1988. In G. J. Anglin (Ed.), *Instructional technology: Past, present, and future.* Englewood, CO: Libraries Unlimited.

Colom, R., Quiroga, M., Shih, P., Martínez, K., Burgaleta, M., Martínez-Molina, A., ... Ramírez, I. (2010). Improvement in working memory is not related to increased intelligence scores. *Intelligence*, *38*(5), 497–505.

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., ... others (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence*, *41*(5), 712–727. doi: 10.1016/j.intell.2013.09.002

Conrad, R. (1964). Acoustic confusions in immediate memory. *British journal of Psychology*, *55*(1), 75–84.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British journal of psychology*, *55*(4), 429–432.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin and review*, *12*(5), 769–786.

Conway, A. R., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183. doi: 10.1016/S0160-2896(01)00096-4

Corsi, P. M. (1972). *Human memory and the medial temporal region of the brain* (Unpublished doctoral dissertation). McGill University, Montreal.

Cowan, N. (1995). *Attention and memory: An integrated framework.* Oxford University Press, Oxford.

Cowan, N. (1999). An embedded-process model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory. mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge, UK: Cambridge University Press.

Cowan, N., Cartwright, C., Winterowd, C., & Sherk, M. (1987). An adult model of preschool children's speech memory. *Memory & cognition*, *15*(6), 511–517. doi: 10.3758/BF03198385

Cowan, N., Elliott, E., Scott Saults, J., Morey, C., Mattox, S., Hismjatullina, A., & Conway, A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive psychology*, *51*(1), 42–100.

Cowan, N., Towse, J. N., Hamilton, Z., Saults, J. S., Elliott, E. M., Lacey, J. F., ... Hitch, G. J. (2003). Children's working-memory processes: A response-timing analysis. *Journal of Experimental Psychology: General*, *132*(1), 113–132. doi: 10.1037/0096-3445.132.1.113

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior*, *11*(6), 671–684.

Craik, F. I., & Watkins, M. J. (1973). The role of rehearsal in short-term memory. *Journal of verbal learning and verbal behavior*, *12*(6), 599–607.

Cumming, G. (2014). The new statistics: why and how. *Psychological science*, *25*(1), 7–29. doi: 10.1177/0956797613504966

Dahlin, E., Neely, A., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*(5882), 1510–1512.

Dahlin, E., Nyberg, L., Bäckman, L., & Neely, A. (2008). Plasticity of executive functioning in young and older adults: Immediate training gains, transfer, and long-term maintenance. *Psychology and Aging*, *23*(4), 720–730.

Dahlin, K. (2011). Effects of working memory training on reading in children with special needs. *Reading and Writing*, *24*(4), 479–491.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, *19*(4), 450–466. doi: 10.1016/S0022-5371(80)90312-6

Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *The Quarterly Journal of Experimental Psychology*, *60*(9), 1227–1245. doi: 10.1080/17470210600926075

Dunning, D. L., & Holmes, J. (2014). Does working memory training promote the use of strategies on untrained working memory tasks? *Memory & cognition*, *42*(6), 854–862. doi: 10.3758/s13421-014-0410-5

Ebbinghaus, H. (1913). Memory (ha ruger & ce bussenius, trans.). *New York: Teachers College.(Original work published 1885).*

Ericcson, K., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, *208*(4448), 1181–1182.

Flavell, J. H., Friedrichs, A. G., & Hoyt, J. D. (1970). Developmental changes in memorization processes. *Cognitive psychology*, *1*(4), 324–340.

Francis, G. (1883). *Inquiries into human faculty and its development.* London.

Friedman, N. P., & Miyake, A. (2005). Comparison of four scoring methods for the reading span test. *Behavior Research Methods*, *37*(4), 581–590. doi: 10.3758/BF03192728

Geary, D. C. (1990). A componential analysis of an early learning deficit in mathematics. *Journal of experimental child psychology*, *49*(3), 363–383. doi: 10.1016/0022-0965(90)90065-G

Gibson, B., Gondoli, D., Johnson, A., Steeger, C., Dobrzenski, B., & Morrissey, R. (2011). Component analysis of verbal versus spatial working memory training in adolescents with adhd: A randomized, controlled trial. *Child Neuropsychology*, *17*(6), 546–563.

Gibson, B., Gondoli, D. M., Kronenberger, W. G., Johnson, A. C., Steeger, C. M., & Morrissey, R. A. (2013). Exploration of an adaptive training regimen that can target the secondary memory component of working memory capacity. *Memory & cognition*, *41*(5), 726–737. doi: 10.3758/s13421-013-0295-8

Green, C., Long, D., Green, D., Iosif, A., Dixon, J., Miller, M., … Schweitzer, J. (2012). Will working memory training generalize to improve off-task behavior in children with attention-deficit/hyperactivity disorder? *Neurotherapeutics*, 1–10.

Guttentag, R. E. (1984). The mental effort requirement of cumulative rehearsal: A developmental study. *Journal of Experimental Child Psychology*, *37*(1), 92–106. doi: 10.1016/0022-0965(84)90060-2

Hambrick, D., Oswald, F., Darowski, E., Rench, T., & Brou, R. (2010). Predictors of multitasking performance in a synthetic work paradigm. *Applied Cognitive Psychology*, *24*(8), 1149–1167. doi: 10.1002/acp.1624

Hitch, G. J., Towse, J. N., & Hutton, U. (2001). What limits children's working memory span? theoretical accounts and applications for scholastic development. *Journal of Experimental Psychology: General*, *130*(2), 184–. doi: 10.1037/0096-3445.130.2.184

Holmes, J., Gathercole, S., & Dunning, D. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science*, *12*(4), F9–F15.

Holmes, J., Gathercole, S., Dunning, D., Hilton, K., Elliott, J., et al. (2010). Working memory deficits can be overcome: Impacts of training and medication on working memory in children with adhd. *Applied Cognitive Psychology*, *24*(6), 827–836. doi: 10.1002/acp.1589

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, *82*(3), 472–481.

Jacobs, J. (1887). Experiments on "prehension". *Mind*(45), 75–79.

Jacola, L. M., Willard, V. W., Ashford, J. M., Ogg, R. J., Scoggins, M. A., Jones, M. M., … Conklin, H. M. (2014). Clinical utility of the n-back task in functional neuroimaging studies of working memory. *Journal of clinical and experimental neuropsychology*, *36*(8), 875–886. doi: 10.1080/13803395.2014.953039

Jaeggi, S., Buschkuehl, M., Jonides, J., & Perrig, W. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829–6833.

Jaeggi, S., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short-and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086.

Jaeggi, S., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the n-back task as a working memory measure. *Memory*, *18*(4), 394–412. doi:

10.1080/09658211003702171

Jaeggi, S., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & cognition*, *42*(3), 464–480. doi: 10.3758/s13421-013-0364-z

Jaeggi, S., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? load-dependent processing in single and dual tasks: functional magnetic resonance imaging study. *NeuroImage*, *19*(2), 210–225.

Jaeggi, S., Studer-Luethi, B., Buschkuehl, M., Su, Y., Jonides, J., & Perrig, W. (2010). The relationship between n-back performance and matrix reasoning—implications for training and transfer. *Intelligence*, *38*(6), 625–635.

Jensen, A. R. (1969). How much can we boost iq and scholastic achievement. *Harvard educational review*, *39*(1), 1–123.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger Publishers.

Johnston, C. D., & Jenkins, J. J. (1971). Two more incidental tasks that differentially affect associative clustering in recall. *Journal of Experimental Psychology*, *89*(1), 92–95.

Kane, M. J., Conway, A. R., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. Conway, C. E. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory.* Oxford: Oxford University Press.

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the n-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 615–622.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: the contributions of goal neglect, response competition, and task set to stroop interference. *Journal of experimental psychology: General*, *132*(1), 47–70.

doi: 10.1037/0096-3445.132.1.47

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217. doi: 10.1037/0096-3445.133.2.189

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of verbal learning and verbal behavior*, *1*(3), 153–161.

Kerns, K., Eso, K., & Thomson, J. (1999). Investigation of a direct intervention for improving attention in young children with adhd. *Developmental neuropsychology*, *16*(2), 273–295.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352–358.

Kleider, H., Parrott, D., & King, T. (2010). Shooting behaviour: How working memory and negative emotionality influence police officer shoot decisions. *Applied Cognitive Psychology*, *24*(5), 707–717. doi: 10.1002/acp.1580

Klein, K., & Boals, A. (2001). The relationship of life event stress and working memory capacity. *Applied Cognitive Psychology*, *15*(5), 565–579. doi: 10.1002/acp.727

Klingberg, T., Fernell, E., Olesen, P., Johnson, M., Gustafsson, P., Dahlström, K., ... Westerberg, H. (2005). Computerized training of working memory in children with adhd-a randomized, controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, *44*(2), 177–186.

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with adhd. *Journal of Clinical and Experimental Neuropsychology*, *24*(6), 781–791.

Kronenberger, W., Pisoni, D., Henning, S., Colson, B., & Hazzard, L. (2011). Working memory training for children with cochlear implants: A pilot study. *Journal of Speech, Language, and Hearing Research*, *54*(4), 1182–1196.

Kuntsi, J., Oosterlaan, J., & Stevenson, J. (2001). Psychological mechanisms in hyperactivity: I response inhibition deficit, working memory impairment, delay aversion, or something else? *Journal of Child Psychology and Psychiatry*, *42*(2), 199–210.

Langer, N., von Bastian, C. C., Wirz, H., Oberauer, K., & Jäncke, L. (2013). The effects of working memory training on functional brain network efficiency. *Cortex*, *49*(9), 2424–2438.

Li, S., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., & Lindenberger, U. (2008). Working memory plasticity in old age: Practice gain, transfer, and maintenance. *Psychology and aging*, *23*(4), 731–742.

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology*, *18*(1), 62–78. doi: 10.1080/09297049.2011.575772

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology: General*, *130*(2), 199–207. doi: 10.1037/0096-3445.130.2.199

Mäntylä, T. (1986). Optimizing cue effectiveness: Recall of 500 and 600 incidentally learned words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 66–71.

Mäntylä, T., & Nilsson, L.-G. (1988). Cue distinctiveness and forgetting: Effectiveness of self-generated retrieval cues in delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 502–509.

McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., & Fisher, P. (2007). The hawthorne effect: a randomised, controlled trial. *BMC medical research methodology*, *7*(30). doi: 10.1186/1471-2288-7-30

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology*, *74*(3), 240–260. doi: 10.1006/jecp.1999.2516

McNamara, D. S., & Scott, J. L. (2001). Working memory capacity and strategy use. *Memory & Cognition*, *29*(1), 10–17. doi: 10.3758/BF03195736

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? a meta-analytic review. *Developmental Psychology*, *49*(2), 270–291. doi: 10.1037/a0028228

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Moody, D. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, *37*, 327–328.

Muennig, P., Robertson, D., Johnson, G., Campbell, F., Pungello, E. P., & Neidell, M. (2011). The effect of an early education program on adult health: the carolina abecedarian project randomized controlled trial. *American Journal of Public Health*, *101*(3), 512–516. doi: 10.2105/AJPH.2010.200063

Murdock, B. B. (1965). Effects of a subsidiary task on short-term memory. *British Journal of Psychology*, *56*(4), 413–419.

Murdock Jr, B. B. (1962). The serial position effect of free recall. *Journal of experimental psychology*, *64*(5), 482–488.

Naus, M. J., Ornstein, P. A., & Aivano, S. (1977). Developmental changes in memory: The effects of processing time and rehearsal instructions. *Journal of Experimental Child Psychology*, *23*(2), 237–251. doi: 10.1016/0022-0965(77)90102-3

Nettelbeck, T., & Burns, N. R. (2010). Processing speed, working memory and reasoning ability from childhood to old age. *Personality and Individual Differences*, *48*(4), 379–384.

Oberauer, K. (2002). Access to information in working memory: exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411–421. doi: 10.1037/0278-7393.28.3.411

Oberauer, K. (2009). Design for a working memory. *Psychology of learning and motivation*, *51*, 45–100. doi: 10.1016/S0079-7421(09)51002-X

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017–1045. doi: 10.1016/S0191-8869(99)00251-2

Olesen, P., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature neuroscience*, *7*(1), 75–79.

O'Connor, T. A., & Burns, N. R. (2003). Inspection time and general speed of processing. *Personality and individual differences*, *35*(3), 713–724.

Pearson, D., & Sahraie, A. (2003). Oculomotor control and the maintenance of spatially and temporally distributed events in visuo-spatial working memory. *The Quarterly Journal of Experimental Psychology: Section A*, *56*(7), 1089–1111. doi: 10.1080/02724980343000044

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of neuroscience methods*, *162*(1), 8–13.

Peng, P., & Fuchs, D. (2015). A randomized control trial of working memory training with and without strategy instruction: Effects on young children's working memory and comprehension. *Journal of learning disabilities*, 1–19. doi: 10.1177/0022219415594609

Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, *58*(3), 193–198.

Postman, L. (1971). Transfer, interference and forgetting. In J. W. Kling & L. A. Riggs (Eds.), *Woodworth and schlosberg's experimental psychology* (pp. 1019–1132). New York: Holt, Reinhardt, & Winson.

Postman, L., & Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly journal of experimental psychology*, *17*(2), 132–138.

Prins, P. J. M., Dovis, S., Ponsioen, A., ten Brink, E., & van der Oord, S. (2011). Does computerized working memory training with game elements enhance motivation and training efficacy in children with adhd? *Cyberpsychology, Behavior, and Social Networking*, *14*(3), 115–122.

Ramey, C. T., & Haskins, R. (1981). The modification of intelligence through early experience. *Intelligence*, *5*(1), 5–19. doi: 10.1016/0160-2896(81)90013-1

Rapport, M. D., Alderson, R. M., Kofler, M. J., Sarver, D. E., Bolden, J., & Sims,

V. (2008). Working memory deficits in boys with attention-deficit/hyperactivity disorder (adhd): the contribution of central executive and subsystem processes. *Journal of Abnormal Child Psychology*, *36*(6), 825–837.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

Rausch, J. R., Maxwell, S. E., & Kelley, K. (2003). Analytic methods for questions pertaining to a randomized pretest, posttest, follow-up design. *Journal of Clinical Child and Adolescent Psychology*, *32*(3), 467–486. doi: 10.1207/S15374424JCCP3203_15

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*(3), 164–171. doi: 10.1027/1015-5759/a000123

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: a meta-analysis. *Psychonomic bulletin & review*, *20*(6), 1102–1113. doi: 10.3758/s13423-013-0453-9

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., … Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *Journal of Experimental Psychology: General*, *142*(2), 359–379. doi: 10.1037/a0029082

Rizopoulos, D. (2006). ltm: An r package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Roid, G. (2003). Stanford-binet intelligence scales (sb5). *Rolling Meadows, IL: Riverside.*

Rudebeck, S., Bor, D., Ormond, A., O'Reilly, J., & Lee, A. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. *PloS ONE*, *7*(11), e50431. doi: 10.1371/journal.pone.0050431

Rueda, M., Rothbart, M., McCandliss, B., Saccomanno, L., & Posner, M. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the national Academy of Sciences of the United States of America*, *102*(41), 14931–14936.

Salamé, P., & Baddeley, A. (1986). Phonological factors in stm: Similarity and the unattended speech effect. *Bulletin of the Psychonomic Society*, *24*(4), 263–265. doi: 10.3758/BF03330135

Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. *Frontiers in Human Neuroscience*, *6*(166). doi: 10.3389/fnhum.2012.00166

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the cogito study. *Frontiers in Aging Neuroscience*, *2*, 1–10.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: an individual differences approach. *Journal of Experimental Psychology: General*, *125*(1), 4–27. doi: 10.1037/0096-3445.125.1.4

Shipstead, Z., Redick, T., & Engle, R. (2010). Does working memory training generalize? *Psychologica Belgica, 50*, *3*(4), 245–276.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*(4), 628–654.

Shute, V. (1991). Who is likely to acquire programming skills? *Journal of Educational Computing Research*, *7*(1), 1–24. doi: 10.2190/VQJD-T1YD-5WVB-RYPJ

Siegler, R. S. (1994). Cognitive variability: A key to understanding cognitive development. *Current directions in psychological science*, 1–5.

Simon, H. A. (1974). How big is a chunk? *Science*, *183*(4124), 482–488.

Söderqvist, S., Nutley, S. B., Ottersen, J., Grill, K. M., & Klingberg, T. (2012). Computerized training of non-verbal reasoning and working memory in children with intellectual disability. *Frontiers in human neuroscience*, *6*.

St Clair-Thompson, H. (2012). Ascending versus randomised list lengths in working memory span tasks. *Journal of Cognitive Psychology*, *24*(3), 335–341. doi: 10.1080/20445911.2011.639760

St Clair-Thompson, H., Stevens, R., Hunt, A., & Bolder, E. (2010). Improving children's working memory and classroom performance. *Educational Psychology*, *30*(2), 203–219. doi: 10.1080/01443410903509259

St Clair-Thompson, H., & Sykes, S. (2010). Scoring methods and the predictive ability of working memory tasks. *Behavior research methods*, *42*(4), 969–975. doi: 10.3758/BRM.42.4.969

Stone, J. M., & Towse, J. N. (2015). A working memory test battery: Java-based collection of seven working memory tasks. *Journal of Open Research Software*, *3*(1), e5. doi: 10.5334/jors.br

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of experimental psychology*, *18*(6), 643–662.

Süß, H., Oberauer, K., Wittmann, W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288.

Terman, L. M. (1916). *The measurement of intelligence.* Houghton Mifflin.

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., . . . others (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS ONE*, *8*(5). doi: 10.1371/journal.pone.0063614

Thorell, L., Lindqvist, S., Bergman Nutley, S., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental science*, *12*(1), 106–113.

Towse, J., Cowan, N., Horton, N., & Whytock, S. (2008). Task experience and children's working memory performance: A perspective from recall timing. *Developmental psychology*, *44*(3), 695.

Towse, J., Hitch, G., Hamilton, Z., Peacock, K., & Hutton, U. (2005). Working memory period: The endurance of mental representations. *The Quarterly Journal of Experimental Psychology Section A*, *58*(3), 547–571.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373.

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language*, *49*(4), 446–468. doi: 10.1016/S0749-596X(03)00095-0

Turner, M., & Engle, R. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. doi: 10.1016/0749-596X(89)90040-5

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling stroop effects by manipulating expectations for color words. *Memory & Cognition*, *20*(6), 727–735.

Tzeng, O. J. (1973). Positive recency effect in a delayed free recall. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 436–439.

Unsworth, N., & Engle, R. W. (2007a). The nature of individual differences in working memory capacity: active maintenance in primary memory and controlled search from secondary memory. *Psychological review*, *114*(1), 104–132.

Unsworth, N., & Engle, R. W. (2007b). On the division of short-term and working memory: an examination of simple and complex span and their relation to higher order abilities. *Psychological bulletin*, *133*(6), 1038–1066.

Unsworth, N., Heitz, R., Schrock, J., & Engle, R. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. doi: 10.3758/BF03192720

Van Der Molen, M., Van Luit, J., Van Der Molen, M., Klugkist, I., & Jongmans, M. (2010). Effectiveness of a computerised working memory training in adolescents with mild to borderline intellectual disabilities. *Journal of Intellectual Disability Research*, *54*(5), 433–447.

von Bastian, C. C., Langer, N., Jäncke, L., & Oberauer, K. (2013). Effects of working memory training in young and old adults. *Memory & cognition*, *41*(4), 611–624. doi: 10.3758/s13421-012-0280-7

von Bastian, C. C., Locher, A., & Ruflin, M. (2013). Tatool: A java-based open-source programming framework for psychological studies. *Behavior research methods*, *45*(1), 108–115. doi: 10.3758/s13428-012-0224-y

von Bastian, C. C., & Oberauer, K. (2013). Distinct transfer effects of training different facets of working memory capacity. *Journal of Memory and Language*, *69*(1), 36–58. doi: 10.1016/j.jml.2013.02.002

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1207–1241. doi: 10.1037/a0020122

Wechsler, D. (2003). Wechsler intelligence scale for children–fourth edition (wisc-iv). *San Antonio, TX: The Psychological Corporation.*

Wechsler, D. (2008). Wechsler adult intelligence scale–fourth edition (wais–iv). *San Antonio, TX: NCS Pearson.*

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Östensson, M., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke-a pilot study. *Brain Injury*, *21*(1), 21–29.

Witmer, L. R. (1935). The association value of three-place consonant syllables. *The Pedagogical Seminary and Journal of Genetic Psychology*, *47*(2), 337–360.

Yerkes, R. M. (1921). *Psychological examining in the united states army: Edited by robert m. yerkes* (Vol. 15). US Government Printing Office.

Zhao, X., Wang, Y., Liu, D., & Zhou, R. (2011). Effect of updating training on fluid intelligence in children. *Chinese Science Bulletin*, *56*(21), 2202–2205.

Appendix A

Working Memory Training Developmental Study One

A.1 Ethics Statement

Ethical approval for all empirical work in this thesis was gained from the Department of Psychology ethics committee at Lancaster University. The studies conducted in chapters 2 and 4 were conducted in schools and permission to do so was gained from the schools after they volunteered their participation. The headteacher of each participating school selected the classes that could participate. Consent from parents was not obtained at the school's request. Chapters 3 and 5 involve the recruitment of adult participants and in each case informed consent was gained.

A.2 Method

A.2.1 Mental Arithmetic Stimuli

Table A.1

Question sets for each block in the Mental Arithmetic task used in the WMT Developmental One study

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Qu1	5 + 3	5 + 7	3 - 2	22 - 14	4 x 3	4 / 2
Qu2	1 + 2	6 + 7	5 - 4	22 - 13	6 x 1	12 / 4
Qu3	1 + 3	7 + 5	2 - 1	12 - 8	6 x 2	12 / 1
Qu4	6 + 2	9 + 1	7 - 3	14 - 5	2 x 4	9 / 3
Qu5	7 + 1	9 + 5	8 - 5	11 - 8	5 x 2	8 / 4
Qu6	2 + 1	7 + 4	6 - 4	18 - 9	11 x 1	12 / 2
Qu7	6 + 1	4 + 6	4 - 3	23 - 14	3 x 4	20 / 10
Qu8	2 + 6	8 + 5	6 - 2	25 - 17	2 x 9	20 / 5
Qu9	4 + 5	8 + 6	8 - 6	25 - 18	4 x 9	16 / 8
Qu10	3 + 4	9 + 2	7 - 6	33 - 15	5 x 5	15 / 5
Qu11	43 + 25	62 + 58	25 - 14	56 - 37	6 x 4	18 / 3
Qu12	88 + 11	43 + 67	47 - 25	23 - 8	5 x 4	40 / 8
Qu13	84 + 12	26 + 84	28 - 16	34 - 16	8 x 6	42 / 6
Qu14	86 + 12	45 + 76	42 - 31	37 - 19	11 x 2	56 / 7
Qu15	58 + 21	24 + 77	84 - 53	36 - 17	6 x 7	28 / 4
Qu16	18 + 61	39 + 32	28 - 17	43 - 26	5 x 10	35 / 5
Qu17	72 + 13	59 + 64	23 - 12	56 - 29	6 x 9	56 / 8
Qu18	81 + 13	45 + 66	58 - 13	38 - 29	8 x 6	63 / 9
Qu19	77 + 21	38 + 85	26 - 15	67 - 39	12 x 6	81 / 9
Qu20	17 + 22	43 + 97	45 - 13	42 - 18	9 x 5	64 / 8
Qu21	553 + 415	358 + 267	453 - 311	438 - 289	7 x 12	72 / 8
Qu22	317 + 141	532 + 189	582 - 251	617 - 438	9 x 7	33 / 11
Qu23	533 + 352	196 + 725	574 - 341	733 - 346	7 x 11	44 / 11
Qu24	421 + 353	296 + 254	534 - 213	426 - 177	9 x 12	36 / 6
Qu25	627 + 112	192 + 619	625 - 413	755 - 367	12 x 4	84 / 12
Qu26	713 + 266	272 + 148	364 - 253	851 - 673	9 x 8	96 / 12
Qu27	341 + 328	169 + 381	843 - 612	731 - 484	12 x 7	70 / 10
Qu28	647 + 112	692 + 189	788 - 532	545 - 368	9 x 10	108 / 12
Qu29	345 + 221	364 + 277	845 - 614	643 - 257	9 x 12	132 / 11
Qu30	357 + 341	448 + 482	756 - 315	762 - 167	11 x 8	100 / 10

A.3 Data Clean Up

A.3.1 WMP2

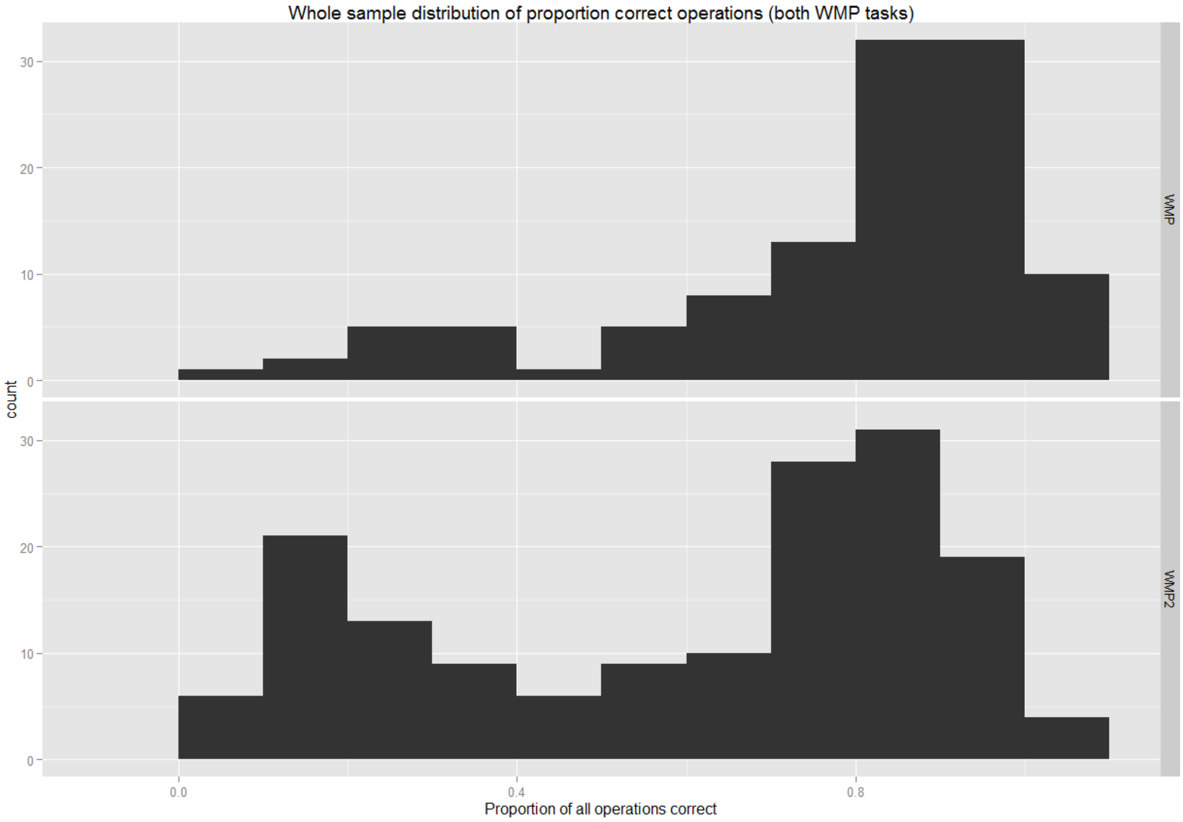


Figure A.1. Distribution of the proportion of correct operations; WMP (top) and WMP2 (bottom)

Specifics regarding removed data points from the WMP2 training data:

- p2 sessions 2, 3
- p5 all 6 sessions (5 were below ops cutoff)
- p6 all 7 sessions (6 were below ops cutoff)
- p7 sessions 2, 3, 4, 5
- p9 all 8 sessions (7 were below ops cutoff)
- p11 all 7 sessions (6 were below cutoff)

- p12 sessions 3, 4, 5
- p19 all 4 sessions (3 were below ops cutoff)
- p25 all 3 sessoins (all below ops cutoff)
- p26 sessions 4, 5
- p28 all sessions (both below cutoff)
- p29 sessions 2, 5

A.3.2 Stroop

- p11 session 7
- p14 session 3
- p17 session 5
- p19 sessions 3,4,6
- p25 sessions 4,5,6
- p28 sessions 3

A.4 Practice Effects

A.4.1 WMP

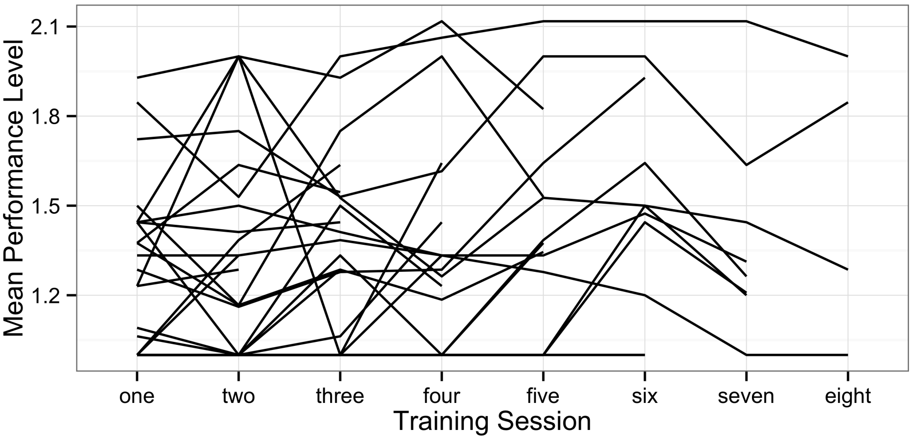


Figure A.2. WMPmean level spaghetti plot

A.4.2 WMP2

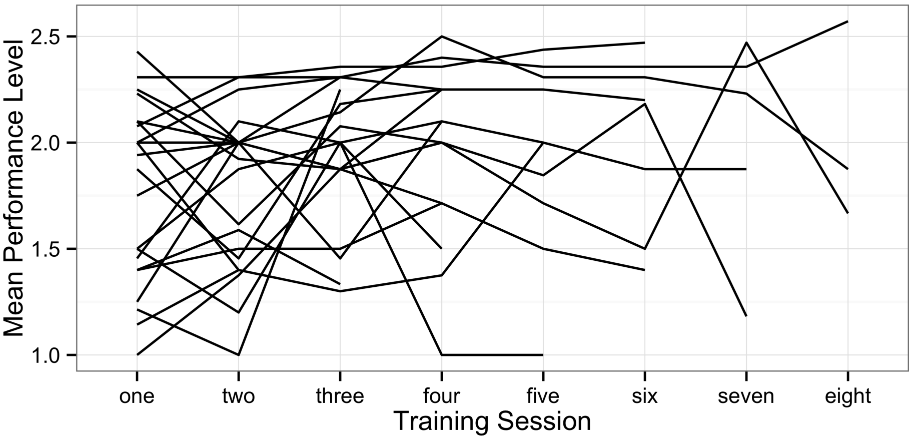


Figure A.3. WMP2 mean level spaghetti plot

A.4.3 Colour Corsi

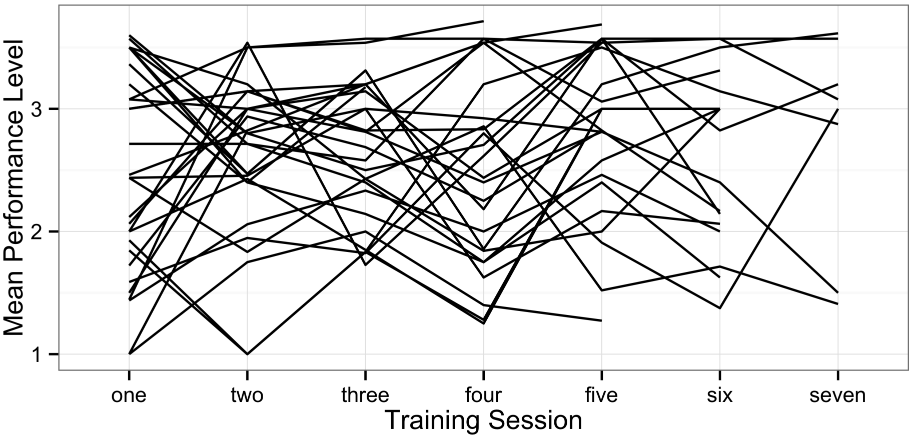


Figure A.4. CC mean level spaghetti plot

A.4.4 Memory Update

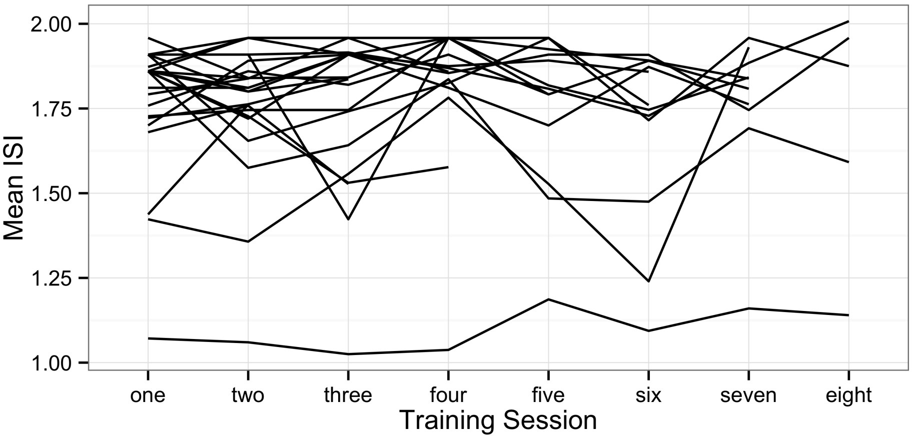


Figure A.5. MU mean ISI spaghetti plot

A.4.5 Stroop

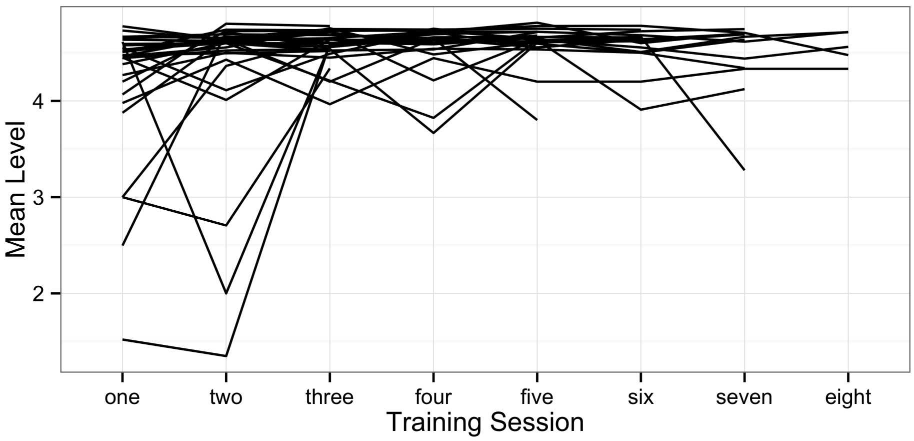


Figure A.6. Stroop mean level spaghetti plot

Appendix B

Working Memory Training Developmental Study Two

B.1 Method

B.1.1 Software Images

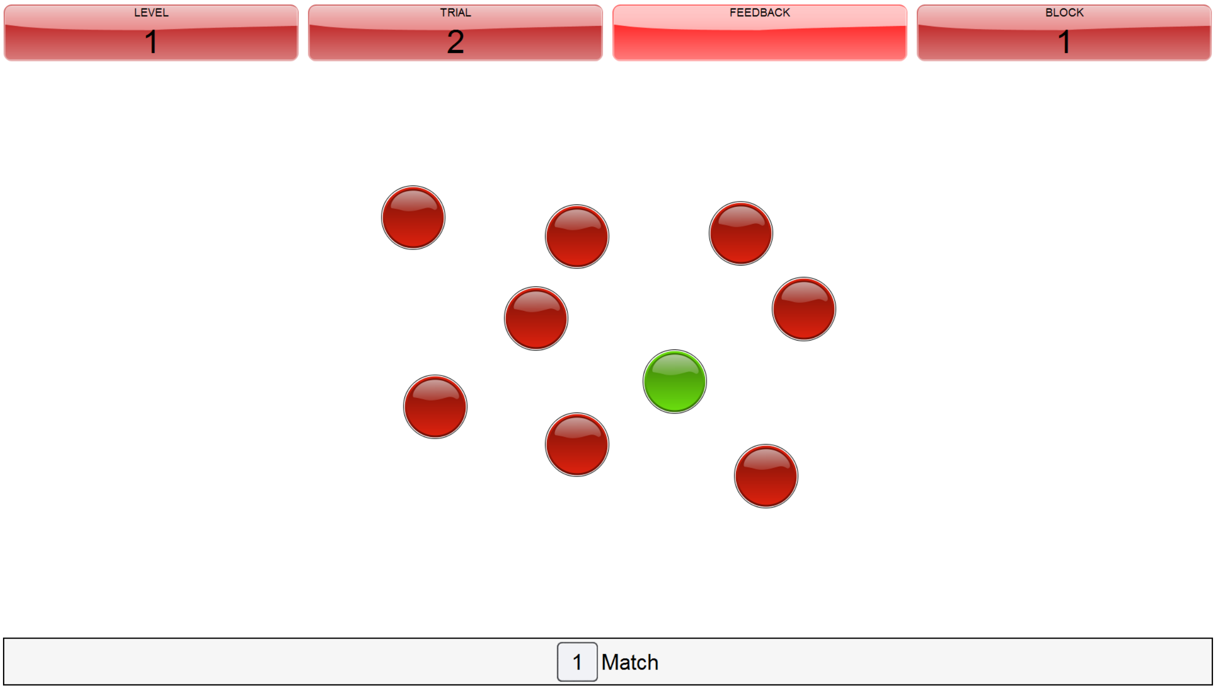
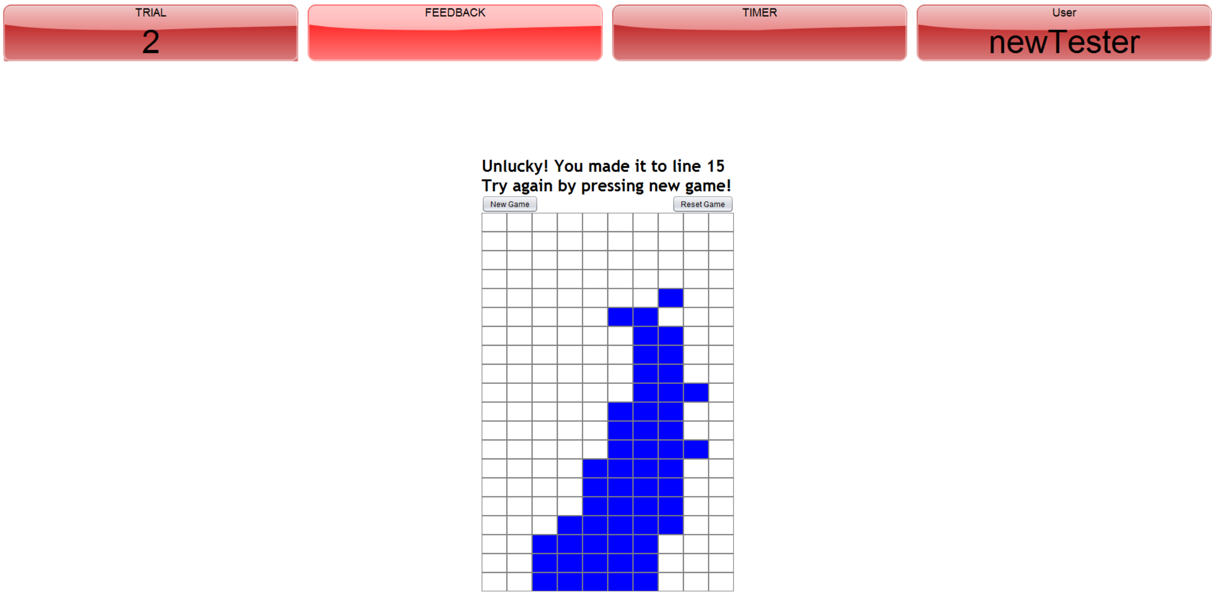


Figure B.1. Screenshot showing the ’look and feel’ of the Spatial N-Back task used in the second developmental training study. This highlights the status panel changes that were altered from that used in the adult study in an effort to make the program look more colourful and appealing to the much younger sample.



Unlucky! You made it to line 15
Try again by pressing new game!

Figure B.2. Screenshot that shows the ’Stacker’ game that was included in the second developmental study as a reward for completing a training session. Participants were given two minutes to play Stacker at the end of each completed session.

B.1.2 Mental Arithmetic Stimuli

Table B.1

Question sets for each block in the Mental Arithmetic task used in the WMT Developmental Two study

	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6
Qu1	5 + 3	5 + 7	3 - 2	11 - 6	4 x 3	4 / 2
Qu2	1 + 2	6 + 7	5 - 4	10 - 4	6 x 1	12 / 4
Qu3	1 + 3	7 + 5	2 - 1	13 - 4	6 x 2	12 / 1
Qu4	6 + 2	9 + 1	7 - 3	22 - 14	2 x 4	9 / 3
Qu5	7 + 1	9 + 5	8 - 5	22 - 13	5 x 2	8 / 4
Qu6	2 + 1	7 + 4	6 - 4	12 - 8	11 x 1	12 / 2
Qu7	6 + 1	4 + 6	4 - 3	14 - 5	3 x 4	20 / 10
Qu8	11 + 8	8 + 5	6 - 2	11 - 8	2 x 9	20 / 5
Qu9	12 + 5	8 + 6	8 - 6	18 - 9	4 x 9	16 / 8
Qu10	10 + 6	9 + 2	7 - 6	23 - 14	5 x 5	15 / 5
Qu11	21 + 6	12 + 19	17 - 5	25 - 17	6 x 4	18 / 3
Qu12	12 + 14	15 + 6	19 - 6	25 - 18	5 x 4	40 / 8
Qu13	14 + 13	23 + 18	19 - 4	33 - 15	8 x 6	42 / 6
Qu14	86 + 12	14 + 19	42 - 31	56 - 37	11 x 2	56 / 7
Qu15	58 + 21	14 + 27	84 - 53	23 - 8	6 x 7	28 / 4
Qu16	18 + 61	39 + 32	28 - 17	34 - 16	5 x 10	35 / 5
Qu17	72 + 13	59 + 64	23 - 12	37 - 19	6 x 9	56 / 8
Qu18	81 + 13	45 + 66	58 - 13	36 - 17	8 x 6	63 / 9
Qu19	77 + 21	38 + 85	26 - 15	67 - 39	12 x 6	81 / 9
Qu20	17 + 22	43 + 97	45 - 13	42 - 18	9 x 5	64 / 8
Qu21	56 + 33	358 + 267	453 - 311	438 - 289	7 x 12	72 / 8
Qu22	28 + 41	532 + 189	582 - 251	617 - 438	9 x 7	33 / 11
Qu23	64 + 12	196 + 725	574 - 341	733 - 346	7 x 11	44 / 11
Qu24	647 + 112	296 + 254	534 - 213	426 - 177	9 x 12	36 / 6
Qu25	345 + 221	192 + 619	625 - 413	755 - 367	12 x 4	84 / 12
Qu26	357 + 341	272 + 148	364 - 253	851 - 673	9 x 8	96 / 12
Qu27	553 + 415	169 + 381	843 - 612	731 - 484	12 x 7	70 / 10
Qu28	317 + 141	692 + 189	788 - 532	545 - 368	9 x 10	108 / 12
Qu29	533 + 352	364 + 277	845 - 614	643 - 257	9 x 12	132 / 11
Qu30	421 + 353	448 + 482	756 - 315	762 - 167	11 x 8	100 / 10

B.2 Practice Effects

B.2.1 WMP

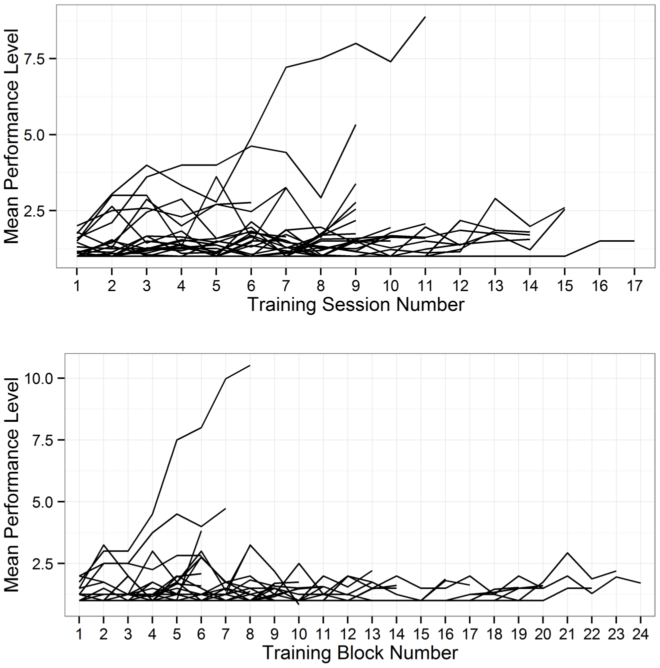


Figure B.3. WMP Training; Top - Performance over all training sessions for each participant as measured by the mean level achieved at each session, Bottom - Split into trial blocks

B.2.2 Colour Corsi

B.2.3 Spatial N-Back

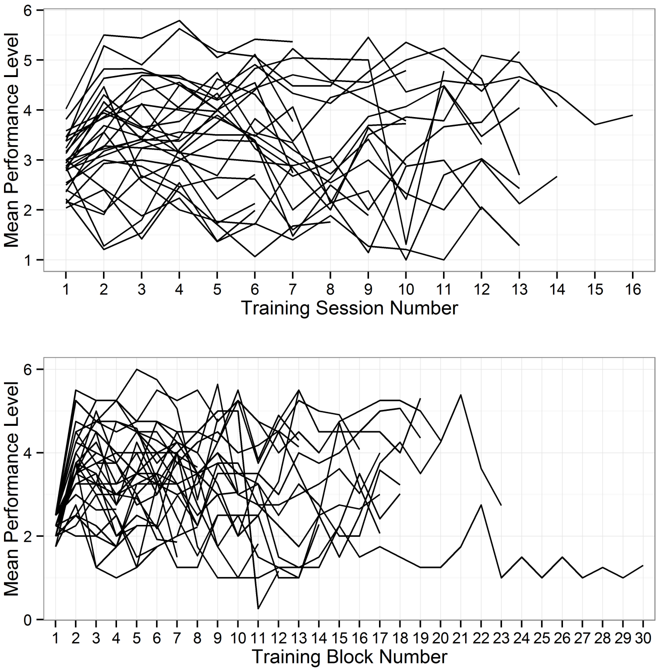


Figure B.4. CC Training; Top - Performance over actual logged training sessions for each participant as measured by the mean level achieved, Bottom - Split into trial blocks

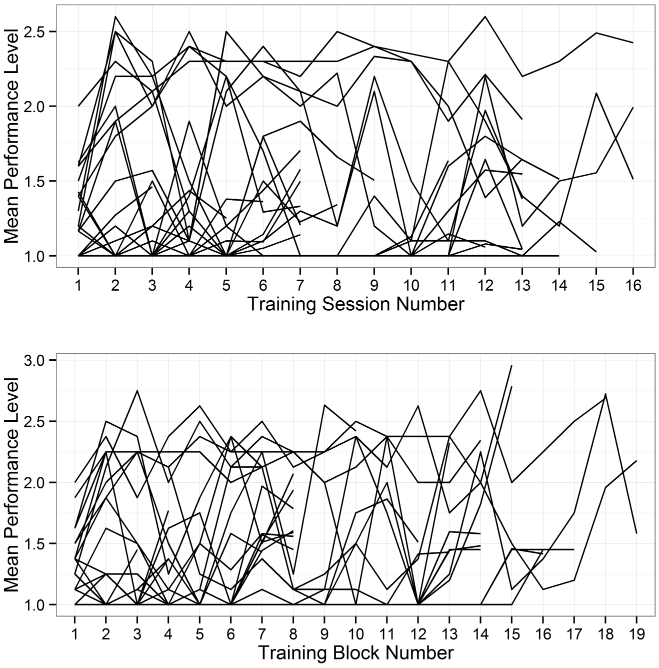


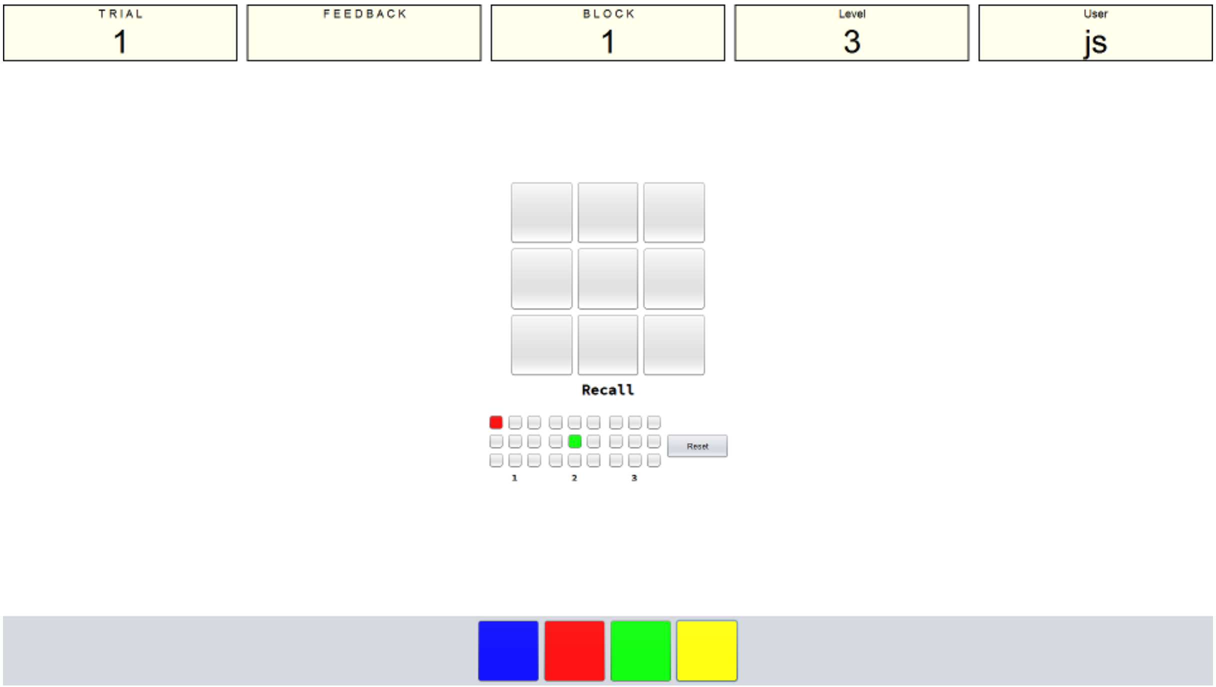
Figure B.5. N-Back Training; Top - Performance over all training sessions for each participant as measured by the mean level (n) achieved at each session, Bottom - Split into trial blocks

Appendix C

Working Memory Training Adult Study

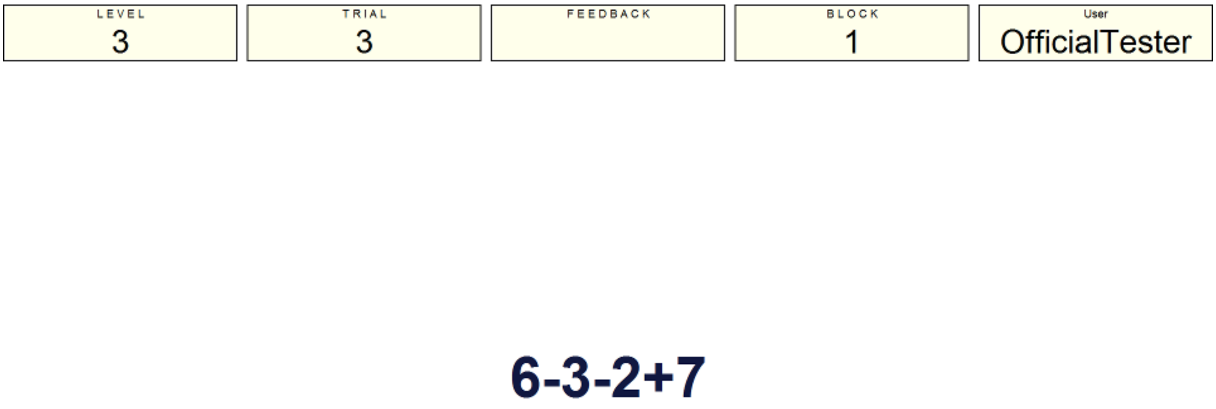
C.1 Method

C.1.1 Software Images



TRIAL	FEEDBACK	BLOCK	Level	User
1		1	3	js

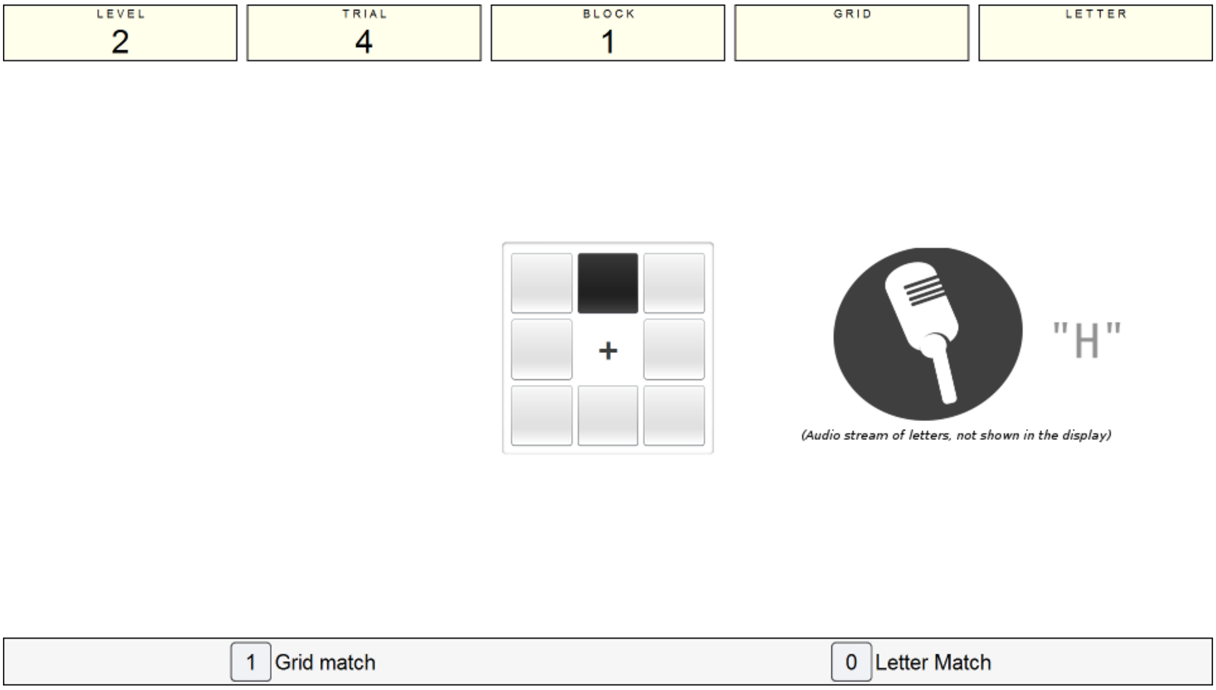
Figure C.1. Screenshot showing the look and feel of the Colour Corsi task as used in the adult WMT study.



LEVEL	TRIAL	FEEDBACK	BLOCK	User
3	3		1	OfficialTester

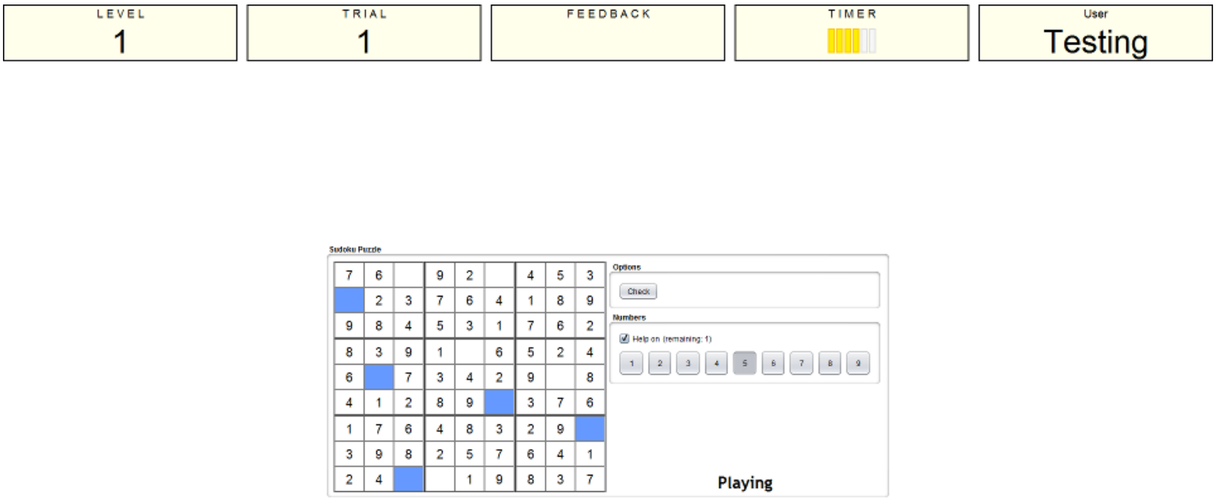
6-3-2+7

Figure C.2. Screenshot showing the look and feel of the Working Memory Period task as used in the adult WMT study.



LEVEL	TRIAL	BLOCK	GRID	LETTER
2	4	1		

Figure C.3. Screenshot showing the look and feel of the dual N-Back task as used in the adult WMT study. The microphone graphic was not visible, it is included in the diagram to show that as grids were presented so were letters in the auditory domain.



LEVEL	TRIAL	FEEDBACK	TIMER	User
1	1			Testing

Sudoku Puzzle

7	6		9	2		4	5	3
🟦	2	3	7	6	4	1	8	9
9	8	4	5	3	1	7	6	2
8	3	9	1		6	5	2	4
6	🟦	7	3	4	2	9		8
4	1	2	8	9	🟦	3	7	6
1	7	6	4	8	3	2	9	🟦
3	9	8	2	5	7	6	4	1
2	4	🟦		1	9	8	3	7

☑ Help on (remaining: 1)

Playing

Figure C.4. Screenshot showing the look and feel of the Sudoku (one example of the Active Control tasks) task as used in the adult WMT study.

C.2 Practice Effects

C.2.1 WMP

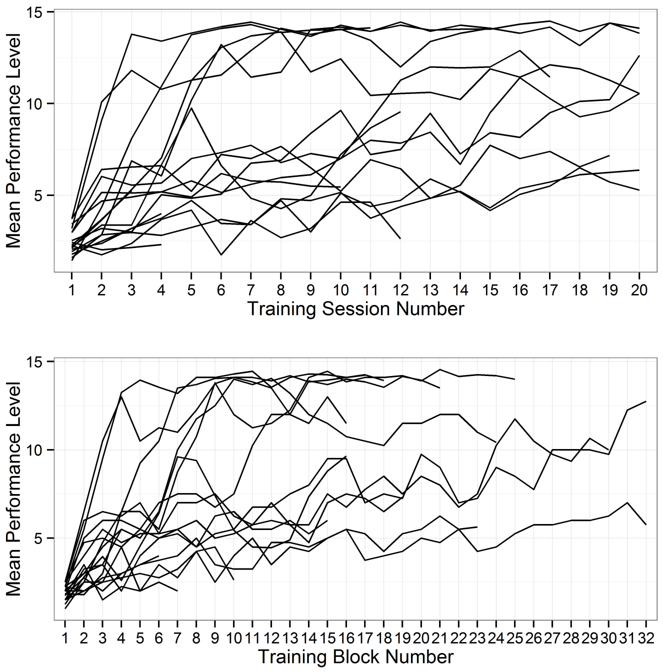


Figure C.5. WMP (Adult Study); Top - Performance aggregate per training 'session' for each participant as measured by the mean level achieved, Bottom - Split into trial blocks

C.2.2 Colour Corsi

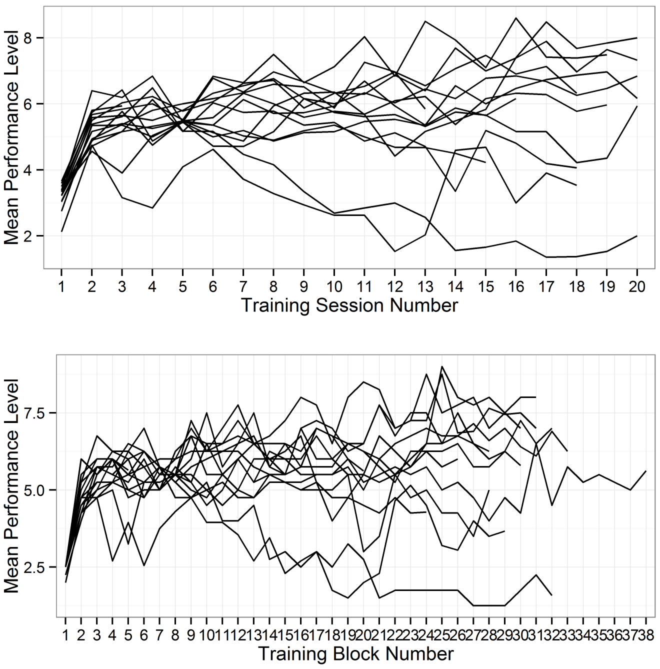


Figure C.6. CC (Adult Study); Top - Performance aggregate per training 'session' for each participant as measured by the mean level achieved, Bottom - Split into trial blocks

C.2.3 Dual N-Back

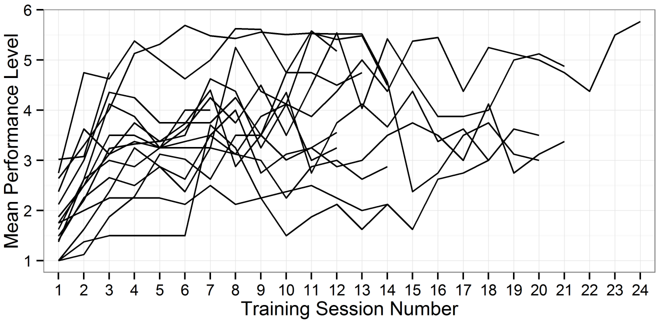


Figure C.7. Dual N-Back Training (Adult Study) - Performance over all training sessions for each participant as measured by the mean level (n) achieved at each session