The Effect of Computerised Adaptive Cognitive Training on Executive Functioning, Fluid Intelligence, and Academic Achievement in Grade 3 and Grade 5 Students: A Randomised Clinical Trial

David Lawrence Hegarty

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
Faculty of Arts and Social Sciences
University of Sydney

2020

Certificate of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.
I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

This trial has been registered with the Australian and New Zealand Clinical Trials Registry (ANZCTR) Trial Id: ACTRN12615000016538 (www.anzctr.org.au).

David Lawrence Hegarty

Acknowledgements

I would like to take this opportunity to acknowledge the people who contributed significantly to my achievement in conducting this research and writing this thesis.

A huge thanks must first go to my primary supervisor Paul Ginns. I am indebted to his kindness, support, patience, and understanding over the eight years it has taken me to complete this thesis. Thanks also to my secondary supervisor Susan Colmar for her feedback on drafts of this thesis.

Thanks to:

- My brother Simon for his help in creating the complex span PEBL tasks. His prodigious coding talent has been of great benefit, not only to me, but for future researchers wanting access to an open source tool for measuring working memory.
- Bruce Wexlar of Yale University for allowing me to use the C8 Sciences cognitive training program to enable a thorough determination of its efficacy.
- The students and staff and the school in which the study was completed. This intervention was a huge commitment for young primary-aged students and their good humour and dedication to completing the assessments and training was appreciated. Teachers are already overloaded with work and their spirit of cooperation to complete behavioural observations above-and-beyond their day to day work was greatly valued.

This thesis would not have eventuated if not for the great hospital staff that assisted me in getting back to full health during my hospital stays for meningitis, pneumonia (with a collapsed lung) and severed fingers, that unfortunately occurred throughout this process. An especially big thanks to the doctor who reattached my fingers on my right hand to help me type again. I am so fortunate to live in a country with free public health and I am indebted to the wonderful nurses and doctors who have kept me alive and fully-functioning.

Finally, an enormous thank you to my family. Thanks to Rosi, Larry, Ros and Kev for helping me in so many ways over the years and enabling me to complete this work. To my dog Miff who is always happy to see me, provides me with great joy, and company when working. To my son Shaun, who provides me with daily smiles and a new perspective on life. It is a privilege watching you grow into an intelligent, considerate and sensitive young boy. I am so proud of you. To Joshua, I will never know you as I would have hoped, but you will forever be in my heart. Lastly, to Anne-Marie. Your patience, courage, support, kindness, and above all, love, have been instrumental in me starting, working on, and completing this thesis. Thank you for all your help with providing advice, your editing work and providing me with sustenance when I was too busy to stop.

Abstract

Adaptive computerised cognitive training programs are increasingly used in an effort to improve the cognitive functioning of children. Although the results of these programs have been mixed, there have also been significant results that hold promise for some form of adaptive computerised cognitive training program, especially around the training of executive processes, also known as executive function (EF). This project set out to test the efficacy of a computerised EF training program whilst addressing many of the previous criticisms of cognitive training research. This study explored whether a school-based, theoretically-aligned EF training program could improve EF as measured by latent factor scores (near-transfer) and, as a result, improve academic achievement, fluid intelligence and academic behaviours (far-transfer). A number of ancillary outcomes were considered, including using an improvement expectancy measure and using a measure of implicit theory of intelligence to determine if it was a potential moderator for training gains.

Normally-functioning primary school children from grades 3 and 5 (Mean age =10 years) were assigned randomly to the EF training program (*n* = 50) or an active control group (*n* = 55). This research program found no evidence for improvement on either near- or far-transfer outcome measures as a result of the EF training program. Although there was some indication of potential moderating effects of the participant's implicit theory of intelligence, the clear lack of linearity and small sample sizes involved in this three-way interaction analysis reduce the generalisability of this finding. Exploratory results showed some weak evidence that those participants in the intervention who improved more on the working memory measures may have a more malleable view of intelligence than those who did not improve. This study provides a unique contribution to the extant literature by rigorously assessing the efficacy of

cognitive training through the use of randomly allocated groups, an active control group, and examining near-transfer using a latent variable approach.

List of Tables

List of Figures

Abbreviations

ADHD	Attention Deficit Hyperactivity Disorder
AIC	Akaike Information Criterion
ANCOVA	Analysis of covariance
BCST	Berg Card Sorting Test
CE	Central executive
CEFI	Comprehensive Executive Function Inventory
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CST	Complex span task
CT	Connections Test
EF	Executive function
fMRI	Functional magnet resonance imaging
gC	Crystallised intelligence
gF	Fluid intelligence
ICSEA	Index of Community Socio-Educational Advantage
IEM	Improvement Expectancy Measure
ITIM	Implicit Theory of Intelligence Measure
LTM	Long-term memory
NAI	Naglieri Ability Index
NAPLAN	National Assessment Program - Literacy and Numeracy
NNAT2	Naglieri Nonverbal Ability Test – Second Edition
OSPAN	Operation span
PAT	Progressive Achievement Test
PEBL	Psychology Experiment Building Language

PFC	Pre-frontal cortex
PPPP	Pirate Pete’s Packing Panic
PS	Processing speed
RMSEA	Root mean squared error of approximation
RSPAN	Reading span
RT	Reaction time
SEM	Structural equation modelling
SRMR	Squared root mean residual
SRT	Simple Reaction Time
SSPAN	Symmetry span
STM	Short-term memory
TLI	Tucker-Lewis index
TMT	Trail Making Test
WCST	Wisconsin Card Sorting Test
WM	Working memory

Table of Contents

1 Introduction

How modifiable are core cognitive processes such as working memory or executive functions at different points over the lifespan? Over the past decade, researchers have tested such questions across a range of interventions for children (see Diamond, 2012 for a review). Adaptive computerised working memory training programs (i.e., where difficulty levels change dependent upon performance) are one form of intervention that have been increasingly used in an effort to improve the cognitive functioning of children (Alloway, Bibile, & Lau, 2013; Beck, Hanson, Puffenberger, Benninger, & Benninger, 2010; Dahlin, 2011; Loosli, Buschkuehl, Perrig, & Jaeggi, 2012; St Clair-Thompson, Stevens, Hunt, & Bolder, 2010). Although meta-analytic results have demonstrated a lack of efficacy for these programs (Aksayli, Sala, & Gobet, 2019; Melby-Lervag & Hulme, 2013), there have also been significant results that hold promise for some form of adaptive computerised cognitive training program, especially around the training of executive processes, also known as executive function (García-Madruga et al., 2013; Gibson et al., 2013; Morrison & Chein, 2011). Given that cognitive training programs have been suggested as an individualised intervention for school psychologists to use with students (Decker, Hale, & Flanagan, 2013) and are actively marketed to school systems (Shipstead, Hicks, et al., 2012), this project endeavoured to test the efficacy of a broader-ranging cognitive training program that addresses many of the previous criticisms of previous cognitive training research.

Many of the training programs to date have focussed upon training of short-term storage capabilities and theoretically, this should not result in improvement in other cognitive capacities (Gibson et al., 2013; Hulme & Melby-Lervag, 2012; Shipstead, Redick, et al., 2012) as it is more complex processing tasks that are related to higher order abilities rather than a link with short-term storage capabilities (Daneman & Carpenter, 1980; Daneman & Merikle, 1996). Therefore, this study aimed to take a more theoretical approach to cognitive training by targeting the training of executive function. As a result of a thorough training paradigm, the transfer of training effects to higher order cognitive abilities was determined and was followed up 1-year post-training. Addressing other methodological issues of previous research relating to the problems with measurement of executive function (Best & Miller, 2010; Miyake, Emerson, & Friedman, 2000), a latent variable approach was used.

This study aimed to test the efficacy of a computerised adaptive cognitive training program by looking at near-transfer to executive function components and far-transfer to higher order cognitive abilities (fluid intelligence and academic achievement). The study addresses previous criticisms of the cognitive training literature by: (a) using a cognitive training program that is more aligned to current executive function theory than previous research, by training executive level processes; (b) using multiple measures for each executive function construct to estimate latent scores of executive functions; (c) using experimental and active control groups that both utilise adaptive computer-based tasks and have the same face-to-face contact time so that a realistic comparison can be made as to the efficacy of the cognitive training program; (d) determining the impact of cognitive training on executive functions, fluid intelligence and academic achievement by following a cohort of children over a period of one year; and (e) assessing the impact of different training expectancies between the experimental and control group and its impact upon training gains. In addition to these student-based measures, behaviourally-based reports from teachers and parents provided additional sources of evidence of the program's effects. Lastly, the potential moderating effect of an individual's implicit theory of intelligence on training gains was assessed.

The following literature review will provide a theoretical overview of the construct of working memory, including the higher-level executive functions that come under the control of the central executive. A review of the research of the links between executive function and academic achievement and intelligence will be provided alongside the development of executive function across childhood. Previous cognitive training research will then be critiqued and implicit theories of intelligence as a potential moderator to the effectiveness of cognitive training will be outlined.

2 Working Memory and Executive Functioning: A Literature Review

2.1 Working Memory

Working memory (WM) is a limited capacity cognitive system that simultaneously stores and actively manipulates information in the course of complex cognitive activities, such as comprehension, learning and reasoning (Baddeley, 2000; Shah & Miyake, 1999). The concept of WM has been around since the early days of modern psychology (James, 1890). Although a number of theoretical models for WM exist, arguably the most influential model for WM was developed by Baddeley and Hitch (1974), which is commonly referred to as Baddeley's WM model.

2.1.1 Baddeley's WM model

Baddeley's (2000) model of WM includes four components: two subsidiary, modality-specific slave systems, the phonological loop and the visuo-spatial sketchpad; the episodic buffer, which provides a link to long-term memory (LTM); and an attentional control system, the central executive (Baddeley, 2002; Baddeley, 2012).

Short-term storage. The phonological loop is a limited-capacity, short-term memory (STM), speech-based store of verbal information (Baddeley, 2012). The phonological loop was proposed to account for the extensive evidence regarding short-term verbal memory by incorporating both a phonological store and an articulatory rehearsal system (Baddeley, 2002). It was deemed a 'loop' because items within the short-term phonological store were shown to decay unless they were kept active, through the use of the articulatory rehearsal system by sub-vocal rehearsal (Gathercole, Adams, & Hitch, 1994). With regard to the educational significance of this component of Baddeley's WM model, it has been argued by Baddeley, Gathercole, and Papagno (1998) that the phonological loop is important for children's acquisition of language.

As visual patterns that are challenging to name must be encoded visually (Dehn, 2008), they must use a different mode of short-term storage within Baddeley's WM model. The STM visual and spatial store that stands in contrast to the verbally based phonological loop is the visuo-spatial sketchpad. The visuo-spatial sketchpad is believed to responsible for temporary maintenance and manipulation of visuo-spatial information, performing an important role in spatial alignment and in the solution of visuo-spatial problems (Baddeley, 2002).

The episodic buffer was proposed to deal with problems the WM model had in dealing with the interaction between the slave systems and LTM (Baddeley, 2003). Its contents are driven both by knowledge stored in LTM and by the contents of the modality-specific temporary stores (Logie, 2011). It is assumed that this is a limited capacity storage system that combines information to form integrated episodes and is controlled by a central executive (Baddeley, 2003).

Theoretically, it is necessary to differentiate between the concepts of STM and WM as this has consequences for the components of a potential cognitive training program (Shipstead, Redick, et al., 2012). Therefore, as Baddeley's model of WM is hierarchical, the central executive is at the top level controlling the aforementioned short-term subcomponents (Dehn, 2008).

Central executive. As outlined by Baddeley (2000), the central executive (CE) acts as a supervisory system controlling the flow of information from and to its slave systems. It mediates the relation between the modality-specific slave systems and retrieval from LTM and is responsible for the regulation and control of cognitive processes. The CE is typically regarded as an attentional system, where the term attention is used to refer to the control processes that operate throughout the WM system (Baddeley & Logie, 1999). As outlined by Shah and Miyake (1999), the terms CE and attention are used almost interchangeably.

The CE has been “the most important but least understood component of WM” (Baddeley, 2003, p. 835). By his own admission, Baddeley’s initial description of the CE was vague in which all the complex cognitive functions that occur when completing even the most simple task could be “stuffed” (Baddeley, 1996, p. 6) and was just a convenient homunculus (Shah & Miyake, 1999). Baddeley and Logie (1999) also acknowledge that the CE became a “ragbag to contain all the phenomena that cannot be readily accounted for otherwise” (p. 39).

In response to these criticisms of the CE, Baddeley (1996) suggested that researchers should consider whether the CE would be better regarded as an individual function or as separable functions. He subsequently suggested that there were separate functions of the CE, which he referred to as dual-task, random generation and selective attention.

Further consideration of the functions of the CE resulted in a number of suggested executive sub-processes such as the capacity to coordinate performance on two separate tasks, to attend selectively to one stimulus and inhibit the disrupting effect of others, to focus and to switch attention, to activate, to hold and to manipulate representations within LTM (Duval, Coyette, & Seron, 2008). Arguably, the most significant suggestion was for the use of a latent variable analysis to determine the functions of the CE (Miyake & Shah, 1999). As a result of Miyake et al.'s (2000) seminal latent variable analysis, what was originally viewed as a single CE involved in controlling attention is now viewed as a range of individual, but related, executive sub-processes that include inhibition, task switching and updating (Friedman & Miyake, 2017; Logie, 2011). These individual components of the CE are outlined in Section 2.2.

2.1.2 Attentional Model of Working Memory

The Baddeley model of WM is not the only one in existence (Dehn, 2008), and an important alternative conception of WM is the attentional view of WM (e.g., Cowan, 2005; Cowan et al., 2005; Engle, 2002; Unsworth & Spillers, 2010). As outlined by Unsworth and Spillers (2010), the attentional view of WM acknowledges that WM is a multidimensional construct, but emphasises that it is attentional control or inhibition that is the salient component in terms of the predictive power of WM. This is referred to as a dual component model of WM as it is the combination of inhibition and memory abilities that results in WM capacity, or the amount of information retained in WM (Unsworth & Spillers, 2010). For example, Cowan (2005) views WM as a system for controlling the focus of attention. Within this perspective, executive attention in WM is thought to activate and maintain memory representations, switch attention between tasks, inhibit irrelevant information and suppress unnecessary response tendencies (Cowan, 2005).

Although competing models of WM are often viewed as a contrast to the Baddeley model of WM (e.g., Dehn, 2008), these theories can also be viewed as reflecting the operation of the executive sub-processes within the CE in the multiple-component model (Logie, 2011). In general, most of the differing models of WM focus upon the executive control aspect of Baddeley's WM model, accepting the contributions of separate visual and verbal STM components (Baddeley, 2012).

2.2 Executive Function

The CE is responsible for the regulation and control of information within the WM system and performing various executive functions (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). The constructs of the CE on the one hand, and executive functions on the other, are very similar (Henry, 2012). Executive function (EF) is essentially what makes WM 'work' (Nee et al., 2012) and WM is a key aspect of EF (Miyake, Friedman, et al., 2000). However, EF is not viewed only as a part of the WM schema, but instead it is considered as a broader umbrella term for higher order cognitive functions that orchestrate cognition (Miyake, Friedman, et al., 2000; Vaughan & Giovanello, 2010). As outlined by Diamond (2013), EF refers to a "family of cognitive processes needed when you have to concentrate and pay attention, when going on automatic or relying on instinct or intuition would be ill-advised, insufficient or impossible" (p. 136). Furthermore, as outlined by Henry (2012), it is helpful to view both the CE and EF as "a broad attentional control space" (p. 22).

EF is a cognitive component that has important social and educational implications (Diamond, 2012; Miyake & Friedman, 2012). Many children who have EF difficulties display poor interpersonal skills and experience difficulties maintaining social relationships (Anderson, 2002) and children's attentional and behavioural problems reflect stable EF difficulties that continue into late adolescence and early adulthood (Friedman et al., 2007; Holmes, Gathercole, Place, Alloway, et al., 2010). As outlined by Friedman et al. (2007), important social and academic behaviours can be predicted by cognitive measures of EF. Furthermore, they argue that "individual differences in EF abilities are important for understanding even normal variation in attention problems, not just clinical extremes" (Friedman et al., 2007, pp. 898–899). In addition, EF abilities may also be protective when it comes to anxiety for some children (White, McDermott, Degnan, Henderson, & Fox, 2011).

It is assumed that much of EF takes place in the pre-frontal cortex (PFC; Kane & Engle, 2002; Miller & Cohen, 2001) through top-down control of behaviour. In contrast to “bottom-up” processing where behaviours are determined by the type of sensory stimuli and ingrained neural pathways that are connected with matching responses, Miller and Cohen (2001) outline that the PFC is important when behaviour must be directed by internal states or intentions. A review by Bressler and Menon (2010) that includes functional magnetic resonance imaging (fMRI) data confirms EF activity in the PFC network. In contrast, a meta-analysis of fMRI data demonstrated that EF is not as localised to the PFC as previously suggested (Nee et al., 2012). However, as argued by Huizinga, Dolan, and van der Molen (2006), these differences may be due to different components of EF relying on different parts of the PFC (e.g., Aron, Robbins, & Poldrack, 2004; Crone, Wendelken, Donohue, & Bunge, 2006; Narayanan et al., 2005). Subsequent research suggests that it is the frontoparietal control network (FPCN) that coordinates activity across other cortical and subcortical areas for EF (Dixon et al., 2018;), but the PFC still appears to play an important role (Engelhardt, Harden, Tucker-Drob, & Church, 2019; Friedman & Miyake, 2017).

According to Henry (2012), there is growing consensus that EF is fractionated into several areas that are nevertheless linked together. However, a significant issue in the research literature is the use of different terminology when referring to cognitive processes that are applicable to EF. As outlined by Miyake, Emerson and Friedman (2000), EF is a challenging topic to study as it is difficult to define and to measure. Researchers have referred to EF as cognitive control (Diamond, 2013; Kray & Ferdinand, 2013; van Muijden, Band, & Hommel, 2012; Weintraub et al., 2013), effortful control (Zhou, Chen, & Main, 2012), executive control (Diamond, 2013), self-control (Moffitt et al., 2011), or self-regulation (Anderson, 2002). Although some of these items, (e.g., self-control), may only apply to one aspect of EF, as will be outlined in Sections

2.2.1 – 2.2.3, many of these aforementioned aspects are included under the umbrella term of EF (Diamond, 2013).

As alluded to earlier, one of the most influential models that has attempted to describe EF is the Miyake et al. (2000) model, which has been derived empirically, firstly in adult (Miyake, Friedman, et al., 2000) and subsequently in child populations (Lehto, Juujarvi, Kooistra, & Pulkkinen, 2003; Rose, Feldman, & Jankowski, 2011; Wu et al., 2011). Using a latent variable approach, Miyake et al. (2000) found that EF consists of three separable but related components:

1. Deliberately suppressing dominant or automatic responses (inhibition).
2. Shifting or switching between different tasks or mental sets (shifting).
3. Maintaining and monitoring working memory representations (updating).

Although there are other potential EF (e.g., dual-tasking and planning), these are typically considered to be higher-level constructs that implicate the three EF previously identified (Miyake & Friedman, 2017) and those examined by Miyake et al. (2000) have largely dominated the literature (Diamond, 2012, 2013; Friedman et al., 2006; Miyake & Friedman, 2012). However, it should be noted that there is still some disagreement concerning whether EF is a unitary or multidimensional construct (Best & Miller, 2010; Karr et al., 2018).

It appears that individual differences in EF functioning is relatively stable throughout a child's development (Miyake & Friedman, 2012) and it has been suggested that EF is highly heritable and almost entirely genetic in origin (Friedman et al., 2008; Miyake & Friedman, 2017). EF components appear to be distinguishable and separable in early childhood (Karr et al., 2018; Senn, Espy, & Kaufmann, 2004) but the independence of individual EFs may change developmentally (Best & Miller, 2010; Karr et al., 2018). The development of EF across childhood will be covered in more detail in Section 2.6.

Miyake and Friedman (2012) posit that individual differences in EF show both unity and diversity. Their research has demonstrated that “different EFs correlate with one another, thus tapping some common underlying ability (unity), but they also show some separability (diversity)” (p. 9). Later research has confirmed this “separable but related” perspective on EF (e.g., Cirino et al., 2018; Karr et al., 2018). Miyake et al. (2000) argue that the maintenance of goal and context information in WM is a common feature in all EF tasks and it has been argued that WM is a key component of all EF measures (Best & Miller, 2010).

As Miyake et al.'s (2000) three-factor model encompassing inhibition, shifting and updating has been replicated in both children and adults (Lehto et al., 2003; Rose et al., 2011; Wu et al., 2011), this is the model that was used for this study and will be expanded upon below.

2.2.1 Inhibition

Inhibition, which is sometimes referred to as selective attention (Fournier-Vicente, Larigauderie, & Gaonac’h, 2008), or inhibitory control (Diamond, 2013), refers to the central, active suppression of information that is irrelevant to the task at hand (Johnson, Im-Bolter, & Pascual-Leone, 2003). As noted by Diamond (2013), a distinction is required between a behavioural inhibition component (self-control) and a cognitive inhibition component (interference control). With regard to the cognitive inhibition distinction, Fournier-Vicente et al. (2008) distinguish between inhibition that refers to the ability to actively supress previously retrieved information while performing a deliberate search and retrieval from LTM and inhibition that refers to the ability to selectively attend to specific targets in the presence of one or more irrelevant distractors, referred to as selective attention. This study focussed upon the cognitive component of inhibition and more specifically, the ability to selectively attend in the face of irrelevant distractors.

2.2.2 Shifting

Shifting is also referred to as cognitive or mental flexibility (Diamond, 2013; Yeniad, Malda, Mesman, van Ijzendoorn, & Pieper, 2013), mental set shifting (Diamond, 2013; Rushworth, Passingham, & Nobre, 2002) , task-switching (Davidson, Amso, Anderson, & Diamond, 2006) or task-shifting (Karbach & Kray, 2009). Shifting refers to changing the mental set that has been learned to a new one (Yeniad et al., 2013). Irrespective of their particular form, all shifting tasks are comprised of two stages that heavily involve other EFs (Garon, Bryson, & Smith, 2008). In the first stage of a shifting task participants create a mental set where an association is made between a particular stimulus and a response (Garon et al., 2008; Yeniad et al., 2013). In forming this set, other EF components are required as participants must focus on the relevant stimuli whilst simultaneously inhibiting distractors and holding the mental set in WM. In the next stage of a shifting task the participant must shift to a new mental set that is different to the first mental set in some way (Garon et al., 2008).

Brain imaging data outlines two types of shifting abilities: attention shifting and intention shifting (Rushworth et al., 2002). Attentional shifting requires an individual to change the rules by which they select between sensory stimuli, whereas intentional shifting requires an individual to change the rules by which they select between motor responses (Rushworth et al., 2002). Garon et al. (2008) refer to this as a distinction between tasks that require a shift in the way a stimulus is perceived as opposed to a shift in response. The present study focusses upon the shift in response.

2.2.3 Updating / WM

Baddeley's model of WM has been explained previously (see Section 2.1.1). Although Miyake et al. (2000) outline that the third EF is referred to as updating as opposed to WM, updating has been interpreted by many authors as the same as WM (Garon et al., 2008). Research by St Clair-Thompson and Gathercole (2006) found that measures of WM share a common association with updating skills in children. Research by McAuley and White (2011) demonstrated that a three-factor model representing processing speed, inhibition and working memory for individuals ranging from 6 to 24 years of age provided a better fit than a four-factor model that separated WM into storage and updating abilities. Furthermore, Martinez et al. (2011) found that WM capacity and updating correlated highly with each other (r = .85) in undergraduate students. In contrast, using a latent variable analysis, Schmiedek, Hildebrandt, Lovden, Wilhelm, and Lindenberger (2009) found complex span measures of WM and updating were not statistically discriminable from each other (r = .96). This apparent similarity between the type of measure of WM and updating is particularly salient as this study used complex span measures of WM as outlined below.

In a cross-sectional study measuring WM in 5-80 year-olds, Alloway and Alloway (2013) found no functional differences within WM (i.e., no distinction between storage and updating). Their findings suggest a functional difference within WM may be due to task choice (Alloway & Alloway, 2013), especially when some confusion may arise between measuring STM or WM, rather than any real functional differences within WM. Given that updating tasks require “constant monitoring and rapid addition/deletion of working memory contents” (Miyake & Friedman, 2012, p. 9), this appears to be synonymous with most WM tasks.

The most recent research on discriminating between WM and updating by Redick et al. (2016) found that although a two-factor model provided a better fit in a confirmatory factor analysis (CFA), there was a high correlation between WM factors (*r* = 0.84). Their conclusion was that, "although statistically the best fit is achieved with the model with two separate WM factors, the high amount of shared variance indicates all of the WM measures tap primarily overlapping processes, with relatively little method-specific variance distinguishing complex span and noncomplex span tasks" (p. 1481).

Therefore, as a result of the above-mentioned findings, the present study took the view that WM and updating are synonymous with each other and any distinction between the two is purely due to either task choice or a misunderstanding with regards to WM and the mistaken use of STM tasks for its measurement. From this point on, the third EF component will simply be referred to as WM.

There is one more aspect of WM, specifically with regard to its role as an EF, which needs to be clarified: the domain-general or domain-specific nature of WM. There has been long-term disagreement regarding whether the executive component of WM is domain-general or domain-specific (Miyake, 2001). The present study specifically considers the executive components of the CE within the WM system. As clearly explained by Baddeley (2000, 2002; 2012), the CE acts as a supervisory system that controls and regulates cognitive processes and, as outlined by Baddeley and Logie (1999), the CE component of WM is typically regarded as an attentional system.

Seminal research by Kane et al. (2004), which used structural equation modelling (SEM) and complex span tasks (CSTs), demonstrated that the executive component of WM is domain-general. It is this researcher's view that WM is the simultaneous storage and processing of information that is measured by CSTs (e.g., Broadway & Engle, 2010; Redick, Broadway, et al., 2012; Unsworth, Heitz, Schrock, &

Engle, 2005; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). CSTs do just what is referred to in Baddeley's (2000) definition of WM: they require an individual to simultaneously store and actively manipulate information in the course of complex cognitive activities, such as comprehension. Not only have many researchers acknowledged that CSTs are the most valid and reliable measures of WM (Conway et al., 2005; Kane, Conway, & Engle, 1999; Miyake, 2001), but it is the domain-general mechanism of WM assessed by CSTs that is most predictive of complex cognition, e.g., fluid intelligence and academic achievement (Hitch, Towse, & Hutton, 2001; Kane et al., 2004; Pardo-Vazquez & Fernandez-Rey, 2012; Swanson, 2011; Unsworth, Redick, et al., 2009). Therefore, it is important to note that this study used a domain-general view of WM when referring to WM as one of the core EFs.

2.3 EF and Academic Achievement

EF is important for just about every aspect of life (Diamond, 2013). Importantly for children, individual differences in EF are among the best cognitive predictors of academic achievement (Walton & Dweck, 2009) and all EF components are required for the successful navigation of the demands of the classroom (McClelland & Cameron, 2011, 2012). For example, a longitudinal study by Jacobson, Williford, and Pianta (2011) found that EF abilities that were assessed prior to starting school significantly predicted sixth grade competence both academically and socially, as rated by both teachers and parents. Interestingly, children’s academic competence was accounted for by EF with almost as much variance as socioeconomic status, a known correlate of academic success (Jacobson et al., 2011; Sirin, 2005). Research by Rosas et al. (2017) also demonstrated EF abilities in kindergarten-aged children were able to predict later achievement in mathematics and reading. Research based upon SEM by Nayfeld, Fuccillo, and Greenfield (2013) found that EF's capacity to predict achievement in mathematics and literacy could be extended to science achievement.

In another longitudinal study, which followed a sample of girls for five years, Miller and Hinshaw (2010) found that childhood EF abilities predicted academic achievement and social functioning (peer acceptance and social skills) at adolescence. Importantly, the predictive nature of EF abilities at childhood was valid for all children, regardless of whether a child had an Attention Deficit Hyperactivity Disorder (ADHD) diagnosis or was typically developing and even when controlling for intelligence. Furthermore, when this cohort of girls was followed for a further five years it was found that childhood EF measures were still predictive of academic functioning and were also predictive of employment status, again independent of diagnostic group status or intelligence (Miller, Nevado-Montenegro, & Hinshaw, 2012). Other longitudinal studies have found similar associations between early EF abilities and later academic achievement (McClelland, Acock, Piccinin, Rhea, & Stallings, 2013), even after individual differences in cognitive ability and achievement in reading were accounted for (Clark, Pritchard, & Woodward, 2010). Shorter-term studies (Locascio, Mahone, Eason, & Cutting, 2010; Roebers, Cimeli, Rothlisberger, & Neuenschwander, 2012; Swanson, 2011) and a meta-analysis (Booth, Boyle, & Kelly, 2010) have also found similar relationships between EF and academic achievement. Interestingly, it has been found that the correlation between EF and academic achievement changes across development and is strongest at approximately 8 to 9 years of age (Best, Miller, & Naglieri, 2011).

Just how this relationship between EF and academic achievement might manifest itself has been suggested by Blair and Razza (2007) with regard to mathematics problem solving. They outline the potential role of each EF in the problem solving process, such as working memory when considering the need to represent information for the problem, shifting when it is necessary to change which component of the problem is being attended to, and inhibition when it is necessary to attend to a part of the problem in the face of irrelevant distractors or a more recent aspect of the problem (Blair & Razza, 2007). Research by Kotsopoulos and Lee (2012) demonstrated the requirements of all EF factors as suggested above for mathematics problem solving in situ. In contrast, LeFevre et al. (2013) suggest that EF is related to mathematical abilities with regard to fluency (latencies in simple mathematical problem-solving) as opposed to knowledge (application of procedures).

It is important to note that the causal relationship between EF and academic performance is not clearly understood (Best, Miller, & Jones, 2009). Best et al. (2009, p. 193) outline a number of hypotheses for this relationship: (a) EF directly affects academic performance; (b) EF affects performance through language skills or reasoning ability; or (c) behaviours in the classroom could mediate the relationship between EF and scores on academic tests. Furthermore, they outline that school may promote EF development through students engaging in everyday classroom activities that necessitate EF practice (Best et al., 2009).

Although some researchers have argued that no specific EF component predicts later academic achievement (Bull, Espy, & Wiebe, 2008; Kotsopoulos & Lee, 2012; Miller & Hinshaw, 2010; Roebers et al., 2012), others suggest that individual EF components rather than EF as a whole have more influence on academic achievement (e.g., Booth et al., 2010) or that individual EF components have varied impacts dependent upon the academic domain (Latzman, Elkovitch, Young, & Clark, 2010). Before outlining the research that highlights the links between individual EF components and academic achievement, it is important to note that the implication of one EF component but not others, typically depends upon the nature of the problem-solving exercise (Kotsopoulos & Lee, 2012) or the EF task being used (e.g., Booth et al., 2010). This last point regarding the type of EF task being used refers to the task-impurity problem for EF measures, which is particularly salient for EF research and is outlined in Section 2.7.

2.3.1 Inhibition and academic achievement

There have been contradictory results with regard to inhibition's relationship with academic achievement. For example, Montoya et al. (2019) found that inhibition predicted achievement in early numeracy and literacy skill outcomes in pre-school-aged children. Similarly, Blair and Razza (2007) found that inhibition was related to both reading and mathematics abilities in kindergarten-aged children and St Clair-Thompson and Gathercole (2006) found that inhibition was related to achievement in English, mathematics and science in 11- and 12-year-old children. The link between inhibition and mathematics and science achievement has also been found in older boys and adolescents (Latzman et al., 2010) and other research has found that inhibition is related to measures of intelligence (Schweizer, Moosbrugger, & Goldhammer, 2005). Neubauer, Gawrilow, and Hasselborn (2012) found that preschool-aged children who performed better on an inhibition task achieved better academic results and had less behavioural problems, as rated by a teacher, at the end of grade 1, even after controlling for intelligence, gender and age. However, it should be noted that this was a behavioural inhibition task that would be more closely aligned with the concept of self-control, as outlined previously. In contrast to the aforementioned findings, however, there has been research that has found that inhibition is not related to academic achievement in children, whereas other EF factors are (van der Ven, Kroesbergen, Boom, & Leseman, 2012).

As mentioned previously, the EF task used has an important impact upon the relationship between a single EF component and academic achievement. Research by Lan, Legare, Ponitz, Li, and Morrison (2011) highlights this fact. In a cross-cultural analysis of the link between EF and academic achievement, they found that inhibition predicted mathematics achievement. Importantly however, this link was no longer evident when WM was taken into account (Lan et al., 2011), as the task used to measure inhibition heavily depended upon WM. This was further complicated by the fact that only one measure was used for inhibition, again referring to the task-impurity problem, which will be addressed in Section 2.7.

2.3.2 Shifting and academic achievement

The link between shifting and academic achievement seems to be somewhat more consistent. There has been a significant positive relationship found between shifting ability and achievement in both English and mathematics in preschool-aged children (Bull et al., 2008; Clark et al., 2010), kindergarten aged children (Coldren, 2013), and older children (Agostino, Johnson, & Pascual-Leone, 2010; Latzman et al., 2010; van der Sluis, de Jong, & van der Leij, 2007). A meta-analysis found that shifting ability was significantly associated with performance in both mathematics and reading (Yeniad et al., 2013). However, Yeniad et al. (2013) also found that intelligence was more strongly associated with academic performance than shifting ability, and that shifting was strongly related to intelligence.

2.3.3 WM and academic achievement

Arguably, the greatest amount of research has focused on WM and its relationship with academic achievement. Given that the executive component of WM and attention are closely related to one another (Baddeley & Logie, 1999; Miyake &

Shah, 1999), it is not surprising that WM may have a significant impact upon academic achievement.

Not only do children with developmental difficulties, such as ADHD, appear to have difficulties with WM (Holmes, Gathercole, & Dunning, 2010; Kasper, Alderson, & Hudec, 2012; Martinussen, Hayden, Hogg-Johnson, & Tannock, 2005), but WM is an excellent predictor of academic achievement in typically developing children (Bull et al., 2008; Gathercole & Brown, 2003; Gathercole, Pickering, Knight, & Stegmann, 2004; Raghubar, Barnes, & Hecht, 2010; Rose et al., 2011; Toll, Van der Ven, Kroesbergen, & Van Luit, 2011).

In the early days of WM research, Daneman and Carpenter (1980) found that complex measures of WM (simultaneous processing and storage) were good predictors of reading comprehension abilities, considerably better than short-term capacity measures (simple span measures). A later meta-analysis by Daneman and Merikle (1996) confirmed this relationship and suggested that complex processing and storage tasks are predictive of mathematics achievement as well. Much of the WM research since then has attempted to disentangle the relationship between WM and achievement in both English and mathematics.

In kindergarten-aged children, WM is related to language abilities with more complex WM measures being related to syntax and comprehension and short-term phonological stores related to vocabulary (de Abreu, Gathercole, & Martin, 2011). Longitudinal research demonstrated that WM abilities for children who had just entered school were predictive of children's subsequent levels of achievement at 7 years of age in English, but not in mathematics (Gathercole & Brown, 2003). Other longitudinal research has also found that WM ability at four and a half years of age was predictive of academic achievement at 15 years of age after controlling for early academic achievement, demographic and home environment variables (Ahmed, Tang, Waters, & Davis-Kean, 2018). Alloway, Gathercole, Kirkwood, and Elliott (2009) found that children 5 to 11 years of age who were identified as having low WM, struggled in both mathematics and English achievement and also received worse behavioural reports from their teachers. Similarly, WM abilities in children 6 to 11 years of age were found to be related to both English and mathematics ability (Gathercole, Alloway, Willis, & Adams, 2006). This was found to be independent of intelligence, verbal abilities, STM or phonological awareness skills (Gathercole et al., 2006).

This relationship between WM and both English and mathematics achievement was also found for children 7 years of age (Gathercole, Pickering, Knight, et al., 2004). However, although this relationship was still evident between WM and mathematics for children 14 years of age, the relationship no longer existed for level of English achievement (Gathercole, Pickering, Knight, et al., 2004). Gathercole et al. (2004) believe that these results substantiate their longitudinal results (Gathercole & Brown, 2003) in that WM is more important for the acquisition of literacy skills (Baddeley et al., 1998), but less so for higher-level conceptual and analytic abilities that are tapped by English assessments at later ages. Given research demonstrating that individual differences in WM predicted comprehension beyond word-attack skills (Swanson & O'Connor, 2009) and a meta-analysis which demonstrated that problems with reading comprehension in both children and adults were strongly related to complex WM tasks (Carretti, Borella, Cornoldi, & De Beni, 2009) it appears that both the academic ability test and WM task used are important factors related to the strength of association between WM and English achievement. This influence of ability test and task choice is similar to problems affecting the relationship between other EF components and academic achievement.

Research has also demonstrated a complex relationship between WM and mathematics achievement. In younger children, poor WM is related to weaker mathematical skills both prior to and after school entry (Preßler, Krajewski, & Hasselhorn, 2013). Using a latent variable approach in a longitudinal study, Toll et al. (2011) found that WM was a good predictor of later mathematical abilities in kindergarten children, even above that of kindergarten mathematical performance. Other longitudinal research has also found that WM at preschool age is predictive of mathematical achievement three years later (Bull et al., 2008) and has also demonstrated a link between mathematical learning difficulties and poor WM (Geary, 2011). When comparing WM to other EF components, research has also shown the link between WM and mathematical development in both children (Kolkman, Hoijtink, Kroesbergen, & Leseman, 2013; Lee et al., 2012; Swanson, 2011; van der Ven et al., 2012) and adolescents (Agostino et al., 2010) above that of either shifting or inhibition. However, it should be noted that Senn et al. (2004) found a more nuanced relationship. They found that inhibition was a better predictor of problem solving than WM in three-year-old children; however, this relationship was reversed in six-year-old children (Senn et al., 2004).

In a review of experimental, cross-sectional and longitudinal research, Raghubar et al. (2010) concluded that WM is related to a variety of mathematical outcomes even when other cognitive and academic factors are taken into account. Subsequently, a meta-analysis by David (2012) found that significant difficulties with mathematics are attributable to WM difficulties.

It appears that the relationship between WM and academic achievement may differ across cultures (Lan et al., 2011). In a cross-cultural study by Lan et al. (2011), WM was strongly related to both reading and mathematics in Chinese students, whereas for students in the USA, WM was more strongly related to mathematics than to reading. Similar to the relationship between language abilities and WM, there appears to be some differences between mathematical achievement and WM dependent upon the type of task determining achievement and the WM task that is being used. As suggested by Raghubar et al. (2010), the relationship may also have some dependency upon what level of mathematics ability is being predicted.

In addition to WM's role in the storage and manipulation of information in the performance of complex cognitive activities, it also plays a crucial role in classroom situations, both in supporting learning and in maintaining focused behaviour (Gathercole, Durling, Evans, Jeffcock, & Stone, 2008). As outlined by Shipstead, Redick, and Engle (2010), WM capacity is crucial to academic learning as it "predicts the ability to engage in appropriate behaviour at the appropriate time in real-world tasks" (p. 246). A review by Hambrick and Meinz (2011) found that WM capacity predicts performance in complex tasks (e.g., playing piano, poker, chess) even in adults with high levels of domain-specific experience and knowledge. This finding is particularly salient, as they found that WM capacity was a positive predictor of performance above and beyond deliberate practice.

In summary, it appears that the link between EF and academic achievement seems to be strong. However, there is relatively little research identifying which individual EF component is more important for domain-specific achievement. Rose et al. (2011) found that WM is the EF most pivotal to academic achievement when comparing prematurely-born to full-term born children when assessed as adolescents. Furthermore, they found that the relationship of both shifting and inhibition to academic achievement disappeared once WM was taken into account. As a result, Rose et al. (2011) suggest that inhibition and shifting are related to achievement only through the variance they share with WM. As mentioned previously, this touches on the important aspect of the task impurity problem in all EF tasks, including WM tasks, which will be addressed in Section 2.7.

The present study endeavoured to determine which of the core EFs are responsible for resultant improvement in two foundational areas of academic achievement, English and mathematics, as a result of cognitive training. By using multiple tasks to measure each EF construct, and because of the experimental nature of the present study, potential causal inferences can be made.

2.4 EF and Intelligence

Although it is clear that there is a strong relationship between general intelligence, or g, and school achievement, with correlations ranging between .77 to .94 (Calvin, Fernandes, Smith, Visscher, & Deary, 2010; Kaufman, Reynolds, Liu, Kaufman, & McGrew, 2012), EF also explains a similar amount of variance in both intelligence and academic achievement (Alloway & Alloway, 2010; Lu, Weber, Spinath, & Shi, 2011). Some researchers suggest that it is EF that is responsible for abilities in g, particularly in relation to fluid intelligence (Duggan & Garcia, 2015; Engle, 2002; Unsworth, Miller, et al., 2009; Uka, Gunzenhauser, Larsen, & von Suchodoletz, 2019; Unsworth & Spillers, 2010).

The distinction between crystallised intelligence and fluid intelligence is commonly made by researchers (Nisbett et al., 2012). As outlined by Nisbett et al. (2012), crystallised intelligence (gC) is an “individual's store of knowledge about the nature of the world and learned operations… which can be drawn on in solving problems” (pp. 131-132), whereas fluid intelligence (gF) “is the ability to solve novel problems that depend relatively little on stored knowledge as well as the ability to learn" (p. 132). Both gC and gF correspond to broad abilities outlined in the Cattell-Horn-Carroll theory of intelligence (McGrew, 2009), an influential theory on the structure of human intelligence (Schneider & McGrew, 2018).

EF is related to, yet distinct from intelligence (Aarnoudse-Moens, Smidts, Oosterlaan, Duivenvoorden, & Weisglas-Kuperus, 2009), but it has been suggested that not all EF components are related to g and that it is WM that plays the most important role (Friedman et al., 2006). Some research has shown that gF in particular is strongly correlated with WM (Jastrzębski, Ciechanowska, & Chuderski, 2018; Martinez et al., 2011), whereas other research has demonstrated a strong relationship between gF and both inhibition and WM (Uka et al., 2019). SEM research by Unsworth and Spillers (2010) found that the relationship between WM and gF reflects variation in attentional control ability (active maintenance of information in WM), variation in the scope of attention (size or capacity of WM) and variation in the ability to retrieve and strategically search information from LTM. Similarly, Engle (2002) argued that WM is related to gF due to the attentional nature of WM. This was supported by experimental research by Chuderski and Necka (2012) that found both attention control (which used a measure of inhibition) and the scope of attention contributed to achievement in gF. However, this could be due to the different use of measurement tasks (complex span vs. lower n versions of the n-back task). Nevertheless, other longitudinal research by Engel de Abreu, Conway and Gathercole (2010) has demonstrated that it appears to be the executive component of WM, rather than the STM components, that is significant regarding the link between WM and gF.

In contrast to research demonstrating the importance of only WM and its link to gF, Arffa (2007) found that a variety of EF measures were significantly correlated with a full-scale intelligence score for children aged 6 to 14 years. Similarly, van der Sluis et al.

(2007), although finding that WM was related to gC, also found that shifting and updating abilities were correlated with both gC and gF. A later study by Brydges, Reid, Fox and Anderson (2012) found that EF was related to both gF and gC in children aged 7 to 9 years of age. Brydges et al. (2012) found that the relationship between EF and both gF and gC were quite high ($r = .89$ and $r = .83$ respectively). They also found that a three-factor model, where EF, gF and gC were all separate, provided the best fit in a CFA, and SEM found that the commonality between gF and gC was completely predicted by EF (Brydges et al., 2012).

A longitudinal study by Richland and Burchinal (2013) found that EF at school entry, as determined by both a composite EF measure and a separate inhibition component, predicted gC ability in adolescence, even when taking other covariates such as parental education and vocabulary into account. Interestingly, they only used one task to determine composite EF and one task to determine inhibition – referring to the common issue regarding construct measurement, where discrepancies between studies may be related to the differing tasks used to measure EF and with the propensity for researchers to use a single task to measure an EF component (Rose et al., 2011).

Although there may be a link between an individual EF component or EF as a whole and both gF and gC as outlined above, many researchers suggest that EF is responsible for variation in gF in particular (Diamond, 2013; Engle, 2002; Martinez et al., 2011; Unsworth, Miller, et al., 2009, 2010; Uka et al., 2019). Therefore, this research also investigated the effect that cognitive training has upon this higher cognitive ability and aimed to investigate further whether it is the improvement in an individual EF component or EF as a whole that may result in an improvement in gF abilities.

2.5 Processing speed

Significantly, many researchers have identified the importance of processing speed (PS) and its potential effect upon cognitive development (Nettelbeck & Burns, 2010). Fry and Hale (1996) proposed the developmental cascade theory for PS – a sequence of processing stages within which the effectiveness of processing at the first stage has a flow-on effect for the next stage, which influences the next and so on. In their cross-sectional research of 7-19 year-olds, they found that increases in PS resulted in improvements in WM that, in turn contributed to improvements in *g*F (Fry & Hale, 1996). This finding was consistent even when age-related differences in speed, WM and gF were statistically controlled (Fry & Hale, 1996). Furthermore, they reported that the magnitude of the correlation between PS and intelligence did not appear to change with age (r = -.61; Fry & Hale, 1996). Subsequent research has also found that cognitive development is substantially mediated by PS in 8 to 14 year-olds (Nettelbeck & Burns, 2010) and 13 to 17 year-olds (Coyle, Pillow, Snyder, & Kochunov, 2011).

Longitudinal research by Rose, Feldman, Jankowski, and Van Rossem (2012), which measured multiple cognitive domains (PS, attention, memory) using a latent variable approach, also supported the developmental cascade theory as both PS and attention at seven months of age predicted memory at 11 years of age. Furthermore, McAuley and White (2011) found that in 6 to 24 year-olds, improvements in WM and response inhibition over time are largely mediated by associated improvements in PS. The authors noted that WM also shows an independent effect of age even after the contribution of PS was taken into account (McAuley & White, 2011). Importantly, Cepeda, Blackwell, and Munakata (2013) outlined the significance of task choice when measuring PS. They found that many speed-related tasks actually measure EF due to the requirement for goal maintenance and manipulation of information in WM (Cepeda et al., 2013). As a result, they suggest using simple PS measures to reduce the influence of task choice upon the conclusions drawn when researching its relationship with EF.

It is interesting to note that not many researchers in the cognitive training field have included measures of PS in their studies, even though it has been found that PS is related to all three EF components (Friedman et al., 2008). Therefore, the importance of including measurements of PS in addition to the core EF is apparent (van der Ven et al., 2012) in order to test its potential role in the development of EF.

2.6 Development of EF across Childhood

The structure of EF may be different in the early years of schooling when compared to later childhood, as the degree of independence of the three EF components may change developmentally (Best & Miller, 2010; Garon et al., 2008; Karr et al., 2018; Miyake & Friedman, 2012). Some CFA research has shown that EF is a unitary ability during the preschool years (ages 2 to 6) (Hughes, Ensor, Wilson, & Graham, 2010; Visu-Petra, Cheie, Benga, & Miclea, 2012; Wiebe et al., 2011). Furthermore, researchers have also suggested that the single factor model of EF is also a better fit for older children (ages seven to nine) (Brydges et al., 2012). However, Miller, Giesbrecht, Muller, McInerney, and Kerns (2012) demonstrated the importance of task choice when they found that a two-factor model of EF, consisting of WM and inhibition, provided a better fit in preschool-aged children than a one-factor model. As they chose tasks that attempted to reduce the overlap in WM and inhibition task demands, they suggested that previous CFA research has only found support for the unitary EF view because of the task-impurity problem (Miller, Giesbrecht, et al., 2012).

The task-impurity problem may also be why other researchers have found support for a two-factor model of EF in young children. Van der Ven et al. (2012) found that although WM was a separate, distinguishable factor in seven- to eight-year-old children, inhibition and shifting were indistinguishable, providing support for a two-factor model of EF. Similarly, Lee et al. (2012) also found the same two-factor model for six-year-old children, and St Clair-Thompson and Gathercole (2006) found in 11 to 12-year-old children that there was no shifting factor as only WM and inhibition factors were evident. However, perhaps indicating that the differences in structure may in fact be developmental, van der Ven et al. (2013) found a two-factor model in six to seven-year-old children even when they attempted to take both task choice and scoring methods of EF tasks into account.

Interestingly, some of the aforementioned studies did not find support for the shifting factor in children. Senn et al. (2004) argued that shifting "may be less differentiated from working memory and inhibition in young children than in older participants" (p. 459). Garon et al. (2008) believes this is because shifting is the most complex of the core EF, builds upon the other two EFs and develops later in childhood. However, Miller, Giesbrecht, et al. (2012) do not rule out the possibility of the shifting factor existing in preschool-aged children, as they believe that there may have been overlap between their tasks for shifting and WM. Again, the influence of task choice may be evident in studies that have found support for the shifting factor. Huizinga et al. (2006) found that shifting and WM were detectable in children as young as 7 years-old, but inhibition was not. Interestingly, they found that this model was consistent from childhood to early adulthood (Huizinga et al., 2006). Similarly, Van der Sluis et al. (2007) found the same results in children aged 9 to 12.

Other researchers have even found support for the three-factor EF model in children. Lehto et al. (2003) used similar, but more developmentally appropriate tasks than Miyake et al. (2000) and found that in children aged 8 to 13 years-old there was support for a three-factor EF model where components were related but separable. Wu et al. (2011) also replicated the Miyake et al. (2000) model of EF in children aged 7 to 14 years of age, and Rose et al. (2011) found the three-factor model of EF was evident in children aged 11 years. They used a different battery of tasks to Miyake et al. (2000), but again, they found that all factors were related but separable. As outlined by Garon et al. (2008), this does seem to point to the potential for a tripartite model of EF in older children.

In summary, most studies find support for either a two-factor (Huizinga et al., 2006; Lee et al., 2012; Miller, Giesbrecht, et al., 2012; St Clair-Thompson & Gathercole, 2006; van der Sluis et al., 2007; van der Ven et al., 2012) or a three-factor (Lehto et al., 2003; Rose et al., 2011; Wu et al., 2011) EF model in children. However, some researchers suggest that this appears to follow a developmental path where EF becomes more separable as a child ages (Best & Miller, 2010; Lee, Bull, & Ho, 2013; Miyake & Friedman, 2012; Rose et al., 2011). Indeed, some of the research to date may suggest that for children aged approximately five to seven, EF are not completely separable and may be better explained by a two-factor model (e.g., Lee et al., 2012; Miller, Giesbrecht, et al., 2012; van der Ven et al., 2012, 2013), whereas once a child reaches approximately 8 to 9 years of age, EF may be better explained by a three-factor model (e.g., Lehto et al., 2003; Rose et al., 2012; Wu et al., 2011).

There does appear to be some support for a clear developmental change in EF occurring at around 7 to 8 years of age. As demonstrated in a large (N=2036) sample of children aged 5 to 17, Best, Miller, and Naglieri (2011) noted a significant increase in performance on EF tasks by 8 years of age, even though performance on all EF tasks did keep developing into adolescence. Later cross-sectional research found a period of rapid EF development around the 5- to 9-year age range (Korkman, Lahti-Nuuttila, Laasonen, Kemp, & Holdnack, 2013). According to Anderson (2002), EF displays rapid development between 7 and 9 years of age, which corresponds with the second growth spurt in the frontal lobe. Anderson (2002) also outlines another growth spurt, between 11 and 13 years-old, the period when all EF components approach maturity. Indeed, as outlined by Anderson (2002), it may be through the development of the PFC that children's full EF approaches maturity and may therefore be dissociable in research studies.

This kind of developmental progression regarding separability of function has been demonstrated in WM. Pickering, Gathercole, and Peaker (1998) found that for five- and eight-year-olds, the phonological loop and the visuo-spatial sketchpad were separable. In six- and seven-year-olds, Gathercole and Pickering (2000) found that the CE and the phonological loop were independent but still related, which is more consistent with the adult model of WM. However, the visuo-spatial sketchpad was not separable from the CE (Gathercole & Pickering, 2000). From 6 years of age onwards, the tripartite model with factors of WM consistent with Baddeley's model of WM were present (Gathercole, Pickering, Ambridge, & Wearing, 2004). It has also been found that the capacity of each component increases linearly from age four to early adolescence (Gathercole & Pickering, 2000).

Although some research does appear to demonstrate differences between male and female EF abilities in older adolescents or adults (Reed, Gallagher, Sullivan, Callicott, & Green, 2017; Singh, 2016; van den Bos, Homberg, & de Visser, 2013), these sex differences do not appear in younger children (i.e. less than 13 years old; Alarcon, Cservenka, Fair, & Nagel, 2014; Cross, Copping, & Campbell, 2011; Derntl, Pintzinger, Kryspin-Exner, & Schopf, 2014; Voyer, Voyer, & Saint-Aubin, 2017). However, even the results for sex differences in EF older adolescents and adults have been questioned recently, with modifications to task design appearing to contribute greatly to differences obtained (Grissom & Reyes, 2019).

Although it does appear logical that EF would develop throughout childhood, especially as in late childhood to early adolescence results on more complex EF measures approach adult levels of achievement (Huizinga et al., 2006; Korkman et al., 2013), it is hard to ascertain the details of EF development when considering the task-impurity problem of EF tasks.

2.7 The Measurement of EF and the Task Impurity Problem

The task-impurity problem has been mentioned at numerous points above, and has been characterised as a serious problem for the assessment of EF (Hedge, Powell, & Sumner, 2018; Miller, Nevado-Montenegro, et al., 2012; Miyake, Emerson, et al., 2000; van der Sluis et al., 2007; van der Ven et al., 2013; Zelazo, Blair, & Willoughby, 2017). As outlined by Miyake et al. (2000), the problem with assessment of EF is that there is no pure measure that taps EF exclusively as, by definition, EF operates on other cognitive processes. EF tasks are impure measures because they "necessarily involve both executive and non-executive processes and individual performance on these tasks may be biased by the non-executive requirements of the task" (Fournier-Vicente et al., 2008, p. 33). For example, the non-executive requirements of a task, such as language comprehension, may play a dominant role, thereby skewing test results (Miyake, Emerson, et al., 2000).

Not only is the problem related to non-executive requirements of a task, but there is also significant overlap with regards to executive components for EF tasks (Best & Miller, 2010). For example, some researchers use measures for inhibition that impose a significant burden upon other EF, where a rule must be remembered or where the participant must shift response (Best & Miller, 2010). The overlap between shifting tasks and WM and inhibition has already been outlined (see Section 2.2.2) and in fact Garon et al. (2008) believe that there is no pure shifting task due to this overlap. Miyake and Friedman (2012) stated that "both non-EF variance and measurement error are substantial, making it difficult to cleanly measure the EF variance of interest" (p. 8). Additionally, another major limitation of the most widely used EF tasks with children are that they are often not sensitive enough to detect individual differences (van der Sluis et al., 2007; Wilbourn, Kurtz, & Kalia, 2012). Therefore, as is evident from the limitations in previous research outlined above, the importance of careful choice of EF task(s) is obvious.

Furthermore, as EF involves both cognitive and behavioural elements (Anderson, 2002), to ensure that the measurement of school-aged children's EF is ecologically valid, it is important to gather parent and teacher observations of children's EF behaviours (Anderson, 2002; Isquith, Crawford, Espy, & Gioia, 2005). As Isquith et al. (2005) argue, "these measures should not be viewed as alternatives to performance-based assessment, but rather as complementary" (p. 210). However, it should be noted that the correlation between ratings and performance indicators of EF has been very low (Toplak, West, & Stanovich, 2013). Toplak et al. (2013) noted that "these two classes of measures should not be interpreted as equivalent, interchangeable, or as types or subcategories of one another" (p. 139). Instead, they argue that they represent different aspects of cognitive and behavioural functioning.

Ultimately, "there is no gold standard for assessment of" EF (Chan, Shum, Toulopoulou, & Chen, 2008, p. 213). However, there does appear to be clear consensus that researchers must use more than one performance measure for each EF construct

(Best et al., 2011; Huizinga et al., 2006; Miller, Giesbrecht, et al., 2012) and having at least three indicators to define an EF construct is desirable (e.g., Kline, 2015).

In summary, given the substantial associations between EF and achievement and EF and intelligence, particularly through gF, generating improvements in children's EF functioning could have a substantial effect upon their educational experiences. The main issue with previous research in this area has been the use of single tasks to measure EF components. Therefore, to determine the efficacy of a computerised cognitive training program on EF or higher-level cognitive abilities (academic achievement and gF), this study used multiple tasks to measure EF components. The following sections review the evidence base for cognitive training to provide the intervention context for the present study.

3 Cognitive Training

The core argument behind cognitive training is that extended practice on one or more cognitively challenging tasks or drills can produce changes in the brain and hence increase cognitive capacity in a manner that can benefit people in their daily lives (O'Connell & Robertson, 2012). Research in the neurosciences has shown that normal associative learning and experience evoke important changes in cortical sensory, synaptic connectivity, dendritic arborisation and axonal sprouting (Kolb, Teskey, & Gibb, 2010). Furthermore, activation of neurons or networks of neurons strengthens the connections between them and improves their efficiency (O'Connell & Robertson, 2012). It is important to note, however, that the degree to which these changes in the brain (i.e., plasticity) are possible can vary between individuals, dependent upon age, genetics, and prior environmental influences (Jolles & Crone, 2012).

The changes in neural activity that follow cognitive training are diverse (Kelly, Foxe, & Garavan, 2006) and it has been established that neurologically healthy people can substantially improve their performance on a given cognitive task (Kelly et al., 2006). However, it remains to be determined whether changes in activity within a particular region reflects flexibility (i.e., changes within the limits of the current functional capacity) or structural plasticity (i.e., changes of those limits, associated with structural brain changes) (Jolles, van Buchem, Rombouts, & Crone, 2012).

However, from an educational point of view, the key question is whether these improvements as a result of cognitive training transfer to untrained domains such as gF and classroom academic achievement. In other words, it is key to ascertain if improved performance reflects a fundamental enhancement of underlying cognitive capacity, or simply the acquisition of a set of skills that is specific to the practised task (Jolles & Crone, 2012; O'Connell & Robertson, 2012).

3.1 WM Training

Although initial views of WM capacity held it to be fixed (Cowan, 2001), research has raised the possibility that it may respond to interventions (Klingberg, Forssberg, & Westerberg, 2002). Since then, many studies have demonstrated that active, computerised training of WM may lead to an increase in WM capacity (Dahlin, 2011; Holmes, Gathercole, Place, Dunning, et al., 2010; Klingberg et al., 2005; Westerberg et al., 2007). However, although studies identifying significant improvement in recall on memory tasks as a result of strategy training have been around for a long time (e.g., Ericsson & Chase, 1982), the ability to produce generalisation and improvement in closely related (near-transfer) and non-related tasks (far-transfer) have not been nearly as successful (Chein & Morrison, 2010). Although some WM strategy training regimens have been found to be successful (St Clair-Thompson et al., 2010; Turley-Ames & Whitfield, 2003), it is believed that process-based training of EF might demonstrate broader generalisation due to it being more domain-general in nature (Morrison & Chein, 2011).

Research has shown that adaptive computerised WM training has the potential to help students with their concentration and learning in a variety of academic subjects (Gathercole, 2008). For instance, studies on adaptive WM training have shown that it leads to improved ability to follow classroom instructions (Holmes, Gathercole, & Dunning, 2009), may lead to improvements in inhibition (Chein & Morrison, 2010; Klingberg et al., 2005, 2002) and classroom behaviour (Halperin et al., 2012) and may lead to improvement in reading comprehension (Chein & Morrison, 2010; Dahlin, 2011; Karbach, Strobach, & Schubert, 2015; Loosli et al., 2012), mathematics achievement (Holmes & Gathercole, 2013; Holmes et al., 2009; Nelwan & Kroesbergen, 2016) and spelling (Alloway et al., 2013). Other researchers have found, using fMRI, that WM training leads to increased brain activity in EF-related areas of the brain (Olesen, Westerberg, & Klingberg, 2004). Importantly, some of this research has demonstrated improvements in normally functioning children within a classroom environment (e.g., Holmes & Gathercole, 2013).

Research by Roughan and Hadwin (2011) found that, given the reciprocal nature of the relationship between academic achievement and negative affect (McCarty et al., 2008), adaptive WM training may be useful in reducing test-related anxiety in children with social, emotional and behavioural difficulties. In a seminal study by Jaeggi, Buschkuehl, Jonides, and Perrig (2008), it was found that adaptive WM training using the n-back task (where participants specify if the item presented matches the item that was presented *n* items back) could improve gF in university-aged students. This was important work as it demonstrated far-transfer to a factor of significant importance to everyday life, and given gF's "predictive power for a large variety of intellectual tasks and professional success" (Jaeggi et al., 2008, p. 6832), this finding could be highly relevant to applications in education (Jaeggi et al., 2008). Additionally, a meta-analysis by Au et al. (2015) demonstrated a small but significant positive effect of n-back training on improving gF. Interestingly, in a double-blinded, randomised controlled experiment that trained gF in addition to WM, Bergman Nutley et al. (2011) found that the training groups improved when compared to the control group on untrained measures of gF and WM. Similarly, Peng et al. (2017) found n-back training improved gF in preschool-aged children when compared to an active control group; importantly, an improvement which was maintained 12 months later. Typically, researchers have found "that the amount of transfer to non-trained tasks within the trained construct was roughly proportionate to the amount of training on that construct" (Bergman Nutley et al., 2011, p. 598).

Although the research appears promising for adaptive WM training and its far-transfer effects to gF, evidence of the alteration of higher-level cognitive functioning as a result of adaptive WM training has been mixed (Rabipour & Raz, 2012; Shipstead et al., 2010). For example, Holmes et al. (2010) found that adaptive WM training leads to improvements in WM capacity, but not gF. Similarly, other researchers have also found no far-transfer to gF as a result of WM training (Chein & Morrison, 2010; Dahlin, Nyberg, Bäckman, & Neely, 2008; Harrison et al., 2013; Thompson et al., 2013; Westerberg et al., 2007). Interestingly, Moody (2009) outlines some methodological issues with the Jaeggi et al. (2008) study in which test protocols were modified for a standardised test of gF (Raven's Standard Progressive Matrices), which may have altered what the test was actually measuring. The methodological concerns regarding this research are not isolated, as will be outlined in Section 3.1.1.

Significantly, there have also been discrepant findings with regard to adaptive WM training's transfer to other EF components, with some research demonstrating successful transfer to inhibition (Chein & Morrison, 2010; Klingberg et al., 2005, 2002), but others failing to show transfer (Dahlin et al., 2008; Foster et al., 2017; Westerberg et al., 2007). The first research to look at impacts of WM training on a range of EF components appears to have been done by Salminen, Strobach, and Schubert (2012). They used a WM training program similar to Jaeggi et al. (2008) with university-aged participants over a three-week period, measuring transfer to four EFs (WM, shifting, inhibition and dual task performance). They found near-transfer to the shifting, inhibition and WM aspects of EF (Salminen et al., 2012). However, this research also shared some of the methodological (e.g., no-contact control group) and measurement (e.g., single measures to determine an EF construct) concerns regarding WM training research that have been alluded to above. It is obvious that more research with regard to near-transfer effects of adaptive WM training is desirable.

Importantly, there have been studies that have questioned the veracity of the claims made by the WM training research to date. For example, in an attempt to replicate the work of Jaeggi et al. (2008), Redick et al. (2012) found that adaptive WM training was not effective in improving either gF, gC, WM capacity or PS in university-aged students. Additionally, another attempted replication study that did not find any transfer effects as a result of n-back training was conducted by Chooi and Thompson (2012). They found that although participants' performance on the WM training task improved, this did not result in any transfer to increased WM capacity or gF.

There have been two studies within an Australian context, and both have limited or null results. Testing a computerised WM training program on six- to seven-year old children with low WM, Roberts et al. (2016) found limited transfer to short-term visuospatial WM at six and twelve months after training, but this improvement was not maintained at twenty-four months. They found no other benefits to other outcomes. A more recent study with typically developing primary school children addressed some methodological concerns of previous research (e.g., lack of active control groups) and found no effect for either near-transfer to other EF components or far-transfer to academic achievement, and no effect at a 3-month follow-up (Hitchcock & Westwell, 2017).

There have been several meta-analyses of WM training studies. The first meta-analysis by Melby-Lervag and Hulme (2013) found that overall, WM training had limited benefit for near-transfer. That is, they found it had limited benefit for WM that uses the visuo-spatial storage component, but found no benefit for WM that uses the phonological short-term storage. Another key finding was the lack of evidence of far-transfer effects. Importantly, as outlined by Shipstead, Hicks, and Engle (2012), "in the absence of near-transfer there is no clear reason why abilities should change" (p. 190). The meta-analysis by Melby-Lervag and Hulme (2013) included 23 studies with clinical and typically developing participants who ranged from children to adults. Even though the results do appear somewhat disappointing for the promise of adaptive WM training, the authors acknowledge some of the limitations of this meta-analysis, such as a non-normal distribution of age-ranges and a wide variety of included clinical conditions (Melby-Lervag & Hulme, 2013). It is important to note that when the authors split their sample into categories for age, there did appear to be better training effects for younger children (under approximately 10 years of age at the time of training) (Melby-Lervag & Hulme, 2013).

A subsequent meta-analysis by Schwaighofer, Fischer, and Bühner (2015) included 47 studies and included training conditions as moderators for transfer effects, which Melby-Lervag and Hulme (2013) did not include. They found that session duration, training frequency, training interval, modality and feedback were not significant moderators of near-transfer effects (Schwaighofer et al., 2015). However, training location was a significant moderator with some inconsistent results. Results demonstrated that training in a school provided a larger near-transfer effect to verbal WM than when training in a laboratory, whereas training in a laboratory provided a larger transfer effect to nonverbal ability than when training in a school. All effect sizes were relatively large (eta squared ($ŋ^2$) = 0.19 - 0.20), according to the descriptors outlined by Watson (2019). They also found that supervision of the training sessions was a significant moderator of near-transfer effects to verbal WM, showing that training that was individually supervised resulted in a larger mean effect size than training where a person was simply present or there was no supervision at all. Again, effect sizes were relatively large ($ŋ^2 = 0.16 - 0.17$). In contrast to Melby-Lervag and Hulme (2013), they found that there was no effect of age on training outcomes (Schwaighofer et al., 2015).

Melby-Lervag, Redick, and Hulme (2016) performed a meta-analysis using 87 publications with 145 experimental comparisons. They found that there were large effects for near-transfer tasks (Hedges' $g = 1.88$) but that the effects were smaller when there was an active control group present (Hedges' $g = 0.80$). There were some far-transfer effects present when assessed immediately after training was completed (e.g. reading comprehension, mathematics, nonverbal abilities). However, these varied depending upon whether an active control group was used, with only reading comprehension (Hedges' $g = 0.15$) showing significant improvement when compared to an active control group. The follow-up post-tests (an average of 5 months after training) showed different results with only mathematics (Hedges' $g = 0.22$) demonstrating significant improvement when compared to an active control group. However, Melby-Lervag et al (2016) outline that for both the immediate and follow-up significant far transfer effects (reading comprehension and mathematics respectively), these significant improvements appeared to be driven by unexplained drops in the active control groups between pre-test and post-test rather than an improvement in the training groups.

A meta-analysis by Aksayli, Sala, and Gobet (2019) analysed the results from 50 WM training studies and importantly included the type of near-transfer task as a moderator. That is, they considered how close the near-transfer task was to the training task itself, which, if the tasks were very similar, they referred to as very-near-transfer. They found that the similarity between the training tasks and the near-transfer tasks was a significant moderator, with a medium effect for the very-near-transfer tasks (Hedges' *g* = 0.45) and a small effect for near-transfer tasks that were more different from the training tasks (Hedges' *g* = 0.25). They stated that this pattern of results "was in line with the hypothesis according to which transfer is a function of the extent to which the trained task and the target task overlap (i.e., share common features)" (Aksayli et al., 2019, p. 237). They found that far-transfer effects as a result of WM training did not exist. Importantly, they concluded that the field of cognitive training would benefit greatly from "the study of the impact of the training on latent factors rather than observed variables" (Aksayli et al., 2019, p. 240).

The most recent meta-analysis by Sala and Gobet (2020) analysed the results from 41 WM training studies where participants were typically developing children. They looked at design qualities (types of control group), age of participants, and near-transfer task similarity as potential moderators. They found that the type of control group was a significant moderator for far-transfer with nonactive control groups providing a larger effect size (Hedges' *g* = 0.14) than active control groups (Hedges' *g* = 0.03). Age was not found to be a significant moderator, but near-transfer task similarity was again a significant moderator with a medium effect for the very-near-transfer (Hedges' *g* = 0.47) and a small effect for near-transfer tasks (Hedges' *g* = 0.26). Again, this seems to confirm the importance of using latent factors, but also seems to indicate the importance of well-designed studies that use active control groups.

3.1.1 Limitations

As alluded to earlier, it appears that WM training studies are plagued by a number of significant issues. The criticisms and weaknesses of WM training studies can be summarised along definitional, methodological and measurement lines.

First, with regard to definitional issues, Shipstead et al. (2012) outline that some studies (e.g., Olesen et al., 2004) have explicitly defined WM "as a storage system that is responsible for retaining small amounts of information over brief intervals of time" (p. 629) and have measured them using simple span tasks. As these simple span tasks are typically used to determine STM capacity, it would appear that incorrect definitions of WM are being used by some researchers (Shipstead, Redick, et al., 2012). Importantly, as noted by Baddeley (1992), the significance of STM as a predictor of complex cognition was resolved in the 1970s, which led to the tripartite model of WM, which included STM as a subcomponent of the larger WM system. Furthermore, if measures that tax the combined processing and storage resources of WM are better predictors of language comprehension performance than are measures that tax only storage (Daneman & Carpenter, 1980; Daneman & Merikle, 1996), then it stands to reason that simple training tasks that focus purely on increasing storage might not be as effective as training that focuses upon both processing and storage (Gibson, Gondoli, Johnson, Steeger, & Morrissey, 2012).

The findings from a longitudinal study by Engel de Abreu, Conway, and Gathercole (2010) confirmed that the executive components of WM rather than the storage component alone are the source of the link with gF. This is significant as many of the WM training tasks to date simply train the short-term storage component (either visuo-spatial or verbal), thus further highlighting definitional issues and the subsequent lack of theoretical justification for transfer (Shipstead, Hicks, et al., 2012). Shipstead, Hicks, et al. (2012) question whether training on short-term storage component tasks simply turns trainees into expert span task performers. Longitudinal data suggests that growth in the executive component of WM, rather than the short-term storage components, is related to mathematical problem solving (Swanson, 2011); therefore, it is difficult to claim that a training program that focuses mainly upon these short-term factors should result in an improvement in academic achievement.

Many WM experimental studies to date suffer from methodological issues that diminish the ability to draw causal inferences from their findings. For example, comparing the intervention group to a no-contact control group is common among studies where near- and far-transfer is found (e.g., Hulme & Melby-Lervag, 2012; Melby-Lervag & Hulme, 2013; Morrison & Chein, 2011; Redick, Shipstead, et al., 2012; Rudebeck, Bor, Ormond, O'Reilly, & Lee, 2012; Salminen et al., 2012). There are even studies where post-test results are determined for the experimental group but not for the control group (e.g., Holmes et al., 2009) or there is no control group at all (e.g., Halperin et al., 2012; Holmes & Gathercole, 2013; Millner, Jaroszewski, Chamarthi, & Pizzagalli, 2012). Sala and Gobet (2017) go so far as to propose that effect sizes for WM training research are "inversely related to the quality of the experimental design" (p. 3).

These methodological issues are significant as they raise the possibility of Hawthorne effects, wherein participants in an experimental group expect improvement in the measures of interest, whereas participants in a no-contact control group do not expect such improvement. Even when studies attempt to address issues around the Hawthorne effect, by including an active control group with similar task features (Alloway et al., 2013), the control group still does not have the same frequency of contact as the cognitive training group, again allowing for the potential of a Hawthorne effect as well as simple (and confounding) differences in practice effects. Furthermore, results where the control group also demonstrates a significant improvement in measures of a construct (e.g., Van der Molen, Van Luit, Van der Molen, Klugkist, & Jongmans, 2010) further confound the findings of some of this experimental research.

Additionally, with regard to the use of control groups in the cognitive training literature, none of the research has addressed the issue of improvement expectancy. As outlined by Boot, Simons, Stothart, and Stutts (2013), active control groups are not sufficient in experimental psychology to determine causation unless the expectancy of how much the intervention will improve results on dependent variables within both the experimental and control group is determined so that placebo effects can be ruled out.

With regard to measurement concerns, "WM training studies often test generalisation from training with only a single task, but interpret observed improvements on that task as reflecting gains in some broadly defined cognitive ability" (Morrison & Chein, 2011, p. 55). Many published WM training studies suffer from measurement limitations, such as single measures of cognitive constructs (e.g., Salminen et al., 2012), and mixed results, such as transfer effects for some tasks but not others and inconsistent transfer effects for the same tasks across studies (Hulme & Melby-Lervag, 2012; Melby-Lervag & Hulme, 2013; Redick, Shipstead, et al., 2012).

Despite the theoretical, methodological and measurement weaknesses outlined above, it is important to note that the ability to change the executive components of WM is still an open question and requires further investigation (Aksayli et al., 2019; Shipstead, Hicks, et al., 2012).

3.2 EF Training

Although it has been acknowledged that computerised training on a wider range of EFs could have significant implications for education (Gibson et al., 2012; Morrison & Chein, 2011; Walton & Dweck, 2009), most of the training to date is related to only computerised WM training (Zinke, Einert, Pfennig, & Kliegel, 2012) which appears to only train shorter-term storage capabilities (Shipstead, Hicks, et al., 2012). When researchers have attempted to train a wider range of EF, there have again been mixed results, with research that is characterised by many of the methodological and measurement criticisms of the WM training literature.

Johnstone et al. (2012) found that a cognitive training program focussed upon both WM and inhibition improved the behaviour of children (aged 7 to 14 years) with and without ADHD. It also demonstrated an improvement in spatial WM, inhibition and sustained attention in children with ADHD and improvements were maintained six weeks following training (Johnstone et al., 2012). In an adult group, Millner et al. (2012) demonstrated transfer to untrained tasks when completing an inhibition training task. This improvement was demonstrated as an improvement in reaction time and accuracy on inhibition tasks.

Some researchers appear to have found successful transfer in training only one EF component, with particular success in computerised shifting training (Karbach & Kray, 2009; Kray, Karbach, Haenig, & Freitag, 2012; Parong et al., 2017; Zinke et al., 2012). For example, Zinke et al. (2012) found that shifting training for adolescents resulted in decreased error rates and reaction times in a related, non-trained, shifting task when compared to a control group. There also appeared to be small, though not statistically significant, improvements in reaction times in an unrelated WM task for the training group. Karbach and Kray (2009) found that shifting training was effective in both near- and far-transfer. The near-transfer was a similar shifting task, whereas the far-transfer was for inhibition, WM and gF. They found that the training effect was greatest for younger children (8 to 10-years-old) and for older adults (62 to 76-years-old). Notably, the effect size for both near- and far-transfer were reasonably large (Cohen's $d = 0.88 – 2.12$). Additionally, Kray, Karbach, Haenig, and Freitag (2012) implemented a task-switching training program for a small sample ($n = 20$) of 8- to 12-

year-old boys with ADHD and found transfer to inhibition and WM, but not gF. Although the findings for shifting-specific training appears promising, once again, what appears to be successful transfer of training tasks is affected by methodological issues such as no-contact control groups (Zinke et al., 2012) and single tasks used to measure a construct (Karbach & Kray, 2009; Kray et al., 2012).

Other researchers have attempted to use computerised programs to train a variety of EFs simultaneously (Kable et al., 2017; Rueda, Rothbart, McCandliss, Saccomanno, & Posner, 2005; Sánchez-Pérez et al., 2018; Thorell, Lindqvist, Nutley, Bohlin, & Klingberg, 2009; van Muijden et al., 2012) with mixed results. For example, a computer training program for children 7- to 12-years-old that trained inhibition, shifting and WM abilities found an improvement in near-transfer inhibition tasks, but more importantly, in transfer to maths fluency and non-verbal intelligence (Sánchez-Pérez et al., 2018). In contrast, Thorell, Lindqvist, Nutley, Bohlin, and Klingberg (2009) found that EF training which targeted both WM and inhibition for typically developing preschool-aged children was effective for WM capacity but not for inhibition. In a randomised control trial Kable et al. (2017) attempted to train inhibition, shifting and WM in young adult participants using an adaptive cognitive training program. However, further elucidating the conflicting results in this research area, they found that both the experimental and the active control groups improved on near-transfer measures, indicating that improvement on post-tests was a result of practice effects.

Using a sample of both younger (20-31 years) and older (65-80 years) adults, Schmiedek, Lovden, and Lindenberger (2010) found that after an extensive schedule (more than 100 days) of computerised cognitive training that included training of PS, WM and episodic memory there was both near- and far-transfer, including far-transfer to gF (Cohen's *d* = .19 - .52). Interestingly, it was found that transfer occurred for younger adults, but not for older adults. Although Schmiedek et al. (2010) used a large battery of tests to determine each construct, their study was limited in its use of non-active control groups for the comparison of training effects.

Some researchers have used a mixture of computerised and non-computerised EF training programs. For example, Mackey, Hill, Stone, and Bunge (2011) found that gF and PS training for children (7 to 10 years-old) was successful in near-transfer to similar tasks. The effect sizes were large for the improvement in both the fluid reasoning (Cohen's d = 1.51) and PS (Cohen's d = 1.15) tasks. Furthermore, the fluid reasoning training group improved (Cohen's d = 0.65) on an untrained task for working memory. This was an important finding as it was demonstrated in a classroom setting (after-school care), with well-validated outcome measures (Mackey et al., 2011). Significantly, there were some unexplained results, such as the fluid reasoning training group significantly improving on a measure of PS whereas the PS training group did not improve on this measure (Mackey et al., 2011).

One of the largest computerised EF training studies to date was conducted by Owen et al. (2010), where over 11,000 adults trained across 12 different computerised training tasks, which included working memory and reasoning tasks, but this study found no evidence of training transfer. However, Klingberg (2010) has criticised this study due to the short training times (10 minutes), unsupervised training and the intermittent days without training.

A meta-analysis on the near- and far-transfer effects of training among children's EF skills by Kassai et al. (2019) found that although there was some evidence of a near-transfer effect (Hedges' g = 0.44), no evidence was found for far-transfer (Hedges' g = 0.11). Notably, due to the consistent lack of far-transfer, Kassai et al. (2019) found that there was limited practical benefits in training single EF components in childhood, especially as there was some difference between the near-transfer effects between WM (Hedges' g = 0.50) and the two other EF components (inhibition: Hedges' g = 0.24, shifting: Hedges' g = 0.37) and this was consistent whether the training program focussed on one or more EF skills at a time. Therefore, if attempting to improve EF in children, Kassai et al. (2019) suggest that it might be advisable to train multiple EF components within both the educational and clinical practice environments.

3.2.1 Non-Computerised Training

As outlined by Diamond and Lee (2011), there are many other non-computerised possibilities for training EF. They include, but are not limited to, aerobic exercise, mindfulness, and additions to classroom curricula as areas of potential application.

Dowsett and Livesey (2000) successfully implemented non-computerised inhibition training in three- to four-year-old children. Through simple practice of EF tasks (card sort and stop-signal tasks), children improved on another inhibition measure (Go/No-Go task).

A number of researchers have targeted preschool and primary school aged children for non-computerised training upon a broader range of EF. Rothlisberger, Neuenschwander, Cimeli, Michel, and Roebers (2012) found that a non-computerised EF training program which focussed upon WM, inhibition and shifting training tasks for preschool-aged children was effective for all three EF components. However, there were specific age-based effects with WM and shifting significantly improving in five-year-olds, whereas inhibition was only improved in six-year-olds (Rothlisberger et al., 2012). Importantly, these were typically functioning children and the delivery of the training program was shared between the researchers and the teacher. Another non-computerised intervention aimed at preschool children by Halperin et al. (2012) trained inhibition, WM, motor control, attention, planning and visuo-spatial abilities, and found significant reduction in parent and teacher reported ratings of ADHD behaviours (Halperin et al., 2012). Significantly, there was no control group for the comparison of training effects. Similarly, research by García-Madruga et al. (2013) using non-computerised EF training for normally functioning children (8 to 9 years-old) found an improvement in reading comprehension, WM, and gF. Traverso et al. (2015) found EF training for WM, inhibition, and shifting was effective in preschool-aged children for many of the near-transfer EF measures. However, they also found some inconsistent results with the no-contact control group improving more than the experimental group in one EF measure and no difference amongst other EF measures (Traverso et al., 2015).

More recent research by Paananen et al. (2018) found improvement in both EF and academic measures for primary school aged children with attention and EF difficulties as a result of non-computerised training. However, they found that improvement differed depending upon the initial severity of the EF problems, as participants with either low or high pre-intervention attention and EF scores (indicating either a weakness or strength in attention and EF) showed no change after the intervention (Paananen et al., 2018).

There have been a number of meta-analyses conducted to determine the effect of aerobic exercise on EF abilities in children and they tend to find quite similar effects. For instance, Ludyga, Gerber, Brand, Holsboer-Trachsler, and Pühse (2016) found a small effect of exercise on speed (Hedges' $g = 0.35$) and accuracy (Hedges' $g = 0.22$) components of EF tasks. Similarly, a meta-analysis by de Greeff, Bosker, Oosterlaan, Visscher, and Hartman (2018) that looked at the effect of acute physical activity on EF in preadolescent children, found a small effect size (Hedges' $g = 0.24$).

A meta-analysis by Zenner, Herrnleben-Kurz, and Walach (2014) found that school-based mindfulness improved cognitive performance, which often included measures of EF, with a large and significant effect size (Hedges' $g = 0.80$). However, a more recent meta-analysis by Dunning et al. (2019) specifically separated EF outcomes for mindfulness studies in schools and found a much smaller effect size (Cohen's $d = 0.30$). Notably, when restricting studies to only those that included an active control group, the effect size was even smaller and was no longer significant (Cohen's $d = 0.10$).

Interestingly, more recent innovative interventions have attempted to embed EF training in everyday activities. In an Australian study, Howard, Powell, Vasseleu, Johnstone, and Melhuish (2017) found that a children's book designed to challenge EF during shared individual reading improved shifting ability and WM in preschool children. This is significant, as it demonstrates the potential for training EF in a way that can be administered by educators or parents, at low cost, and which can be integrated into children's existing school or home routines (Howard et al., 2017). This is unlike many other interventions that are added extras and must be incorporated into the child's day.

3.2.2 Limitations

Although the results of EF training research - both computerised and non-computerised - appear promising, they too suffer from similar methodological problems as computerised WM training research. EF training research also has issues with comparisons with non-active or no-contact control groups (García-Madruga et al., 2013; Mackey et al., 2011; Millner et al., 2012; Rothlisberger et al., 2012; Sánchez-Pérez et al., 2018; Schmiedek et al., 2010; Zinke et al., 2012), no control group at all (Halperin et al., 2012) and single measures for EF constructs (Howard et al., 2017; Karbach & Kray, 2009; Kray et al., 2012; Mackey et al., 2011; Sánchez-Pérez et al., 2018). Despite these significant limitations and many conflicting results in the research to date, as outlined by Diamond and Ling (2016), there are still “burning questions that are calling out for answers” in EF training research (p. 42).

3.3 Individual Differences

Given the wide variety of conflicting results in cognitive training, researchers have attempted to address individual differences in training gains in order to determine who benefits most from cognitive training interventions (Bürki, Ludwig, Chicherio, & de Ribaupierre, 2014; Karbach & Unger, 2014; Könen & Karbach, 2015; Titz & Karbach, 2014; van der Donk et al., 2017).

There are two opposing theories that attempt to explain and describe individual differences in performance gains as a result of cognitive training: the magnification effect and the compensation effect (Karbach & Kray, 2016; Karbach & Unger, 2014; Katz, Jones, Shah, Buschkuehl, & Jaeggi, 2016). The magnification effect proposes that high-achieving individuals are most likely to benefit from cognitive training because they “have more efficient cognitive resources to acquire and implement new strategies and abilities” (Karbach & Unger, 2014, p. 8). As a result, if there is a magnification effect present then pre-test cognitive scores should be positively correlated with the training-related gains (Karbach & Unger, 2014). In contrast, the compensation effect “assumes that high-performing individuals will benefit less from cognitive interventions, because they are already functioning at the optimal level and therefore have less room for improvement” (Karbach & Unger, 2014, p. 8). Thus, if there is a compensation effect present then pre-test cognitive scores should be negatively correlated with training gains (Karbach & Unger, 2014). As EF develops across childhood and adolescence, if there are magnification effects occurring then cognitive training results should see an increase in individual differences and age differences on cognitive measures, whereas if there are compensation effects occurring then results should see a reduction in individual differences and age differences on cognitive measures (Karbach & Unger, 2014).

There has been some research that supports the magnification account of training effects where the younger the participant, the more training gains were made (Bürki et al., 2014). Other research supporting the magnification effect has demonstrated that high-ability participants performed better in WM training than low-ability students (Foster et al., 2017; van der Donk et al., 2017, Wiemers, Redick, & Morrison, 2019). However, in contrast, some research outlines the significance of the type of training task as being whether magnification or compensation effects predominate, with the suggestion that perhaps process-based training (such as cognitive training) as opposed to strategy-based training is more effective for adults when compared to children (Karbach & Verhaeghen, 2014). Interestingly, it has been found that when using typically achieving children who have no apparent difficulties which may impede upon their learning, that training effects are more consistent across the group (Bergman Nutley & Söderqvist, 2017).

3.4 Behavioural Measures of EF

As outlined by Bryck and Fisher (2012), there are other limitations of the current EF training research. They state that “future training studies in children should include pre-post measures of academic achievement and teacher reports of classroom behaviour to allow for an assessment of the potential broad effects training might induce in ecologically valid contexts” (Bryck & Fisher, 2012, p. 97). Given that EF abilities appear to play a crucial role in classroom situations, both in maintaining focused behaviour and in supporting learning (Gathercole, Alloway, et al., 2008) and as

EF abilities predict “the ability to engage in appropriate behaviour at the appropriate time in real-world tasks” (Shipstead et al., 2010, p. 246), then the effect of cognitive training on student behaviour is particularly salient. This is linked to the key question from an educational perspective: do improvements as a result of cognitive training transfer to untrained tasks such as academic achievement and gF? As stated previously, a key additional objective of this thesis is to determine if improved performance reflects a fundamental enhancement of underlying cognitive capacity or simply the acquisition of a set of skills that is specific to the practised task (Jolles & Crone, 2012; O’Connell & Robertson, 2012).

3.5 Implicit Theories of Intelligence

One potential moderating variable that is almost entirely absent from the cognitive training literature are an individual's personal beliefs about intelligence and their impact upon motivation and subsequent training gains. When computerised cognitive training studies do include individual motivation factors as a potential moderating variable, they are typically related to how motivated participants were by the program itself (e.g., Soderqvist, Nutley, Ottersen, Grill, & Klingberg, 2012), not other motivation-related factors. Researchers have suggested that individual differences in cognitive training gain may relate to whether or not individuals believe that intelligence is a malleable or a fixed construct, or whether or not participants are easily frustrated in the face of difficult cognitive tasks (Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Shah, Buschkuehl, Jaeggi, & Jonides, 2012). Therefore, implicit theories of intelligence may be a significant moderating factor in transfer gains as a result of computerised cognitive training.

Individual beliefs about the malleability of intelligence can indirectly affect academic achievement as an individual's achievement is affected by effort or motivation (Blackwell, Trzesniewski, & Dweck, 2007; Dweck & Leggett, 1988; Dweck, 2000; Jones, Wilkins, Long, & Wang, 2012; Weber, Lu, Shi, & Spinath, 2013; Yeager & Dweck, 2012). As outlined by Dweck and Leggett (1988), the idea that ability can be changed through effort orientates those with a view of intelligence as a malleable quality toward more challenging tasks where the goal is skill acquisition and where effort can overcome difficulty. Thus, students that place a strong emphasis upon developing their competence, use more active strategies and put more effort into learning activities (Dupeyrat & Mariné, 2005). Dweck and Leggett (1988) distinguish between an entity (fixed) or incremental (malleable) view of intelligence, which have different resultant goal orientations of performance (entity) and learning (incremental) respectively. While incremental theorists focus on the acquisition of competence, entity theorists focus on the confirmation of competence (Ablard & Mills, 1996). Therefore, individual differences in implicit theories lead people to create distinctive frameworks or meaning systems for interpreting and responding to success and failure (King, McInerney, & Watkins, 2012).

As outlined by Dweck (2012), implicit theories about intelligence are not about stability, but about control. Dweck (2012) states:

> An entity theorist believes that people do not have control over their attributes or the power to change them. However, an entity theorist may believe that intelligence or personality can deteriorate with age. Moreover, an incremental theorist believes that people can change, but not necessarily that most people do change (p. 47-48).

When individuals hold an incremental theory of their intelligence, they tend to focus more upon learning goals, with the aim of increasing their ability (Dweck, 2000; Hong, Chiu, Dweck, Lin, & Wan, 1999). They are also more likely to take remedial action if performance in a task is unsatisfactory (Hong et al., 1999). Significantly, a meta-analysis by Burnette, O'Boyle, VanEpps, Pollack and Finkel (2013), with a large sample size (N=28,217), was consistent with the view that incremental theories are positively related to learning-oriented goals and negatively correlated with performance goals, although they were only small to moderate in magnitude ($r = 0.24$ and $r = -0.20$ respectively). Furthermore, they demonstrated that individuals who hold an incremental view of intelligence are more likely to use goal-directed strategies and avoid anxiety or hopelessness when evaluating their pursuit of a goal (Burnette et al., 2013).

A longitudinal study by Blackwell et al. (2007) found that an incremental view of intelligence predicted more positive motivational patterns, which resulted in an improvement in math scores two years later. Additionally, cross-sectional work by King et al. (2012) demonstrated that students who believe their intelligence is fixed are more likely to feel anger, anxiety, shame, hopelessness and boredom.

A study by Autin and Croizet (2012) demonstrated that interventions on an individual's beliefs about a task (e.g., reframing a difficult task as a learning opportunity) can result in an improvement in EF abilities in 11-year-old children. Although they did not directly address individuals’ implicit theories of intelligence, but addressed the specific task that was being attempted, the similarity of this reframing with an incremental view of intelligence is apparent. Furthermore, they also found that not only did reframing a difficult task affect EF abilities, but it also led to improvements in reading comprehension (Autin & Croizet, 2012).

Importantly for the prospect of cognitive training, it has also been demonstrated in a small sample of children that those with a malleable view of intelligence responded better to a training schedule for children with reading problems (Pepi, Alesi, & Rappo, 2008). One cognitive training study did take into account personality traits and their influence upon cognitive training (Studer-Luethi, Jaeggi, Buschkuehl, & Perrig, 2012) and found that neuroticism and conscientiousness were associated with training outcomes for young adults when engaged in an n-back training protocol. They found that higher neuroticism was associated with lower training scores overall and with lower gains in near- and far-transfer tasks. In contrast, they found that conscientious participants showed better overall training performance and better gains in near-transfer, but not far-transfer, tasks (Studer-Luethi et al., 2012). Given that students with an incremental view of their own intelligence may be expected to be more conscientious on a challenging task (Dupeyrat & Mariné, 2005), the potential for implicit theories of intelligence to moderate training gains is apparent.

As outlined by Redick (2019), there have been a few cognitive training studies that have attempted to determine if implicit theory of intelligence moderates training gains. Although no studies to-date have found evidence of moderation (Redick, 2019), there is one study that has demonstrated some promising results. Jaeggi et al. (2014) found that an incremental view of intelligence was associated with greater training gains in a sample of university students who completed cognitive training. Although the implicit view of intelligence-by-intervention interaction was not significant, they did believe that this was not significant due to effects in the active control group (where they showed a placebo effect). However, they highlight the importance of assessing motivation and beliefs about intelligence when conducting training studies.

It is evident that a malleable or fixed view about one's own abilities could be significant when undertaking a challenging cognitive training task over a period of weeks. It may be that a student who has an incremental view of their own intelligence may put in more effort into an adaptive cognitive training task, as they believe that their abilities are malleable. However, the converse may be true of someone with an entity view of their own intelligence when faced with a challenging cognitive training task.

4 Rationale

It has been demonstrated that EF has important long-term implications for children's long-term well-being. The Dunedin Multidisciplinary Health and Development Study, a longitudinal study of a complete birth cohort of over 1,000 children born in one city in a single year in New Zealand, followed childhood levels of EF and subsequent health, wealth and crime outcomes over thirty years (Moffitt et al., 2011). They found that childhood EF predicts physical health, substance dependence, personal finances, and criminal offending outcomes, following a gradient of self-control. Most significantly, they found that outcomes followed a gradient across the full distribution of self-control in the population, thus outlining that "the observed gradient implies room for better outcomes even among the segment of the population whose childhood self-control skills were somewhat above average" (Moffitt et al., 2011, p. 2697). Given that this relationship between EF abilities and outcomes are potentially stronger for boys (Schoemaker, Mulder, Deković, & Matthys, 2013), this may be a particularly important demographic to target. Thus, being able to change EF abilities of children, even those who have average EF abilities, could have important well-being outcomes over the longer term across a range of functional domains (Miller, Nevado-Montenegro, et al., 2012).

Cognitive training may be a way that EF abilities can be changed in children (Kassai et al., 2019). However, the efficacy of cognitive training is not simply a scientific curiosity as they are now actively marketed to school systems (Shipstead, Hicks, et al., 2012) and school psychologists (Decker et al., 2013) as an intervention that works. Given that childhood might be a unique period during which cognitive training has specific effects (Jolles & Crone, 2012) and that when younger participants engage in cognitive training it may lead to a wider transfer of training effects (Wass, Scerif, &

Johnson, 2012; Witt, 2011), it may be that school is the best time to implement a cognitive training regime (Schmiedek et al., 2010). As it has been suggested that computerised cognitive training may benefit children of 8- to 12-years-old more than younger children (Diamond & Lee, 2011), a potential critical period for implementation is apparent. However, given the variability in research findings and the methodological weaknesses of much of the cognitive training research, it is essential that cognitive training programs accomplish tangible gains in the classroom that go beyond improvements in specific training-related tasks (Rabipour & Raz, 2012). Given that it may take a significant amount of time for cognitive training's far-transfer to academic achievement (Holmes et al., 2009), the importance of more long-term research looking at far-transfer to higher cognitive abilities is obvious. Importantly, as outlined by Conway and Getz (2010), none of the current research on cognitive training has demonstrated whether gains in gF or learning abilities are durable or if they are just transient practice effects.

A latent variable analysis is a versatile statistical tool that has the potential to make a significant contribution to scholarship on EF (Aksayli et al., 2019; McAuley & White, 2011; Protzko, 2017). Given that any single measure of a theoretical construct will be contaminated by error variance, Hulme and Melby-Lervåg (2012) argue that “multiple measures of key constructs should be used at pre-test and post-test allowing latent variable models to estimate possible changes in true score variance after training” (p. 199). Additionally, the use of latent variables for construct measurement can rule out explanations based upon task-specific abilities or processes (Aksayli et al., 2019; Shipstead, Redick, et al., 2012). This study will use a latent variable approach to the measurement of EF.

A number of researchers outline their key recommendations for future cognitive training research (Aksayli et al., 2019; Gathercole, Dunning, & Holmes, 2012; Gibson et al., 2012; Hulme & Melby-Lervag, 2012; Karbach & Kray, 2016; Redick, Shipstead, Wiemers, Melby-Lervåg, & Hulme, 2015; Shipstead, Hicks, et al., 2012; Simons et al., 2016), stating that research should:

(a) employ theoretically motivated training programs;
(b) engage higher-order control processes (such as EF);
(c) employ outcome measures with multiple measures of each construct;
(d) test the degree of generalisation;
(e) randomly allocate participants to condition;
(f) include an active comparison intervention that control for placebo effects and all potential treatment confounds.

The aim of this study was to address these cognitive training recommendations. Its main aim was to determine if a theoretically based cognitive training program could improve EF abilities (near-transfer) and, as a result, improve higher cognitive abilities, such as gF and academic achievement (far-transfer). Additionally, it attempted to determine if these changes could be sustained or improved over time (1 year). This study also aimed to determine if cognitive training improved student behaviour as viewed by parents and teachers. Finally, attempts were made to determine if an implicit theory of intelligence moderated any training gains made.

4.1 Research Questions

(a) Can a theoretically based cognitive training program improve EF abilities (near-transfer) and, as a result, improve higher cognitive abilities, such as gF and academic achievement (far-transfer)?

(b) Can potential improvements in EF abilities be sustained over time?

(c) Does cognitive training improve student behaviour as viewed by parents and teachers?

(d) Does an implicit theory of intelligence moderate potential training gains?

4.2 Hypotheses

Hypothesis 1: Firstly, it is expected that the group engaging in the theoretically based cognitive training program should improve significantly more in their latent EF scores (near-transfer) when compared to the control group. If the expected improvements are durable gains rather than transient practice effects, then these differences between groups should still exist at 1-year post-training.

Hypothesis 2: As outlined previously (see Sections 2.3 and 2.4), there are substantial associations between EF and academic achievement and EF and intelligence, particularly through gF. Therefore, if EF is functioning as a rate-limiting factor (e.g., Blair & Razza, 2007; Kotsopoulos & Lee, 2012), then generating improvements in participant's EF functioning as a result of the cognitive training program should see an improvement in participant's higher-level cognitive abilities (academic achievement and gF; i.e., far-transfer). Furthermore, given that EF abilities play a crucial role in academic behaviours (see Section 3.4), if improved performance on latent EF scores truly reflects a fundamental enhancement of underlying cognitive capacity then the experimental group should improve more than the control group on teacher and parent reports of EF behaviours.

Hypothesis 3: Finally, as outlined in Section 3.5, an individual's beliefs about the malleability of intelligence can indirectly affect academic achievement as an individual's achievement is affected by effort or motivation. Therefore, if the participant's implicit theory of intelligence is moderating potential training gains, then a significant interaction should be present between the change in participant's latent EF scores, their level of belief about the malleability of intelligence, and their intervention group.

5 Method

5.1 Ethics Statement

In accordance with the Declaration of Helsinki, written informed consent was obtained from the parent or guardian of each participant. This study was approved by the Ethics Committee of The University of Sydney. It was registered with the Australian and New Zealand Clinical Trials Registry (ACTRN 12615000016538) on 14 January 2015 and was conducted between April 2015 and August 2018.

5.2 Pilot Study

We first briefly describe the procedure for a pilot study that assisted in the development of parameters for the experiment. The pilot included 24 third-grade students (mean age = 8.6 years; SD = 0.3) who were randomly assigned to either the experimental or control group and assessed pre- and post-training. Only third-grade students were selected for the pilot study to ensure all vocabulary, training and assessment tasks would be appropriate. The pilot group participants were from the same school as the target experiment and were subsequently unable to participate in the target experiment.

The pilot study was conducted between April 2015 and June 2015 and the findings guided some modifications of the assessment phase of the research. As a result of the pilot study, changes were made to many of the EF measurement tasks. Firstly, as many participants were engaged for well over one hour on each of the EF assessment occasions, changes were made in an attempt to reduce susceptibility to confounds due to boredom and exhaustion. Complex span tasks were shortened significantly to only present one set each instead of the original three, with these shortened tasks still providing a reliable measure of WM (Foster et al., 2014). The Go/No-Go task was also shortened from two conditions of 320 trials to two conditions of 160 trials as participants appeared to become bored and exhausted with the length of this task in its original form. Although there is no reliability data for the shortening of this task, it was deemed important to reduce to susceptibility of participants to potential confounds.

Changes were also made to the functional running of some EF tasks. These changes mostly related to adding more details to instructional screens, changing some vocabulary to make it more age-appropriate, providing additional feedback in practice tasks, and adding confirmation buttons to proceed from one instructional screen to another as many participants inadvertently skipped some instructions. Importantly, the results from the reaction time task showed that many participants were 'button mashing' in an attempt to obtain quicker reaction times (which were displayed to the participant on the screen). This was evident as a large proportion of their responses were invalid (pressed before the stimulus appeared) or was less than 150 milliseconds (ms), and even less than 10 ms, clearly indicating a successful guess. This resulted in a change to the reaction time measure used. This change allowed for a slight pause in the task where feedback was provided to the participant that they needed to wait for the stimulus to appear before hitting the button.

With regard to the training tasks (experimental and control), there were no changes made from the pilot study to the main research experiment.

5.3 Participants

In a meta-analysis, Melby-Lervag et al. (2016) found that the average effect size (Hedge's *g*) for near-transfer in working memory training studies was 0.8, $CI_{95\%}$ [0.62, 0.97], for those studies that compared to a control group. Therefore, G-Power (Faul, Erdfelder, Lang, & Buchner, 2007) was used to calculate the required sample size to achieve a Cohen's *f* effect size of around 0.3 which is similar to the lower end of the confidence interval outlined by Melby-Lervag et al. (2016). As a result, a minimum total sample size of 90 was aimed for in order to detect a Cohen's *f* effect size of 0.30 at 80% power.

As part of the main research study, data was collected from 111 boys from a boys-only independent school in a large metropolitan city in Australia. This was a convenience sample as the researcher worked at the school where the study was performed. Participants were from third-grade and fifth-grade and were aged from 7.9 to 11.5 years of age (mean = 10.1 years; SD = 1.1) at the time of the first assessment. No exclusion criteria were implemented to maximise the generalisability of findings. After collection of consent forms, participants were listed alphabetically and then randomly allocated using a computer-generated random number table (https://www.random.org/sequences/) to either the experimental or active control group.

The main study took place between July 2016 and August 2018 within an independent boys-only school that educates from grade three through twelve. In the school there was a total enrolment of 1236 students, 330 of whom were enrolled at the junior campus (grade three through six). Data from the My School website (www.myschool.edu.au) indicates that students in the school could be classified as advantaged, with 83% of students in the top quartile and 16% of students in the second top quartile of the Index of Community Socio-Educational Advantage (ICSEA). The ICSEA is calculated using family background information provided to the school directly by families, including parental occupation, and the school education and non-school education levels they achieved (ACARA, 2013). According to data from the My Schools website, 16% of students at the school who sat the National Assessment Program - Literacy and Numeracy (NAPLAN), had a language background other than English.

Results for the grade 3 and grade 5 2012 NAPLAN were substantially above the Australian average for all tests. The My School website reported an overall student attendance rate of 95%.

5.4 Materials

Measures were used to assess participants' EFs, PS, gF, implicit theory of intelligence and academic achievement. The EF tests and PS test were completed individually using computer-based assessments in a small group situation, with a maximum group size of 15 and with the researcher supervising administration. Similarly, the gF, academic achievement and implicit theory of intelligence measures were administered in small group situations. Cognitive measures have been administered previously in small group settings and have shown no change to the reliability or validity of these tests (Brannigan & Brannigan, 1995; Golden, 1975).

To ascertain ecologically valid measures of EF, parent and teacher reports using behavioural rating scales were also used. All parents completed these behaviour-rating scales, whereas teachers completed these measures for a random sample of participants.

To individually measure EF, nine assessments were used. The participants were assessed using three measures each for inhibition, shifting and WM. Due to their high reliability for gathering EF data, computer-based tasks were used for assessment purposes (Bailey, 2012; McCabe, 2010; Piper et al., 2012). These EF tasks were either existing measures from the Psychology Experiment Building Language (PEBL; Mueller, 2012c) battery, slightly modified versions of existing measures, or were newly developed specifically for this study using PEBL. The PEBL files which include all parameters used in the target experiment for all nine EF measures and the one reaction time measure are available on the Open Science Framework (https://osf.io/qd79n/files/). All EF tasks were administered on identical iMac computers with 24” displays with a resolution of 1920 x 1080. Input was either by keyboard, numeric keypad or mouse.

5.4.1 Measures of EF

Inhibition tasks

Flanker task. The Flanker task has frequently been used as a measure of the inhibition component of EF in children (e.g., Huizinga et al., 2006; Lee et al., 2012). The PEBL Flanker task (Mueller, 2011b) was used as a measure for inhibition. Some modifications to the original task, as described by Stins, Polderman, Boomsma, and de Geus (2007), were made. Consistent with the protocol outlined by Stins et al. (2007), in the Flanker task participants were first presented with a white fixation cross for 500 ms, which was immediately followed by a horizontal array of five equally sized and spaced white arrows for 800 ms. The array was 10.5 cm wide. Participants were instructed to attend to the central arrow and ignore the arrows on either side of the target. Participants pressed the left shift key for a left facing central arrow and the right shift key for a right facing central arrow. The instructions for which key to press were on the screen throughout all trials. The arrows on either side of the target either all pointed in the same direction as the target arrow (e.g., “←←←←←”) or they all pointed in the opposite direction (e.g., “←← → ←←”) (see Figure 1). Congruent trials were when the arrows on either side of the target pointed in the same direction as the target arrow and incongruent trials were when they pointed in the opposite direction. Unlike the Flanker task used by Stins et al. (2007), the PEBL Flanker task also included neutral trials, where there was just one arrow pointed in either direction (e.g., "←"). Participants received a total of 60 trials (20 congruent, 20 incongruent and 20 neutral) in a random order, requiring an equal number of left or right responses. Prior to the experiment participants received 12 practice trials with corrective feedback.

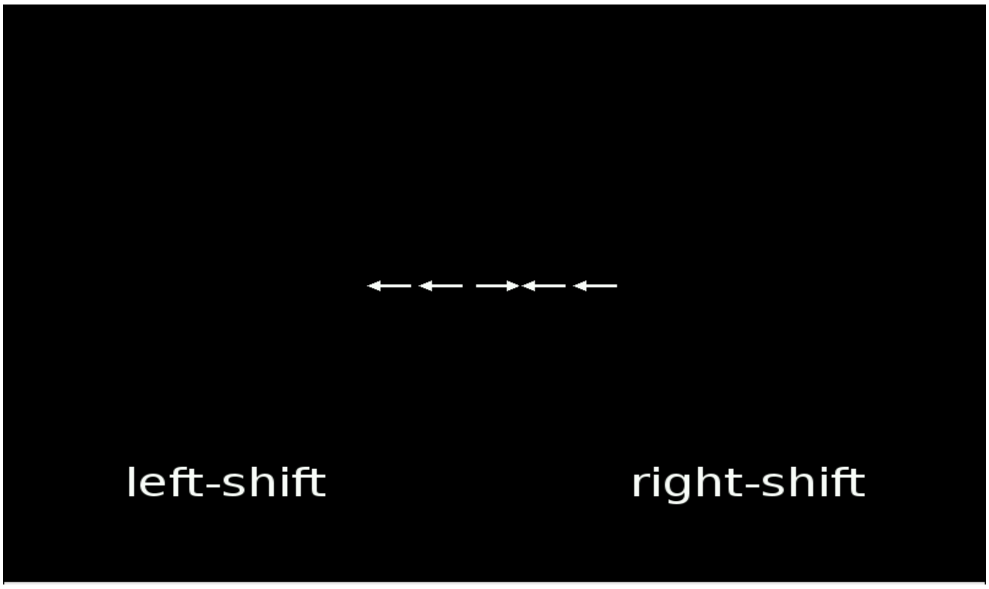


Figure 1. Flanker Task demonstrating an incongruent trial.

Accuracy and reaction times (RTs) for all congruence types were determined. Wöstmann et al. (2013) found that the test-retest reliability for the RT difference between the congruent-incongruent conditions for the flanker task was good (.94), but that test-retest reliability for the accuracy score was only moderate (.69). The internal reliability for the RT measure in the Flanker task has previously been found to be good (Cronbach's alpha = 0.86; Wöstmann et al., 2013). However, as Lee et al. (2012) found that accuracy provided a better indication of individual differences in EF for young children than does RT, the dependent measures used for the PEBL Flanker task was both the difference between the congruent-incongruent RTs and mean Flanker accuracy on the incongruent items.

Go/No-Go task. The PEBL Go/No-Go task (Mueller, 2011c) is a response inhibition task where a motor response must either be executed or inhibited. This task is commonly used as an inhibition component measure for EF (e.g., Miller, Nevado-Montenegro, et al., 2012; Rose et al., 2011; Willoughby, Wirth, Blair, & Family Life Project Investigators, 2012).

The PEBL task was based upon the instructions outlined by Bezdjian, Baker, Lozano, and Raine (2009). Consistent with the protocol outlined by Bezdjian et al. (2009), during this task, participants were required to focus upon a sequential presentation of letters and respond to a target letter by pressing the shift key on the keyboard. The task began with a 2 x 2 array with four stars (one in each square of the array). A single letter (P or R) was then presented in one of the squares for 500 ms duration with an inter-stimulus interval of 1,500 ms. In the first condition, participants were asked to press the shift key in response to the target letter P and to not respond to the non-target letter R (see Figure 2). The ratio of targets to non-targets was 80:20. The first condition consisted of 80 trials. A second, reversal condition was then administered, and participants were then asked to make a response to the target letter R and withhold their response to the non-target letter P. The ratio of targets to non-targets stays exactly the same during both response conditions. Together, the two conditions consisted of 160 trials in total. During both conditions, the instruction to press the shift key for the applicable letter was at the top of the screen. Prior to each of the conditions, the participants were administered a brief practice session of 10 trials with corrective feedback to ensure the task was fully comprehended.

The dependent measures were the proportion of correct No-Go trials (a high score indicating better inhibition) and the mean RT in the Go condition. In a computerised Go/No-Go task, Wöstmann et al. (2013) found that test-retest reliability for both Go condition RTs and proportion of correct No-Go trials has been found to be satisfactory (.78 and .84 respectively). They found good internal reliability (Cronbach's α = 0.92 and 0.87 respectively) for the Go/No-Go task (Wöstmann et al., 2013).

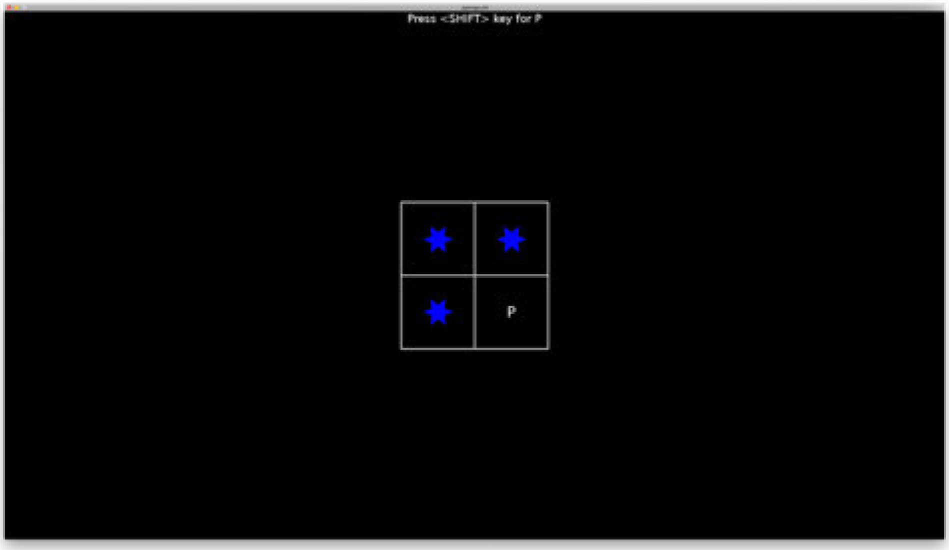
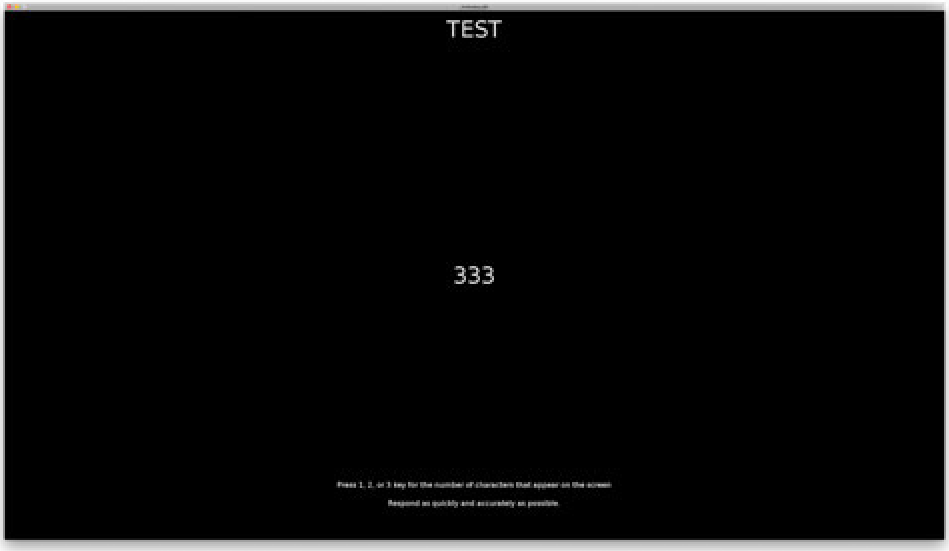


Figure 2. Go / No-Go task demonstrating the Go condition.

Numerical Stroop task. The third measure of inhibition was the PEBL Numerical Stroop task (Mueller, 2011d). Stroop tasks have been used extensively for the measurement of inhibition in childhood EF research (e.g., Agostino et al., 2010; e.g., Huizinga et al., 2006; Miller, Giesbrecht, et al., 2012; van der Sluis et al., 2007; van der Ven et al., 2012; Wu et al., 2011). The Numerical Stroop was adapted from the original colour word Stroop (Stroop, 1935) and has been used in previous research to measure the inhibition component of EF (Fournier-Vicente et al., 2008). Previous research with children has indicated that the Numerical Stroop task involves the same processes as the original Stroop task and achievement on both Stroop tasks is developmentally similar (Demetriou, Spanoudis, Christou, & Platsidou, 2001).

The PEBL Numerical Stroop version used in this study was based upon instructions outlined by Hernández, Costa, Fuentes, Vivas, and Sebastián-Gallés (2010).

Consistent with the protocol outlined by Hernández et al. (2010, p. 318), participants were requested to specify, as accurately and as quickly as possible, how many items appeared in each trial. The number of items ranged from 1 to 3, and participants had to press the keys 1, 2 or 3 on the keyboard with the index, middle and ring fingers of their dominant hand, respectively. There was a distractor variable with three conditions: (a) alphabetic characters (neutral condition: e.g., Z, GGG, MM); (b) digits whose value matched the number of items (congruent condition: 1, 22, 333) (see Figure 3); and (c) digits whose value did not match the number of items (incongruent condition: e.g., 2, 33, 111, etc.) (see Figure 4). The experiment consisted of two blocks of 84 trials each (preceded by a training block of 24 trials with corrective feedback). The three distractor conditions were represented the same number of times in each block in a random fashion. On each trial, a central fixation cross appeared for 1,000 ms, immediately followed by the target, which was presented for a maximum of 2,000 ms or until a response was given.

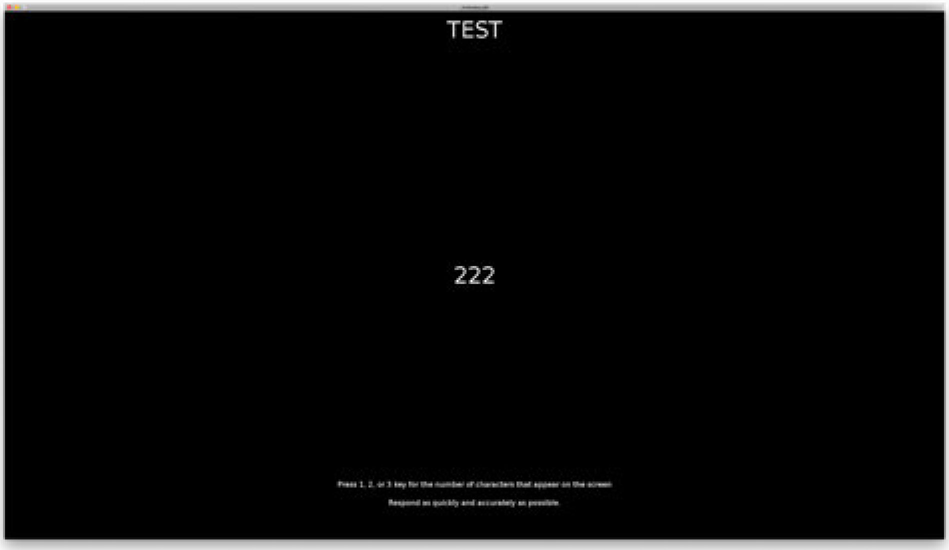


Press 1, 2, or 3 key for the number of characters that appear on the screen.

Respond as quickly and accurately as possible.

Figure 3. The Numerical Stroop demonstrating a congruent condition.

]



Press 1, 2, or 3 key for the number of characters that appear on the screen

Respond as quickly and accurately as possible.

Figure 4. Numerical Stroop demonstrating an incongruent condition.

The total time to name the quantity of items in each condition was recorded. The dependent measures were the accuracy for the incongruent condition and the mean RT for the incongruent condition. The only reliability data available for the Numerical Stroop was for internal reliability with good results for the mean RT for the incongruent condition (Cronbach's α = 0.84) (Lee, Ng, & Ng, 2009).

Shifting tasks

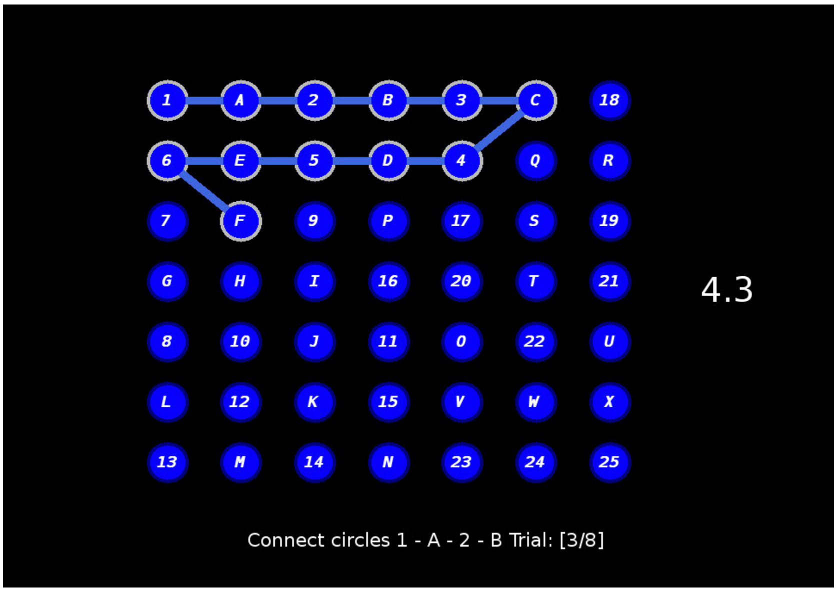
Connections test. It is commonly recognised that the Trail Making Test (TMT; Reitan, Sep-Oct 1955) is a reliable indicator of shifting ability (e.g., Atkinson, Ryan, Kryza, & Charette, 2011; Rose et al., 2011; Sanchez-Cubillo et al., 2009; van der Ven et al., 2012). Sanchez-Cubillo et al. (2009) argued that version B of the TMT, a more complex version which requires shifting between following a trail of letters and numbers in sequence, measures both WM and shifting ability. However, if determining a score by finding the difference between the score on version A, a simple trail following numbers only, and the score on version B, it was found that this provided a good indicator of shifting ability (Sanchez-Cubillo et al., 2009). However, before using the

TMT as a measure for shifting there are a number of limitations that should be noted, such as differing spatial arrangements of targets between versions and the influence of motor coordination and visual search abilities (Salthouse et al., 2000).

The Connections Test (CT; Salthouse et al., 2000) is a variant of the TMT that attempts to address many of the shortcomings of the TMT, whilst still measuring the same construct of shifting (Atkinson et al., 2011). As outlined by Salthouse (2011):

> [T]he differences between the Connections Test and the standard trail making test are: (a) irrelevant influences of visual search and hand movements are minimised because successive target elements are in adjacent locations rather than scattered around the page; (b) a counter-balanced presentation order of simple (A) and alternating (B) trials eliminates a confounding in which the alternating condition always follows the simple condition; (c) both letters and numbers are presented in the simple condition instead of only numbers, which eliminates a confounding of type of material and condition; (d) multiple trials are presented in each condition to increase the precision of the performance measures and allow reliability to be assessed; and (e) administration is efficient because a limit of 20 s is allowed to work on each page instead of monitoring the time until all items are completed. (p. 223)

As computerised TMTs have been used successfully with children as young as five years of age (Piper et al., 2012), the CT was a computerised version on PEBL (Mueller, 2012a). The CT consisted of a set of trials containing 49 blue circles in a 7 x 7 matrix, with either a white number or a letter within each circle and displayed on a black background. Similarly to the procedures outlined by Salthouse (2011), the task for the participant was to use the mouse to connect lines as quickly as possible to connect the elements in sequence, with different trials involving numeric, alphabetic, or alternating numeric and alphabetic sequences (see Figure 5). To reduce non-trivial motor movement, as the CT is a matrix-based trail-making test, each next response is adjacent to the previous response (either vertically, horizontally or diagonally). There were two practice trials on a 4 x 4 matrix for both numeric and alternating numeric and alphabetic sequences. There were eight trials consisting of two alphabetic sequences, two numeric sequences, two numeric-alphabetic sequences and two alphabetic-numeric sequences. The completion of simple and more complex sequences was counter-balanced. Throughout all practice sessions and trials there was a countdown clock (counting down from 20s) placed to the right-hand side of the matrix and instructions about the type of test being completed appeared underneath the matrix.



1	A	2	B	3	C	18
6	E	5	D	4	Q	R
7	F	9	P	17	S	19
G	H	I	16	20	T	21
8	10	J	11	O	22	U
L	12	K	15	V	W	X
13	M	14	N	23	24	25

Connect circles 1 - A - 2 - B Trial: [3/8]

Figure 5. The Connections task demonstrating the alternating condition.

Measurement consisted of switching completion counts; however, only estimated internal reliability measures are reported by Salthouse et al. (2003) for the CT with moderate results for the difference scores (Cronbach's α = 0.68). However, the CT has shown good construct validity (Atkinson & Ryan, 2008; Atkinson et al., 2010).

Berg card sorting task. Card sorting tests have been used to test the shifting component of EF in children (e.g., Miller, Giesbrecht, et al., 2012) and the Wisconsin

Card Sorting Test (WCST; Berg, 1948) has been confirmed as loading upon the shifting component of EF (Godinez, Friedman, Rhee, Miyake, & Hewitt, 2012; Miyake, Emerson, et al., 2000; Miyake, Friedman, et al., 2000). PEBL's Berg Card Sorting Task (BCST; Mueller, 2011a) is a computerised version of the WCST. The shortened version of this task displaying 64 cards instead of 128 cards was used for this study. This shortened PEBL version has been validated as a suitable alternative to the full version on participants 6-74 years of age (Fox, Mueller, Gray, Raber, & Piper, 2013).

The BCST consists of four stimulus cards from a set of 64 response cards that depict four forms (circle, crosses, triangles, and stars), four colours (red, yellow, blue, and green), and four numbers (one, two, three, and four). The test required participants to identify the correct sorting rule and preserve that rule across changing stimulus conditions. Four stimulus cards were presented to participants who then had to match the response cards with one of the four stimulus cards that they thought it matched (see Figure 6). Feedback was provided on the screen so the participant could see if their response was correct or incorrect. No other information was given to help figure out the correct matching principle. The rule (colour, shape or number) could switch as quickly as every fifth trial. The BCST was modified slightly to be a speeded version (a 3 second deadline for each trial) as per the instructions outlined in Kimberg, D'Esposito and Farah (1997) to generate greater variance in a nonclinical sample (Godinez et al., 2012).

Click pile to sort card

Figure 6. Berg Card Sorting Task. This shows the four stimulus cards (top row) and the response card (bottom centre).

The dependent measure in a card sorting test is typically the number of perseverative errors, which is defined as an incorrect response when the category has changed but the response would have been correct for the immediately preceding category (Piper et al., 2012). However, as outlined by Lin et al. (2013), the proportion of conceptual level responses (CLR; i.e., the total number of consecutive correct responses that occur in runs of three or more) is also a useful dependent variable as it indicates insight into the correct sorting rule. Both perseverative errors and CLR were used as dependent variables.

There was no reliability data available for a computerised version of the WCST; however, Piper et al. (2012) found that the pattern of age differences in the BCST was congruent with prior reports for the non-computerised WCST. Given that this same modified (time-limit of 3s per response) and computerised version of the WCST was used in the seminal work of Miyake et al. (2000), it was considered to be an appropriate measure for this study.

Switcher task. The third shifting task was the Switcher task. This was a computerised version of a category-switch task that was developed on PEBL (Mueller, 2012b). Similar to other category-switching tasks (e.g., Friedman et al., 2008), this task required participants to switch between responses based upon a presented rule that would change.

As outlined by Anderson et al. (2012), participants were presented with ten coloured shapes (with five distinct shapes and five distinct colours), each one having one of five letters on it (see Figure 7). Each object matched a single other object on only one dimension (colour, shape, or letter). There was one object circled and at the top of the screen and the participant was prompted to select a matching object based on the displayed characteristic of shape, colour or letter. Once the participant had correctly matched the object, a different feature was specified, to which they must "switch" and attempt to find the object that matched the current object based on that feature. Following a practice round, the Switcher task consisted of nine different configurations where the participants made ten responses. For the first three configurations, participants switched between two of the three features repeatedly. For the next three configurations, participants switched between all three features in a consistent order that differed for each configuration. For the final three configurations, features were switched randomly after each response, so that the next feature could not be anticipated prior to making the response. The variables recorded were time to complete each test, as well as the number of errors made.

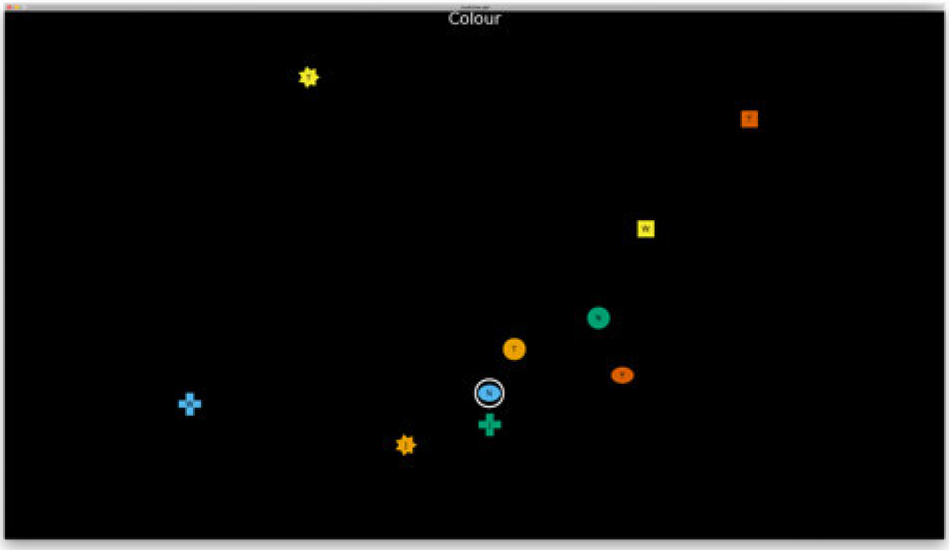


Figure 7. The Switcher task. This shows the one circled object and the prompt at the top of the screen to select a matching object based on colour.

The dependent measures in the Switcher task were an error cost (a difference between the number of errors in the 3 feature and 2 feature rule conditions) and the number of errors made in the 2 feature rule condition.

There is no psychometric data available for this specific Switcher task, although a similar task, the Category-Switch task, had an internal (split-half) reliability of .85 (Friedman et al., 2008). It is believed that the Switcher task conformed to the requirements of a shifting task, where a stimulus must be categorised according to one stimulus, but then the criterion for the response has changed to some other feature of the stimulus.

Working memory

Complex Span Tasks. As mentioned previously (see Section 2.2.3), WM should be measured using CSTs (Broadway & Engle, 2010; Redick, Broadway, et al., 2012; Unsworth et al., 2005). These measures stand in contrast to simple span tasks, which measure a STM capacity that involves storage only (Redick, Broadway, et al., 2012).

There was no need to break up the WM tasks into domain specific tasks (i.e., verbal and visuo-spatial) as CFA and SEM analysis has indicated that CSTs reflect a domain-general factor (Kane et al., 2004). As outlined by Kane et al. (1999), if the goal is to measure the domain-free executive component of working memory, it is important to administer a battery of WM tasks that differ in the domain-specific processes required to perform the task. This study used such tasks.

CSTs have been used to assess the WM construct of EF in children (e.g., Willoughby et al., 2012; Willoughby, Wirth, Blair, & Greenberg, 2010). The tasks used in this study were based upon those developed by Unsworth et al. (2005), Unsworth, Redick, Heitz, Broadway, and Engle (2009), and Broadway and Engle (2010). The three complex WM span tasks, which were programmed for use with PEBL and created specifically for this study, were operation span (OSPAN), reading span (RSPAN) and symmetry span (SSPAN). All CSTs have previously been found to be reliable and valid measures of WM (as shown in Table 1).

Table 1.
Test-Retest and Internal Reliabilities of Complex Span Tasks (CSTs) Partial Score.

	OSPAN	RSPAN	SSPAN
Test-retest reliability	0.83	0.82	0.77
Internal reliability	0.84	0.86	0.76

Note. OSPAN = operation span; RSPAN = reading span; SSPAN = symmetry span. Reliability data from Kane et al. (2004) and Unsworth et al. (2005; 2009).

As outlined by Redick, Broadway, et al. (2012), for all automated CSTs there were three practice conditions before starting the real task: (a) storage-task only, (b) processing-task only, and (c) processing and storage tasks interleaved. An upper limit on processing time is calculated during the processing-task only practice trials– the participant’s mean response time plus 2.5 SDs (Redick, Broadway, et al., 2012). This upper limit is used because CSTs with unlimited processing times do not predict higher-order cognition when compared to CSTs in which there is a processing time limit (Friedman & Miyake, 2004).

The design and administration of all CSTs followed the descriptions and protocols outlined by Redick et al. (2012), Unsworth et al. (2005), and Unsworth et al. (2009), although shortened versions (only one set of 3 - 7 for RSPAN and OSPAN and one set of 2 - 5 for SSPAN) of these tasks were used to reduce the time demands on participants (Foster et al., 2014).

Figure 8. Storage retrieval task component of operation span (OSPAN) and reading span (RSPAN).

Operation span task. In the OSPAN task, participants attempted to mentally solve mathematical problems while also remembering letters (F, H, J, K, L, N, P, Q, R, S, T, Y) in serial order for subsequent memory tests (see Figure 8). In all experimental conditions, letters remained on-screen for 1,000 ms. At recall, participants saw a 4 x 3 matrix of letters and recall consisted of each letter in the correct order. Participants were presented with three to seven items on each trial, with one trial at each list length, for a total maximum score of 25 letters recalled in the correct serial position across trials.

Reading span task. In the RSPAN task, participants were required to read sentences while trying to remember the same set of unrelated letters as OSPAN. The difference was that participants were required to read the sentence and determine whether it made sense instead of solving mathematical problems. Sentences were shortened as is common practice for using the RSPAN with children (Cowan et al., 2003; Delaloye, Ludwig, Borella, Chicherio, & de Ribaupierre, 2008; Gonthier, Aubry, & Bourdin, 2017). There were 3-10 words in each sentence (Mean = 5.5, SD = 1.6). There was one trial of each set-size with list length ranging from three to seven.

Symmetry span task. In the SSPAN task, participants were required to perform a symmetry-judgement task while also remembering sequences of red squares within a 4 x 4 matrix. The same timing parameters used in the OSPAN and RSPAN were used. Participants decided whether the presented matrix was symmetrical and then were immediately presented with a 4 x 4 matrix with one of the cells filled in red for 650 ms. For the recall component of the task, participants selected the sequence of red-square locations in the preceding displays, in the order they appeared by clicking on the cells of an empty matrix. There was one trial of each set-size with list length ranging from two to five.

The same scoring procedure for all CSTs was used. As outlined by Redick et al. (2012), five scores are determined for each of the CSTs: (1) absolute storage score, which is the sum of all trials in which all items were recalled in the correct serial order; (2) partial storage score, which is the sum of items recalled in the correct serial position, regardless of whether the entire trial was recalled correctly; (3) processing errors, which are the total number of errors made on the processing task; (4) speed errors which are the number of processing problems that were not answered before the individualised time limit; and (5) accuracy errors, which are the number of processing problems that were answered incorrectly (note that processing errors = speed errors + accuracy errors). Research indicates that the psychometric properties of the partial storage scoring method are the best (Redick, Broadway, et al., 2012) so this was the dependent variable used for each CST.

5.4.2 Processing speed

Research by Cepeda, Blackwell, and Munakata (2013) has found that a simple RT measure is preferred for measuring PS because it requires minimal EF, motor control and pre-existing knowledge. A modified version of the PEBL Perceptual Vigilance Task (Mueller, 2007) was used to create a simple RT (SRT) task to measure PS, which has previously been used to measure PS with children (e.g., McAuley & White, 2011). The SRT task presented a single stimulus (a red circle) at a specified delay (250-3,000 ms at 250 ms intervals) from the previous response and the participant had to press the spacebar as quickly as possible after the stimulus appeared. There were 6 sets at each specified stimulus delay, presented randomly, giving a total of 72 trials. Response times considered too fast (< 150 ms) or too slow (> 3,000 ms) were not considered in the reaction time statistics and corrective feedback was provided for responses that were too fast or too slow. The RT was measured by the interval between stimulus onset and response. RT data was captured for each of the stimulus delay intervals, too fast (< 150 ms) responses, lapses (> 500 ms) and too slow responses (> 3,000 ms).

SRT measures have been found to have reasonable test-retest reliability in a wide range of ages from very young children (*r* = 0.79; Weissberg, Ruff, & Lawson,

1990) to young adults (*r* = 0.96; Baker, Maurissen, & Chrzan, 1986) to older adults (*r* = 0.9; Lemay, Bédard, Rouleau, & Tremblay, 2004).

5.4.3 Fluid intelligence

The Naglieri Nonverbal Ability Test – Second Edition (NNAT2; Naglieri, 2008) is a nonverbal measure of general cognitive ability that is available in grade-based levels for students in grades K-12. This study used three of these grade-based levels (levels D – E). Age-based norms are available for the NNAT2 that were developed from a nationally representative (U.S. population) sample of more than 100,000 students (Naglieri, 2011).

The NNAT2 is a group-administered paper and pencil test that uses 48 figure matrix items at all levels and takes 30 minutes to complete (Naglieri, 2011). The NNAT2 provided an age-based Naglieri Ability Index (NAI), which has a mean of 100 and a standard deviation of 16 (Naglieri, 2011).

As the NNAT2 uses progressively more difficult matrices which are believed to measure inductive reasoning, it is believed to be a good measure of gF since inductive tests typically serve as markers of the fluid factor of intelligence (Klauer, Willmes, & Phye, 2002).

The test-retest reliability of the NNAT2 varies from 0.75 - 0.78 for the levels used in this study, whereas the internal reliability (K-R 20) varies from 0.84 – 0.87 (Naglieri, 2011).

5.4.4 Implicit theory of intelligence

An Implicit Theory of Intelligence Measure (ITIM), as outlined by Blackwell, Trzesniewki, and Dweck (2007) and Dweck (2000), was used to measure participants' implicit theories of intelligence (see Appendix D- Section 9.4.1). This measure has been used previously to measure primary school-aged children's implicit theory of intelligence (Miele, Son, & Metcalfe, 2013). The ITIM contains items on a 6-point Likerttype scale from 1 (*Agree Strongly*) to 6 (*Disagree Strongly*). The scale consists of six items of which three are entity theory statements and three are incremental theory statements. The three entity theory statements are: (a) "You have a certain amount of intelligence and you really can't do much to change it"; (b) "Your intelligence is something about you that you can't change very much"; and (c) "You can learn new things, but you can't really change your basic intelligence." The three incremental theory statements are: (a) "No matter who you are, you can change your intelligence a lot"; (b) "You can always greatly change how intelligent you are"; (c) "No matter how much intelligence you have, you can always change it quite a bit". The incremental theory items are reverse-scored and a mean theory of intelligence score is calculated for the six items, with lower scores reflecting an entity theory and higher scores reflecting an incremental theory.

The internal reliability of the ITIM measure ranges from 0.78 (Blackwell et al., 2007) to 0.98 (Dweck, Chiu, & Hong, 1995). Da Fonseca et al. (2007) found that Cronbach's α depends upon which items are being measured, with the incremental items (0.81) being slightly more reliable than the entity items (0.78). Test-retest reliability for the ITIM over a 2-week period ranges from 0.77 (Blackwell et al., 2007) to 0.80 (Dweck et al., 1995).

5.4.5 Improvement expectancy

As outlined by Boot et al. (2013), placebo effects in experimental psychology are pervasive. Unless the expectancies of improvement in the dependent variable of interest in both the experimental and control groups are similar, then causal conclusions about the intervention are limited (Boot et al., 2013).

A simple Improvement Expectancy Measure (IEM), as outlined by Boot et al. (2013) was used to measure improvement expectancy in both the experimental and control groups (see Appendix D – Section 9.4.2). Questions were: (a) “Do you think that repeated practice with your training task will help you improve in mathematics (This might be any part of mathematics e.g., addition, subtraction, problem-solving)?” (IEM Maths); (b) “Do you think that repeated practice with your training task will help you improve in English (This might be any part of English e.g., comprehension, spelling)?” (IEM English); (c) “Do you think that repeated practice with your training task will help you improve in your general concentration in class (This might be being able to resist distraction, being able to move from one task to the next or being able to have better memory)?” (IEM Conc.). These measures were designed to determine potential equivalency for participants’ expected improvement on far-transfer (measured by (a) and (b)) and near-transfer (measured by (c)) measures.

If participants answered “Yes” to each of the above questions, they were then asked to indicate on a scale of 1 (“*A Little*”) to 6 (“*A Lot*”) how much they think that improvement would occur. A score was then determined from 0 (indicating participants believe there will be no improvement) to 6 (“*A Lot*” of improvement) for each of the questions.

5.4.6 Parent / teacher behavioural reports of EF

The Comprehensive Executive Function Inventory (CEFI; Naglieri & Goldstein, 2013) is a 100-item rating scale with nine separate scales designed to measure behaviours associated with EF in children and youths aged 5 through 18 years. Only three of the scales were used for this study, each of which match the potential near-transfer effect of cognitive training on behavioural aspects of the core EFs of inhibition, shifting and WM respectively: (a) Attention Scale – avoidance of distractions, concentration on tasks, and sustaining attention; (b) Flexibility Scale – adaptability to circumstances; and (c) Working Memory Scale – keeping important information in mind and remembering important instructions and steps. The Attention Scale consists of 12 items, the Flexibility Scale consists of 7 items, and the Working Memory Scale consists of 11 items, resulting in 30 items in total. The rating scale was completed by a parent and by a random sample of teachers. The CEFI also has a self-report measure for children older than 12 years of age, but this version was not used for this study.

The CEFI was normed on a nationally representative (U.S. Population) sample of teachers (N=1400) and parents (N=1400). Internal reliability (Cronbach's α) for the parent report is 0.92 (Attention Scale), 0.84 (Flexibility Scale), and 0.88 (Working Memory Scale; Naglieri & Goldstein, 2013). The teacher report internal reliability is 0.96 (Attention Scale), 0.90 (Flexibility Scale), and 0.94 (Working Memory Scale; Naglieri & Goldstein, 2013). Test-retest reliability over a 1- to 4-week interval is good, ranging from 0.80 - 0.89 for the parent reported scales to 0.82 - 0.91 for the teacher reported scales (Naglieri & Goldstein, 2013). As reported by Naglieri and Goldstein (2013), inter-rater reliability is moderate for teacher-teacher ratings (range 0.61 - 0.63) and good for parent-parent ratings (range 0.76 - 0.86).

5.4.7 Academic achievement

Two versions of the Australian norm-referenced Progressive Achievement Test (PAT) were used to assess academic achievement in reading comprehension (PAT Reading; Australian Centre for Educational Research, 2008) and mathematics (PAT Maths; Australian Centre for Educational Research, 2013).

PAT Reading. The purpose of this test was to measure comprehension, including both factual and inferential comprehension of written text. The PAT Reading consists of a series of passages from a variety of text types, which the participant is required to read and then answer related questions. Depending on the version of the sub-test, there are 30 to 35 multiple-choice items that takes up to 40 minutes to administer. There were four different versions of the test used across the timeframe of this study.

The PAT Reading results in a norm-referenced score that ranges from 18.9 to 189.5. As this same scaled score is used across all test versions, this allows for a comparison of achievement over time. The PAT Reading was normed on a total of over 9,200 Australian students from Years two to ten (Australian Centre for Educational Research, 2008). The students were from all states and territories of Australia and from all types of school (Australian Centre for Educational Research, 2008).

PAT Maths. This is a test of understanding and skill in mathematics. The test consists of 35 to 39 multiple choice questions, which assess mathematical skills within the areas of number, space, measurement, chance and data, and algebra. There were four different versions of the test used across the timeframe of this study.

The PAT Maths results in a norm-referenced score that ranges from 52 to 189. As this same scaled score is used across all test versions, this allows for a comparison of achievement over time. The PAT Maths was normed on a total of 12,296 Australian students from grade three to ten (Australian Centre for Educational Research, 2005). The students in the norming sample were from all states and territories of Australia and from all types of school (Australian Centre for Educational Research, 2005).

With regards to the external validity of the PAT, in a high school sample, the correlations between PAT scores and subject grades has been reported to be 0.65 - 0.78 for the PAT Reading and 0.65 - 0.87 for the PAT Maths (Australian Centre for Educational Research, 2007). Both tests have been found to be predictive of school performance in their relevant subject areas (Australian Centre for Educational Research, 2007). There was no validity data available for the PAT on primary school students.

5.4.8 Cognitive training task

A number of different commercially available cognitive training programs were considered during the scoping phase of this study to determine an appropriate program (e.g. CogMed (www.cogmed.com), Lumosity (www.lumosity), Jungle Memory (lb.junglememory.com)). However, none of these claimed to train EF and they appeared to focus predominantly on short-term storage training, which, as outlined previously, is not related to higher order functioning (Daneman & Carpenter, 1980; Daneman & Merikle, 1996).

Ultimately, the training program chosen for this study was the commercially available C8 Sciences program (www.c8sciences.com). This computerised web-based program claims to train EF (Wexler et al., 2020), which includes (a) sustained attention, (b) response inhibition, (c) speed of information processing, (d) cognitive flexibility and control, (e) multiple simultaneous attention, (f) working memory, (g) category formation, and (h) pattern recognition and inductive thinking (C8 Sciences, 2011). Although these C8 identified EFs are more than the core EF abilities referred to in this research program, they clearly match the three core EF inhibition (e.g., response inhibition), shifting (e.g., cognitive flexibility and control) and WM components (see Sections 2.2.1 and 2.2.2 for different terminology referring to both inhibition and shifting). In addition to the core EF, this training program also targeted PS (e.g., speed of information processing).

To target these cognitive capacities, C8 has six training programs in the form of games: (a) Treasure Trunk; (b) The Magic Lens; (c) Pirate Pete’s Packing Panic; (d) Ducks!; (e) Grub Ahoy!; and (f) Monkey Trouble. Notably, both Treasure Trunk and The Magic Lens exercise the same cognitive skills and work very similarly with only the playing experience differing. The same is the case with both Grub Ahoy! and Monkey Trouble.

Treasure Trunk and The Magic Lens. In both Treasure Trunk and The Magic Lens training programs participants have to track an object as it moves around the screen and click on it when it changes into a specified target and before it disappears again. When the participant gets enough correct responses, the object begins to move faster and they can move up levels. If several mistakes are made the target will move slower to make it easier. For Treasure Trunk the participant is tracking a light that changes into a jewel, whereas in The Magic Lens the participant is tracking a lens that can reveal a monkey trapped inside a crate. As the participant moves up levels the game becomes more complex and demanding, e.g., only clicking on certain targets and ignoring others as well as choosing the matching missing half of a target. There are 163 task levels in both tasks. The first 27 levels all have only one target on the screen at a time. Levels 28 to 55 go through the same task variations but with two targets on the screen at once. Once the participant completes all the 2-target levels, the program goes back to the most difficult of the 1-target levels. This is to allow for consolidation of skill development and mastery of additional elements present but not mastered when first working on the easier task. After returning to the 1-target levels, the participant works through the 2-target levels again before moving on to 3-target levels. After completing these, the participant returns to the difficult 2-target levels and works their way up through the levels again. Throughout all levels, feedback is provided to ensure participants are following task requirements.

Treasure Trunk and The Magic Lens both exercise inhibition beginning in level 4 by requiring participants to ignore the moving target when it turns into a specified non-target but click on it as soon as possible when it turns into a specified target. The need for inhibition is increased and a shifting component is introduced in level 8 by having the target switch back and forth between items, so targets that were salient to click on just moments before have to be ignored. WM is exercised beginning in level 16 when the participant has to click on a target only if it matches the same feature as the one just before it. Demands on WM and shifting are more challenging at 2 (levels 28 – 55; 68 – 95; 124 – 135) and 3 (levels 95 – 123; 136 – 163) target levels, because each target has its own history and the current aim for one target may be different for another target.

Pirate Pete's Packing Panic. In Pirate Pete's Packing Panic (PPPP) training program, participants have to identify items that belong to different categories. As the items move across the screen, participants have to click on the items that belong in the specified category. As the training progresses, the categories change more quickly, growing more abstract, or begin rotating sequences that have to be remembered. The items carry signs with numbers, letters or pictures on them and participants have to click on signs that match a specified category before the item moves off the screen. For example, if the category was numbers, then participants have to click on all the items with numbers on them. If the category was animals, they have to click on all items with pictures of animals. The items move faster when the participants make several correct answers, and slow down if they make mistakes. The number of items on the screen increases gradually up to a maximum of six.

There are 186 levels of PPPP, which are defined by the number of items on the screen, the nature of the target category and additional cognitive tasks added to the required categorisations. PPPP introduces the concept of categories and then exercises the use and formation of categories at all levels of the game. Shifting and inhibition are required when the target category changes and pictures that were previously targets are no longer (e.g., levels 82-84, "Plants but not Trees"). Shifting and working memory are exercised when there are multiple items on the screen and the participant has to search for more than one target category at the same time (e.g., levels 46-48, "Plants or Animals"). Shifting and WM are important when the participant is required to move rapidly from one category to another (e.g., levels 94-96, "Food alternating with Clothes") and to find pairs of things in the same category without a specified target category (e.g., levels 118-120, "Find Two in the Same Category"). Shifting and WM are also necessary when the target category moves back and forth between two or rotates among three possibilities (e.g., levels 154-156, "Furniture, Sports, Machines in Rotation").

Ducks! In Ducks! training program, participants are asked to identify patterns based on information from three symbols, and then choose the correct option to complete the pattern. The target patterns progress through shapes, colours, numbers, images, clocks, and geometric patterns.

In Ducks! participants see three objects in a row with a blank slot in the fourth position. In a second row below the first they see three more objects. Their task is to select an object from the second row that matches the pattern created by the objects in the first row. Participants are given a limited amount of time to respond. A small timer in the upper right of the screen informs how much time is left. When participants click on the correct object it moves to the top row to complete the pattern. If a mistake is made the program shows the correct response.

Instructions are provided at the beginning of each new level. As participants make correct responses, the time allowed for response becomes shorter; as they make mistakes they are given more time. When they get enough correct answers at fast speeds, or when the program determines they had reached the fastest speeds and accuracies of which they were capable, the program moves them to different types and more complex objects and patterns. For example, at the beginning of the training all three objects in the top row may be identical, so the participant is required to click on that same object on the second row. This teaches them the basic principles and mechanics of the game. In the next phase of training, the objects may be the same shape but differ in colour, so the participant has to choose the same shape from the bottom row even if it is different in colour from others on the first row.

There are 75 levels in Ducks!. Starting at level 9 participants are required to complete sequences (e.g., ABAB or BABA) based upon colours or shapes. At level 18 number patterns are introduced and geometric shape patterns are introduced at level 24. At level 30 pictures are introduced and participants have to complete the pattern based upon the categories (e.g., food, clothing, food, clothing). There are then clock patterns introduced at level 41 and the remaining levels are a variety of combinations of all previously introduced tasks at increasing levels of difficulty (e.g., sequences based on shapes or colours, geometric shapes, complex picture category puzzles, complex number patterns and clock position patterns in 15, 20, 30 and 45 minute intervals).

In Ducks! shifting and inhibition are engaged frequently as the same response options are included in many trials, but one that is correct for one trial is wrong for the next. Some response options also share some features with items in the top pattern but are not actually the correct choice; the shared features draw the incorrect option to attention and response to it has to be inhibited while searching continues for the correct answer. The initial levels of WCN introduce one pattern principle at a time and move to more complex principles or patterns at subsequent levels. At more advanced levels, trials with different types of patterns, stimuli and sorting principles are intermixed, demanding more shifting ability.

Grub Ahoy! and Monkey Trouble. These games are classified as spatial working memory training exercises, much like the types of activities required in many working memory training programs (Shipstead, Hicks, et al., 2012). The participant sees a visual pattern appear on screen, suggested by the path of a group of pirates asking for grub to eat (Grub Ahoy!) or a mischievous monkey (Monkey Trouble). If participants get the pattern correct, they repeat the exercise with a longer pattern. If they fail to remember the correct pattern, the pattern grows shorter. After a certain time the patterns will require a reverse response where the participant needs to match the correct pattern in reverse.

For all of the games described above, the difficulty of the C8 training program is adaptive in that it adjusts its difficulty every 10 seconds according to the participant's ability. It provides corrective messaging to help the participant learn faster and its algorithms also track progress to pick up plateaus in participant achievement, which it then adapts to help accelerate progress. Time spent training is controlled by the researcher within the program, with participants required to train for five minutes per game and four games per training session for a total of 20 minutes training per day. There is a countdown timer present in the top right of the screen, providing a continual indication of how much time is left to train on each game. Although the participant has a choice in what game they choose to play, if they do not play enough of one game then algorithms in the game force the choice for them.

Importantly, the computerised training program is not the only component of the C8 Sciences program. There is also an aerobic exercise component that includes a complete physical exercise curriculum that is meant to be administered at the same time as the computerised cognitive training program (“C8 Sciences: Physical Exercise Programs”, n.d.). Although the efficacy of exercise programs ability to improve EF is outlined in Section 3.2.1, this aspect of the C8 Sciences program was not used in this study. The main reason that the exercise component of the C8 Sciences program was not included in this study was due to comparability. Keeping this study as a computerised cognitive training task allowed for comparability with other computerised training studies.

5.4.9 Control group task

TypingClub (www.typingclub.com) is an online program that allows participants to learn touch-typing. Participants were allocated tasks by the researcher that were completed during the same time as the cognitive training tasks. Although not game-based, the tasks are visually engaging with points and stars allocated for task completion (typing accuracy and speed) that encourage the participants throughout the program.

As soon as participants logged into the TypingClub program, they see the tasks that are available to them. Participants see what tasks they have completed and what achievements they have for those tasks. The tasks available to participants are graded according to difficulty introducing one or two keys at a time. All participants had similar experiences within the TypingClub program. Participants worked on the TypingClub activities for the same time as the cognitive training program participants.

5.5 Procedure

Participants from grade three and grade five were randomly allocated to one of two experimental or two control groups in the two years of the experiment, resulting in four groups for each. Control and experimental group pairings then formed four waves for the experiment.

Participants from each wave were initially assessed on all measurements. The PEBL EF battery of tests (see Table 2), which also included the PS measure (Simple Reaction Time), was completed in groups in a computer lab located on the school premises.

Table 2.
Summary of Executive Function (EF) Tasks.

	Executive Function Domain		
	Inhibition	Shifting	Working Memory
PEBL Task	(1) Flanker	(2) Connections Test	(3) Operation Span
	(4) Go/No-Go	(5) Berg Card Sorting	(6) Reading Span
	(7) Numerical Stroop	(8) Category Switch	(9) Symmetry Span

Note. Numbers indicate order that tests were administered – see Table 3 below. There was also a tenth test, to measure processing speed: (10) the Simple Reaction Time task.

In order to prevent any order or sequence effects, the order of EF battery tests were counter-balanced using a balanced Latin-square design (see Table 3). Participants were randomly allocated to one of 10 groups that determined the order of EF (including PS) battery administration.

Table 3.
Balanced Latin-Square Design.

Group 1	1	2	10	3	9	4	8	5	7	6
Group 2	2	3	1	4	10	5	9	6	8	7
Group 3	3	4	2	5	1	6	10	7	9	8
Group 4	4	5	3	6	2	7	1	8	10	9
Group 5	5	6	4	7	3	8	2	9	1	10
Group 6	6	7	5	8	4	9	3	10	2	1
Group 7	7	8	6	9	5	10	4	1	3	2
Group 8	8	9	7	10	6	1	5	2	4	3
Group 9	9	10	8	1	7	2	6	3	5	4
Group 10	10	1	9	2	8	3	7	4	6	5

There were two separate individual assessment sessions required to measure EF ability, which were broken up into two separate sessions with an equal number of assessments in each session. Each testing session lasted approximately half an hour.

After the EF battery was completed, participants were administered the NNAT2, ITIM, PAT Reading and PAT Maths in groups. These assessments were completed in a classroom with the researcher providing instructions and supervising administration. The NNAT2 and ITIM were completed together on one day, followed by the PAT Reading and PAT Maths which were completed on separate days in the order listed. Each of the three sessions took approximately 30 – 40 minutes. The IEM was completed after two weeks of training and again at the end of the training period.

The CEFI was sent home for completion by a parent / guardian at the beginning of the assessment phase. All completed parent / guardian CEFI forms were returned prior to the beginning of the training regimen. A random 50% of the experimental and control groups were allocated to have the CEFI assessment completed by their teacher. This same student then had a CEFI completed by a teacher for the subsequent assessment occasions (2 - 4).

Each wave started on different occasions throughout the first year of the study. Wave 1 started at the beginning of the year and progressed through one assessment time (Assessment Time 1) and then began the training program (Training Regimen 1) immediately afterwards, followed by another assessment (Assessment Time 2).

Prior to the beginning of the training program, both the experimental and control groups were provided with an instructional session to ensure treatment fidelity, referring to both structure (i.e., adherence, duration) and process (i.e., quality of delivery) (Mowbray, Holter, Teague, & Bybee, 2003). In addressing structural fidelity, the training tasks in the experimental and control groups were explained. Additionally, the duration component of structural fidelity was addressed within the training programs as the researcher could control time limits for each training session. Lastly, adherence was checked by the researcher on a day-to-day basis as both the experimental and control group training programs allowed the researcher to determine whether participants were completing the assigned training tasks. If participants were absent for a particular day's training, they would simply keep going from where they were up to the following day.

With regard to process fidelity, the participants were given a number of guidelines for where, when and how the training programs should take place. These were summarised as follows:

(a) Environment: the location for training would be in the school computer lab and there would be no distractions (e.g., no talking).
(b) Timing: training would take place early in the morning, just prior to the start of the school day.
(c) Breaks: participants would endeavour to complete the training in one sitting without extended breaks (although short breaks were allowed).
(d) Effort: participants should give all tasks their best effort.

The training program went for six and a half weeks with both the experimental and control groups training at school. Training went for 20 minutes each morning and was completed five days per week for the full six-and-a-half-week training period. The six-and-a-half-week training period was used in an attempt to have all participants complete a minimum of nine hours of training, the minimum amount of training found to function as a significant moderator (Melby-Lervag & Hulme, 2013). Immediately after training, both experimental and control groups were assessed again (Assessment Time

2). Wave 2 began the same process at the beginning of the following term. Waves 3 and 4 followed the same process in the following academic year. One year after beginning Assessment Time 2, each wave completed another assessment (Assessment Time 3).

On each assessment occasion (1 to 3), the participants completed the EF battery, NNAT2, PAT Reading and PAT Maths as outlined above. On each assessment occasion every parent/guardian completed the three CEFI scales and teachers completed the selected CEFI scales on the random sample of participants as outlined above. IEM data was collected two weeks after beginning the training intervention so that participants had an idea of what the training was going to be like. IEM data was collected again on the last training intervention session to determine if there was a change in improvement expectancy as a result of training. An overview of the assessments completed can be seen in Table 4.

Table 4.
Assessments Completed for Times 1 - 3.

Individual Participant	Teacher	Parent / Guardian
EF Battery (incl. PS)	CEFI (random allocation)	CEFI
NNAT2		
PAT Reading		
PAT Maths		
ITIM[a]		
IEM[b]		

Note. EF = executive function; PS = processing speed; NNAT2 = Naglieri Nonverbal Achievement Test – Second Edition; PAT = Progressive Achievement Test; ITIM = Implicit Theory of Intelligence Measure; IEM = Improvement Expectancy Measure; CEFI = Comprehensive Executive Function Inventory.

[a] ITIM only completed at Time 1.

[b] IEM only completed two weeks after starting training and again before the last training session.

Overall, for each wave, the training and assessment period went for just over one year from the beginning of Assessment Time 1 to the completion of Assessment Time 3. Treatment took place at school, and all sessions were completed during the morning in school hours. Training periods were planned to be completed within a single school term, to ensure that training sessions would not be interrupted for a longer period of time due to holiday periods. Post-treatment assessment took place within one week after the last training session and one-year post assessment took place one-year after the last training session (within one week).

5.6 Statistical Analyses

As outlined in the Rationale (Section 4), the use of single EF measures will be contaminated by error variance. Therefore, the multiple measures for EF will be used to construct a latent variable model which will then be used to create factor scores.

To determine if the intervention is successful for near-transfer (*Hypothesis 1*) and far-transfer (*Hypothesis 2*), an analysis of covariance (ANCOVA) will be used to determine if scores at the completion of training and scores one year after training completion are different between the control and experimental groups, with baseline scores entered as a covariate.

To determine if the participant's implicit theory of intelligence is moderating potential training gains (*Hypothesis 3*), a three-way interaction will be performed with post-training factor scores as the outcome variable, pre-training factor scores as the predictor, and both ITIM and experimental group as moderator.

6 Results

6.1 Data Collection

Between April 2015 and August 2018, a total of 111 participants (57 control group; 54 experimental group) received the intervention and were assessed at baseline (T1), at training completion (T2) and again one year after training completion (T3). The CONSORT (Consolidated Standards of Reporting Trials; Moher et al., 2001) flow diagram (Figure 9) shows the flow of participants through the study.

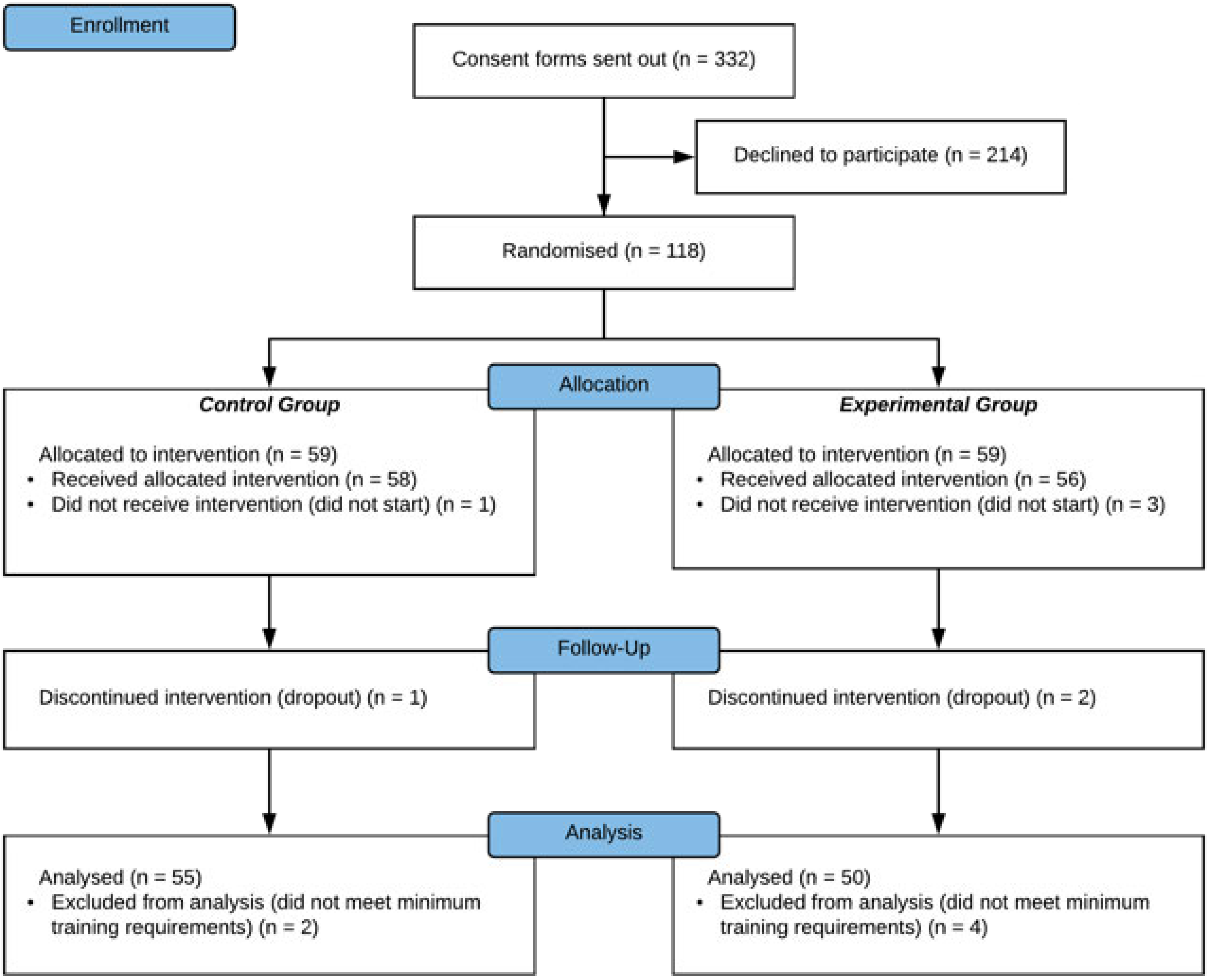


Figure 9. Consolidated Standards of Reporting Trials (CONSORT) flow diagram.

6.2 Data Summary

Table 5 presents summary data for each EF measure, achievement tests and behavioural report measures at T1. There was a small amount of missing data at each time point for EF data, which was due to unexplained computer errors and data not being recorded. There was less than 0.7% missing EF data at T1 and T2 and less than 0.5% missing EF data at T3. There was one data point in the BCST perseverative error that was Winsorized (placed just beyond the next nearest value and so maintaining rank ordering), similar to criteria used elsewhere (Cirino et al., 2018). This was due to the participant's poor performance on the T1 card sorting task (where no categories were achieved, thus no perseverative errors). Data analysis was performed in RStudio (Version 1.2.1335; RStudio Team, 2018) using R (Version 3.6.0; R Core Team, 2019) and both the data and the data analysis script are freely available (see Appendix E - Section 9.5).

6.3 Data Tidying

As per the normal administration of automated CSTs (Unsworth et al., 2005), data was removed due to poor performance on the distractor task. Although other researchers who use CSTs to measure WM do not mention the removal of CST data based upon the distractor task accuracy (e.g., Chooi & Thompson, 2012; Gonthier et al., 2017; Miyake, Friedman, et al., 2000; Schmiedek et al., 2009), it was evident that some participants were sacrificing accuracy on the distractor to enhance performance on the recall tasks. As this effectively changes the meaning of the task and potentially what it is measuring, it was decided to remove these participant's data, as per the original design of the tasks (Unsworth et al., 2005). Although the initial research by Unsworth et al. (2005) specified an accuracy criterion of 85%, this removed a significant amount of the data in this study (51%, 24% and 38% of data across T1 - T3 on operation, reading and symmetry tasks respectively). Reducing the accuracy criterion for CSTs has been used previously in cognitive training research (e.g., Blacker, Negoita, Ewen, & Courtney, 2017), so it was determined to reduce the accuracy criterion for the OSPAN, RSPAN and SSPAN tasks to 72%, 84% and 71% respectively. This approximated the 15% of data removal obtained by Unsworth et al. (2005) in their development of the automated OSPAN task. This procedure appeared to remove the obvious cases where participants had sacrificed distractor accuracy to enhance recall accuracy.

Data was also removed for those participants who did not achieve the minimum set number of hours of training. As Melby-Lervag and Hulme (2013) found a greater effect size for training greater than nine hours, this was determined to be the cutoff. This resulted in the removal of six participants' data (four experimental and two control group) who had not completed a minimum of nine hours. After the completion of data removal, the final number of participants was 55 in the control group and 50 in the experimental group. This final sample size achieved 80% power to detect a Cohen's *f* effect size of 0.28 using ANCOVA (Faul et al., 2007).

6.4 Multivariate Normality

Multivariate and univariate normality was tested in R (R Core Team, 2019) using the MVN package (Korkmaz, Goksuluk, & Zararsiz, 2014). Missing data was omitted when running normality tests and the results demonstrated that the shifting and inhibition measures were not considered normally distributed with skewness being the main issue. WM data appeared to be multivariate normal. Data transformation using a

Table 5.
Summary Data for Time 1.

	Control Group								Experimental Group							
	n	Mean	SD	Min.	Max.	Skew	Kurtosis	SE	n	Mean	SD	Min.	Max.	Skew	Kurtosis	SE
Flanker RT Conflict Cost[a]	54	33.39	29.35	-26.25	126.45	0.7	1.13	3.99	48	33.93	38.13	-39.35	139.45	0.5	0.44	5.5
Flanker Incongruent Accuracy	54	0.74	0.19	0.3	1	-0.7	-0.37	0.03	48	0.75	0.18	0.2	1	-0.94	0.65	0.03
No-Go Accuracy	55	0.79	0.1	0.55	0.96	-0.63	-0.41	0.01	50	0.76	0.1	0.54	0.9	-0.57	-0.64	0.01
Go RT Average[a]	53	512.7	49.5	420.11	655.04	0.57	0.19	6.8	50	519.47	60.54	367.81	656.12	0.02	-0.07	8.56
Stroop RT Incongruent[a]	55	773.16	114.56	536.54	1015.77	0.24	-0.77	15.45	50	807.68	151.31	463.2	1180	0.05	-0.46	21.4
Stroop Incongruent Accuracy*	55	0.84	0.09	0.57	0.96	-0.76	0.12	0.01	50	0.8	0.09	0.59	0.95	-0.8	0.16	0.01
Switching Completion	55	15.97	4.08	4	24.25	-0.52	0.57	0.55	50	14.45	4.2	7.75	26	0.3	-0.47	0.59
BCST Perseverative Errors[a]	55	26.34	7.76	7.81	43.75	0.25	-0.54	1.05	50	25.35	6.75	14.06	43.75	0.69	0.39	0.95
BCST Conceptual Level Responses	55	45.21	15.13	14.1	70.3	-0.42	-0.86	2.04	50	44.51	14.27	15.6	68.8	-0.43	-0.67	2.02
Switcher Errors[a]	55	3.53	5.42	0	21	2	3.02	0.73	49	4.12	5.48	0	21	1.68	2.04	0.78
Switcher Error Difference[a]	55	0.13	6.14	-18	20	0.09	2.48	0.83	49	-0.57	6.49	-17	20	0.4	2.29	0.93
OSPAN Partial Score**	48	0.67	0.22	0.12	1	-0.55	-0.68	0.03	39	0.54	0.2	0.12	0.92	-0.1	-0.85	0.03
RSPAN Partial Score*	46	0.67	0.19	0.24	1	-0.58	0.07	0.03	44	0.57	0.23	0.08	1	-0.33	-0.76	0.03
SSPAN Partial Score	47	0.59	0.22	0.14	0.93	-0.06	-0.99	0.03	43	0.52	0.2	0.14	0.93	0.1	-0.9	0.03
Mean RT (SRT task)	55	376.85	54.6	291.07	528.86	0.66	-0.4	7.36	50	393.56	56.99	287.45	552.83	0.97	0.66	8.06
NAI (gF)*	55	115.96	13.27	91	146	0.18	-0.85	1.79	50	110.2	11.48	81	147	0.1	1.35	1.62
PAT Maths*	55	134.21	8.74	118.8	161	0.51	0.04	1.18	50	130.27	11	102.5	153.2	-0.09	-0.61	1.56
PAT Reading	55	133.84	12.35	104.3	169	0.07	0.29	1.67	50	130.01	12.63	95.5	158.7	-0.26	-0.14	1.79
CEFI Parent - Inhibition	55	100.16	12.2	78	135	0.47	0.1	1.65	50	96.46	13.71	69	128	0.21	-0.83	1.94
CEFI Parent - Shifting	55	97.29	11.91	74	130	0.43	0.03	1.61	50	92.86	12.67	72	121	0.14	-0.94	1.79
CEFI Parent - WM	55	98.71	11.9	72	126	-0.08	-0.65	1.6	50	96.04	13.63	68	129	0.6	-0.15	1.93
CEFI Teacher - Inhibition**	29	109.52	14.66	79	128	-0.26	-1.25	2.72	26	96.73	13.41	75	128	0.26	-0.5	2.63
CEFI Teacher - Shifting*	29	110.55	14.62	77	134	-0.55	-0.55	2.71	26	100.69	13.94	77	122	0.07	-1.2	2.73
CEFI Teacher - WM**	29	108.24	14.12	79	131	-0.38	-0.89	2.62	26	97.35	12.7	76	119	-0.01	-1.04	2.49

Note. RT = reaction time; BCST = Berg Card Sorting Task; OSPAN = operation span; RSPAN = reading span; SSPAN = symmetry span; SRT = simple reaction time; NAI = Naglieri Ability Index; gF = fluid intelligence; PAT = Progressive Achievement Test; CEFI = Comprehensive Executive Function Inventory; WM = working memory. Differences between groups are indicated by * $p < 0.05$ and ** $p < 0.01$.

[a] Items were later reversed to make higher scores equate to a better response.

variety of methods was attempted on the shifting and inhibition data but the results were unsatisfactory as is quite often reported (Wilcox, 2017). Data with tail weight and symmetry consistent with a normal distribution is extremely rare in psychological data (Field & Wilcox, 2017) and the validity of testing for normality has been questioned too (Wilcox, 2017). As seen in Table 6, the data collected in this study had a mixture of no tails or heavy tails and the distributions were generally skewed and there were a significant number of outliers; as is quite often the case in applied data (Micceri, 1989). As a result, it was decided that robust methods of data analysis would be used for initial data analysis.

Table 6.
Multivariate Normality Statistics for Executive Function (EF) Measures.

				Skewness		Kurtosis	
		n	Outliers	Statistic	p-value	Statistic	p-value
Control	Shifting	55	15	161.90	< .001	2.65	< .001
	Inhibition	52	7	80.48	.018	0.63	.527
	WM	36	0	4.12	.942	-1.20	.229
Exp.	Shifting	49	16	79.33	< .001	1.17	.242
	Inhibition	48	9	93.29	.001	1.39	.166
	WM	29	1	13.31	.207	-0.66	.512

Note. WM = working memory.

6.5 Baseline Comparisons

To compare the starting values of the experimental and control groups at T1 using robust methods, Welch's *t*-test was used, as suggested by Delacre, Lakens, and Leys (2017). The stats package in R (R Core Team, 2019) was used for robust data analysis and Cohen's *d* effect sizes were calculated using the compute.es package (Del Re, 2013). The groups did not differ with regard to the age of participants, Welch's

$t(100.7) = -0.74$, $p = .46$, $d = -0.14$, $\text{CI}_{95\%}[-0.53, 0.25]$, or the number of hours they participated in training, Welch's $t(94.0) = 0.47$, $p = .641$, $d = 0.11$, $\text{CI}_{95\%}[-0.28, 0.49]$. Even though participants were randomly allocated to groups, there were significant differences between the experimental and control groups, with the control group performing significantly better on the number Stroop incongruent accuracy (Welch's $t(102.1) = -2.05$, $p = .043$, $d = -0.44$, $\text{CI}_{95\%}[-0.84, -0.05]$), OSPAN (Welch's $t(83.7) = -2.66$, $p = .009$, $d = -0.62$, $\text{CI}_{95\%}[-1.05, -0.18]$), RSPAN (Welch's $t(83.2) = -2.44$, $p = .017$, $d = -0.48$, $\text{CI}_{95\%}[-0.90, -0.05]$), NAI (Welch's $t(102.8) = -2.39$, $p = .019$, $d = -0.46$, $\text{CI}_{95\%}[-0.86, -0.07]$), PAT Maths (Welch's $t(93.5) = -2.02$, $p = .047$, $d = -0.40$, $\text{CI}_{95\%}[-0.79, -0.01]$), and all teacher CEFI reports (p's $< .014$). Even after adjusting for multiple comparisons using Holm's method (Holm, 1979), the teacher CEFI report for inhibition (Welch's $t(52.97) = -3.38$, $p_{adj} = .003$, $d = -0.91$, $\text{CI}_{95\%}[-1.48, -0.34]$), the teacher report for shifting (Welch's $t(52.79) = -2.56$, $p_{adj} = .013$, $d = -0.69$, $\text{CI}_{95\%}[-1.25, -0.13]$), and the teacher report for WM (Welch's $t(52.99) = -3.01$, $p_{adj} = .008$, $d = -0.81$, $\text{CI}_{95\%}[-1.37, -0.25]$) were different between groups with the control group scoring higher on all T1 reports.

6.6 Correlations

Correlations among the EF measures were low to moderate, with Spearman correlations among individual EF measures in T1 ranging from -0.18 to 0.63 (see Table 7). Table 7 also shows the correlations amongst all measures for T1 data.

6.7 Confirmatory Factor Analysis

To make use of the multiple measures for each EF construct the initial primary analytic technique was to consider the constructs using a CFA measurement model.

Table 7.

Correlations Among All Measures.

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
1. Flanker RT Conflict Cost	n	—	102	102	100	102	102	102	102	102	101	101	84	88	88	102	102	102	102	102	102	102	54	54	54
	r	—	0.03	0.00	-0.01	-0.07	0.01	-0.08	-0.07	-0.03	-0.14	-0.18	0.05	-0.06	-0.11	0.02	0.07	-0.05	-0.07	-0.01	-0.07	-0.04	-0.02	0.05	-0.01
2. Flanker Incongruent Accuracy	n		—	102	100	102	102	102	102	102	101	101	84	88	88	102	102	102	102	102	102	102	54	54	54
	r		—	0.21*	-0.04	-0.07	0.34***	0.21*	-0.11	0.23*	-0.23*	-0.10	0.12	0.21	0.13	-0.24*	0.29**	0.27**	0.24*	0.12	0.14	0.10	0.17	0.34*	0.24
3. No-Go Accuracy	n			—	103	105	105	105	105	105	104	104	87	90	90	105	105	105	105	105	105	105	55	55	55
	r			—	0.16	-0.08	0.27**	0.11	-0.16	0.33***	-0.18	-0.10	0.03	0.12	0.14	-0.17	0.34***	0.25**	0.23*	0.14	0.00	0.13	0.29*	0.36**	0.31*
4. Go RT Average	n				—	103	103	103	103	103	102	102	85	90	89	103	103	103	103	103	103	103	53	53	53
	r				—	0.52***	-0.02	-0.49***	0.06	-0.16	0.16	0.20*	-0.04	-0.26*	-0.14	0.55***	-0.07	-0.45***	-0.23*	-0.11	-0.03	-0.01	0.03	-0.03	0.01
5. Stroop RT Incongruent	n					—	105	105	105	105	104	104	87	90	90	105	105	105	105	105	105	105	55	55	55
	r					—	0.09	-0.54***	0.10	-0.20*	0.17	0.18	-0.19	-0.30**	-0.23*	0.46***	-0.14	-0.47***	-0.36***	-0.12	-0.09	-0.09	-0.08	-0.20	-0.10
6. Stroop Incongruent Accuracy	n						—	105	105	105	104	104	87	90	90	105	105	105	105	105	105	105	55	55	55
	r						—	0.05	0.08	0.13	-0.16	-0.03	0.26*	0.08	0.02	-0.12	0.20*	0.16	0.02	0.03	0.03	-0.05	0.17	0.21	0.25
7. Switching Completion	n							—	105	105	104	104	87	90	90	105	105	105	105	105	105	105	55	55	55
	r							—	-0.17	0.31**	-0.10	-0.04	0.23*	0.38***	0.34**	-0.54***	0.29**	0.73***	0.46***	0.29**	0.15	0.26**	0.22	0.37**	0.20
8. BCST Perseverative Errors	n								—	105	104	104	87	90	90	105	105	105	105	105	105	105	55	55	55
	r								—	-0.63***	0.08	0.07	-0.01	-0.17	-0.18	0.14	-0.11	-0.27**	-0.18	-0.08	0.00	-0.05	-0.05	-0.35**	-0.15
9. BCST Conceptual Level Responses	n									—	104	104	87	90	90	105	105	105	105	105	105	105	55	55	55
	r									—	-0.08	-0.05	0.13	0.17	0.18	-0.39***	0.34***	0.33***	0.24*	0.19	-0.03	0.19	0.21	0.43***	0.29*
10. Switcher Errors	n										—	104	86	89	89	104	104	104	104	104	104	104	54	54	54
	r										—	0.57***	0.02	-0.11	-0.13	0.09	-0.08	-0.21*	-0.14	0.06	-0.08	-0.02	-0.31*	-0.28*	-0.33*
11. Switcher Error Difference	n											—	86	89	89	104	104	104	104	104	104	104	54	54	54
	r											—	0.06	-0.12	-0.07	0.07	-0.01	-0.17	-0.12	0.09	-0.04	0.03	-0.08	-0.17	-0.10
12. OSPAN Partial Score	n												—	76	75	87	87	87	87	87	87	87	41	41	41
	r												—	0.37***	0.40***	-0.21	0.18	0.29**	0.24*	-0.06	0.01	-0.03	0.36*	0.54***	0.32*
13. RSPAN Partial Score	n													—	78	90	90	90	90	90	90	90	44	44	44
	r													—	0.27*	-0.29**	0.24*	0.43***	0.37***	0.13	0.06	0.06	0.31*	0.31*	0.32*
14. SSPAN Partial Score	n														—	90	90	90	90	90	90	90	46	46	46
	r														—	-0.23*	0.12	0.38***	0.22*	-0.03	-0.01	0.01	-0.05	0.15	0.04
15. Mean RT (SRT task)	n															—	105	105	105	105	105	105	55	55	55
	r															—	-0.19	-0.47***	-0.37***	-0.07	-0.01	-0.04	-0.03	-0.19	-0.07
16. NAI (gF)	n																—	105	105	105	105	105	55	55	55
	r																—	0.41***	0.39***	0.16	0.05	0.05	0.41**	0.43**	0.38**
17. PAT Maths	n																	—	105	105	105	105	55	55	55
	r																	—	0.65***	0.27**	0.02	0.22*	0.26	0.44***	0.28*
18. PAT Reading	n																		—	105	105	105	55	55	55
	r																		—	0.25*	0.00	0.16	0.32*	0.44***	0.28*
19. CEFI Parent - Inhibition	n																			—	105	105	55	55	55
	r																			—	0.63***	0.73***	0.21	0.35**	0.27*
20. CEFI Parent - Shifting	n																				—	105	55	55	55
	r																				—	0.52***	0.03	0.15	0.13
21. CEFI Parent - WM	n																					—	55	55	55
	r																					—	0.20	0.31*	0.28*
22. CEFI Teacher - Inhibition	n																						—	55	55
	r																						—	0.72***	0.89***
23. CEFI Teacher - Shifting	n																							—	55
	r																							—	0.76***
24. CEFI Teacher - WM	n																								—
	r																								—

Note. EF = executive function; RT = reaction time; BCST = Berg Card Sorting Task; SRT = simple reaction time; OSPAN = operation span; RSPAN = reading span; SSPAN = symmetry span; NAI = Naglieri Ability Index; gF = fluid intelligence; PAT = Progressive Achievement Test; CEFI = Comprehensive Executive Function Inventory; WM = working memory; n = sample size for pairwise correlations.
All correlation coefficients are Spearman's rho. * $p < .05$, ** $p < .01$, *** $p < .001$

Models were run in R (R Core Team, 2019) using lavaan (Rosseel, 2012) with standardised variables using robust full-information maximum-likelihood estimation to establish model fit and full information maximum-likelihood estimation for missing data. Model fit was evaluated at both the global and local level. For global fit, common measures and accepted cut-offs were used (Hu & Bentler, 1999): Comparative fit index (CFI) and Tucker–Lewis index (TLI) > .95; squared root mean residual (SRMR < .08); and root mean squared error of approximation (RMSEA < .06). For relative fit, the Akaike Information Criterion (AIC) was used. The AIC controls for degrees of freedom and favours simpler models, with a smaller AIC indicating a better fit (Kline, 2015). For local fit, parameters such as factor loadings, standard errors, z-scores and residuals were examined. In all models, first indicators were free to vary and prior to model fit, indicators were standardised and scaled so that higher scores were associated with better performance.

As CFA can be an iterative process (as long as changes made are theory-based), whereby modifications are indicated in the initial results (Schreiber, Nora, Stage, Barlow, & King, 2006; Streiner, 2006; Whittaker, 2012), a number of different models were attempted. Through the use of robust global and local fit indices, including modification indices, factor loadings, and multiple squared correlations, inappropriate indicators were dropped and the model simplified until a model with satisfactory fit indices was found. All changes made to the models made theoretical sense and, at least initially, all indicators were still within their originally aligned EF construct. As there were multiple items from the same EF measure, it was expected that some of these indicators might need to be dropped as they might cause problems with the model.

To begin with, there were 14 indicators from the nine EF measures (as outlined in the methodology), which were included in the model and model specification was initially attempted on a three-factor model. This initial model provided a poor fit (χ^2 (74) = 281.21, $p < .001$; AIC = 3937.14; CFI = 0.39; TLI = 0.24; RMSEA = 0.15; SRMR = 0.13). High modification indices suggested the dropping of three indicators (Switcher Errors, Card Sorting Perseverative Errors, Go RT Average) as these items shared residual correlations with the other items from within the same measure (Switcher, Card Sorting and Go/No-Go respectively). Other high modification indices did not make theoretical sense to change (e.g., changing Switching Completion to WM or inhibition) and so were not done. After dropping these three items, the subsequent model was still a poor fit (χ^2 (41) = 62.78, $p = .016$; AIC = 3069.02; CFI = 0.81; TLI = 0.75; RMSEA = 0.07; SRMR = 0.08) and made a model with indistinguishable latent factors with a correlation between shifting and inhibition greater than one. In an attempt to keep a three-factor model, factor loadings in this model were looked at and Switcher Error Difference (0.05; SE 0.16; $p = .74$), Stroop Incongruent Accuracy (0.13; SE 0.10, $p = .32$), and Flanker RT Conflict Cost (0.10; SE 0.10; $p = .32$) were removed as these were the lowest factor loadings. The next model provided a satisfactory fit (χ^2 (17) = 19.88, $p = .28$; AIC = 2174.61; CFI = 0.97; TLI = 0.95; RMSEA = 0.04; SRMR = 0.05) but it still made a model with indistinguishable latent factors with a correlation between shifting and inhibition of 1.27. Therefore, it was decided to combine the shifting and inhibition into one combined EF factor, which is theoretically justifiable as previous research has found a two-factor EF model with WM and a separate but combined shifting and inhibition factor in children (Karr et al., 2018; Lee et al., 2012; van der Ven et al., 2012, 2013). This solved the problem of indistinguishable latent factors and resulted in slightly improved model fit (χ^2 (19) = 21.11, $p = .33$; AIC = 2171.983; CFI = 0.98; TLI = 0.97; RMSEA = 0.03; SRMR = 0.05) and all remaining factor loadings were statistically significant ($p < .05$). Figure 10 shows the path model of the final two-factor model for T1.

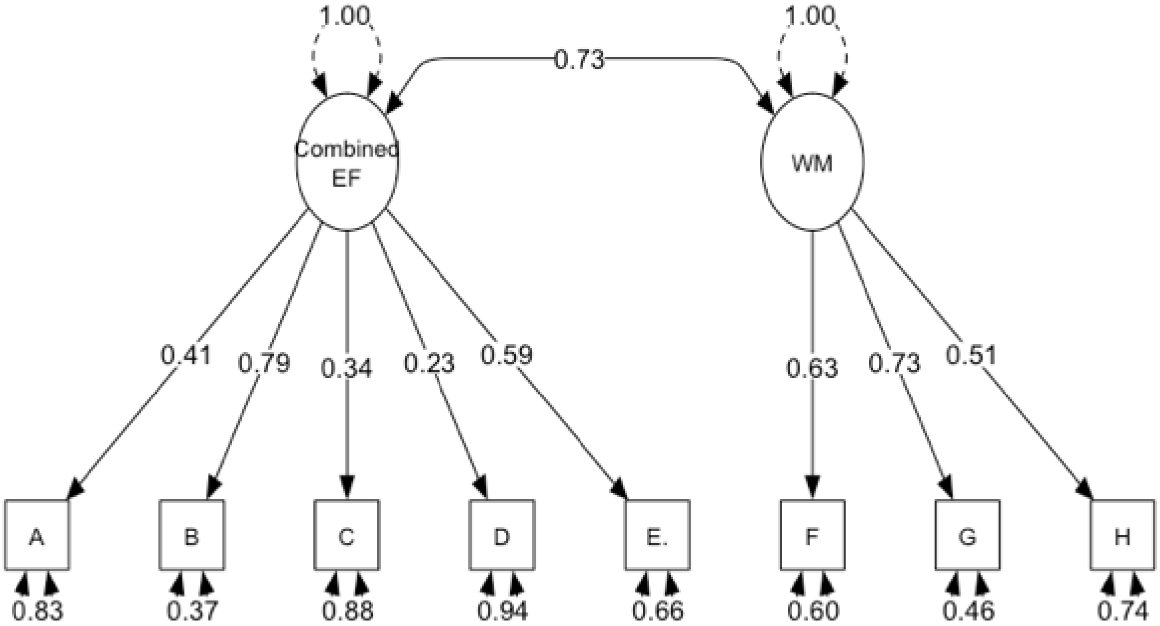


Figure 10. Path model for the final two-factor executive function (EF) model for Time 1. A = Card Sorting Conceptual Level Responses; B = Switching Completion; C = Flanker Incongruent Accuracy; D = No-Go Accuracy; E = Stroop RT Incongruent; F = Operation Span Partial Score; G = Reading Span Partial Score; H = Symmetry Span Partial Score; Combined EF = Shifting and Inhibition latent factor; WM = working memory latent factor. All factor loadings are significant ($p < .05$).

After identification, this two-factor model was used to compare with alternative two-factor models, a potential one-factor model and the originally proposed three-factor model across all time periods (T1, T2 and T3) to see if this identified two-factor model still provided the best fit. The two-factor model provided a better fit than the one-factor model at T1 (χ^2 (1) = 11.99, $p < .001$) and T2 (χ^2 (1) = 23, $p < .001$), but marginally did not provide a better fit at T3 (χ^2 (1) = 3.81, $p = .051$). Table 8 presents the fit results for the proposed factor models. The two-factor model that combined shifting and inhibition tasks and with a separate but correlated WM factor was the best fitting model across T1 - T3.

Table 8.
Confirmatory Factor Analysis (CFA) Fit Criteria.

	Absolute / Predictive Fit		Comparative Fit		Other	
Models	χ^2	AIC	CFI	TLI	RMSEA (90% CI)	SRMR
Time 1 Models (pre-test)						
Three-Factor[#]	19.88[*]	2174.61	0.97	0.95	.04 (.00 - .10)	.05
One-Factor	29.45[*]	2177.66	0.91	0.87	.07 (.00 - .11)	.06
Combined EF + WM	**21.10[*]**	**2171.98**	**0.98**	**0.97**	**.03 (.00 - .09)**	**.05**
Combined EF + Shifting	28.95[*]	2179.31	0.90	0.86	.07 (.00 - .12)	.06
Combined EF + Inhibition[#]	29.15[*]	2178.77	0.90	0.86	.07 (.00 - .12)	.06
Time 2 Models (post-test)						
Three-Factor[#]	21.13	2150.59	0.98	0.96	.05 (.00 - .10)	.05
One-Factor	33.16	2155.98	0.92	0.89	.08 (.02 - .12)	.06
Combined EF + WM	**21.05[*]**	**2147.29**	**0.99**	**0.98**	**.03 (.00 - .09)**	**.05**
Combined EF + Shifting[#]	33.59	2157.93	0.91	0.87	.08 (.03 - .13)	.06
Combined EF + Inhibition	29.57[*]	2154.35	0.94	0.91	.07 (.00 - .12)	.06
Time 3 Models (1-year post-test)						
Three-Factor[#]	17.86	2204.99	0.99	0.99	.02 (.00 - .10)	.06
One-Factor	23.97[*]	2205.87	0.96	0.95	.05 (.00 - .10)	.06
Combined EF + WM	**20.45[*]**	**2204.25**	**0.99**	**0.98**	**.03 (.00 - .10)**	**.06**
Combined EF + Shifting[#]	23.06[*]	2206.70	0.96	0.95	.05 (.00 - .10)	.06
Combined EF + Inhibition	23.94[*]	2207.56	0.96	0.93	.05 (.00 - .11)	.06

Note. EF = executive function; WM = working memory; AIC = Akaike information criterion; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square Error of approximation; SRMR = standardised root mean square residual. Bolded models indicate best fit. All model statistics are robust measures.
[#] Indicates overfitted model (factor correlation > 1).
[*] p > .05;

The final five indicators in the combined inhibition-shifting factor (combined EF factor) were the Card Sorting Conceptual Level Responses, Switching Completion, Flanker Incongruent Accuracy, No-Go Accuracy and Stroop RT Incongruent measures. Importantly, the CFA iteration process removed all measurement items from the Switcher task but each of the other five EF measures still had one indicator present in the final model. The final indicators in the WM factor were the OSPAN, RSPAN and SSPAN measures.

The simsem package (Jorgensen, Pornprasertmanit, Miller, & Schoemann, 2018) was used to run a Monte Carlo simulation for model fit evaluation and power analysis for the CFA model for T1. Model fit simulations were run with the same sample size and were run 500 times which resulted in 496 converged replications and all factor loadings were statistically significant at $p < .05$ level. All indicators had greater than 88% power to detect if the parameter estimate was significantly different from zero except for No-Go Accuracy which only had 51% power. A sample size of 188 would have been required for the No-Go Accuracy parameter to have 80% power to detect if the parameter estimate was significantly different from zero.

6.8 Factor Scores

The creation of latent factor scores in lavaan (Rosseel, 2012) using the indicators and factor structure proposed by the CFA was attempted using a latent curve model with decreasing constraints (Beaujean, 2014). Even though the final unconstrained model fit was satisfactory (χ^2 (250) = 292.73, $p = .033$; AIC = 6012.54; CFI = 0.95; TLI = 0.95; RMSEA = 0.04; SRMR = 0.07) and better fitting than more constrained models (χ^2 (1) = 46.12, $p < .001$), latent correlations were consistently greater than one. Changing the latent factor structure did not remedy these covariance issues. Therefore, latent factor scores were created using lavaan (Rosseel, 2012) for T1-T3 separately for both the combined EF factor (shifting and inhibition) and the WM factor. It was expected that these latent factor scores should be positively correlated with the far-transfer measures (maths, English and gF measures) as previous research has demonstrated the link between EF and academic achievement. As seen in Table 9, both the combined EF and WM latent factor scores were significantly correlated with the gF, PAT Maths and PAT Reading measures, thus appearing to demonstrate convergent validity for the latent factor scores.

Table 9.

Correlations Between Latent Factor Scores and Far-Transfer Measures.

Variable	1	2	3	4	5
1. EF Factor	—	$.87^{**}$ [.81, .91]	$.36^{**}$ [.18, .52]	$.73^{**}$ [.63, .81]	$.51^{**}$ [.35, .64]
2. WM Factor		—	$.35^{**}$ [.17, .50]	$.64^{**}$ [.51, .74]	$.47^{**}$ [.31, .61]
3. NAI (gF)			—	$.42^{**}$ [.25, .57]	$.42^{**}$ [.25, .56]
4. PAT Maths				—	$.65^{**}$ [.53, .75]
5. PAT Reading					—

Note. EF = executive function; WM = working memory; NAI = Naglieri Ability Index; gF = fluid intelligence; PAT = Progressive Achievement Test. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).
* $p < .05$. ** $p < .01$.

Results of Pearson correlational analysis indicated that there was a significant negative association between PS and both the combined EF factor score, (r(103) = -.59, $CI_{95\%}$[-.70, -.45], $p < .001$), and the WM factor score, (r(103) = -.47, $CI_{95\%}$[-.60, -.30], $p < .001$). This indicated that those who had a faster reaction time, as indicated by the PS, scored higher in both the combined EF and WM factor scores.

6.9 Near-Transfer

6.9.1 Executive Function

As latent factor scores for T1 - T3 appeared to be valid it was possible to determine if the intervention was effective in improving near-transfer to EF (*Hypothesis 1*). Summary data can be seen in Table 10 for the latent factor scores and Shapiro-Wilk normality tests indicated that T2 data were non-normal.

Table 10.
Summary Data for Latent Factor Scores.

		Combined EF Factor Score			WM Factor Score		
		T1	T2	T3	T1	T2	T3
Control	N	55	55	55	55	55	55
	Mean	0.18	0.31	0.23	0.23	0.37	0.20
	Minimum	-1.67	-1.51	-1.25	-1.47	-1.37	-1.41
	Maximum	1.66	1.69	1.85	1.65	1.47	1.74
	Skewness	-0.44	-0.37	0.27	-0.26	-0.69	0.01
	Kurtosis	-0.29	0.11	-0.88	-0.27	0.19	-0.50
	Shapiro-Wilk *p*-value	0.25	0.690	0.138	0.349	0.037	0.372
Exp.	N	50	50	50	50	50	50
	Mean	-0.20	-0.34	-0.25	-0.25	-0.41	-0.23
	Minimum	-2.11	-2.49	-2.64	-2.31	-2.51	-2.58
	Maximum	1.90	0.97	1.44	1.68	1.22	1.34
	Skewness	0.01	-0.53	-0.29	-0.16	-0.55	-0.54
	Kurtosis	-0.63	-0.63	-0.64	-0.43	-0.50	-0.21
	Shapiro-Wilk *p*-value	0.781	0.021	0.177	0.821	0.043	0.107

Note. EF = executive function; WM = working memory.

It is important to note that there was a significant difference between the control and experimental groups on T1 combined EF latent scores, Welch's $t(98.2) = -2.28$, $p = .025$, $d = -0.44$, $CI_{95\%}[-0.83, -0.05]$ and T1 WM latent scores, Welch's $t(99) = -3.00$, $p = .003$, , $d = -0.59$, $CI_{95\%}[-0.99, -0.19]$, with the control group scoring higher on both measures. Plots for the T1-T3 factor scores can be seen for the combined EF score in Figure 11 and for WM in Figure 12.

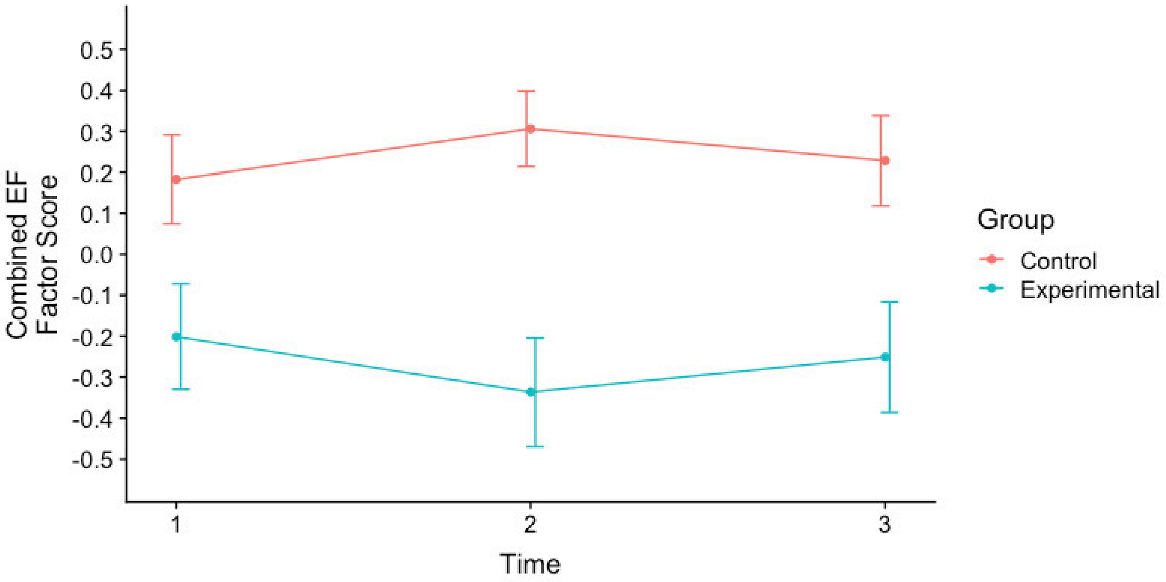


Figure 11. Combined executive function (EF) factor score plots. Error bars represent standard errors.

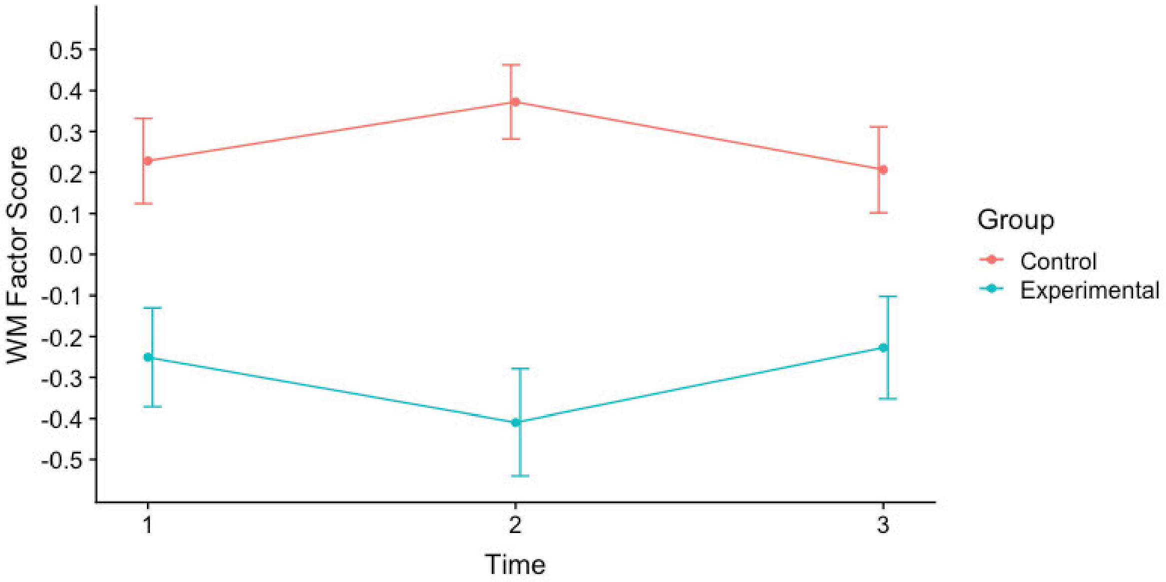


Figure 12. Working memory (WM) factor score plots. Error bars represent standard errors.

Differences between groups on pre-test scores in a randomised controlled trial is referred to as “unhappy randomisation” (Kenny, 1979, p. 269). Importantly, data analysis using ANCOVA provides unbiased effect estimates when pre-test covariates are not balanced because "the difference between adjusted means is an unbiased estimate of what the difference between group means would be if each group had a covariate mean equal to the grand covariate mean" (Huitema, 2011, p. 201). Thus, pre-test covariate adjustment with ANCOVA reduces differences between conditions at the start, increasing the ability to explore the true treatment differences (Read, Kendall, Carper, & Rausch, 2013). As stated by Kenny (1979), when there is unhappy randomisation, data analysis using ANCOVA "is not only appropriate but necessary" (p. 269).

Parametric ANCOVAs were attempted in R (R Core Team, 2019) and all models satisfied the assumption of homogeneity of slopes and all standardised residuals appeared to be distributed normally as demonstrated by Q-Q plots. However, T2 data for both the combined EF score ($F(1,103) = 7.30$, $p = .008$) and the WM score ($F(1,103) = 7.61$, $p = .007$) violated the assumptions of homogeneity of variance. Consequently, robust ANCOVAs using WRS2 (Mair & Wilcox, 2018) were used to determine if there was a significant difference between intervention groups using T2 scores while controlling for T1 scores. Robust ANCOVAs compare trimmed means using Yuen's tests and no parametric assumptions are made about the form of the regression lines. It uses design points which are specific scores where the nonparametric regression lines can be compared (scores of the covariate for which the relationship between the outcome and covariate is roughly the same in both groups). The design points were not specified, and the level of the interval smoother was not set so the WRS2 program automatically chose five points for comparison and set the interval smoothing parameter. Where appropriate, partial eta-squared (η_p^2) effect sizes were calculated using the sjstats package (Lüdecke, 2019).

Using robust ANCOVA there was no significant effect of the intervention group on T2 EF latent scores after controlling for T1 EF latent scores, as all confidence intervals for trimmed mean differences included zero (all p's $> .05$). There was a significant effect of the intervention group on T2 WM latent scores after controlling for T1 WM latent scores, at T1 scores of -0.24 (difference = 0.64, $p = .003$, $CI_{95\%}[0.18, 1.10]$, $n = 65$), 0.25 (difference = 0.40, $p = .041$, $CI_{95\%}[-0.02, 0.82]$, $n = 67$), 0.52 (difference = 0.39, $p = .041$, $CI_{95\%}[-0.02, 0.81]$, $n = 59$), and 1.3 (difference = 0.48, $p = .009$, $CI_{95\%}[0.09, 0.87]$, $n = 29$). Figure 13 shows a scatterplot for the WM latent scores for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points. At the T1 scores where there was a significant difference between groups, the effect of the intervention group was the opposite of what was expected. That is, the control group scored higher on T2 WM scores after controlling for T1 scores.

T3 data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 EF scores after controlling for the T1 scores on EF, $F(1, 102) = -2.28$, $p = .135$, ($\eta_p^2 = 0.02$, $CI_{90\%}[0.00, 0.09]$) with the covariate, T1 scores, significantly related to the T3 scores, $F(1, 102) = 235.10$, $p < .001$, ($\eta_p^2 = 0.72$, $CI_{90\%}[0.64, 0.77]$). There was no significant effect of the intervention group on T3 WM latent scores after controlling for T1 WM scores, $F(1, 102) = -0.29$, $p = .59$, ($\eta_p^2 = 0.003$, $CI_{90\%}[0.00, 0.04]$) with the covariate, T1 scores, significantly related to the T3 scores, $F(1, 102) = 143.72$, $p < .001$, ($\eta_p^2 = 0.61$, $CI_{90\%}[0.51, 0.68]$).

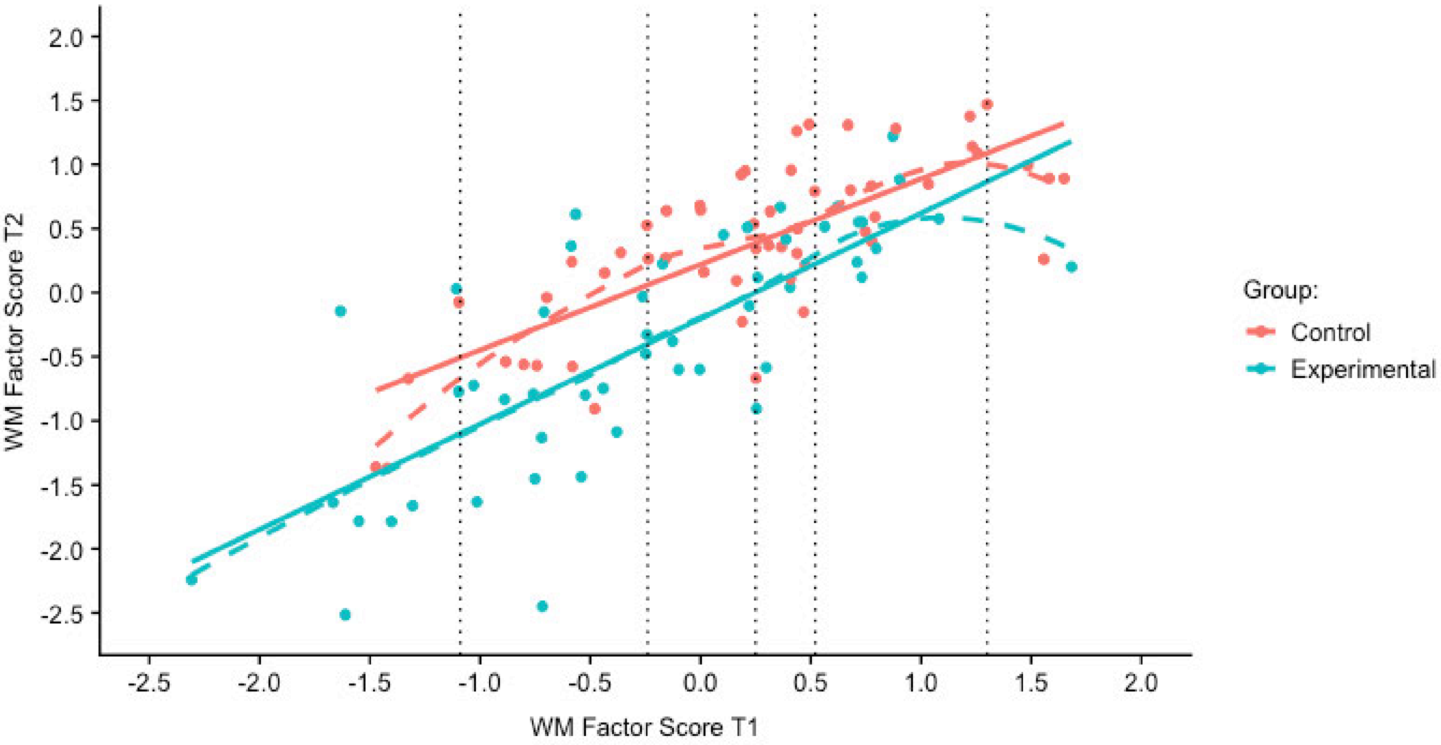


Figure 13. Working memory (WM) factor scores demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

In order to ensure that outliers did not play an important role in the findings, a secondary analysis was completed with outliers removed. It is important to note that when multivariate outliers were removed from the T2 combined EF factor score (final n_{exp}= 44; n_{ctrl}= 51), all parametric ANCOVA assumptions were met. Using parametric ANCOVA, there was a significant effect of group on T2 scores after controlling for the T1 scores on EF, $F(1, 92) = -9.14$, $p = .003$, (η_p^2= 0.09, $CI_{90\%}$[0.02, 0.19]) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 92) = 130.33$, $p < .001$, (η_p^2= 0.63, $CI_{90\%}$[0.52, 0.69]). However, this difference is the opposite of what was expected: that is, the control group scored higher at T2 after controlling for T1 scores (difference = -0.30, $t = -3.02$, p = .003, $d = -0.63$, $CI_{95\%}$[-1.05, -0.21]). Even with multivariate outliers removed, the WM factor score did not meet assumptions for running a parametric ANCOVA due to violating the assumption of homogeneity of variance ($F(1, 93) = 5.15$, p

= .026). A robust ANCOVA showed a significant effect of the intervention group on T2 WM latent scores after controlling for T1 WM latent scores, across all T1 points of comparison: -0.74 (difference = 0.58, p = .043, $\text{CI}_{95\%}$[-0.07, 1.22], n = 36), 0 (difference = 0.47, p = .030, $\text{CI}_{95\%}$[0.02, 91], n = 61), 0.32 (difference = 0.41, p = .043, $\text{CI}_{95\%}$[-0.02, 0.84], n = 61), 0.52 (difference = 0.37, p = .043, $\text{CI}_{95\%}$[-0.04, 0.78], n = 56), and 1.23 (difference = 0.48, p = .017, $\text{CI}_{95\%}$[0.08, 0.88], n = 28). Again, the results were the opposite of what was expected, with the control group scoring higher than the experimental group.

With outliers removed, T3 data for both EF and WM met all parametric ANCOVA assumptions. There was no effect of group on T3 scores for EF, $F(1, 92) = -1.31$, $p = 0.26$, (η_p^2= 0.01, $\text{CI}_{90\%}$[0.00, 0.08]), or WM, $F(1, 92) = 0.31$, $p = 0.58$, (η_p^2= 0.003, $\text{CI}_{90\%}$[0.00, 0.05]), with the covariate, T1 scores, significantly related to the T2 scores for both EF, $F(1, 92) = 185.15$, $p < 0.001$, (η_p^2= 0.69, $\text{CI}_{90\%}$[0.60, 0.75]), and WM, $F(1, 92) = 120.03$, $p < 0.001$, (η_p^2= 0.59, $\text{CI}_{90\%}$[0.48, 0.67]).

Although a central aim of this study was to use a latent variable model to assess near-transfer, the CFA iteration process resulted in the removal of a number of individual shifting and inhibition tasks due to the creation of a combined EF factor. Additionally, as the WM measures were very similar (all CSTs) when compared to the varied EF measures assessing shifting and inhibition, it could be challenging to separate method from construct variance with regard to the final endorsed two-factor CFA model (T. Redick, personal communication, 13 May 2020). To ensure that nothing was missed with regard to potential near-transfer, near-transfer was assessed at the individual EF task level too. There were no significant findings found in the expected direction for near-transfer when comparing the experimental to the control group at the individual EF task level. This analysis is included in Appendix G (Section 9.7). As there were no meaningful changes in near-transfer measures as a result of the intervention, test-retest reliability for the far-transfer measures are presented in Appendix H (Section 9.8).

6.9.2 Processing Speed

Parametric ANCOVA demonstrated there was no significant effect of intervention group on T2 PS after controlling for the T1 PS, $F(1, 102) = 2.52$, $p = .12$, $\eta_p^2 = 0.02$, $CI_{90\%}[0.00, 0.09]$. A robust paired t-test using the WRS2 package (Mair & Wilcox, 2018) showed that PS slowed from T1 to T2, $t_y(62) = -4.28$, $p < .001$, $d = -0.3$ but that there was then no significant difference between T1 and T3, $t_y(62) = 1.54$, $p = .13$, $d = 0.11$.

6.10 Far-Transfer

For the far-transfer measures of PAT Maths, PAT Reading and the gF measure (NAI), parametric ANCOVAs were attempted (*Hypothesis 2*). All assumptions for running a parametric ANCOVA were met apart from the T2 NAI which violated the assumption of homogeneity of slopes as there was an interaction present. As a result, robust ANCOVAs using WRS2 (Mair & Wilcox, 2018) was used for T2 NAI.

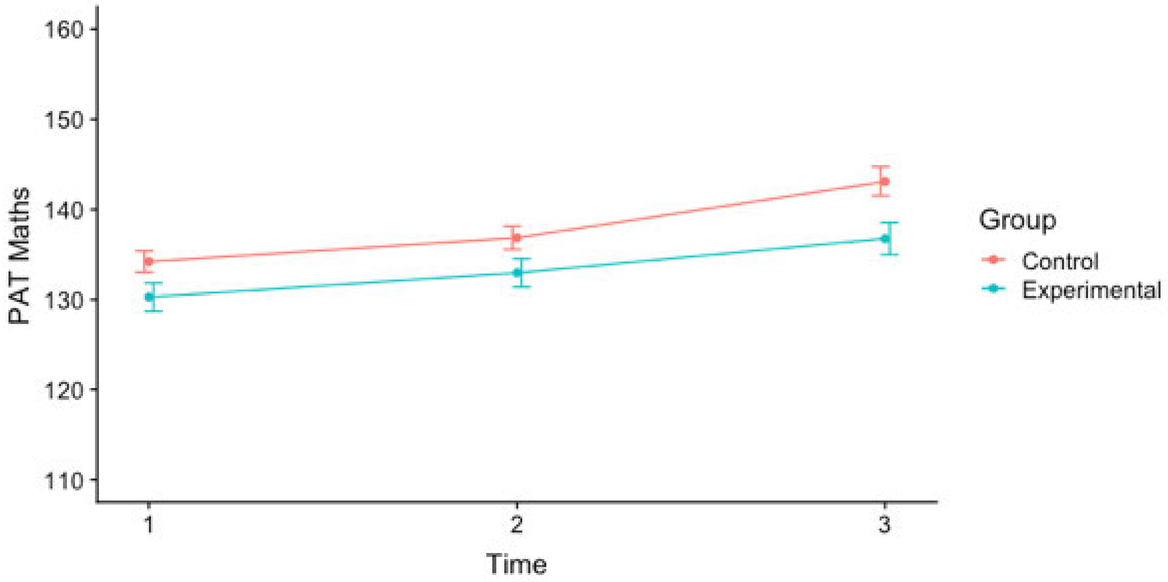


Figure 14. Progressive Achievement Test (PAT) Maths plot. Error bars represent standard errors.

6.10.1 PAT Maths

Scores across all time periods (T1-T3) are presented in Figure 14 for the PAT Maths. Parametric ANCOVA showed no significant effect of group on T2 scores after controlling for the T1 scores on PAT Maths, *F*(1, 102) = 0.15, *p* = .70, (η_p^2= 0.001, $CI_{90\%}$[0.00, 0.04]) with the covariate, T1 scores, significantly related to the T2 scores, *F*(1, 102) = 254.99, *p* < .001, (η_p^2= 0.72, $CI_{90\%}$[0.65, 0.77]). Similarly, for the T3 PAT Maths scores, there was no significant effect of intervention group after controlling for the T1 scores, *F*(1, 102) = 2.84, *p* = .095, (η_p^2= 0.03, $CI_{90\%}$[0.00, 0.10]) with the covariate, T1 scores, significantly related to the T2 scores, *F*(1, 102) = 130.11, *p* < .001, (η_p^2= 0.58, $CI_{90\%}$[0.48, 0.66]).

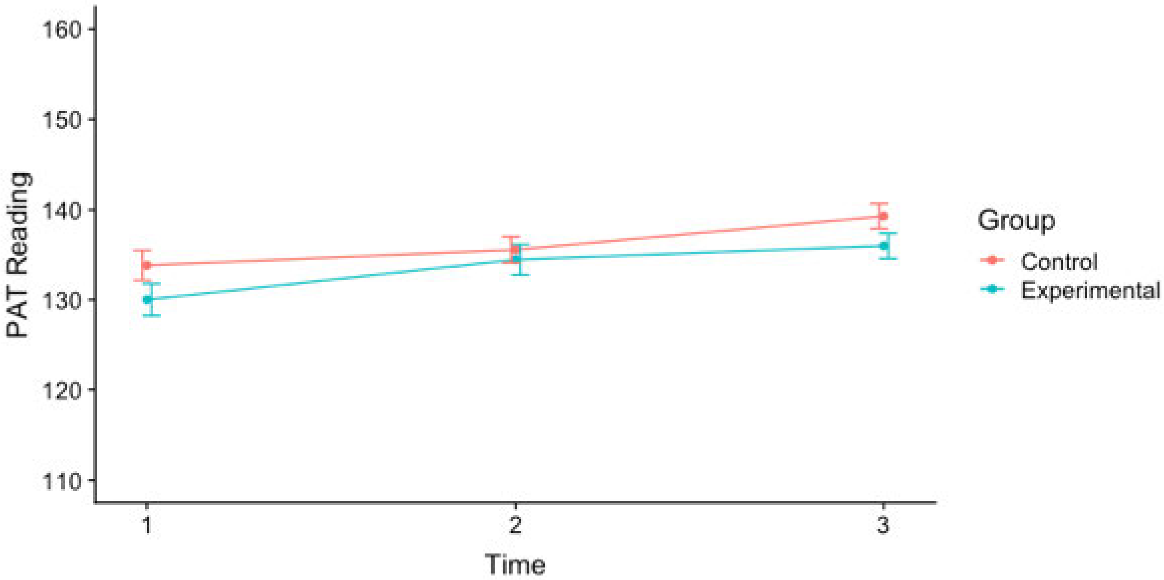


Figure 15. Progressive Achievement Test (PAT) Reading plot. Error bars represent standard errors.

6.10.2 PAT Reading

Scores across all time periods (T1-T3) are presented in Figure 15 for the PAT Reading. Parametric ANCOVA showed no significant effect of group on T2 scores after controlling for the T1 scores on PAT Reading, $F(1, 102) = 1.04$, $p = .31$, ($\eta_p^2 = 0.01$, $CI_{90\%}[0.000\ 0.06]$) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 102) = 132.11$, $p < .001$, ($\eta_p^2 = 0.56$, $CI_{90\%}[0.46, 0.64]$). Similarly, for the T3 PAT Reading scores, there was no significant effect of intervention group after controlling for the T1 scores, $F(1, 102) = 0.77$, $p = .38$, ($\eta_p^2 = 0.008$, $CI_{90\%}[0.00, 0.06]$) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 102) = 61.55$, $p < .001$, ($\eta_p^2 = 0.39$, $CI_{90\%}[0.27, 0.49]$).

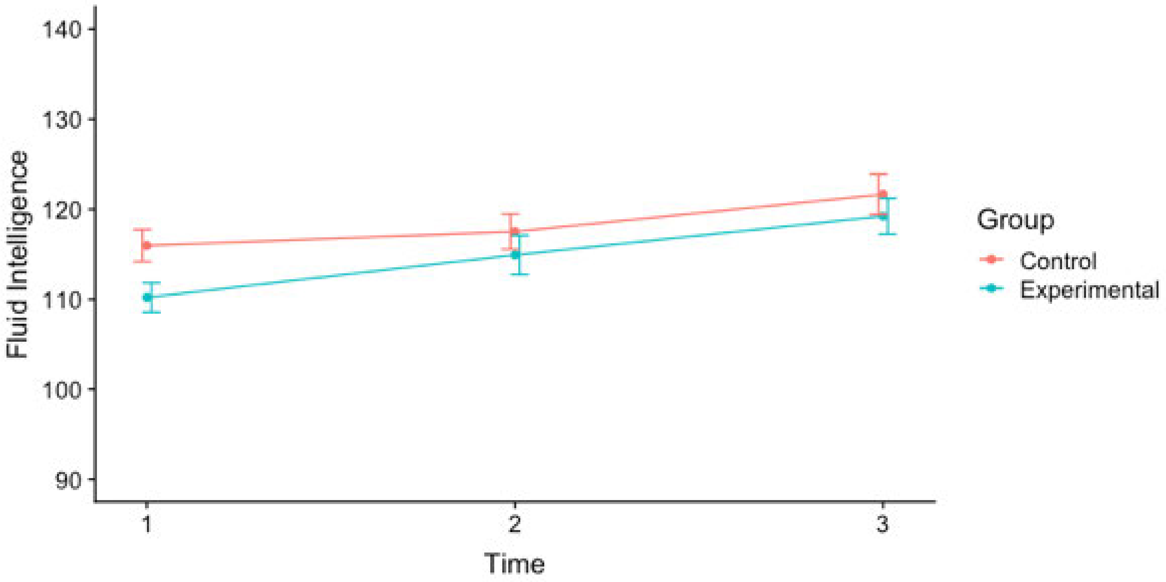


Figure 16. Fluid intelligence plot. The fluid intelligence measure was the Naglieri Ability Index (NAI) from the Naglieri Nonverbal Achievement Test – Second Edition (NNAT2). Error bars represent standard errors.

6.10.3 Fluid intelligence

Scores across all time periods (T1-T3) are presented in Figure 16 for the gF measure (i.e., the NAI from the NNAT2). Robust ANCOVA showed there was no significant effect of the intervention group on the T2 scores of the NAI after controlling for T1 scores at all T1 points of comparison: 94 (difference = 2.61, $p = .442$, $CI_{95\%}[-7.02, 12.23]$, $n = 31$), 104 (difference = -1.05, $p = .699$, $CI_{95\%}[-8.34, 6.24]$, $n = 62$), 112 (difference = -0.21, p = .939, CI$_{95\%}$[-7.43, 7.01], n = 76), 122 (difference = 0.29, p = .936, CI$_{95\%}$[-9.55, 10.13], n = 58), and 130 (difference = 3.65, p = .495, CI$_{95\%}$[-11.55, 18.85], n = 37). As there was an interaction present (as indicated by the results of the parametric ANCOVA), Figure 17 presents a scatterplot of the T1 and T2 NAI scores with a linear and non-linear slope for each intervention group. In general, it can be seen that there is no real effect of the experimental group on the T2 NAI score, but the interaction appears to be at the control group level with those participants who scored lower in T1 seemingly improving, whereas those who scored higher at T1 seemingly performing worse at T2. Parametric ANCOVA showed no significant effect of group on T3 scores after controlling for the T1 scores on the NAI, F(1, 102) = 1.01, p = .32, (η_p^2= 0.01, CI$_{90\%}$[0.00, 0.06]) with the covariate, T1 scores, significantly related to the T2 scores, F(1, 102) = 81.49, p < .001, (η_p^2= 0.45, CI$_{90\%}$[0.33, 0.54]).

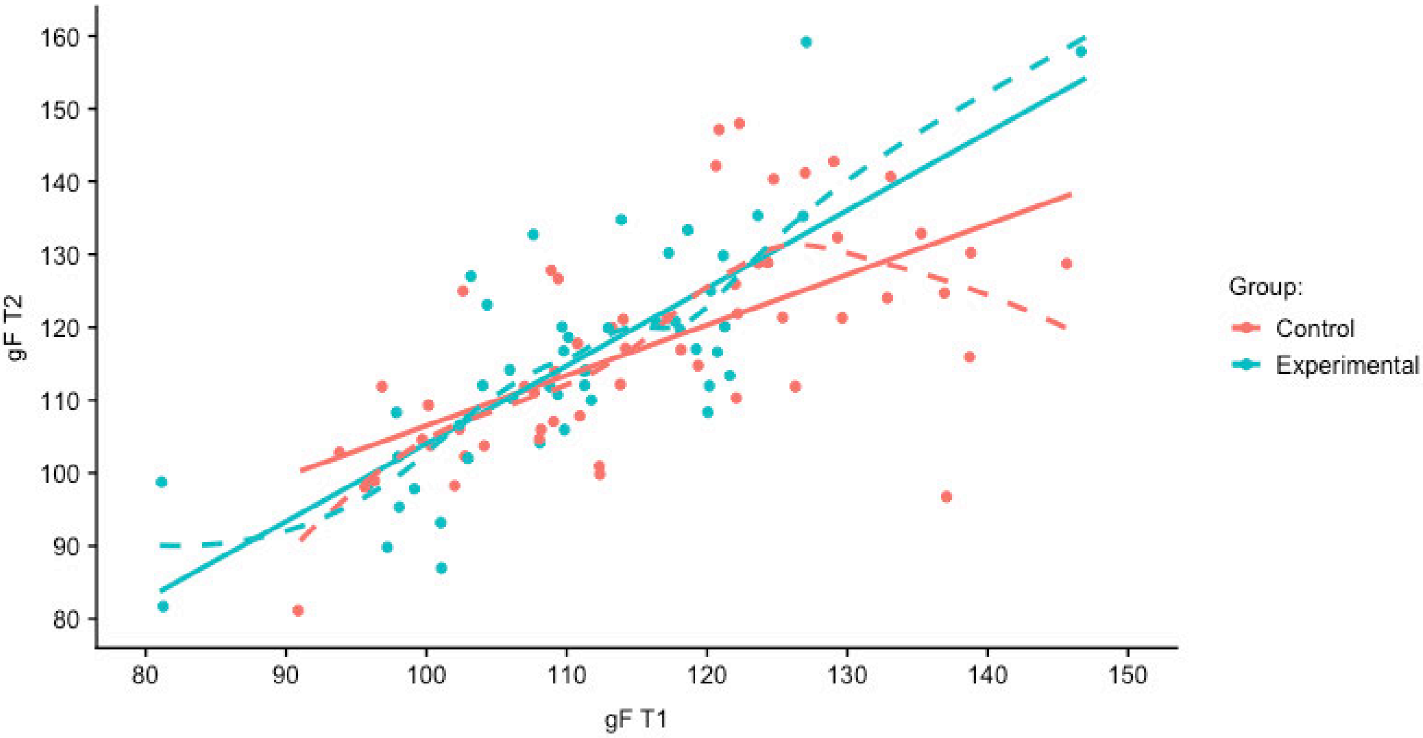


Figure 17. Interaction plot for fluid intelligence (gF).

6.10.4 Parent behavioural measures

Summary data for parent behavioural measures are presented in Table 11. For the CEFI behavioural observation measures for inhibition, shifting and WM behaviours that were completed by parents, all assumptions of running a parametric ANCOVA were met for all time periods.

Inhibition. Parametric ANCOVA showed no significant effect of intervention group on T2 scores after controlling for the T1 scores on parent behavioural reports on inhibition, $F(1, 99) = 1.31$, $p = .26$, ($\eta_p^2 = 0.01$, $CI_{90\%}[0.00, 0.07]$) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 99) = 34.17$, $p < .001$, ($\eta_p^2 = 0.25$, $CI_{90\%}[0.13, 0.36]$). Similarly, for the T3 parent observed inhibition behaviours, there was no significant effect of intervention group after controlling for the T1 observations, $F(1, 96) = 0.06$, $p = .81$, ($\eta_p^2 = 0.001$, $CI_{90\%}[0.00, 0.03]$) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 96) = 29.18$, $p < .001$, ($\eta_p^2 = 0.24$, $CI_{90\%}[0.12, 0.35]$).

Shifting. Parametric ANCOVA showed no significant effect of intervention group on T2 scores after controlling for the T1 scores on parent behavioural reports on shifting, $F(1, 99) = 1.15$, $p = .29$, ($\eta_p^2 = 0.01$, $CI_{90\%}[0.00, 0.07]$) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 99) = 24.11$, $p < .001$, ($\eta_p^2 = 0.19$, $CI_{90\%}[0.08, 0.30]$). Similarly, for the T3 parent observed shifting behaviours, there was no significant effect of intervention group after controlling for the T1 observations, $F(1, 96) = 0.30$, $p = .58$, ($\eta_p^2 = 0.003$, $CI_{90\%}[0.00, 0.05]$) with the covariate, T1 scores, significantly related to the T2 scores, $F(1, 96) = 22.12$, $p < .001$, ($\eta_p^2 = 0.19$, $CI_{90\%}[0.08, 0.30]$).

WM. Parametric ANCOVA showed no significant effect of intervention group on T2 scores after controlling for the T1 scores on parent behavioural reports on WM, $F(1,$

99) = 0.02, p = .90, (η_p^2= 0.00, $CI_{90\%}$[0.00, 0.01]) with the covariate, T1 scores, significantly related to the T2 scores, F(1, 99) = 46.82, p < .001, (η_p^2= 0.32, $CI_{90\%}$[0.20, 0.43]). Similarly, for the T3 parent observed WM behaviours, there was no significant effect of intervention group after controlling for the T1 observations, F(1, 96) = 0.34, p = .56, (η_p^2= 0.003, $CI_{90\%}$[0.00, 0.05]) with the covariate, T1 scores, significantly related to the T2 scores, F(1, 96) = 39.95, p < .001, (η_p^2= 0.30, $CI_{90\%}$[0.18, 0.41]).

6.10.5 Teacher behavioural measures.

Summary data for teacher behavioural measures are presented in Table 12. For the CEFI behavioural observation measures for inhibition, shifting and WM behaviours that were completed by teachers, all assumptions of running a parametric ANCOVA were met for all time periods.

Inhibition. Parametric ANCOVA showed no significant effect of intervention group on T2 scores after controlling for the T1 scores on parent behavioural reports on inhibition, F(1, 51) = 0.31, p = .58, (η_p^2= 0.006, $CI_{90\%}$[0.00, 0.08]) with the covariate, T1 scores, significantly related to the T2 scores, F(1, 51) = 55.45, p < .001, (η_p^2= 0.59, $CI_{90\%}$[0.43, 0.68]). Similarly, for the T3 parent observed inhibition behaviours, there was no significant effect of intervention group after controlling for the T1 observations, F(1, 49) = 1.13, p = .29, (η_p^2= 0.02, $CI_{90\%}$[0.00, 0.13]) with the covariate, T1 scores, significantly related to the T2 scores, F(1, 49) = 37.63, p < .001, (η_p^2= 0.51, $CI_{90\%}$[0.34, 0.62]).

Table 11.
Parent Behavioural Observation Summary Data

	Time 1						Time 2						Time 3					
	Inhibition		Shifting		WM		Inhibition		Shifting		WM		Inhibition		Shifting		WM	
	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.
n	55	50	55	50	55	50	53	49	53	49	53	49	53	46	53	46	53	46
Mean	100.16	96.46	97.29	92.86	98.71	96.04	102.02	102.41	99.64	99.94	102.59	101.31	103.19	101.07	101.06	100.00	101.55	99.07
SD	12.20	13.71	11.91	12.67	11.90	13.63	11.88	10.57	12.06	10.60	11.42	11.17	12.52	12.14	12.32	13.83	10.88	12.63
Min.	78	69	74	72	72	68	75	78	77	77	82	75	78	76	80	74	68	68
Max.	135	128	130	121	126	129	123	122	130	121	131	129	128	120	130	130	124	127
Skew	0.49	0.23	0.46	0.15	-0.08	0.64	-0.04	-0.09	0.80	0.17	0.32	0.26	-0.04	-0.27	0.33	0.31	-0.60	0.10
Kurtosis	0.35	-0.69	0.27	-0.82	-0.50	0.10	-0.45	-0.68	0.35	-0.55	-0.20	0.00	-0.95	-0.81	-0.47	-0.80	0.73	-0.02
SE	1.65	1.94	1.61	1.79	1.60	1.93	1.63	1.51	1.66	1.52	1.57	1.60	1.72	1.79	1.69	2.04	1.49	1.86

Note. WM = working memory; Ctrl = Control group; Exp. = Experimental group.

Table 12.
Teacher Behavioural Observation Summary Data

	Time 1						Time 2						Time 3					
	Inhibition		Shifting		WM		Inhibition		Shifting		WM		Inhibition		Shifting		WM	
	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.	Ctrl	Exp.
n	29	26	29	26	29	26	29	25	29	25	29	25	28	25	28	25	28	25
Mean	109.52	96.73	110.55	100.69	108.24	97.35	110.86	99.84	114.21	102.32	110.69	99.80	106.50	93.56	106.46	94.76	104.21	96.12
SD	14.66	13.41	14.62	13.94	14.12	12.70	15.31	12.61	14.10	12.66	14.07	12.46	15.98	14.90	13.81	14.01	14.01	15.07
Min.	79	75	77	77	79	76	78	76	89	73	82	67	74	64	80	77	74	69
Max.	128	128	134	122	131	119	129	124	138	126	129	122	128	123	124	122	126	128
Skew	-0.29	0.29	-0.61	0.08	-0.42	-0.01	-0.50	0.22	-0.44	-0.20	-0.72	-0.20	-0.37	0.21	-0.36	0.78	-0.70	0.42
Kurtosis	-1.10	-0.09	-0.21	-1.02	-0.64	-0.80	-0.90	-0.47	-0.82	0.24	-0.42	0.87	-1.18	-0.29	-1.17	-0.46	0.15	-0.48
SE	2.72	2.63	2.71	2.73	2.62	2.49	2.84	2.52	2.62	2.53	2.61	2.49	3.02	2.98	2.61	2.80	2.65	3.01

Note. WM = working memory; Ctrl = Control group; Exp. = Experimental group.

Shifting. Parametric ANCOVA showed no significant effect of intervention group on T2 scores after controlling for the T1 scores on parent behavioural reports on shifting, *F*(1, 51) = 3.14, $p = .06$, (η_p^2= 0.06, $CI_{90\%}$[0.00, 0.18]) with the covariate, T1 scores, significantly related to the T2 scores, *F*(1, 51) = 33.17, $p < .001$, (η_p^2= 0.48, $CI_{90\%}$[0.31, 0.59]). Similarly, for the T3 parent observed shifting behaviours, there was no significant effect of intervention group after controlling for the T1 observations, *F*(1, 49) = 3.30, $p = .06$, (η_p^2= 0.06, $CI_{90\%}$[0.00, 0.19]) with the covariate, T1 scores, significantly related to the T2 scores, *F*(1, 49) = 19.60, $p < .001$, (η_p^2= 0.36, $CI_{90\%}$[0.18, 0.50]).

WM. Parametric ANCOVA showed no significant effect of intervention group on T2 scores after controlling for the T1 scores on parent behavioural reports on WM, *F*(1, 51) = 1.06, $p = .31$, (η_p^2= 0.02, $CI_{90\%}$[0.00, 0.12]) with the covariate, T1 scores, significantly related to the T2 scores, *F*(1, 51) = 52.19, $p < .001$, (η_p^2= 0.57, $CI_{90\%}$[0.42, 0.67]). Similarly, for the T3 parent observed WM behaviours, there was no significant effect of intervention group after controlling for the T1 observations, *F*(1, 49) = 0.42, $p = .52$, (η_p^2= 0.008, $CI_{90\%}$[0.00, 0.09]) with the covariate, T1 scores, significantly related to the T2 scores, *F*(1, 49) = 15.86, $p < .001$, (η_p^2= 0.30, $CI_{90\%}$[0.13, 0.44]).

As there were no meaningful changes in far-transfer measures as a result of the intervention, test-retest reliability for the far-transfer measures are presented in Appendix H (Section 9.8).

6.11 Improvement Expectancy

Given that results reported so far are contrary to the initial expectations of the study, it is important to look at the IEM to see if both groups expected the same improvement in near- and far-transfer from the experimental and control group interventions. A data summary of the IEM can be seen in Table 13 and Shapiro-Wilk normality tests indicated that all data were non-normal.

For expected improvement on near-transfer measures, there was no significant difference between groups on their T1 IEM Conc. scores, Welch's $t(99.91) = 0.22$, $p = .83$, $d = 0.04$, $CI_{95\%}[-0.35, 0.43]$. For expected improvement on far-transfer measures (English and Maths academic achievement), there was a significant difference between the control and experimental groups on T1 IEM English scores, Welch's $t(86.4) = 5.56$, $p < .001$, $d = 1.10$, $CI_{95\%}[0.69, 1.52]$, with the control group expecting to improve more in English. There was not a significant difference between the T1 IEM Maths scores, Welch's $t(102.8) = 1.83$, $p = .07$, $d = 0.35$, $CI_{95\%}[-0.04, 0.74]$.

For T2 IEM scores, ANCOVAs were run to determine if there was a change in the IEM as a result of the intervention group while controlling for T1 IEM scores. Parametric ANCOVAs were attempted but IEM Maths violated the assumption of homogeneity of slopes as there was an interaction present, and both IEM Conc. ($F(1, 103) = 4.48$, $p = .037$) and IEM English ($F(1, 103) = 20.17$, $p < .001$) violated the assumption of homogeneity of variance. As a result, robust ANCOVAs were run using WRS2 (Mair & Wilcox, 2018).

As there was an interaction present for IEM Maths, a scatterplot is presented in Figure 18 showing the T1 and T2 IEM Maths scores with a linear and non-linear slope for each intervention group. There is no real meaning applied to this implied interaction due to the clear lack of linearity and the apparent lack of a meaningful pattern in this data.

Table 13.

Improvement Expectancy Measure (IEM) Summary Data.

	Control Group								Experimental Group							
	n	Mean	SD	Min.	Max.	Skew	Kurtosis	SE	n	Mean	SD	Min.	Max.	Skew	Kurtosis	SE
IEM Maths T1	55	2.55	2.02	0	6	-0.18	-1.57	0.27	50	3.22	1.75	0	6	-0.69	-0.65	0.25
IEM English T1	55	3.98	1.30	0	6	-0.62	0.54	0.18	50	2.22	1.87	0	6	0.02	-1.46	0.26
IEM Conc. T1	55	3.95	1.69	0	6	-0.73	-0.16	0.23	50	4.02	1.83	0	6	-0.96	-0.13	0.26
IEM Maths T2	55	2.24	2.16	0	6	0.21	-1.56	0.29	50	2.82	1.90	0	6	-0.25	-1.21	0.27
IEM English T2	55	4.22	1.20	0	6	-0.61	1.03	0.16	50	2.60	1.99	0	6	-0.05	-1.32	0.28
IEM Conc. T2	55	3.67	1.72	0	6	-0.67	-0.33	0.23	50	3.60	2.15	0	6	-0.61	-1.08	0.30

Note. IEM = Improvement Expectancy Measure; Conc. = concentration. T1 IEM data was collected two weeks after beginning the training intervention and T2 IEM data was collected on the last training intervention session.

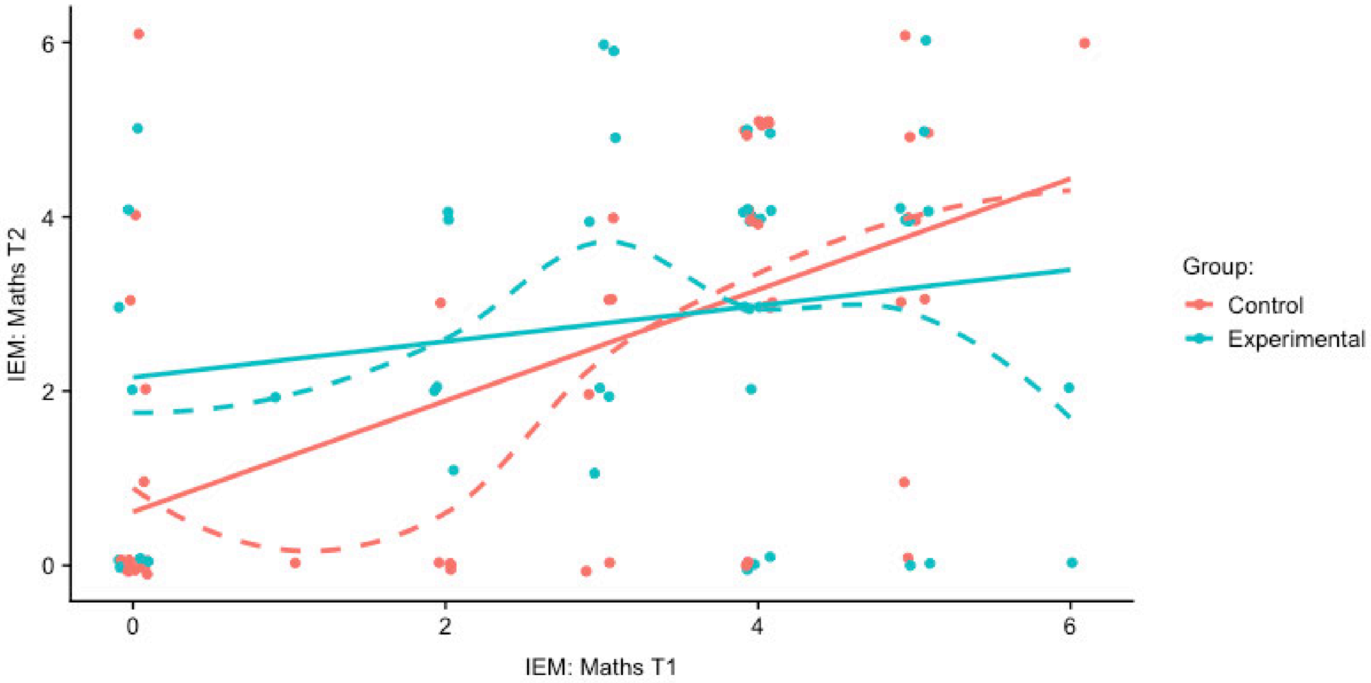


Figure 18. Interaction plot for Improvement Expectancy Measure (IEM) for maths.

Robust ANCOVAs showed that there was no significant effect of the intervention group on T2 IEM Maths or Conc. scores after controlling for T1 IEM scores, as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). There was a significant effect of the intervention group on T2 IEM English scores after controlling for T1 IEM English scores, but only for the T1 score of 4 (difference = 1.12, *p* = .048, $CI_{95\%}$[0.02, 2.22], *n* = 76). This showed that there was a possible increase in the control group's expectancy of improvement in English due to completing their intervention. Figure 19 presents a scatterplot of the T1 and T2 IEM English scores with a linear and non-linear slope for each intervention group and a vertical dotted line representing the point where a significant difference occurred.

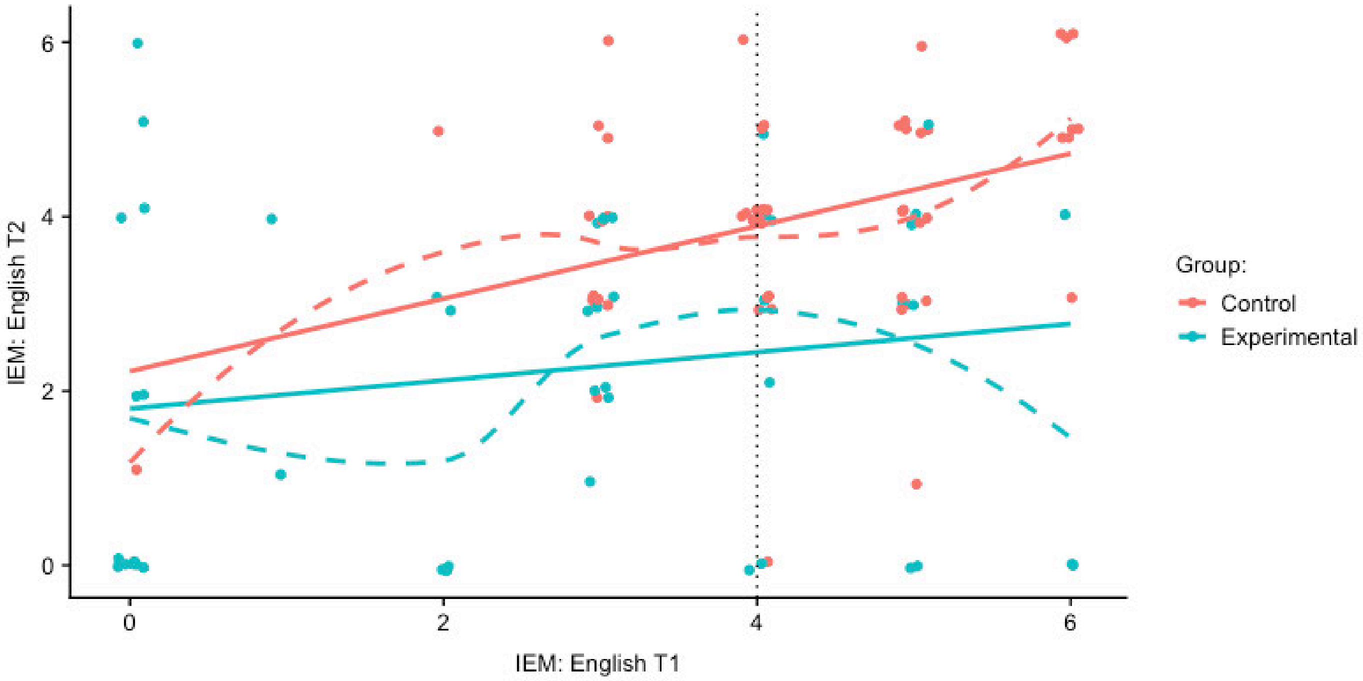


Figure 19. Improvement Expectancy Measure (IEM) for English demonstrating the robust analysis of covariance (ANCOVA) point of comparison.

6.12 Implicit Theory of Intelligence

Summary data for the Implicit Theory of Intelligence Measure (ITIM) is presented in Table 14. There was no difference at baseline between the two intervention groups according to Welch's *t*-test, *t*(96.4) = 0.302, *p* = .763, *d* = -0.06, $CI_{95\%}$[-0.45, 0.33]. Firstly, to see if there was any difference between those with a high ITIM (malleable) and low ITIM (fixed) on changes in latent factor scores, participants were separated dichotomously using a median split (e.g., Jaeggi et al., 2014)[1]. As a result, there was no difference in gain scores between those that believed intelligence was fixed and those that believed it was malleable for either EF, *t*(103) = -1.14, *p* = .26,

---

[1] The weaknesses of this form of data analysis is recognised (e.g. Tidwell et al., 2014), but this was included as a comparison to the Jaeggi et al. (2014) study that used this method and was referred to on p. 56 (Section 3.5).

or WM, *t*(103) = -0.25, *p* = 0.80. All correlations between ITIM and factor score gains for each intervention group were small (largest *r* = -0.09) and non-significant (all *p*'s > .50).

Table 14.
Implicit Theory of Intelligence Measure (ITIM) Summary Data

	Control	Experimental
n	55	50
Mean	26.53	26.18
SD	5.35	6.34
Min.	11	10
Max.	36	36
Skewness	-0.48	-0.75
Kurtosis	0.12	0.18
SE	0.72	0.90

To test the hypothesis that ITIM moderates training gains, a two-way interaction was performed with T2 factor scores as the outcome variable with T1 factor score as the predictor and ITIM as moderator. Analysis was performed in R (R Core Team, 2019) using a robust linear model from the MASS package (Venables & Ripley, 2002). As seen in Table 15, there was no significant interaction present between ITIM and the relevant factor score.

To determine if the ITIM moderated training gains dependent upon group, a three-way interaction was tested with T2 factor scores as the outcome variable, T1 factor scores as the predictor, and both ITIM and group as moderator (*Hypothesis 3*). As seen in Table 16, the three-way interaction between ITIM, WM Factor Score and Group was not significant (*b* = 0.04, t(97) = 1.79, *p* = .075) but the three-way interaction between ITIM, EF Factor Score, and Group was significant (*b* = 0.06, t(97) = 2.46, *p* = .016). Plots of the three-way interactions for WM and EF can be seen in Figure 20 and Figure 21 respectively.

Table 15.
Two-Way Interaction Model Output.

	Dependent variable:	
	EF Factor Score T2	WM Factor Score T2
ITIM	0.0002	0.01
	(0.01)	(0.01)
EF Factor Score T1	0.90^{**}	
	(0.28)	
ITIM:EF Factor Score T1	-0.003	
	(0.01)	
WM Factor Score T1		0.88^{**}
		(0.33)
ITIM:WM Factor Score T1		-0.002
		(0.01)
Constant	-0.003	-0.14
	(0.25)	(0.28)
Observations	105	105
Residual Std. Error (df = 101)	0.50	0.56

Note: ITIM = Implicit Theory of Intelligence Measure; EF = executive function; WM = working memory.
$^{**}p<0.01$

Table 16.
Three-Way Interaction Model Output.

	Dependent variable:	
	EF Factor Score T2	WM Factor Score T2
ITIM	0.006 (0.01)	0.004 (0.01)
EF Factor Score T1	1.67^{**} (0.49)	
WM Factor Score T1		1.29^{**} (0.49)
Group	-0.39 (0.49)	-0.86 (0.49)
ITIM:EF Factor Score T	-0.04^{*} (0.02)	
ITIM:WM Factor Score T1		-0.02 (0.02)
ITIM:Group	-0.001 (0.02)	0.01 (0.02)
EF Factor Score T1:Group	-1.22^{*} (0.59)	
ITIM:EF Factor Score T1:Group	0.06^{*} (0.02)	
WM Factor Score T1:Group		-0.90 (0.61)
ITIM:WM Factor Score T1:Group		0.04 (0.02)
Constant	0.07 (0.36)	0.15 (0.35)
Observations	105	105
Residual Std. Error (df = 97)	0.42	0.45

Note: ITIM = Implicit Theory of Intelligence Measure; EF = executive function; WM = working memory.
$^{*}p < .05$ $^{**}p < .01$

Figure 20 and Figure 21 demonstrate that the pattern of the three-way interaction was similar for both the WM and combined EF factor scores respectively; however, only the EF interaction was statistically significant. This appears to show that the effect of ITIM on T2 combined EF score differs between groups depending upon the initial T1 combined EF score. This means that when a participant in the experimental group scored lower on their T1 combined EF score, but had a higher ITIM score (which means a more malleable view of their intelligence), they tended to improve more on their T2 combined EF score. In contrast, someone who scored lower but had a lower ITIM score (which means a more fixed view of their intelligence), they tended to improve less. The results were the opposite for the control group.

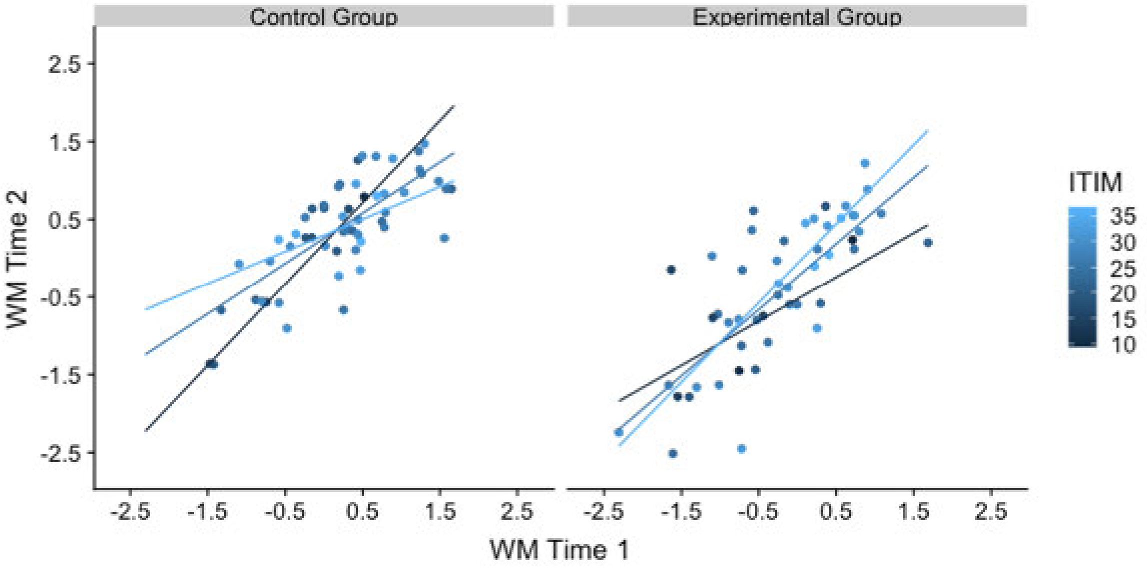


Figure 20. Working memory (WM) factor score three-way interaction with Implicit Theory of Intelligence Measure (ITIM) and group.

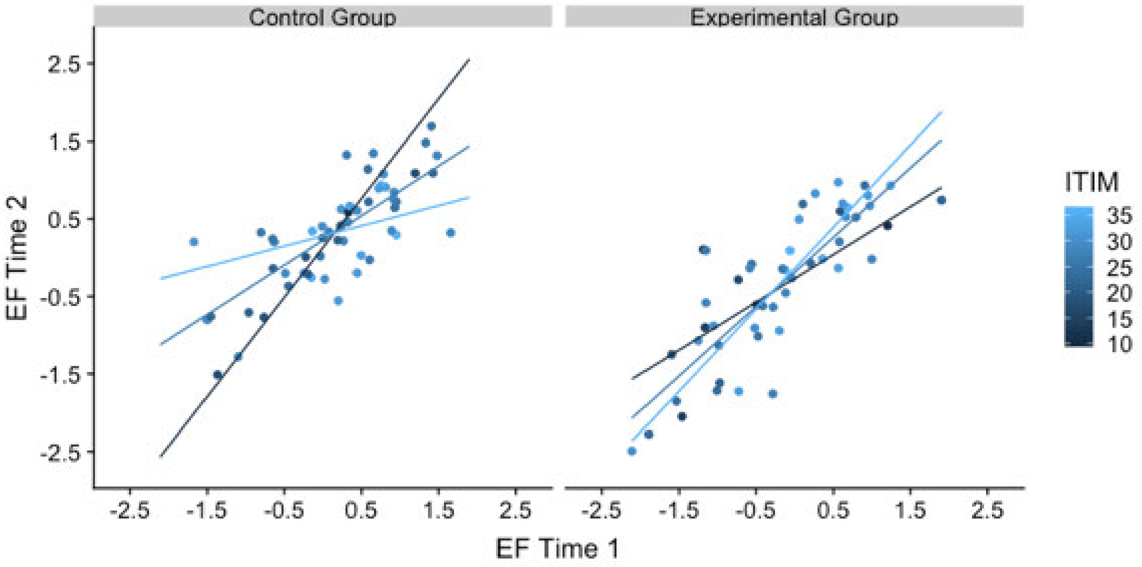


Figure 21. Combined executive function (EF) factor score three-way interaction with Implicit Theory of Intelligence Measure (ITIM) and group.

However, it is important to determine if linearity assumptions are correct when looking at interaction plots (Hainmueller, Mummolo, & Xu, 2018). Using the jtools package (Long, 2019) in R (R Core Team, 2019), it can be seen in Figure 22 that the EF data does not appear to fit assumptions of linearity when the ITIM (centred data, hence ITIMc) is split into terciles. Importantly, the small sample sizes that determine this interaction can be seen in Figure 22 and sample sizes required to identify a reliable three-way interaction may be significantly higher than were present in this study (Gelman, 2018; Heo & Leon, 2010).

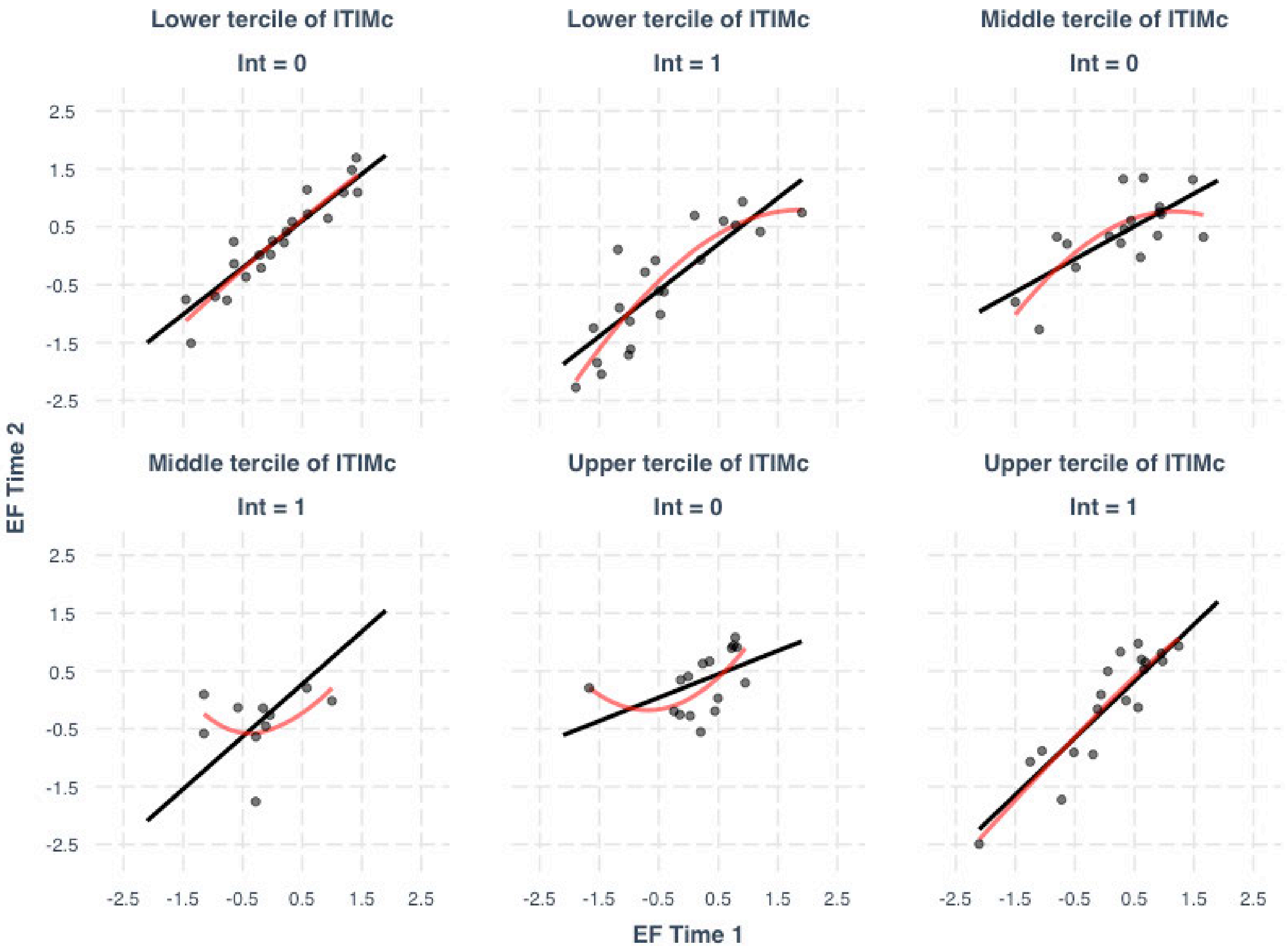


Figure 22. Linearity plots for combined executive function (EF) factor score three-way interaction.

6.13 Treatment Fidelity

Treatment fidelity adherence was obtained within both the experimental and control tasks. A variety of results were obtained from the experimental task, depending upon the training game being played. These results included scores (where a high score indicated better performance in the game and movement to higher levels of difficulty), plateaus (a count of the number of times a participant had to be moved up a level due to not improving sufficiently), longest sequence, and average sequence length. In the control task, results were in the form of a score (where a high score indicated better performance in the typing program and movement to higher levels of difficulty), average typing accuracy (as a percentage), the average typing speed (measured in words per minute), and attempts (the number of attempts made overall at each typing task).

For the control group (see Table 17) there was a significant negative correlation between the difference in the WM factor score and the overall score ($r(48) = -.33$, $p = .02$, $CI_{95\%}[-.56, -.05]$) and the average typing speed ($r_{\tau}(48) = .25$, $p < .05$, $CI_{95\%}[-.42, -.07]$). These results are in the opposite direction of what would be expected if the typing program was expected to improve performance on the EF measures. That is, those who were performing worse on the typing program (i.e., progressing less as indicated by a lower overall score and a lower average typing speed) seemed to improve more on the WM measure. This seems to indicate a spurious correlation as the same pattern was not seen for the combined EF measure.

Table 17.
Treatment Fidelity Data for the Control Group.

Typing Club Variable	EF Difference Score (Time 2 – Time 1)	WM Difference Score (Time 2 – Time 1)
Score	-.02[a] [-.30, .27]	-.33*[a] [-.56, -.05]
Average Typing Accuracy	.17 [-.08, .39]	.16 [-.08, .40]
Average WPM	-.04 [-.23, .15]	-.25* [-.42, -.07]
Attempts	.09 [-.13, .32]	.02 [-.19, .22]

Note. EF = executive function; WM = working memory; WPM = words per minute. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).
[a] Pearson's *r* (all other correlations are Kendall's *tau*). All Kendall's *tau* confidence intervals are bootstrapped.
* $p < .05$;

To determine if those participants who displayed better treatment adherence improved more as a result of their training, correlations were calculated between a latent factor score difference (T2 - T1) and each of the treatment adherence measures. The stats package (R Core Team, 2019) was used for Pearson and Kendall's *tau* correlations and the NSM3 package (Schneider, Chicken, & Becvarik, 2018) was used for Kendall's *tau* correlation confidence intervals. As seen in Table 18, there was a significant negative correlation between the number of plateaus in both the Treasure Trunk game ($r_\tau(50)= -.28$, $p = .007$, $\mathrm{CI}_{95\%}[-.46, -.07]$) and the Magic Lens game ($r_\tau(50) = -.24$, $p = .025$, $\mathrm{CI}_{95\%}[-.44, -.02]$) and improvement on the combined EF factor score. On Monkey Trouble there was a significant positive correlation between the longest sequence recalled and improvement on the combined EF factor score ($r_\tau(50) = .25$, $p = .024$, $\mathrm{CI}_{95\%}[.03, .44]$) and significant positive correlations between the average length recalled and improvement on both the combined EF factor score ($r_\tau(50) = .26$, $p = .012$, $\mathrm{CI}_{95\%}[.08, .43]$) and the WM factor score ($r_\tau(50) = .22$, $p = .028$, $\mathrm{CI}_{95\%}[-.003, .43]$). Importantly, after correcting for multiple comparisons using Holm's method (1979), only the correlation between the number of plateaus in the Treasure Trunk Game was still significant ($p = .042$).

Table 18.

Treatment Fidelity Data for the Experimental Group.

Game	Variable	EF Difference Score (Time 2 – Time 1)	WM Difference Score (Time 2 – Time 1)
Pirate Pete's Packing Panic	Panic Points	.23[a] [-.05, .48]	.26[a] [-.02, .50]
	Plateaus	-.13 [-.37, .09]	-.15 [-0.35, .06]
Ducks!	Points	-.07[a] [-.34, .22]	.11[a] [-.18, .37]
	Plateaus	-.06 [-.27, .15]	-.08 [-.30, .15]
Treasure Trunk	Points	.19 [.01, .35]	.14 [-.07, .34]
	Plateaus	-.28** [-.46, -.07]	-.15 [-.33, .05]
The Magic Lens	Points	.11 [-.09, .31]	-.04 [-.22, .16]
	Plateaus	-.24* [-.44, -.02]	-.21 [-.42, .03]
Monkey Trouble	Points	.02[a] [-.26, .30]	-.12[a] [-.39, .16]
	Longest Sequence	.25* [.03, .44]	.18 [-.02, .36]
	Average Length	.26* [.08, .43]	.22* [-.00, .43]
Grub Ahoy!	Points	.17[a] [-.11, .43]	.01[a] [-.27, .29]
	Longest Sequence	.16 [-.05, .37]	.08 [-.15, .30]
	Average Length	.10 [-.09, .29]	.17 [-.07, .39]

Note. EF = executive function; WM = working memory. Points: Higher points equates to more progress made and better performance within the game; Plateaus: The number of times a participant's level was increased as a result of a plateau in achievement or performance. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).

[a] Pearson's *r* (all other correlations are Kendall's *tau*). All Kendall's *tau* confidence intervals are bootstrapped.

* $p < .05$; ** $p < .01$

6.14 Individual Differences

To test the hypothesis of either magnification or compensation effects in explaining training-related performance gains for the experimental group, correlations were calculated between baseline performance on the academic achievement and gF measures and difference scores for the near-transfer measures (e.g., Karbach, Strobach & Schubert, 2015). As Shapiro-Wilk normality tests indicated that all data were normal, Pearson correlations were calculated. None of the correlations were significant in either the positive or negative direction, appearing to rule out both magnification and compensation effects respectively (see Table 19).

Table 19.
Individual Differences Correlations.

		Pearson Correlation [95% CI]	
		Combined EF Score Difference (Time 2 – Time 1)	WM Score Difference (Time 2 – Time 1)
Academic Achievement	PAT Maths	-0.16 [-0.42, 0.13]	-0.08 [-0.35, 0.20]
	PAT Reading	-0.11 [-0.38, 0.17]	-0.03 [-0.31, 0.25]
Fluid Intelligence	NNAT2	0.05 [-0.23, 0.33]	0.13 [-0.15, 0.40]

Note. EF = executive function; WM = working memory; PAT = Progressive Achievement Test; NNAT2 = Naglieri Nonverbal Achievement Test – Second Edition.

However, it is important to note that the sample used in this study was much better performing than a normal sample, thereby reducing the ability to ascertain if compensation effects were occurring. An indication of the difference of this sample of students when compared to a normal sample can be seen in performance on the gF measure. The 55 participants in the control group had an average score of 115.96 (*SD* = 13.27) and the 50 participants in the experimental group had an average of 110.2 (*SD* =

11.48). The NNAT2 was standardised on 100,000 students with an average of 100 (SD = 16; Naglieri, 2011) and a one way analysis of variance showed that there was a significant difference between the normed, control, and experimental groups, $F(2, 100102) = 37.50$, $p < .001$. Tukey HSD post-hoc tests showed that there was a significant difference between the normed sample and both the control (difference = -15.96, CI$_{95\%}$[-21.02, -10.90], $p < .001$, $d = -1.0$, CI$_{95\%}$[-1.26, -0.73]) and experimental (difference = -10.20, CI$_{95\%}$[-15.50, -4.90], $p < .001$, $d = -0.64$, CI$_{95\%}$[-0.91, -0.36]) groups but no difference between the control and experimental groups (difference = 5.76, CI$_{95\%}$[1.57, 13.09], $p = 0.16$, $d = 0.46$, CI$_{95\%}$[0.07, 0.86]). Therefore, it is difficult to categorically rule out compensation effects as the participants in this study were comparatively high-achieving.

Given that this was a high-ability group, it was decided to look closer at potential magnification effects. In a post-hoc attempt to determine if there were cognitive (or motivational) characteristics of the experimental group that demonstrated why they showed improvement when compared to those who did not, participants were separated dichotomously using a median split on their change scores in EF and WM (T2 - T1) resulting in a group that improved more (High) and a group that improved less (Low). They did not differ on the age of the participant for either the EF groups, Welch's $t(47.89) = 0.32$, $p = .75$, , $d = 0.09$, CI$_{95\%}$[-0.48, 0.66], or WM groups, Welch's $t(47.39) = -0.02$, $p = .99$, $d = -0.01$, CI$_{95\%}$[-0.58, 0.56].

As the alternative hypothesis of magnification effects is that those in the group that improved more should have higher baseline scores (i.e., High > Low), a one-sided Welch's *t*-test was performed to determine baseline differences between groups. As there was very low power to find differences between the groups using conventional *p*-values (54% power to detect a Cohen's *d* effect size of 0.5 using $p = .05$, (Faul et al., 2007)) and due to this investigation being exploratory, Bayes Factors with default priors were also presented to give an indication of likelihood for the null hypothesis (BF_{01}) which is that there is no difference in baseline measures between groups. The only exception for the alternative hypothesis was for the PS baseline measure, where it was expected that the High group might be less than the Low group (i.e., faster PS). As nearly all measures were non-parametric as determined by Shapiro-Wilk, Bayes Factors were calculated in JASP (JASP Team, 2018) as this was the only way the researcher could do Mann-Whitney non-parametric calculations.

As seen in Table 20, it would appear that magnification effects are not occurring within this sample of participants as there were no apparent differences on baseline measures of cognitive characteristics (e.g., maths achievement, reading comprehension achievement, or g*F*) between the groups that improved more or less on near-transfer measures. The null hypothesis ranged from being 1.44 times more likely (NAI; WM) to 4.98 times more likely (PAT Reading; EF) with this data on the baseline cognitive measures. However, there was weak evidence (as determined by the descriptions of Bayes Factors by Raftery [1995]) that there were different motivational or expectation effects between those who improved more on the WM near-transfer measure and those who did not improve. The alternative hypothesis, that the group that improved more on WM had higher baseline scores on the ITIM measure (a more malleable view of intelligence), was 1.72 times more likely than the null hypothesis, whereas the hypothesis that the group that improved more on WM thought they would improve more with regards to concentration was 1.39 times more likely than the null hypothesis. There was equivocal evidence for a difference between the WM groups on the IEM Maths (1.02) and PS (1.05).

Table 20.
Post-Hoc Magnification Effects Data.

		Mean (SD)				
		Low[a]	High[a]	Cohen's *d* [CI$_{95\%}$]	*p*-value	BF$_{01}$
PAT Maths	EF	130.51 (10.35)	130.03 (11.82)	0.04 [-0.53, 0.61]	.561	3.26
	WM	129.27 (10.82)	131.27 (11.31)	-0.18 [-0.75, 0.39]	.263	1.86
PAT Reading	EF	131.06 (12.37)	128.96 (13.06)	0.17 [-0.40, 0.73]	.719	4.98
	WM	129.63 (14.03)	130.38 (11.35)	-0.06 [-0.63, 0.51]	.418	3.09
NAI (gF)	EF	109.96 (10.31)	110.44 (12.75)	-0.04 [-0.61, 0.53]	.442	3.65
	WM	108.16 (12.03)	112.24 (10.75)	-0.36 [-0.93, 0.22]	.106	1.44
PS	EF	394.82 (64.24)	392.29 (49.99)	0.04 [-0.52, 0.61]	.439	3.57
	WM	405.33 (65.28)	381.79 (45.6)	0.42 [-0.16, 0.99]	.073	1.05
ITIM	EF	26.16 (5.67)	26.2 (7.06)	-0.01 [-0.57, 0.56]	.491	2.57
	WM	24.60 (6.90)	27.76 (5.40)	-0.51 [-1.09, 0.07]	.027	0.58
IEM Conc.	EF	4.20 (1.58)	3.84 (2.08)	0.19 [-0.38, 0.76]	.753	4.83
	WM	3.52 (2.06)	4.52 (1.45)	-0.56 [-1.14, 0.02]	.027	0.72
IEM Maths	EF	3.16 (1.82)	3.28 (1.72)	-0.07 [-0.64, 0.50]	.406	2.87
	WM	2.80 (1.96)	3.64 (1.44)	-0.49 [-1.07, 0.09]	.045	1.02
IEM English	EF	1.84 (1.93)	2.60 (1.76)	-0.41 [-0.99, 0.16]	.076	1.10
	WM	1.96 (1.74)	2.48 (1.98)	-0.28 [-0.85, 0.29]	.165	1.58

Note. EF = executive function; WM = working memory; PAT = Progressive Achievement Test; NAI = Naglieri Ability Index; gF = fluid intelligence; PS = processing speed; ITIM = Implicit Theory of Intelligence Measure; IEM Conc. = Improvement Expectancy Measure (concentration question); IEM Maths = Improvement Expectancy Measure (mathematics question); IEM English = Improvement Expectancy Measure (English question).
[a]*n* = 25 for each group. Low / High = the change scores for EF and WM (T2 – T1) separated dichotomously using a median split into low improving (Low) and high improving (High) groups.
p-values were calculated using Welch's *t*-test. BF$_{01}$ = Bayes Factor as support for null hypothesis (H0) over the alternative hypothesis (H1), with a higher Bayes Factor indicating more support for the null. All Bayes Factors calculated using the Mann-Whitney test in JASP (JASP Team, 2018).

6.15 Differences Between Grades

Despite random allocation, the groups were different with regard to their baseline ability and therefore some post-hoc tests were run to investigate where these differences were occurring. When groups were split into grade 3 and grade 5 (see Table

21), it appeared that the differences between the groups came from the grade 3 control group performing significantly better than the grade 3 experimental group on baseline measures. There were no significant differences between the groups for the grade 5 baseline measures.

Given that power was very low when splitting the samples up into their grades (47% for Grade 3 and 65% for Grade 5, to detect a Cohen's *f* effect size of 0.3 using ANCOVA (Faul et al., 2007)), it was not useful to split the sample up for analysis. However, graphs demonstrating the changes in latent scores clearly show there would be no change to the outcomes of this study: see Figure 23 and Figure 24 below.

Table 21.
Summary Data by Grade.

	Mean (SD)			
Grade 3	Control (n = 20)	Experimental (n = 22)	Difference	Cohen's *d* [95% CI]
Combined EF	-0.26 (0.65)	-0.93 (0.63)	-0.67**	-1.07 [-1.74, -0.40]
WM	-0.22 (0.66)	-0.91 (0.63)	-0.69**	-1.07 [-1.74, -0.40]
PAT Maths	128.03 (6.25)	122.31 (8.15)	-5.72*	-0.78 [-1.43, -0.13]
PAT Reading	127.13 (12.22)	121.77 (12.35)	-5.36	-0.44 [-1.07, 0.20]
NAI (gF)	118.65 (12.47)	108.00 (10.00)	-10.65**	-0.95 [-1.61, -0.29]
Grade 5	Control (n = 35)	Experimental (n = 28)	Difference	Cohen's *d* [95% CI]
Combined EF	0.44 (0.78)	0.37 (0.65)	-0.07	-0.1 [-0.60, 0.41]
WM	0.48 (0.72)	0.27 (0.62)	-0.21	-0.31 [-0.82, 0.20]
PAT Maths	137.74 (8.02)	136.52 (8.71)	-1.22	-0.15 [-0.65, 0.36]
PAT Reading	137.67 (10.83)	136.48 (8.51)	-1.19	-0.12 [-0.63, 0.39]
NAI (gF)	114.43 (13.65)	111.93 (12.43)	-2.5	-0.19 [-0.70, 0.32]

Note. EF = executive function; WM = working memory; PAT = Progressive Achievement Test; NAI = Naglieri Ability Index; gF = fluid intelligence.
* $p < .05$; ** $p < .01$

For the combined EF score in Figure 23 it is evident that there is no improvement in the grade 5 experimental group at T2 with the variance of the scores appearing to reduce. For the grade 5 control group, there does not appear to be much change at all across all three time periods. The grade 3 experimental group appears to show increasing variance in scores over time in the combined EF score.

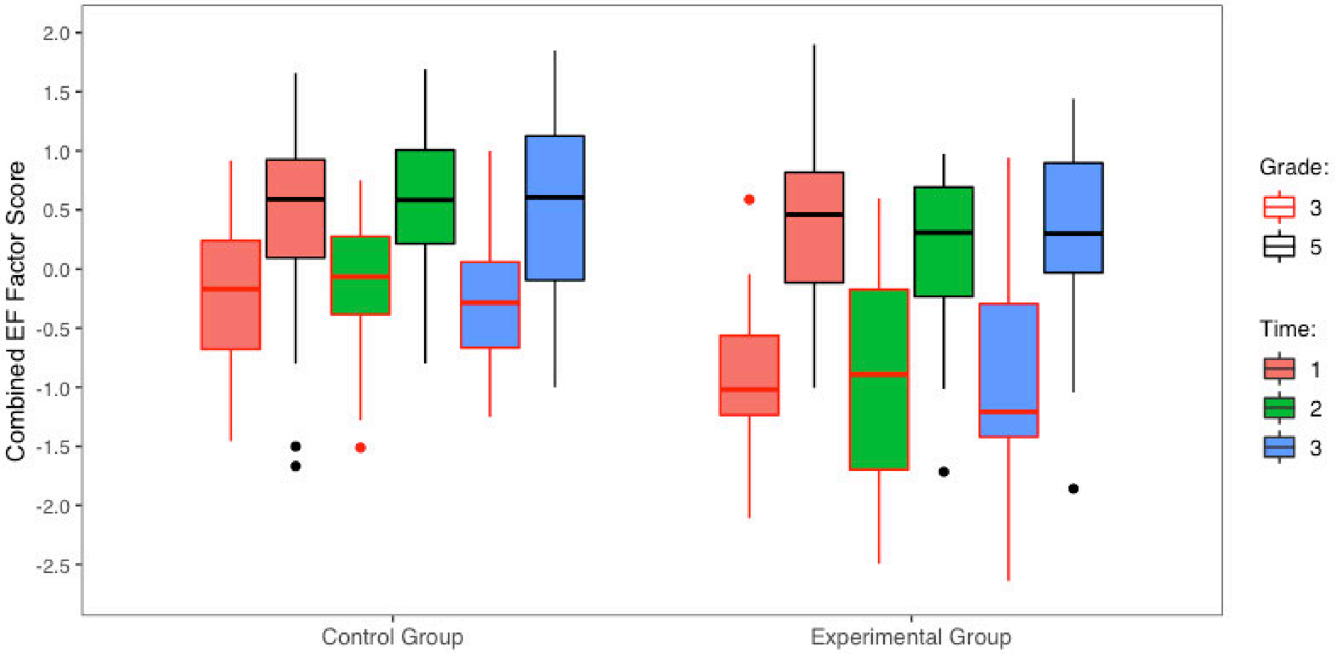


Figure 23. Combined executive function (EF) factor score split by grade.

For the WM score in Figure 24 it can be seen that the only group that appears to improve from T1 to T2 is the grade 3 control group. All other T1 to T2 scores appear somewhat equal.

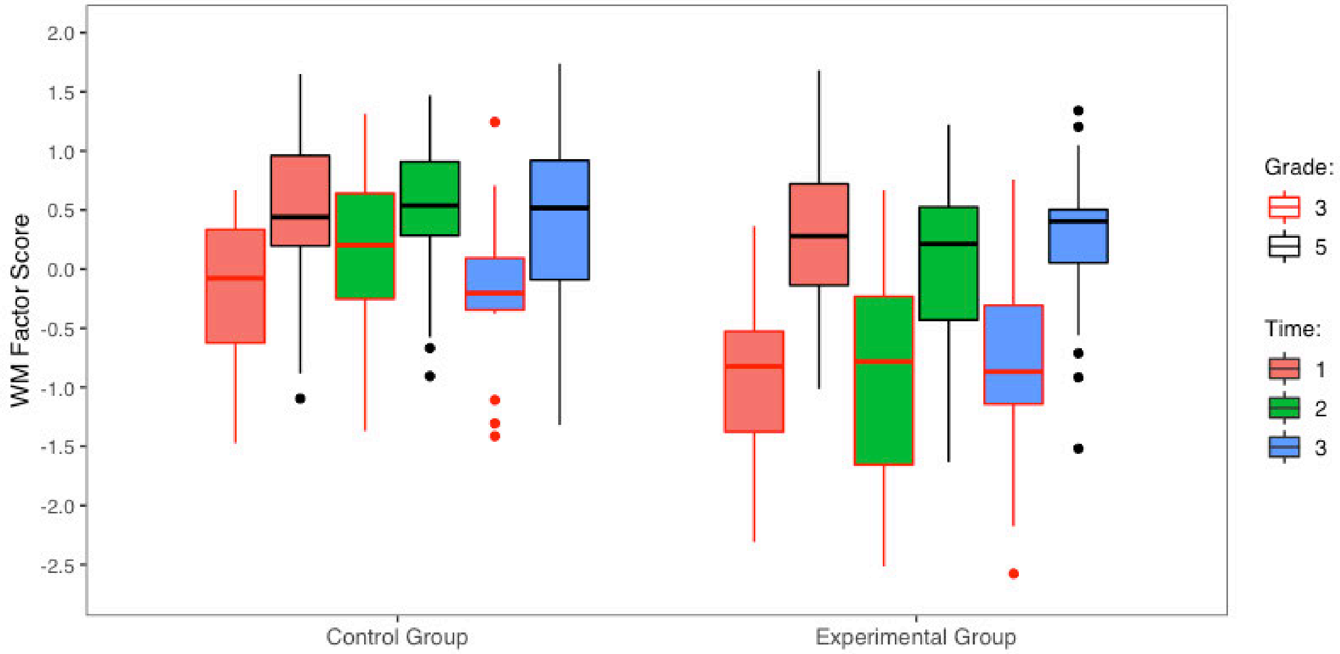


Figure 24. Working memory (WM) factor score split by grade.

7 Discussion

7.1 Introduction

This study investigated whether school-based computerised adaptive EF training improved near-transfer EF measures using a latent variable approach, as well as whether the training impacted academic and behavioural aspects of school performance, and gF in normally functioning children. Importantly, this study addressed many of the criticisms of past cognitive training research by using a training program that is more aligned to current EF theory, using multiple measures for each EF construct, and using randomly allocated experimental and active control groups.

The following section provides an overview of findings as they relate to the study's research questions, any noteworthy conclusions relevant to existing theories or previous research, and implications for attempting to ameliorate EF difficulties in schools. Limitations of the current study and suggestions for future research are also presented.

7.2 Near-Transfer

7.2.1 Executive Function

Most of the cognitive training research to-date has focussed upon single measures for near-transfer (Aksayli et al., 2019). Given that any single measure of a theoretical construct will be contaminated by error variance (Hulme & Melby-Lervåg, 2012) and the use of latent variables for construct measurement can exclude explanations based upon task-specific processes or abilities (Shipstead, Redick, et al., 2012), the use of latent variables to determine near-transfer for EF abilities was a methodological strength of this study.

The EF model determined by this study provided a good fit to the data even though the sample size was not large with regard to conducting a CFA. A two-factor model of EF with a combined inhibition-shifting component and separate but related WM component was found. Although a three-factor model proposed by Miyake et al. (2000) was expected, this two-factor model has been found in other EF research in children (Lee et al., 2013, 2012; van der Ven et al., 2013) and is consistent with a recent review of latent variable studies (Karr et al., 2018), where there is a reported increasing multidimensionality of EF as a child develops. This study is also consistent with previous findings of a non-separable shifting factor earlier in a child's development (Garon et al., 2008) and the view that the shifting factor emerges as an independent construct later in a child's development (Müller & Kerns, 2015). This study also confirmed the high correlations reported between EF constructs (Friedman et al., 2008; Salthouse, Atkinson, & Berish, 2003).

Importantly, similar to results in past research, the factor scores that were determined as a result of the CFA were highly correlated with other academic measures (e.g., Blair & Razza, 2007; Bull et al., 2008; St Clair-Thompson & Gathercole, 2006; Toll et al., 2011; Yeniad et al., 2013) and gF (e.g., Brydges et al., 2012; Richland & Burchinal, 2013; van der Sluis et al., 2007), and they improved upon the reliability of individual EF measures. Therefore, the latent scores appeared to be valid and reliable measures for each EF construct.

Most significantly, relating to the core aim of this research program, this study did not show any improvement in T2 or T3 WM or combined EF latent factor scores due to completing the cognitive training program (*Hypothesis 1*). In fact, the only significant differences were found on T2 WM latent scores where the control group scored higher than the experimental group after controlling for T1 scores. Results were similar if outliers were removed from the data.

Although the central aim of this study provided a null result, this is a significant finding as numerous meta-analyses of the cognitive training literature that have been undertaken since this study started (e.g., Aksayli et al., 2019; Kassai et al., 2019; Melby-Lervåg et al., 2016; Sala et al., 2018; Sala & Gobet, 2020; Soveri, Antfolk, Karlsson, Salo, & Laine, 2017) are unconvinced of the claims of cognitive training programs, especially around the use of single items for measurement of near-transfer and also regarding the long-term outcomes. Given that optimistic results from cognitive training studies appear to be related to poor experimental design (Blacker et al., 2017; Sala & Gobet, 2017; Simons et al., 2016), the present study makes a substantial contribution to that corpus of evidence through careful consideration of design issues.

The use of latent factors in cognitive training research is particularly important (Protzko, 2017). As outlined by Aksayli et al. (2019) in their recent meta-analysis:

> Improvements on a latent factor extracted from a broad set of memory tasks would represent far more compelling evidence of cognitive enhancement than that often provided in the reviewed primary studies, which are based on few observed measures. Such an experimental design would dramatically contribute to settling the debate regarding the true significance of near transfer induced by… [a] cognitive-training program (p. 240).

Therefore, as no improvement on near-transfer latent scores (which appeared to valid and reliable measures of EF) was found as a result of the cognitive training program, this study contributes towards an understanding of the potential true (null) effects of this type of computerised cognitive training.

7.2.2 Processing speed

This study found that PS was correlated with each of the factor scores for EF and is consistent with findings that EFs are related to PS (Friedman et al., 2008; Hedden & Yoon, 2006; Salthouse et al., 2003). The size of the correlations found in this study were similar to those reported by Friedman et al. (2008) and Hedden and Yoon (2006). The results from this study with regard to PS are also consistent with Fry and Hale's (1996) developmental cascade theory for PS. According to the developmental cascade theory for PS, an improvement in PS should see an improvement in WM. Given that there was no apparent improvement in PS from T1 to T2, an increase in WM should not be expected and was not seen in this study.

7.3 Far-Transfer

A related aim of this study was to determine if higher cognitive abilities, such as gF and academic achievement, could be improved as a result of the improvement in near-transfer EF abilities (*Hypothesis 2*). Additionally, if there was near-transfer as a result of the cognitive training program, it would be expected that teacher and parent observations of behaviour would show related improvement. Given that there was no improvement in near-transfer EF abilities as a result of the cognitive training program, it would be unexpected if there was any concomitant improvement in far-transfer abilities.

As expected, this study found no evidence for improvement on any of the far-transfer outcome measures (academic, gF, or behavioural) as a result of the cognitive training program. This is consistent with much of the cognitive training literature, where far-transfer does not occur (e.g., Aksayli et al. 2019; Chein & Morrison, 2010; Chooi & Thompson, 2012; Dahlin et al., 2008; Harrison et al., 2013; Hitchcock & Westwell, 2017; Redick, Shipstead, et al., 2012; Roberts et al., 2016; Thompson et al., 2013; Westerberg et al., 2007).

7.4 Ancillary Measures

7.4.1 Improvement expectancy

A number of ancillary measures were investigated, including using a measure of improvement expectancy to rule out Hawthorne or placebo effects, as suggested by Boot et al. (2013). The results showed that both groups were equivalent on their expected improvement on near-transfer measures and their expected improvement in Maths achievement as a result of completing their training program. However, there was a difference between groups in the expected improvement on English achievement, with the control group expecting to improve more as a result of their training program. There was also a possible increase in the control group's expectancy of improvement in English due to completing their intervention. In hindsight, both of these findings are not surprising, given the extensive use of letters, words and spelling as part of the control group's typing program. In contrast, the cognitive training program did not have such a heavy demand on the use of language. Importantly however, both groups expected similar improvement on near-transfer tasks, and this would have allowed causal conclusions to be made about the intervention, had any near-transfer effects occurred.

The results from the IEM highlight the strength of the control task (i.e. engaging and motivating), especially when compared to many of the non-active or non-challenging tasks often chosen in cognitive training studies (e.g., Hulme & Melby-Lervag, 2012; Melby-Lervag & Hulme, 2013; Morrison & Chein, 2011; Redick, Shipstead, et al., 2012; Rudebeck, Bor, Ormond, O'Reilly, & Lee, 2012; Salminen et al., 2012). As outlined by Boot et al. (2013), when participants expect to benefit at a similar level or even more from the control task, then any observable improvement for the experimental group has causal potency. Significantly, this study clearly addressed a key suggestion of Boot et al. (2013), that "researchers, reviewers, and editors should no longer accept inadequate control conditions, and causal claims should be rejected unless a study demonstrably eliminates differential placebo effects" (p. 452).

7.4.2 Implicit theory of intelligence

A measure of implicit theory of intelligence was used to determine if it was a potential moderator for training gains (*Hypothesis 3*). Although there was some indication of potential moderating effects of the participant's implicit theory of intelligence, the clear lack of linearity and small sample sizes involved in this three-way interaction analysis reduce the generalisability of this finding. Exploratory results showed some weak evidence that those participants in the experimental group who improved more on the WM measure have a higher ITIM (a malleable view of intelligence) than those who did not improve.

Past research has demonstrated a link between a malleable view of intelligence and training gains (Jaeggi et al., 2014), but this study did not find an association between the ITIM measure and improvement in near-transfer, either overall or specifically for the experimental group. However, the link that Jaeggi et al. (2014) found between a malleable view of intelligence and training gains was only found in their active control group and the correlation between a malleable view of intelligence and gains in the treatment group was low ($r < .05$), similar to what was found in this study. Interestingly, this study found a difference on the ITIM measure between those participants in the experimental group who improved more on the WM measure when compared to those who did not, with those who improved having a more malleable view of intelligence. Yet these results should be interpreted with caution as they were exploratory in nature and the strength of the evidence was relatively weak (BF_{01} = 0.58).

Jaeggi et al. (2014) found a non-significant effect of ITIM as a moderator but this study was essentially inconclusive with regards to this interaction. There was a three-way interaction picked up and the pattern was consistent across both latent variables for near-transfer and statistically significant for the interaction between the combined EF measure, ITIM and the intervention group. However, the sample sizes used were considerably smaller than necessary to definitively say whether this moderation effect of ITIM was real or spurious. It is clear that a larger sample size would be required to determine the real nature of this effect, but it is certainly an area that may be worth further investigation.

7.5 Treatment Fidelity

Those participants who performed better in some aspects of the cognitive training program did appear to show more improvement on the near-transfer measures. This pattern was consistent between the EF and WM scores for participants completing the cognitive training program, even if the correlations for some did not reach statistical significance. This pattern was not apparent for participants completing the control task. These results appear to indicate that those participants who performed better in the cognitive training program, improved slightly on some near-transfer measures. Furthermore, there were significant (but small) negative correlations between plateaus (where levels were moved up due to the participant not improving sufficiently) for two of the games in the cognitive training program, indicating that a lack of plateaus leads to more improvement on near-transfer. This may mean that someone who is more motivated or participating more meaningfully in the training program and has fewer plateaus in performance will show a greater rate of improvement.

This finding appears promising as it seems to indicate that there is some level of effectiveness for this cognitive training program, at least for some participants. The challenge is finding out which participants it is useful for and this does appear to be the direction that cognitive training should be heading. As outlined by Katz, Shah, and Meyer (2018), cognitive training research needs to stop asking the simple question of whether cognitive training works more generally, but rather focus on why it might work, under what conditions, and for whom.

7.6 Individual Differences

This study did attempt to answer the question of who cognitive training might work for as it used participants who were relatively high-achieving to determine if there was a magnification effect. The magnification effect assumes that cognitive training will most benefit high-achieving individuals because their cognitive resources are more efficient which allows them to implement new strategies and acquire new abilities. Given that there was no overall effect, this study does appear to rule out magnification effects. Furthermore, within the experimental group, there were no magnification effects as baseline cognitive performance at pre-test was not significantly positively correlated with the training-related gains. This would appear to be consistent with the compensation effect which assumes that individuals with higher baseline performance are already functioning at the optimal level and are closer to ceiling performance at the task, which leaves less room for improvement (Titz & Karbach, 2014). However, if compensation effects are occurring, baseline cognitive performance should be negatively correlated with training gains and this study did not find any support for compensation effects within the experimental group. Nevertheless, it is difficult to categorically rule out a compensation effect for this sample of participants, given their overall high-achieving levels at baseline.

The results of this study are in agreement with past research and reviews that have concluded cognitive training is not beneficial for higher achieving participants (Hardy et al., 2015; Titz & Karbach, 2014). It could be, as proposed by Karbach and Unger (2014), that compensation effects are more likely in cognitive training as opposed to magnification effects which might be more likely in strategy training. However, this is clearly an area that requires more targeted research.

7.7 Theoretical Considerations

As this study was primarily focussed upon the efficacy of cognitive training, this study was similar to other research where there was no effect generated from a well-designed study applied within a school setting using an active control group (Hitchcock & Westwell, 2017). Similar to Hitchcock and Westwell's (2017) results, this study found no near-transfer effects and no far-transfer to academic achievement or teacher-reported behaviours. The added benefit of this study is that it used latent factor scores for near-transfer measures.

As outlined by Aksayli et al. (2019), small sample sizes, the lack of an active control group (or any control group), and non-random allocation of participants to conditions are some of the major flaws that may bias the results of cognitive training studies to-date. As this study had sample sizes that were quite large compared to many other cognitive training studies, an active control group, and random allocation to conditions, it may be that this study has picked up the true (null) effects of this type of computerised cognitive training. It may simply be that training on one task has limited transferability to other tasks, commonly known as the 'curse of specificity' (Strobach & Karbach, 2016). Studies identifying significant improvement in specific tasks as a result of training have been around for a long time (e.g., Ericsson & Chase, 1982) but the goal is reliable near-transfer and, as a result, far-transfer. As the whole aim of cognitive training research is to ascertain if improved performance reflects a fundamental enhancement of underlying cognitive capacity or simply the acquisition of a set of skills that is specific to the practised task (Jolles & Crone, 2012; O'Connell & Robertson, 2012), only the use of multiple measures for near-transfer tasks (e.g., latent scores) can truly answer this question. As outlined by Katz et al. (2018), what some researchers refer to as near-transfer, may actually mean 'superficial transfer'. This may be why many cognitive training studies show near-transfer but do not show a subsequent far-transfer: the (usually individual) near-transfer measures are simply too similar to the training tasks. If these training studies do show near transfer and these improved skills are rate-limiting factors or bottle-necks for performance for more complex cognitive task performance, then it is somewhat puzzling that far-transfer is not found (Katz et al., 2018).

Alternatively, the lack of far-transfer seen in the present study and much of the cognitive training literature could also relate to ideas around the structure of intelligence itself. Protzko (2017) argues that as intelligence is a hierarchical factor structure, this necessarily prevents transfer from subfactors back up to the higher order (g) factor. Regarding the hierarchical structure of intelligence, Protzko (2017) notes:

> If this is the true structure, targeted cognitive training interventions will fail to increase intelligence not because intelligence is immutable, but simply

> because there is no causal connection between, say, working memory and intelligence. Seeing the structure of intelligence for what it is, a causal measurement model, allows us to focus testing on the presence and absence of causal links. If we can increase subfactors without transfer to other facets, we may be confirming the correct causal structure more than testing malleability (p. 1022).

It is also critical to acknowledge that there is still no clear causal association between EF and academic achievement (Jacob & Parkinson, 2015). As outlined extensively in Section 2.3, there are clear correlations between EF and academic achievement and this study clearly supports this view. However, as outlined by Redick et al. (2015), these associations can be interpreted in at least two ways: (a) good EF can cause academic success, or (b) individuals exhibit good EF performance as a consequence of their academic success. Therefore, cognitive training research needs to re-focus upon the proper measurement of the near-transfer components. Only when these have been measured properly, preferably using latent factors, can we assess the bigger picture questions of (a) whether or not cognitive training will work in general, due to the structure of intelligence itself; and (b) whether there is, in fact, a causal relationship between EF and academic achievement. Therefore, this study is one small step in that process by demonstrating, in this case, that ‘real’ near-transfer was not possible as a result of this cognitive training regimen. Whether a null result is more likely when near-transfer is measured using multiple measures, it is clearly too early to say on the basis of this study alone.

7.8 Practical Implications

The results of this study suggest that, at this stage, the use of school-based computerised cognitive training programs may not be an efficacious approach to improving EF abilities, academic achievement, gF, or important academic behaviours, at least not in relatively high-ability students around the 8- to 10-year-old range. Importantly, these results are similar to two other Australian studies that have found limited (Roberts et al., 2016) or null results (Hitchcock & Westwell, 2017) when using a commercially available computerised cognitive training program (CogMed) with children with low WM or with typically developing children respectively.

A null result is an important finding as cognitive training programs are actively marketed to schools for use with students (Decker et al., 2013; Shipstead, Hicks, et al., 2012) and the time spent training on these programs could be better used on more efficacious and targeted academic interventions. Given the greater capital reserves of independent private schools in Australia, this is one context where computerised cognitive training companies might target their marketing and where the use of cognitive training may be more likely. Therefore, it is particularly important to inform these schools and their school psychologists that such interventions may not be particularly effective, at least until it is better known why, when and for whom these interventions work.

Given that computerised cognitive training programs are added extras, their use takes away valuable curriculum time or are perhaps used in place of more justifiable interventions. At this stage, a more justifiable intervention to improve EF abilities in children within a school context are similar to those tested by Howard et al. (2017). As this type of intervention, where a children's book designed to challenge EF is used during shared individual reading, is low to no cost and can be incorporated into children's existing school routine, it may be less intrusive.

Although there are many other possibilities for training EF (Diamond & Lee, 2011), there are still many methodological issues in these other areas of EF training research (Dunning et al., 2019; Takacs & Kassai, 2019), similar to those encountered in the cognitive training literature, which makes it difficult to comprehensively determine their efficacy. As many EF interventions are implemented to help students who are struggling with certain aspects of their academic achievement, it may be more pragmatic to simply focus upon these areas of difficulty directly. Given that far-transfer effects are rarely achieved (Kassai et al., 2019), it seems counter-productive at this point in time to target these areas of difficulty indirectly. This is particularly relevant considering the large effects that some direct interventions, such as one-to-one remedial tutoring (Bloom, 1984) or small group interventions on specific academic weaknesses (Wheldall, Wheldall, Madelaine, Reynolds, & Arakelian, 2017), can have upon achievement.

7.9 Limitations and Directions for Future Research

The present findings are to be viewed in light of several shortcomings, which present potential directions for future research on cognitive training. Although the key recommendations by previous researchers (Aksayli et al., 2019; Gathercole et al., 2012; Gibson et al., 2012; Hulme & Melby-Lervag, 2012; Karbach & Kray, 2016; Redick et al., 2015; Shipstead, Hicks, et al., 2012; Simons et al., 2016) are still applicable to any future cognitive training research, some additional recommendations as a result of the limitations of this study are discussed below.

7.9.1 Sample characteristics

It is quite clear that, despite random allocation, the two groups appear to be different at their baseline measures for gF, maths achievement, teacher reports of behaviour, and for both latent EF factor scores. Having different groups at baseline after random allocation is not unusual in cognitive training (e.g., Dunning, Holmes, & Gathercole, 2013; Jaeggi et al., 2014), but it made interpretation of some aspects of this study challenging. For instance, it is difficult to know if the study did not have any main effects because this cognitive training program is not effective or if because the two groups were substantively different and it was the characteristics of the experimental group that led to the training not being effective. Given that there was a significant lack of improvement in the post-training factor scores for the experimental group while the control group did improve, certainly contributes to the idea that something unusual was going on with the experimental group. Whether this was unmeasured cognitive characteristics or due to unmeasured motivational or personality effects (e.g., feeling like they were being over-tested), is hard to determine. Indeed, some researchers have found that personality traits and motivation can play an important part in whether training is effective or not (Studer-Luethi et al., 2012; Zhao, Xu, Fu, & Maes, 2018) and this is an important aspect to consider in future cognitive training research.

Importantly, the idea that this study might not have worked for this particular sample is not especially surprising if we take the perspective that individuals with higher baseline performance may be closer to ceiling performance at the task and that training is not going to be as effective for relatively high-ability individuals. It could even be that near-transfer EF measures lack the sensitivity to detect changes for those who perform well on these tasks. Other research has found that high-ability participants did not improve as much in near-transfer (Hardy et al., 2015). Hardy et al.'s (2015) study was particularly large (N = 4,715) and their overall effect size was 0.26 (Cohen's *d*). Given this was the overall effect size and they found that high-ability participants did not improve as much, then the efficacy of cognitive training for high-ability participants is questionable due to potential ceiling effects. Consequently, far-transfer would then be questionable for high-ability participants as they would presumably not have the rate-limiting factors or bottle-necks for performance for more complex cognitive task performance that would be expected of lower-performing participants. In their review of the literature, Titz and Karbach (2014) found that compensation effects were more likely in cognitive training, whereas magnification effects were more likely in strategy training. This seems consistent with what was found in this study and the idea that ceiling effects would impact potential gains from cognitive training for high-ability participants. The results of this study were quite clear in this regard. Even post-hoc exploratory analysis did not reveal any evidence for magnification effects for participants of this age-group.

Although there are other limitations of this study that might explain why no overall effect was found for this study using these participants, future researchers should be cautious about using computerised cognitive training programs with high-achieving participants. It is reasonable, based on this study and other recent research (e.g. Hardy et al., 2015), that the efficacy of cognitive training for high-ability participants is questionable due to potential ceiling effects.

As this study used a convenience sample that was boys-only, then the findings should be generalised with caution. Although meta-analyses seem to indicate that sex differences in EF are not present in this age group (Cross et al., 2011; Voyer et al., 2017), it would be prudent to be cautious in generalising these results to girls, especially considering the unique cognitive (i.e. high achieving) and social (i.e. high SES) characteristics of the sample.

7.9.2 Training program

With regards to why this cognitive training program might not have worked, it may simply be that the training program was not training EF appropriately or effectively. Of the six games that were available in the C8 program, two of the games are said to train 'spatial working memory' (*Monkey Trouble* and *Grub Ahoy!*). In reality, these games are similar to the ones that other computerised working memory training programs target (e.g., CogMed) and it is clear that these are simple span tasks, which are not targeting working memory (Shipstead, Hicks, et al., 2012). Given that one-third of the time spent training for the experimental group was spent on tasks that theoretically might not improve EF, then it may not surprising that EF did not improve to an appreciable extent. However, it is notable that improvement in near-transfer measures was significantly associated with performance on these two simple span task games (as outlined in Section 6.13). This was unexpected. It is possible that the improvement indicated by performance on short-term storage tasks in the cognitive training program are indicative of a confounding and unmeasured variable, perhaps some kind of personality or motivation factor, similarly indicated by the number of plateaus as reported in treatment fidelity outcomes. As suggested by other researchers, these personality or motivational factors may be critical as some participants may be easily discouraged and thus do not benefit from a cognitive training intervention (Katz et al., 2018).

Although the possible weaknesses outlined regarding the cognitive training program used in this study are unfortunate, it is important to note that not all aspects of the training program were used (as outlined in Section 5.4.8); specifically, the aerobic exercise component of the C8 Sciences program was not included in the present study. However, with regard to the cognitive training program specifically, in the scoping phase of this study it proved quite difficult to find an adaptive computerised training program that did not spend a considerable amount of time training on simple span tasks. Therefore, the use of the training program selected was a trade-off in attempting to train multiple EF while accepting that some simple span training would occur. Ultimately, a training program that is fully aligned with EF theory would be ideal, but the researcher is unaware of any at this stage and the program used in this study would still appear to be the best option.

It may be that the removal of the aerobic exercise component of the program limited the potential synergistic effects provided when completing both the exercise and cognitive training aspects simultaneously. There are case studies provided by C8 Sciences for its efficacy ("C8 Sciences: Case Studies", n.d.), and some limited research on the efficacy of combined exercise and computerised cognitive training programs for adults (e.g. Daugherty, 2018) and older adults (e.g. Bamidis, 2015). However, there appears to be a dearth of peer-reviewed research on the efficacy of combined programs in children. Therefore, there is a clear opportunity for further research to investigate the efficacy of each component of a combined aerobic exercise and computerised cognitive training program in children.

Importantly, as there was some evidence that participants who appeared to perform better in the computerised cognitive training program improved slightly on some near-transfer measures, it is plausible that this program in its existing form and used alone could still be effective. It may be that some changes to the administration of training, such as increasing the level of supervision when participants are completing the training program, could further improve the overall efficacy of this program.

7.9.3 Level of supervision

Even though this study did supervise the testing and training phases of the study, it was done in groups and it is not possible to get the same level of supervision as when testing and training is done individually. This could be an important component linked to a lack of near-transfer. This hypothesis is certainly in line with other research where transfer was not found when assessment and training was done unsupervised (e.g., Owen et al., 2010). In contrast, much of the cognitive training literature where transfer has occurred with large effect sizes appears to incorporate individually administered assessment and individually supervised training (e.g., Dahlin, 2011; Klingberg et al., 2005; Westerberg et al., 2007), with the concomitant costs that this would incur. Given the small, but promising results for treatment fidelity adherence (see Section 7.5), it could be that individual supervision of the training program would further enhance training effectiveness.

Training being supervised individually could be an important factor in training fidelity, as when done individually, it may ensure that training is at maximum effort, something that is not possible when supervising training in a group. In fact, a meta-analysis by Schwaighofer, Fischer, and Bühner (2015) found that individual supervision for cognitive training was a significant moderator. However, it can be difficult to tease supervision out as an important factor due to the theoretical, methodological and measurement weaknesses in cognitive training research outlined previously. Importantly, when both assessment and training have been individually supervised in a well-designed study (e.g., Hitchcock & Westwell, 2017), transfer has not been successful. Nevertheless, individual supervision of assessment and training could potentially reduce the likelihood of confounding factors in future research. However, this supervision would need to be monitored closely to ensure that both implicit and explicit encouragement of the participant by the supervisor is tightly controlled, otherwise, this may just introduce further confounding factors.

With regard to group testing, given that all EF tasks were administered automatically by computer, there is the potential for error to be introduced, regardless of what safe-guards are put in place within the test battery (as was done as a result of the pilot test – see Section 5.2). As outlined previously (Section 6.2), there was one data point in the card sorting perseverative error that was Winsorized due to the participant achieving no categories - something that could have been rectified with individual supervision of test administration. This was an obvious error that could be fixed at the data tidying stage but more subtle errors, due to the nature of group testing, may be more difficult to pick up. For example, a participant may take multiple short breaks within a test without being noticed by the researcher and without the severe performance costs that may be picked up in data analysis. Nevertheless, their performance is impacted and these are introduced errors, they are not obvious, and they cannot be corrected. Therefore, it is possible that there is "noise" in the near-transfer measures, hence reducing the likelihood that any real changes in near-transfer are picked up in data analysis. There is the potential for individual administration of all EF tasks to ameliorate these potential issues.

7.9.4 Over-testing

The large number of assessments completed by participants is a potential problem in this study. As outlined by Katz et al. (2018), testing on a large number of factors can affect the quality of assessment implementation. As they state, "outcome measures, when given to participants in rapid succession, shortened for time constraints, and administered over several hours, may be less reliable than ideal" (Katz et al., 2018, p. 9900). Although the test-retest reliability of the measures in this study appeared to be satisfactory, there could be issues with regard to participants being able to perform at their best when presented with 14 challenging tests and two surveys, as was the case with each assessment session in this study. Participants facing so many assessments may be "more susceptible to confounds due to boredom and exhaustion (i.e., the quality of the assessment data is diminished over the assessment session)" (Morrison & Chein, 2011, p. 56). As a result, this may introduce confounding personality, motivational and other extraneous cognitive factors that were not measured. This is a major challenge in cognitive training research: on the one hand, ensuring that near-transfer measures are 'real'; on the other hand, ensuring that other confounds are not introduced as a result of over-testing.

It is believed that personality and motivational factors could have played a role in this study due to the large number of tests that participants had to sit. Given that there was some limited evidence of the impact of motivational factors as indicated by the potential moderating effects of the participant's implicit theory of intelligence and exploratory results demonstrating weak evidence of an impact of an implicit theory of intelligence on training gains, it is evident that motivational factors appear to play some part in a long-term training program like this one. Therefore, it is plausible that there are also potential motivational effects throughout the testing phase too. One might even consider these to be more significant given that each assessment session for EF measures was considerably longer and potentially more cognitively demanding than a training session.

As outlined earlier, the number one priority for cognitive training research needs to be ensuring that 'real' near-transfer is being measured. Only once this issue has been addressed can cognitive training research really move towards *why* it might work, under *what conditions*, and for *whom* (Katz et al., 2018). Since Hulme and Melby-Lervåg

(2012) first advocated for the importance of using multiple measures of key constructs for pre- and post-test measures to allow latent variable models, there has been a dearth of cognitive training research addressing this issue. However, addressing this issue for the measure of EF can be quite an exhaustive process for participants and introduce confounding factors, such as personality and motivational factors, as may have been the case in this study. One solution to this issue may be to focus upon the assessment of one EF at a time. Although it has been acknowledged that computerised training on a wider range of EF could have significant implications for education (Gibson et al., 2012; Morrison & Chein, 2011; Walton & Dweck, 2009), it is believed, as a result of this study, that assessing all EF simultaneously is just not tenable, especially for younger participants.

7.9.5 EF measures

With regard to the assessments, there were some issues with the measurement items chosen for individual EF measures, with the Switcher task dropped completely due to it not fitting the proposed model in the CFA. In hindsight, it is not a complete surprise that the Switcher task was removed from the analysis. Throughout the EF assessment part of this study the researcher observed some potential shortcomings of the Switcher task. For example, many participants appeared to show frustration with the task as they were having to visually search over the whole computer screen for the next appropriate target. It is believed that this may have introduced some potential confounding factors into this task. A similar task within the shifting battery, the Connections Test (Salthouse et al., 2000), was created to address similar shortcomings in the Trail Making Test, such as reducing “irrelevant influences of visual search and hand movements” (Salthouse, 2011, p. 223) by moving successive targets to adjacent locations. As this was obviously an issue for the Trail Making Test in previous research, it is not surprising that the Switcher task would have the same issues. Additionally, the Connections Test was modified to have “a counter-balanced presentation order of simple (A) and alternating (B) trials [which] eliminates a confounding in which the alternating condition always follows the simple condition” (Salthouse, 2011, p. 223). The Switcher task also made the same mistake as the Trail Making Test, where the task always started off with more simple switching trials and moved to more complex and random switching trials, perhaps introducing the confounding effects the Connections Test was designed to remove. It is clear that the Switcher task was a poor choice for a shifting measure and its removal as a result of the CFA was warranted. However, this is the clear benefit of using a CFA approach to develop latent measures: inappropriate measures are dropped that would have otherwise been included in the analysis where the potential for spurious findings increases.

The availability of open-source measures, such as those used for this study using PEBL, especially the complex WM span tasks specifically developed for this study, is a significant factor in promoting better measurement of near-transfer. The use of easily available measures that can be shared amongst research projects and available for use on multiple operating systems, makes PEBL a tool that could further improve cognitive training research. However, it is critical that a third measure for shifting be developed or that Shifter is modified in a similar manner to how the Trail Making Test was changed to remove confounding factors to result in the Connections Test.

It is important to note that the dependent measures chosen for the Numerical Stroop task were the accuracy for the incongruent condition and the mean RT for the incongruent condition. Although these measures have been used in previous research using a Stroop task (e.g. Arffa, 2007; Dadon & Henik, 2017), it must be acknowledged that the more traditional dependent variable for use in the Numerical Stroop task has arguably been a RT difference between the incongruent and congruent conditions (e.g. Hedge, Powell, & Sumner, 2018). This is a limitation of the study, and it is unknown if the RT for incongruent conditions, which was the only dependent measure from the Numerical Stroop included in the CFA model, is an accurate measure of inhibition by itself. However, as outlined by Khng and Lee (2014), the reliability of the more frequently used Stroop interference RT measure has been found to be problematic as a measure of inhibition and as they suggest, it may be beneficial in future research to use scores that include both RT and accuracy (i.e. inverse efficiency scores that attempt to reconcile speed with accuracy performance). Importantly, it is noted that the RT for incongruent conditions used in this study still provided a suitable fit within the combined inhibition and shifting factor as a result of the CFA, which included more traditional measures of both inhibition and shifting.

7.9.6 Missing data

Although it has already been highlighted as a significant strength of this study, it is important to recognise some potential issues that could have arisen in the creation of latent factor scores. As was explained in Section 6.3, there was some participants' CST data removed due to them not achieving the required accuracy criteria for the distractor task. This is the normal procedure for completing CSTs (Unsworth et al., 2005) and is justified to ensure that participants are not sacrificing accuracy on the distractor to increase performance on the recall tasks. If data were not removed it is assumed that it would no longer function as a WM task, but a simple STM recall task, as participants are simply focussing on the recall task alone and not the recall and distractor tasks simultaneously. However, the removal of data potentially creates problems for the creation of latent factor scores, more specifically, the WM factor score that was derived from the CSTs.

As part of the CFA process, lavaan used full information maximum likelihood methods to impute missing data. This is acceptable if the data is missing at random or missing completely at random, but may provide biased estimates of the parameters of interest, such as means, regression coefficients or standard errors if the data is missing not at random (MNAR; Galimar, Chevret, Protopopescu, & Resche-Rigon, 2016). Given that the data was removed in this study due to performance issues for the participants, the data is clearly MNAR and this could potentially be problematic for the creation of the WM factor score. Nevertheless, it was decided to proceed with this potentially biased option of creating latent factor scores, due to the methodological benefits outlined previously (see Section 2.7). Another option could have been the removal of cases listwise for all three CSTs simultaneously and then performing a CFA on the remaining data. However, this was not a viable option as this would have reduced the sample size to 41 ($n_{exp} = 14$, $n_{ctrl} = 27$), and potentially implemented its own biases (Pepinsky, 2018).

Importantly, there is no apparent difference in the patterns of outcomes for this study if the WM measures are analysed individually instead of as a factor score and these trends are similar whether data is removed or not (see Appendix F – Section 9.6). However, these analyses do identify a potential weakness in the use of the CSTs with this cohort of participants. As seen in Appendix F (Section 9.6.1), it was younger participants who had more difficulty ensuring distractor accuracy throughout this assessment. Younger participants appeared to focus more on the recall aspect of this test and allowed accuracy for the distractor task to drop significantly – potentially changing the meaning of this test for younger participants. This may be a limitation of using these complex WM span tasks with younger participants or it may require an increased level of supervision, as mentioned Section 7.9.3, to ensure that younger participants are focusing on both the distractor and recall tasks equally. It is recommended that future research take the level of supervision into consideration if using the PEBL CSTs with younger participants. This way, it is more likely that having to remove data for the CSTs can be avoided if the individual supervision provides ongoing reminders to participants to be focussing on both the distractor and recall tasks simultaneously.

7.9.7 Sample size

Lastly, the sample size for this study was not sufficient to determine if the three-way interaction between the combined EF score, ITIM and the intervention group was anything more substantial than a spurious finding (Gelman, 2018; Heo & Leon, 2010). It is plausible that ITIM would interact with both the baseline measure and intervention group as this could impact motivational factors for participants and lead them to either engage or disengage from the training. Indeed, it is believed that there was some kind of motivational factor playing a part in the efficacy of cognitive training as this was detected in the post-hoc exploratory analysis in the experimental group too. This did appear to indicate that those with a more malleable view of intelligence improved more. However, the sample sizes in this study are not large enough to detect such complex interactions.

Increasing sample sizes in this type of research can be challenging, especially if the earlier recommendations of individual supervision of participants is implemented. Given that suggestions of sample sizes to detect effects for three-way interactions is sixteen-times (Gelman, 2018; Heo & Leon, 2010) the size required to detect simple effects, the scale of the numbers required can be seen. Based upon these recommendations, it would require a sample size of approximately 1600 in this study to reliably detect any potential three-way interaction between the combined EF measure, ITIM and the intervention group. Given such large sample size requirements, the only feasible solution might be multi-site collaborative research.

7.10 Conclusion

Computerised adaptive cognitive training may be a way that EF abilities can be changed in children (Kassai et al., 2019). However, the efficacy of cognitive training is not simply a scientific curiosity as they are actively marketed to school systems (Shipstead, Hicks, et al., 2012) and school psychologists (Decker et al., 2013) as an intervention that works. Given the variability in research findings and the methodological weaknesses of much of the cognitive training research, the present study sought to address these gaps in knowledge and research.

This study aimed to test the efficacy of a computerised adaptive cognitive training program by looking at near-transfer to EF components and far-transfer to higher order cognitive abilities (gF and academic achievement). The study addressed previous criticisms of the cognitive training literature by using a program that is more aligned to current EF theory than previous research, by training executive level processes. Significantly, it used multiple measures for each EF construct to estimate latent scores of EFs. Randomly allocated experimental and active control groups that both utilised adaptive computer-based tasks and had the same face-to-face contact time were used so that a realistic comparison could be made as to the efficacy of the cognitive training program. To determine the longer-term impact of cognitive training, participants were followed over a period of one year. Ancillary measures, such as improvement expectancies and a measure of participant's implicit theory of intelligence, were also collected to determine expected improvement equivalency and potential moderating effects. In addition to these student-based measures, behaviourally-based reports from teachers and parents provided additional sources of evidence of the program's effects.

This research program found no evidence for improvement on either near- or far-transfer outcome measures (academic, gF, or behavioural) as a result of the computerised adaptive EF training program. Importantly, this study measured near-transfer using a latent variable approach with five measures contributing towards a combined shifting and inhibition latent EF score and three measures towards a latent WM score. Although there were no definitive results demonstrating the impact of an implicit theory of intelligence on training gains, post-hoc exploratory analysis revealed there may be more to investigate regarding motivation and its interaction with training gains.

As outlined by Protzko (2017), "attempting to enhance our cognitive lives through targeted training is a most noble pursuit" (p. 1029), and one fully endorsed by this researcher. However, this is a challenging field with so many prohibitive factors to performing rigorous research and thus, determining definitive answers with regard to efficacy. Therefore, future research needs to move more carefully by accurately measuring near-transfer before cognitive training is implemented in schools, as we are still unsure as to why it might work, under what conditions, and for whom. The risks are that children are engaging in an intervention where we are unsure as to what outcomes may be when, instead, they could be involved in more robust evidence-based interventions targeted to where their academic difficulties may actually lie.

8 References

Aarnoudse-Moens, C. H., Smidts, D., Oosterlaan, J., Duivenvoorden, H., & Weisglas-Kuperus, N. (2009). Executive function in very preterm children at early school age. *Journal of Abnormal Child Psychology*, *37*, 981–993. https://doi.org/10.1007/s10802-009-9327-z

Ablard, K. E., & Mills, C. J. (1996). Implicit theories of intelligence and self-perceptions of academically talented adolescents and children. *Journal of Youth and Adolescence*, *25*, 137–148. https://doi.org/10.1007/bf01537340

Agostino, A., Johnson, J., & Pascual-Leone, J. (2010). Executive functions underlying multiplicative reasoning: Problem type matters. *Journal of Experimental Child Psychology*, *105*, 286–305. https://doi.org/10.1016/j.jecp.2009.09.006

Ahmed, S. F., Tang, S., Waters, N. E., & Davis-Kean, P. (2018). Executive function and academic achievement: Longitudinal relations from early childhood to adolescence. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000296

Aksayli, N. D., Sala, G., & Gobet, F. (2019). The cognitive and academic benefits of Cogmed: A meta-analysis. *Educational Research Review*, *27*, 229–243. https://doi.org/10.1016/j.edurev.2019.04.003

Alarcon, G., Cservenka, A., Fair, D. A., & Nagel, B. J. (2014). Sex differences in the neural substrates of spatial working memory during adolescence are not mediated by endogenous testosterone. *Brain Research, 1593,* 40-54. http://dx.doi.org/10.1016/j.brainres.2014.09.057

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. *European Psychologist*, *22*(4), 207–218. https://doi.org/10.1027/1016-9040/a000298

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Alloway, T. P., & Alloway, R. G. (2013). Working memory across the lifespan: A cross-sectional approach. *Journal of Cognitive Psychology* , *25*, 84–93. https://doi.org/10.1080/20445911.2012.748027

Alloway, T. P., Bibile, V., & Lau, G. (2013). Computerized working memory training: Can it lead to gains in cognitive skills in students? *Computers in Human Behavior*, *29*, 632–638. https://doi.org/10.1016/j.chb.2012.10.023

Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2009). The cognitive and behavioral characteristics of children with low working memory. *Child Development*, *80*, 606–621. https://doi.org/10.1111/j.1467-8624.2009.01282.x

Anderson, K., Deane, K., Lindley, D., Loucks, B., & Veach, E. (2012). The effects of time of day and practice on cognitive abilities: The PEBL Tower of London, Trail-making and Switcher tasks. Retrieved 26 April 2018, from http://sites.google.com/site/pebltechnicalreports/home/2012/pebl-technical-report-2012-04

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *8*, 71–82. https://doi.org/10.1076/chin.8.2.71.8724

Arffa, S. (2007). The relationship of intelligence to executive function and non-executive function measures in a sample of average, above average, and gifted youth. *Archives of Clinical Neuropsychology: The Official Journal of the National*

Academy of Neuropsychologists, *22*, 969–978. https://doi.org/10.1016/j.acn.2007.08.001

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*, 170–177. https://doi.org/10.1016/j.tics.2004.02.010

Atkinson, T. M., & Ryan, J. P. (2008). The use of variants of the Trail Making Test in serial assessment: A construct validity study. *Journal of Psychoeducational Assessment*, *26*, 42–53. https://doi.org/10.1177/0734282907301592

Atkinson, T. M., Ryan, J. P., Kryza, M., & Charette, L. M. (2011). Using versions of the Trail Making Test as alternate forms. *The Clinical Neuropsychologist*, *25*, 1193–1206. https://doi.org/10.1080/13854046.2011.589410

Atkinson, T. M., Ryan, J. P., Lent, A., Wallis, A., Schachter, H., & Coder, R. (2010). Three trail making tests for use in neuropsychological assessments with brief intertest intervals. *Journal of Clinical and Experimental Neuropsychology*, *32*, 151–158. https://doi.org/10.1080/13803390902881934

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. https://doi.org/10.3758/s13423-014-0699-x

Australian Centre for Educational Research. (2007). *Research on the Progressive Achievement Tests and academic achievement in secondary schools*. Retrieved from Australian Council for Educational Research website: www.acer.edu.au/documents/ACERPress_PAT_Supp.pdf

Australian Centre for Educational Research. (2008). *Progressive Achievement Tests in reading: Comprehension, vocabulary and spelling: Fourth Edition* (4th ed.). Camberwell: Victoria: ACER Press.

Australian Centre for Educational Research. (2013). *Progressive Achievement Tests in mathematics: Fourth Edition* (4th ed.). Camberwell: Victoria: ACER Press.

Autin, F., & Croizet, J. C. (2012). Improving working memory efficiency by reframing metacognitive interpretation of task difficulty. *Journal of Experimental Psychology: General*, *141*, 610–618. https://doi.org/10.1037/a0027478

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology Section A - Human Experimental Psychology*, *49*, 5–28. https://doi.org/10.1080/027249896392784

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423. https://doi.org/10.1016/s1364-6613(00)01538-2

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist*, *7*, 85–97. https://doi.org/10.1027//1016-9040.7.2.85

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews. Neuroscience*, *4*, 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Retrieved from http://www.sciencedirect.com/science/article/pii/S0079742108604521

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*, 158–173. https://doi.org/10.1037/0033-295x.105.1.158

Bailey, H. (2012). Computer-paced versus experimenter-paced working memory span tasks: Are they equally reliable and valid? *Learning and Individual Differences*, *22*, 875–881. https://doi.org/10.1016/j.lindif.2012.06.004

Baker, S., Maurissen, J., & Chrzan, G. (1986). Simple reaction-time and movement time in normal human volunteers - A long-term reliability study. *Perceptual and Motor Skills*, *63*, 767–774. https://doi.org/10.2466%2Fpms.1986.63.2.767

Bamidis, P. D., Fissler, P, Papageorgiou, S. G., Zilidou, V., Konstantinidis, E. I., Billis, A. S., Romanopoulou, E., Karagianni, M., Beratis, I., Tsapanou, A., Tsilikopoulou, G., Grigoriadou, E., Ladas, A., Kyrillidou, A., Tsolaki, A., Frantzidis, C., Sidiropoulos, E., Siountas, A., Matsi, A., Papatriantafyllou, J., Margioti, E., Nika, A., Schlee, W., Elbert, T., Tsolaki, M., Vivas, A. B., & Kolassa, I. (2015). Gains in cognition through combined cognitive and physical training: the role of training dosage and severity of neurocognitive disorder. *Frontiers in Aging Neuroscience,* 7(152), 1-15. https://doi.org/10.3389/fnagi.2015.00152

Beaujean, A. A. (2014). *Latent Variable Modeling Using R*. New York: Routledge.

Beck, S. J., Hanson, C. A., Puffenberger, S. S., Benninger, K. L., & Benninger, W. B. (2010). A controlled trial of working memory training for children and adolescents with

ADHD. *Journal of Clinical Child and Adolescent Psychology, 39*, 825–836. https://doi.org/10.1080/15374416.2010.517162

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *The Journal of General Psychology*, *39*, 15–22. https://doi.org/10.1080/00221309.1948.9918159

Bergman Nutley, S., & Söderqvist, S. (2017). How is working memory training likely to influence academic performance? Current evidence and methodological considerations. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00069

Bergman Nutley, S., Söderqvist, S., Bryde, S., Thorell, L. B., 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*, 591–601. https://doi.org/10.1111/j.1467-7687.2010.01022.x

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child Development*, *81*, 1641–1660. https://doi.org/10.1111/j.1467-8624.2010.01499.x

Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review: DR*, *29*, 180–200. https://doi.org/10.1016/j.dr.2009.05.002

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and Individual Differences*, *21*, 327–336. https://doi.org/10.1016/j.lindif.2011.01.007

Bezdjian, S., Baker, L. A., Lozano, D. I., & Raine, A. (2009). Assessing inattention and impulsivity in children during the Go/NoGo task. *The British Journal of*

Developmental Psychology, *27*, 365–383. https://doi.org/10.1348/026151008X314919

Blacker, K. J., Negoita, S., Ewen, J. B., & Courtney, S. M. (2017). N-back Versus Complex Span Working Memory Training. *Journal of Cognitive Enhancement,* 1–21. https://doi.org/10.1007/s41465-017-0044-1

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*, 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*, 647–663. https://doi.org/10.1111/j.1467-8624.2007.01019.x

Bloom, B. S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership*, 41(8), 4-17. https://doi.org/10.3102/0013189X013006004

Booth, J. N., Boyle, J. M. E., & Kelly, S. W. (2010). Do tasks make a difference? Accounting for heterogeneity of performance of children with reading difficulties on tasks of executive function: Findings from a meta-analysis. *The British Journal of Developmental Psychology*, *28*, 133–176. https://doi.org/10.1348/026151009x485432

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science: A Journal of the*

Association for Psychological Science, *8*, 445–454. https://doi.org/10.1177/1745691613491271

Brannigan, G., & Brannigan, M. (1995). Comparison of individual versus group administration of the modified version of the Bender-Gestalt Test. *Perceptual and Motor Skills*, *80*, 1274–1274.

Bressler, S. L., & Menon, V. (2010). Large-scale brain networks in cognition: Emerging methods and principles. *Trends in Cognitive Sciences*, *14*, 277–290. https://doi.org/10.1016/j.tics.2010.04.004

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*, 563–570. https://doi.org/10.3758/BRM.42.2.563

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. *The American Psychologist*, *67*, 87–100. https://doi.org/10.1037/a0024657

Brydges, C. R., Reid, C. L., Fox, A. M., & Anderson, M. (2012). A unitary executive function predicts intelligence in children. *Intelligence*, *40*, 458–469. https://doi.org/10.1016/j.intell.2012.05.006

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, *33*, 205–228. https://doi.org/10.1080/87565640801982312

Bürki, C. N., Ludwig, C., Chicherio, C., & de Ribaupierre, A. (2014). Individual differences in cognitive plasticity: an investigation of training curves in younger and older adults. *Psychological Research*, *78*(6), 821–835. https://doi.org/10.1007/s00426-014-0559-3

Burnette, J. L., O'Boyle, E. H., VanEpps, E. M., Pollack, J. M., & Finkel, E. J. (2013). Mind-sets matter: A meta-analytic review of implicit theories and self-regulation. *Psychological Bulletin*, *139*, 655–701. https://doi.org/10.1037/a0029531

C8 Sciences. (2011). *The 8 core cognitive capacities*. Retrieved from http://www.c8sciences.com/about/8ccc

C8 Sciences: Case Studies. (n.d.). Retrieved from http://www.c8sciences.com/category/case-studies/

C8 Sciences: Physical Exercise Programs. (n.d.). Retrieved from http://www.c8sciences.com/about/exercise

Calvin, C. M., Fernandes, C., Smith, P., Visscher, P. M., & Deary, I. J. (2010). Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in England. *Intelligence*, *38*, 424–432. https://doi.org/10.1016/j.intell.2010.04.005

Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and Individual Differences*, *19*, 246–251. https://doi.org/10.1016/j.lindif.2008.10.002

Cepeda, N. J., Blackwell, K. A., & Munakata, Y. (2013). Speed isn't everything: Complex processing speed measures mask individual differences and developmental changes in executive control. *Developmental Science*, *16*, 269–286. https://doi.org/10.1111/desc.12024

Chan, R. C. K., Shum, D., Toulopoulou, T., & Chen, E. Y. H. (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of*

Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists, *23*, 201–216. https://doi.org/10.1016/j.acn.2007.08.010

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review*, *17*, 193–199. https://doi.org/10.3758/pbr.17.2.193

Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*, 531–542. https://doi.org/10.1016/j.intell.2012.07.004

Chuderski, A., & Necka, E. (2012). The contribution of working memory to fluid reasoning: Capacity, control, or both? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *38*(6), 1689–1710. https://doi.org/10.1037/a0028465

Cirino, P. T., Ahmed, Y., Miciak, J., Taylor, W. P., Gerst, E. H., & Barnes, M. A. (2018). A framework for executive function in the late elementary years. *Neuropsychology*, *32*(2), 176–189. https://doi.org/10.1037/neu0000427

Clark, C. A. C., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental Psychology*, *46*, 1176–1191. https://doi.org/10.1037/a0019672

Coldren, J. T. (2013). Cognitive control predicts academic achievement. *Mind, Brain and Education: The Official Journal of the International Mind, Brain, and Education Society*, *7*, 40–48. https://doi.org/10.1111/mbe.12006

Conway, A. R. A., & Getz, S. J. (2010). Cognitive ability: Does working memory training enhance intelligence? *Current Biology: CB*, *20*, R362–R364. https://doi.org/10.1016/j.cub.2010.03.001

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*, 769–786. https://doi.org/10.3758/BF03196772

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*, 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2005). *Working memory capacity*. New York, NY : Psychology Press.

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*, 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

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

Coyle, T. R., Pillow, D. R., Snyder, A. C., & Kochunov, P. (2011). Processing speed mediates the development of general intelligence (g) in adolescence. *Psychological Science*, *22*, 1265–1269. https://doi.org/10.1177/0956797611418243

Crone, E. A., Wendelken, C., Donohue, S. E., & Bunge, S. A. (2006). Neural evidence for dissociable components of task-switching. *Cerebral Cortex* , *16*, 475–486. https://doi.org/10.1093/cercor/bhi127

Cross, C. P., Copping, L. T., & Campbell, A. (2011). Sex differences in impulsivity: A meta-analysis. *Psychological Bulletin, 137 (1)*, 97-130. http://dx.doi.org/10.1037/a0021591

Dadon, G. & Henik, A. (2017). Adjustment of control in the numerical Stroop task. *Memory & Cognition, 45*, 891-902. https://doi.org/10.3758/s13421-017-0703-6

Da Fonseca, D., Schiano-Lomoriello, S., Cury, F., Poinso, F., Rufo, M., & Therme, P. (2007). Validition study of the implicit theories of intelligence scale. *Encephale-Revue De Psychiatrie Clinique Biologique Et Therapeutique*, *33*, 579–584. https://doi.org/10.1016/s0013-7006(07)92056-4

Dahlin, E., Nyberg, L., Bäckman, L., & Neely, A. S. (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. https://doi.org/10.1037/a0014296

Dahlin, K. (2011). Effects of working memory training on reading in children with special needs. *Reading and Writing*, *24*, 479–491. https://doi.org/10.1007/s11145-010-9238-y

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466. https://doi.org/10.1016/s0022-5371(80)90312-6

Daneman, M., & Merikle, P. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*, 422–433. https://doi.org/10.3758/bf03214546

Daugherty, A. M., Zwilling, C., Paul, E. J., Sherepa, N., Allen, C., Kramer, A. F., Hillman, C. H., Cohen, N. J., & Barbey, A. K. (2018). Multi-modal fitness and cognitive training to enhance fluid intelligence. *Intelligence, 66*, 32-43. https://doi.org/10.1016/j.intell.2017.11.001

David, C. V. (2012). Working memory deficits in Math learning difficulties: a meta-analysis. *International Journal of Developmental Disabilities*, *58*, 67–84. https://doi.org/10.1179/2047387711y.0000000007

Davidson, M. C., Amso, D., Anderson, L. C., & Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia*, *44*, 2037–2078. https://doi.org/10.1016/j.neuropsychologia.2006.02.006

Decker, S. L., Hale, J. B., & Flanagan, D. P. (2013). Professional practice issues in the assessment of cognitive functioning for educational applications. *Psychology in the Schools*, *50*, 300–313. https://doi.org/10.1002/pits.21675

Dehn, M. J. (2008). *Working memory and academic learning: Assessment and intervention*. Hoboken, NJ: John Wiley & Sons.

de Greeff, J. W., Bosker, R. J., Oosterlaan, J., Visscher, C., & Hartman, E. (2018). Effects of physical activity on executive functions, attention and academic performance in preadolescent children: a meta-analysis. *Journal of Science and Medicine in Sport,* 21 (5), 501-507. https://doi.org/10.1016/j.jsams.2017.09.595

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology,* 30(1), 92-101. https://doi.org/10.5334/irsp.82

Delaloye, C., Ludwig, C., Borella, E., Chicherio, C., & de Ribaupierre, A. (2008). L'empan de lecture comme épreuve mesurant la capacitéde mémoire de travail: Normes basées sur une population franco-phone de 775 adultes jeunes et âgés. *Revue*

Européenne DePsychologie Appliquée, *58* (2), 89-103. https://doi.org/10.1016/j.erap.2006.12.004

Del Re, A. C. (2013). compute.es: Compute Effect Sizes. *R Package*. Retrieved from http://cran.r-project.org/web/packages/compute.es

Demetriou, A., Spanoudis, G., Christou, C., & Platsidou, M. (2001). Modeling the Stroop phenomenon: Processes, processing flow, and development. *Cognitive Development*, *16*, 987–1005. https://doi.org/10.1016/S0885-2014(02)00069-2

Derntl, B., Pintzinger, N., Kryspin-Exner, I., & Schopf, V. (2014). The impact of sex hormone concentrations on decision-making in females and males. *Frontiers in Neuroscience, 8*, 1-11. http://dx.doi.org/10.3389/fnins.2014.00352

Diamond, A. (2012). Activities and programs that improve children's executive functions. *Current Directions in Psychological Science*, *21*, 335–341. https://doi.org/10.1177/0963721412453722

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, A., & Lee, K. (2011). Interventions shown to aid executive function development in children 4 to 12 years old. *Science*, *333*, 959–964. https://doi.org/10.1126/science.1204529

Diamond, A., & Ling, D. S. (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental Cognitive Neuroscience*, *18*, 34–48. https://doi.org/10.1016/j.dcn.2015.11.005

Dixon, M. L., De La Vega, A., Mills, C., Andrews-Hanna, J., Spreng, R. N., Cole, M. W., & Christoff, K. (2018). Heterogeneity within the frontoparietal control network and its relationship to the default and dorsal attention networks. *PNAS, 115(7)*, E1598-E1607. http://dx.doi.org/10.1073/pnas.1715766115

Dowsett, S. M., & Livesey, D. J. (2000). The development of inhibitory control in preschool children: Effects of 'executive skills' training. *Developmental Psychobiology*, *36*, 161–174.

Duggan, E. C. & Garcia-Barrera, M. A. (2015). Executive functioning and intelligence. In Goldstein, S., Princiotta, D., & Naglieri, J. A. (Eds.), *Handbook of Intelligence*, 435-458. New York: Springer.

Dunning, D. L., Griffiths, K., Kuyken, W., Crane, C., Foulkes, L., Parker, J., & Dalgleish, T. (2019). Research review: The effects of mindfulness-based interventions on cognition and mental health in children and adolescents – a meta-analysis of randomized controlled trials. *Journal of Child Psychology and Psychiatry, and Allied Disciplines,* 60(3), 244-258. https://doi.org/10.1111/jcpp.12980

Dunning, D. L., Holmes, J., & Gathercole, S. E. (2013). Does working memory training lead to generalized improvements in children with low working memory? A randomized controlled trial. *Developmental Science*, *16*(6), 915–925. https://doi.org/10.1111/desc.12068

Dupeyrat, C., & Mariné, C. (2005). Implicit theories of intelligence, goal orientation, cognitive engagement, and achievement: A test of Dweck's model with returning to school adults. *Contemporary Educational Psychology*, *30*, 43–59. https://doi.org/10.1016/j.cedpsych.2004.01.007

Duval, J., Coyette, F., & Seron, X. (2008). Rehabilitation of the central executive component of working memory: A re-organisation approach applied to a single case. *Neuropsychological Rehabilitation*, *18*, 430–460. https://doi.org/10.1080/09602010701573950

Dweck, C. S. (2000). *Self-theories : their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Dweck, C. S. (2012). Implicit theories. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (Vol. 2). Thousand Oaks, CA: Sage.

Dweck, C. S., Chiu, C. Y., & Hong, Y. Y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry*, *6*, 267–285. https://doi.org/10.1207/s15327965pli0604_1

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256. Retrieved from http://psycnet.apa.org/journals/rev/95/2/256/

Engel de Abreu, P. M. J., Conway, A. R. A., & Gathercole, S. E. (2010). Working memory and fluid intelligence in young children. *Intelligence*, *38*(6), 552–561. https://doi.org/10.1016/j.intell.2010.07.003

Engel de Abreu, P. M. J., Gathercole, S. E., & Martin, R. (2011). Disentangling the relationship between working memory and language: The roles of short-term storage and cognitive control. *Learning and Individual Differences*, *21*, 569–574. https://doi.org/10.1016/j.lindif.2011.06.002

Engelhardt, L. E., Harden, K. P., Tucker-Drob, E. M., & Church, J. A. (2019). The neural architecture of executive functions is established by middle childhood. *NeuroImage, 185*, 479-489. http://dx.doi.org/10.1016/j.neuroimage.2018.10.024

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23. https://doi.org/10.1111/1467-8721.00160

Ericsson, K. A., & Chase, W. G. (1982). Exceptional memory. *American Scientist*, *70*, 607–615. Retrieved from https://www.jstor.org/stable/27851732

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/17695343

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy*, *98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Foster, J. L., Harrison, T. L., Hicks, K. L., Draheim, C., Redick, T. S., & Engle, R. W. (2017). Do the effects of working memory training depend on baseline ability level? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(11), 1677–1689. https://doi.org/10.1037/xlm0000426

Foster, J. L., Shipstead, Z., Harrison, T. L., Hicks, K. L., Redick, T. S., & Engle, R. W. (2014). Shortened complex span tasks can reliably measure working memory capacity. *Memory & Cognition*, *43*(2), 226–236. https://doi.org/10.3758/s13421-014-0461-7

Fournier-Vicente, S., Larigauderie, P., & Gaonac'h, D. (2008). More dissociations and interactions within central executive functioning: A comprehensive latent-variable analysis. *Acta Psychologica*, *129*, 32–48. https://doi.org/10.1016/j.actpsy.2008.04.004

Fox, C. J., Mueller, S. T., Gray, H. M., Raber, J., & Piper, B. J. (2013). Evaluation of a short-form of the Berg Card Sorting Test. *PloS One*, *8*(5), 1–4. https://doi.org/10.1371/journal.pone.0063885

Friedman, N. P., Haberstick, B. C., Willcutt, E. G., Miyake, A., Young, S. E., Corley, R. P., & Hewitt, J. K. (2007). Greater attention problems during childhood predict poorer executive functioning in late adolescence. *Psychological Science*, *18*, 893–900. https://doi.org/10.1111/j.1467-9280.2007.01997.x

Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*, 136–158. https://doi.org/10.1016/j.jml.2004.03.008

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex, 86*, 186-204. http://dx.doi.org/10.1016/j.cortex.2016.04.023

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*, 172–179. https://doi.org/10.1111/j.1467-9280.2006.01681.x

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology. General*, *137*, 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*, 237–241. https://doi.org/10.1111/j.1467-9280.1996.tb00366.x

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. Retrieved from https://CRAN.R-project.org/package=irr

García-Madruga, J. A., Elosúa, M. R., Gil, L., Gómez-Veiga, I., Vila, J. Ó., Orjales, I., … Duque, G. (2013). Reading comprehension and working memory's executive processes:

An intervention study in primary school students. *Reading Research Quarterly*, *48*, 155–174. https://doi.org/10.1002/rrq.44

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin*, *134*, 31–60. https://doi.org/10.1037/0033-2909.134.1.31

Gathercole, S. E. (2008). Working memory in the classroom. *The Psychologist*, *21*, 382–385.

Gathercole, S. E., Adams, A.-M., & Hitch, G. (1994). Do young children rehearse? An individual-differences analysis. *Memory & Cognition*, *22*, 201–207.

Gathercole, S. E., Alloway, T. P., Kirkwood, H. J., Elliott, J. G., Holmes, J., & Hilton, K. A. (2008). Attentional and executive function behaviours in children with poor working memory. *Learning and Individual Differences*, *18*, 214–223. https://doi.org/10.1016/j.lindif.2007.10.003

Gathercole, S. E., Alloway, T. P., Willis, C., & Adams, A. M. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology*, *93*, 265–281. https://doi.org/10.1016/j.jecp.2005.08.003

Gathercole, S. E., & Brown, L. (2003). Working memory assessments at school entry as longitudinal predictors of National Curriculum attainment levels. *Educational and Child Psychology*, *20*, 109–122. Retrieved from https://www.researchgate.net/profile/Lorna_Bourke/publication/250928054_The_relationship_between_working_memory_and_early_writing_at_the_word_sentence_and_text_level/

Gathercole, S. E., Dunning, D. L., & Holmes, J. (2012). Cogmed training: Let's be realistic about intervention research. *Journal of Applied Research in Memory and Cognition*, *1*, 201–203. https://doi.org/10.1016/j.jarmac.2012.07.007

Gathercole, S. E., Durling, E., Evans, M., Jeffcock, S., & Stone, S. (2008). Working memory abilities and children's performance in laboratory analogues of classroom activities. *Applied Cognitive Psychology*, *22*, 1019–1037. https://doi.org/10.1002/acp.1407

Gathercole, S. E., & Pickering, S. J. (2000). Assessment of working memory in six- and seven-year-old children. *Journal of Educational Psychology*, *92*, 377–390. https://doi.org/10.1037//0022-0663.92.2.377

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*, 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology*, *18*, 1–16. https://doi.org/10.1002/acp.934

Geary, D. C. (2011). Consequences, characteristics, and causes of mathematical learning disabilities and persistent low achievement in mathematics. *Journal of Developmental and Behavioral Pediatrics: JDBP*, *32*, 250–263. https://doi.org/10.1097/DBP.0b013e318209edef

Gelman, A. (2018). You need 16 times the sample size to estimate an interaction than to estimate a main effect. Retrieved 26 June 2019, from Statistical Modeling, Causal Inference, and Social Science website: https://statmodeling.stat.columbia.edu/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Gibson, B. S., Gondoli, D. M., Johnson, A. C., Steeger, C. M., & Morrissey, R. A. (2012). The future promise of Cogmed working memory training. *Journal of Applied*

Research in Memory and Cognition, *1*, 214–216. https://doi.org/10.1016/j.jarmac.2012.07.003

Gibson, B. S., 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*, 726–737. https://doi.org/10.3758/s13421-013-0295-8

Godinez, D. A., Friedman, N. P., Rhee, S. H., Miyake, A., & Hewitt, J. K. (2012). Phenotypic and genetic analyses of the Wisconsin Card Sort. *Behavior Genetics*, *42*, 209–220. https://doi.org/10.1007/s10519-011-9502-1

Golden, C. J. (1975). A group version of the Stroop Color and Word Test. *Journal of Personality Assessment*, *39*, 386–388. https://doi.org/10.1207/s15327752jpa3904_10

Gonthier, C., Aubry, A., & Bourdin, B. (2017). Measuring working memory capacity in children using adaptive tasks: Example validation of an adaptive complex span. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0916-4

Grissom, N. M. & Reyes, T. M. (2019). Let's call the whole thing off: evaluating gender and sex differences in executive function. *Neuropsychopharmacology, 44,* 1-11. https://doi.org/10.1038/s41386-018-0179-5

Hainmueller, J., Mummolo, J., & Xu, Y. (2018). *How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice*. https://doi.org/10.2139/ssrn.2739221

Halperin, J. M., Marks, D. J., Bedard, A.-C. V., Chacko, A., Curchack, J. T., Yoon, C. A., & Healey, D. M. (2012). Training executive, attention, and motor skills: A proof-of-concept study in preschool children with ADHD. *Journal of Attention Disorders*. https://doi.org/10.1177/1087054711435681

Hambrick, D. Z., & Meinz, E. J. (2011). Limits on the predictive power of domain-specific experience and knowledge in skilled performance. *Current Directions in Psychological Science*, *20*, 275–279. https://doi.org/10.1177/0963721411422061

Hardy, J. L., Nelson, R. A., Thomason, M. E., Sternberg, D. A., Katovich, K., Farzin, F., & Scanlon, M. (2015). Enhancing Cognitive Abilities with Comprehensive Training: A Large, Online, Randomized, Active-Controlled Trial. *PloS One*, *10*(9), e0134467. https://doi.org/10.1371/journal.pone.0134467

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science*, *24*(12), 2409–2419. https://doi.org/10.1177/0956797613492984

Hedden, T., & Yoon, C. (2006). Individual differences in executive processing predict susceptibility to interference in verbal working memory. *Neuropsychology*, *20*(5), 511–528. https://doi.org/10.1037/0894-4105.20.5.511

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50.* 1166-1186. http://dx.doi.org/10.3758/s13428-017-0935-1

Henry, L. (2012). *The development of working memory in children*. London, England: SAFE Publishing.

Heo, M., & Leon, A. C. (2010). Sample sizes required to detect two-way and three-way interactions involving slope differences in mixed-effects linear models. *Journal of Biopharmaceutical Statistics*, *20*(4), 787–802. https://doi.org/10.1080/10543401003618819

Heo, M., Xue, X., & Kim, M. Y. (2014). Sample size requirements to detect a two- or three-way interaction in longitudinal cluster randomized clinical trials with second-level randomization. *Clinical Trials* , *11*(4), 503–507. https://doi.org/10.1177/1740774514532724

Hernández, M., Costa, A., Fuentes, L. J., Vivas, A. B., & Sebastián-Gallés, N. (2010). The impact of bilingualism on the executive control and orienting networks of attention. *Bilingualism: Language and Cognition*, *13*, 315–325. https://doi.org/10.1017/S1366728909990010

Hitchcock, C., & Westwell, M. S. (2017). A cluster-randomised, controlled trial of the impact of Cogmed Working Memory Training on both academic performance and regulation of social, emotional and behavioural challenges. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *58*(2), 140–150. https://doi.org/10.1111/jcpp.12638

Hitch, G., 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*, 184–198.

Holmes, J., Gathercole, S., & Dunning, D. L. (2010). Poor working memory: Impact and interventions. In J. Holmes (Ed.), *Advances in child development and behavior* (pp. 1–43). https://doi.org/10.1016/B978-0-12-374748-8.00001-9

Holmes, J., & Gathercole, S. E. (2013). Taking working memory training from the laboratory into schools. *Educational Psychology Review*, 1–11. https://doi.org/10.1080/01443410.2013.797338

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science*, *12*, F9–F15. https://doi.org/10.1111/j.1467-7687.2009.00848.x

Holmes, J., Gathercole, S. E., Place, M., Alloway, T. P., Elliott, J. G., & Hilton, K. A. (2010). The diagnostic utility of executive function assessments in the identification of ADHD in children. *Child and Adolescent Mental Health*, *15*, 37–43. https://doi.org/10.1111/j.1475-3588.2009.00536.x

Holmes, J., Gathercole, S. E., Place, M., Dunning, D. L., Hilton, K. A., & Elliott, J. G. (2010). Working memory deficits can be overcome: Impacts of training and medication on working memory in children with ADHD. *Applied Cognitive Psychology*, *24*, 827–836. https://doi.org/10.1002/acp.1589

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, Theory and Applications*, *6*(2), 65–70. Retrieved from http://www.jstor.org/stable/4615733

Hong, Y. Y., Chiu, C. Y., Dweck, C. S., Lin, D. M. S., & Wan, W. (1999). Implicit theories, attributions, and coping: A meaning system approach. *Journal of Personality and Social Psychology*, *77*, 588–599. https://doi.org/10.1037//0022-3514.77.3.588

Howard, S. J., Powell, T., Vasseleu, E., Johnstone, S., & Melhuish, E. (2017). Enhancing preschoolers' executive functions through embedding cognitive activities in shared book reading. *Educational Psychology Review,* 29(1), 1-22. https://doi.org/10.1007/s10648-016-9364-4

Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2010). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology*, *35*, 20–36. https://doi.org/10.1080/87565640903325691

Huitema, B.E. (2011). The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies (2nd Edition). Hoboken, NJ: John Wiley & Sons, Inc.

Huizinga, M., Dolan, C. V., & van der Molen, M. W. (2006). Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia*, *44*, 2017–2036. https://doi.org/10.1016/j.neuropsychologia.2006.01.010

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hulme, C., & Melby-Lervag, M. (2012). Current evidence does not support the claims made for Cogmed working memory training. *Journal of Applied Research in Memory and Cognition*, *1*, 197–200. https://doi.org/10.1016/j.jarmac.2012.06.006

Isquith, P. K., Crawford, J. S., Espy, K. A., & Gioia, G. A. (2005). Assessment of executive function in preschool-aged children. *Mental Retardation and Developmental Disabilities Research Reviews*, *11*, 209–215. https://doi.org/10.1002/mrdd.20075

James, W. (1890). *The Principles of Psychology*. New York: Holt & Co. https://doi.org/10.1037/10538-000

Jackson, J. J., Hill, P. L., Payne, B. R., Roberts, B. W., & Stine-Morrow, E. A. L. (2012). Can an old dog learn (and want to experience) new tricks? Cognitive training increases openness to experience in older adults. *Psychology and Aging*, *27*(2), 286–292. https://doi.org/10.1037/a0025918

Jacob, R., & Parkinson, J. (2015). The Potential for School-Based Interventions That Target Executive Function to Improve Academic Achievement: A Review.

Review of Educational Research, *85*(4), 512–552. https://doi.org/10.3102/0034654314561338

Jacobson, L. A., Williford, A. P., & Pianta, R. C. (2011). The role of executive function in children's competent adjustment to middle school. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *17*, 255–280. https://doi.org/10.1080/09297049.2010.535654

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 6829–6833. https://doi.org/10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 10081–10086. https://doi.org/10.1073/pnas.1103228108

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

JASP Team. (2018). *JASP (Version 0.10.0)[Computer software]*. Retrieved from https://jasp-stats.org/

Jastrzębski, J., Ciechanowska, I., & Chuderski, A. (2018). The strong link between fluid intelligence and working memory cannot be explained away by strategy use. *Intelligence*, *66*, 44–53. https://doi.org/10.1016/j.intell.2017.11.002

Johnson, J., Im-Bolter, N., & Pascual-Leone, J. (2003). Development of mental attention in gifted and mainstream children: The role of mental capacity, inhibition, and speed of processing. *Child Development*, *74*, 1594–1614. https://doi.org/10.1046/j.1467-8624.2003.00626.x

Johnstone, S. J., Roodenrys, S., Blackman, R., Johnston, E., Loveday, K., Mantz, S., & Barratt, M. F. (2012). Neurocognitive training for children with and without AD/HD. *ADHD Attention Deficit and Hyperactivity Disorders*, *4*, 11–23. https://doi.org/10.1007/s12402-011-0069-8

Jolles, D. D., & Crone, E. A. (2012). Training the developing brain: a neurocognitive perspective. *Frontiers in Human Neuroscience*, *6*, 1–13. https://doi.org/10.3389/fnhum.2012.00076

Jolles, D. D., van Buchem, M. A., Rombouts, S. A. R. B., & Crone, E. A. (2012). Practice effects in the developing brain: A pilot study. *Developmental Cognitive Neuroscience*, *2*(S1), S180–S191. https://doi.org/10.1016/j.dcn.2011.09.001

Jones, B. D., Wilkins, J. L. M., Long, M. H., & Wang, F. H. (2012). Testing a motivational model of achievement: How students' mathematical beliefs and interests are related to their achievement. *European Journal of Psychology of Education*, *27*, 1–20. https://doi.org/10.1007/s10212-011-0062-9

Jorgensen, T. D., Pornprasertmanit, S., Miller, P., & Schoemann, A. (2018). simsem: SIMulated Structural Equation Modeling (Version 0.5-14) [R]. Retrieved from https://cran.r-project.org/web/packages/simsem/simsem.pdf

Kable, J. W., Caulfield, M. K., Falcone, M., McConnell, M., Bernardo, L., Parthasarathi, T., … Lerman, C. (2017). No Effect of Commercial Cognitive Training on Brain Activity, Choice Behavior, or Cognitive Performance. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(31), 7390–7402. https://doi.org/10.1523/JNEUROSCI.2832-16.2017

Kane, M. J., Conway, A. R. A., & Engle, R. W. (1999). What do working-memory tests really measure? *The Behavioral and Brain Sciences*, *22*, 101–102. https://doi.org/10.1017/S0140525X99291789

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, *9*, 637–671. https://doi.org/10.3758/bf03196323

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*, 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Karbach, J., & Kray, J. (2009). How useful is executive control training? Age differences in near and far transfer of task-switching training. *Developmental Science*, *12*, 978–990. https://doi.org/10.1111/j.1467-7687.2009.00846.x

Karbach, J., & Kray, J. (2016). Executive Functions. In T. Strobach & J. Karbach (Eds.), *Cognitive Training: An Overview of Features and Applications* (pp. 93–103). https://doi.org/10.1007/978-3-319-42662-4

Karbach, J., Strobach, T., & Schubert, T. (2015). Adaptive working-memory training benefits reading, but not mathematics in middle childhood. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *21*(3), 285–301. https://doi.org/10.1080/09297049.2014.899336

Karbach, J., & Unger, K. (2014). Executive control training from middle childhood to adolescence. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00390

Karbach, J., & Verhaeghen, P. (2014). Making Working Memory Work: A Meta-Analysis of Executive-Control and Working Memory Training in Older Adults. *Psychological Science*, *25*(11), 2027–2037. https://doi.org/10.1177/0956797614548725

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L., & Garcia-Barrera, M. A. (2018). The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin*. https://doi.org/10.1037/bul0000160

Kasper, L. J., Alderson, R. M., & Hudec, K. L. (2012). Moderators of working memory deficits in children with attention-deficit/hyperactivity disorder (ADHD): A meta-analytic review. *Clinical Psychology Review*, *32*, 605–617. https://doi.org/10.1016/j.cpr.2012.07.001

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near- and far-transfer effects among children's executive function skills. *Psychological Bulletin*, *145*(2), 165–188. https://doi.org/10.1037/bul0000180

Katz, B., Jones, M. R., Shah, P., Buschkuehl, M., & Jaeggi, S. M. (2016). Individual Differences and Motivational Effects. In T. Strobach & J. Karbach (Eds.), *Cognitive Training: An Overview of Features and Applications* (pp. 157–166). https://doi.org/10.1007/978-3-319-42662-4

Katz, B., Shah, P., & Meyer, D. E. (2018). How to play 20 questions with nature and lose: Reflections on 100 years of brain-training research. *Proceedings of the National*

Academy of Sciences of the United States of America, *115*(40), 9897–9904. https://doi.org/10.1073/pnas.1617102114

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (Mar-Apr 2012). Are cognitive g and academic achievement g one and the same g? An exploration on the Woodcock-Johnson and Kaufman tests. *Intelligence*, *40*, 123–138. https://doi.org/10.1016/j.intell.2012.01.009

Kelly, C., Foxe, J. J., & Garavan, H. (2006). Patterns of normal human brain plasticity after practice and their implications for neurorehabilitation. *Archives of Physical Medicine and Rehabilitation*, *87*, 20–29. https://doi.org/10.1016/j.apmr.2006.08.333

Kenny, D. (1979). *Correlation and Causality.* New York: Wiley.

Khng, K.H. & Lee, K. (2014). The relationship between Stroop and stop-signal measures of inhibiton in adolescents: Influences from variations in context and measure estimation. *PLOS One, 9(7)*, e101356. https://doi.org/10.1371/journal.pone.0101356

Kimberg, D. Y., D'Esposito, M., & Farah, M. J. (1997). Effects of bromocriptine on human subjects depend on working memory capacity. *Neuroreport*, *8*, 3581–3585. https://doi.org/10.1097/00001756-199711100-00032

King, R. B., McInerney, D. M., & Watkins, D. A. (2012). How you think about your intelligence determines how you feel in school: The role of theories of intelligence on academic emotions. *Learning and Individual Differences*, *22*, 814–819. https://doi.org/10.1016/j.lindif.2012.04.005

Klauer, K. J., Willmes, K., & Phye, G. D. (2002). Inducing inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology*, *27*, 1–25. https://doi.org/10.1006/ceps.2001.1079

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th Edition). New York: Guilford Press.

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences*, *14*, 317–324. https://doi.org/10.1016/j.tics.2010.05.002

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlstrom, K., … Westerberg, H. (2005). Computerized training of working memory in children with ADHD - A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*, 177–186. https://doi.org/10.1097/00004583-200502000-00010

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, *24*, 781–791. https://doi.org/10.1076/jcen.24.6.781.8395

Kolb, B., Teskey, G. C., & Gibb, R. (2010). Factors influencing cerebral plasticity in the normal and injured brain. *Frontiers in Human Neuroscience*, *4*, 1–12. https://doi.org/10.3389/fnhum.2010.00204

Kolkman, M. E., Hoijtink, H. J. A., Kroesbergen, E. H., & Leseman, P. P. M. (2013). The role of executive functions in numerical magnitude skills. *Learning and Individual Differences*, *24*, 145–151. https://doi.org/10.1016/j.lindif.2013.01.004

Komarraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the Big Five personality traits in predicting college students' academic motivation and achievement. *Learning and Individual Differences*, *19*(1), 47–52. https://doi.org/10.1016/j.lindif.2008.07.001

Könen, T., & Karbach, J. (2015). The benefits of looking at intraindividual dynamics in cognitive training data. *Frontiers in Psychology*, *6*, 615. https://doi.org/10.3389/fpsyg.2015.00615

Korkman, M., Lahti-Nuuttila, P., Laasonen, M., Kemp, S. L., & Holdnack, J. (09/2013). Neurocognitive development in 5- to 16-year-old North American children: A cross-sectional study. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *19*, 516–539. https://doi.org/10.1080/09297049.2012.705822

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, Vol. 6, pp. 151–162. Retrieved from https://journal.r-project.org/archive/2014-2/korkmaz-goksuluk-zararsiz.pdf

Kotsopoulos, D., & Lee, J. (2012). A naturalistic study of executive function and mathematical problem-solving. *The Journal of Mathematical Behavior*, *31*, 196–208. https://doi.org/10.1016/j.jmathb.2011.12.005

Kray, J., & Ferdinand, N. K. (2013). How to improve cognitive control in development during childhood: Potentials and limits of cognitive interventions. *Child Development Perspectives*, *7*, 121–125. https://doi.org/10.1111/cdep.12027

Kray, J., Karbach, J., Haenig, S., & Freitag, C. (2012). Can task-switching training enhance executive control functioning in children with attention deficit/-hyperactivity disorder? *Frontiers in Human Neuroscience*, *5*, 1–9. https://doi.org/10.3389/fnhum.2011.00180

Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology*, *108*, 677–692. https://doi.org/10.1016/j.jecp.2010.11.001

Latzman, R. D., Elkovitch, N., Young, J., & Clark, L. A. (2010). The contribution of executive functioning to academic achievement among male adolescents.

Journal of Clinical and Experimental Neuropsychology, *32*, 455–462. https://doi.org/10.1080/13803390903164363

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, 1–21. https://doi.org/10.1111/cdev.12096

Lee, K., Ng, E. L., & Ng, S. F. (2009). The contributions of working memory and executive functioning to problem representation and solution generation in algebraic word problems. *Journal of Educational Psychology*, *101*, 373–387. https://doi.org/10.1037/a0013843

Lee, K., Ng, S. F., Pe, M. L., Ang, S. Y., Hasshim, M., & Bull, R. (2012). The cognitive underpinnings of emerging mathematical skills: Executive functioning, patterns, numeracy, and arithmetic. *The British Journal of Educational Psychology*, *82*, 82–99. https://doi.org/10.1111/j.2044-8279.2010.02016.x

LeFevre, J.-A., Berrigan, L., Vendetti, C., Kamawar, D., Bisanz, J., Skwarchuk, S.-L., & Smith-Chant, B. L. (2/2013). The role of executive attention in the acquisition of mathematical skills for children in Grades 2 through 4. *Journal of Experimental Child Psychology*, *114*, 243–261. https://doi.org/10.1016/j.jecp.2012.10.005

Lehto, J. E., Juujarvi, P., Kooistra, L., & Pulkkinen, L. (2003). Dimensions of executive functioning: Evidence from children. *The British Journal of Developmental Psychology*, *21*, 59–80. https://doi.org/10.1348/026151003321164627

Lemay, S., Bédard, M.-A., Rouleau, I., & Tremblay, P.-L. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical Neuropsychologist*, *18*, 284–302. https://doi.org/10.1080/13854040490501718

Lin, S.-H., Liu, C.-M., Hwang, T.-J., Hsieh, M. H., Hsiao, P.-C., Faraone, S. V., … Chen, W. J. (2013). Performance on the Wisconsin Card Sorting Test in families of schizophrenia patients with different familial loadings. *Schizophrenia Bulletin*, *39*(3), 537–546. https://doi.org/10.1093/schbul/sbs141

Locascio, G., Mahone, E. M., Eason, S. H., & Cutting, L. E. (2010). Executive dysfunction among children with reading comprehension deficits. *Journal of Learning Disabilities*, *43*, 441–454. https://doi.org/10.1177/0022219409355476

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Current Directions in Psychological Science*, *20*, 240–245. https://doi.org/10.1177/0963721411415340

Long, J. A. (2019). *jtools: Analysis and Presentation of Social Scientific Data*. Retrieved from https://cran.r-project.org/package=jtools

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology: A Journal on Normal and Abnormal Development in Childhood and Adolescence*, *18*(1), 62–78. https://doi.org/10.1080/09297049.2011.575772

Lu, L., Weber, H. S., Spinath, F. M., & Shi, J. (2011). Predicting school achievement from cognitive and non-cognitive variables in a Chinese sample of elementary school children. *Intelligence*, *39*, 130–140. https://doi.org/10.1016/j.intell.2011.02.002

Lüdecke, D. (2019). *sjstats: Statistical Functions for Regression Models (Version 0.17.5)*. https://doi.org/10.5281/zenodo.1284472

Ludyga, S., Gerber, M., Brand, S., Holsboer-Trachsler, E., & Pühse, U. (2016). Acute effects of moderate aerobic exercise on specific aspects of executive function in different age and fitness groups: A meta-analysis. *Psychophysiology,* 53(11), 1611-1626. https://doi.org/10.1111/psyp.12736

Mackey, A. P., Hill, S. S., Stone, S. I., & Bunge, S. A. (2011). Differential effects of reasoning and speed training in children. *Developmental Science*, *14*, 582–590. https://doi.org/10.1111/j.1467-7687.2010.01005.x

Mair, P., & Wilcox, R. (2018). *WRS2: Wilcox robust estimation and testing*.

Martinez, K., Burgaleta, M., Roman, F. J., Escorial, S., Shih, P. C., Quiroga, M. A., & Colom, R. (2011). Can fluid intelligence be reduced to ‘simple’ short-term storage? *Intelligence*, *39*, 473–480. https://doi.org/10.1016/j.intell.2011.09.001

Martinussen, R., Hayden, J., Hogg-Johnson, S., & Tannock, R. (2005). A meta-analysis of working memory impairments in children with attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*, 377–384. https://doi.org/10.1097/01.chi.0000153228.72591.73

McAuley, T., & White, D. A. (2011). A latent variables examination of processing speed, response inhibition, and working memory during typical development. *Journal of Experimental Child Psychology*, *108*, 453–468. https://doi.org/10.1016/j.jecp.2010.08.009

McCabe, D. P. (2010). The influence of complex working memory span task administration methods on prediction of higher level cognition and metacognitive control of response times. *Memory & Cognition*, *38*, 868–882. https://doi.org/10.3758/mc.38.7.868

McCarty, C. A., Mason, W. A., Kosterman, R., Hawkins, J. D., Lengua, L. J., & McCauley, E. (2008). Adolescent school failure predicts later depression among girls. *Journal of Adolescent Health Care: Official Publication of the Society for Adolescent Medicine*, *43*, 180–187. https://doi.org/10.1016/j.jadohealth.2008.01.023

McClelland, M. M., Acock, A. C., Piccinin, A., Rhea, S. A., & Stallings, M. C. (2013). Relations between preschool attention span-persistence and age 25 educational outcomes. *Early Childhood Research Quarterly*, *28*, 314–324. https://doi.org/10.1016/j.ecresq.2012.07.008

McClelland, M. M., & Cameron, C. E. (2011). Self-regulation and academic achievement in elementary school children. *New Directions for Child and Adolescent Development,* 133, 29-44. https://doi.org/10.1002/cd.302

McClelland, M. M., & Cameron, C. E. (2012). Self-regulation in early childhood: Improving conceptual clarity and developing ecologically valid measures. *Child Development Perspectives*, *6*, 136–142. https://doi.org/10.1111/j.1750-8606.2011.00191.x

McGrew, K.S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence,* 37, 1-10. https://doi.org/10.1016/j.intell.2008.08.004

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

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of 'far transfer': Evidence from a meta-analytic review. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(4), 512–534. https://doi.org/10.1177/1745691616635612

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*. Retrieved from http://doi.apa.org/journals/bul/105/1/156.html

Miele, D. B., Son, L. K., & Metcalfe, J. (2013). Children’s naive theories of intelligence influence their metacognitive judgments. *Child Development*, 1–8. https://doi.org/10.1111/cdev.12101

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. https://doi.org/10.1146/annurev.neuro.24.1.167

Miller, M. R., Giesbrecht, G. F., Muller, U., McInerney, R. J., & Kerns, K. A. (2012). A latent variable approach to determining the structure of executive function in preschool children. *Journal of Cognition and Development: Official Journal of the Cognitive Development Society*, *13*, 395–423. https://doi.org/10.1080/15248372.2011.585478

Miller, M. R., & Hinshaw, S. P. (2010). Does childhood executive function predict adolescent functional outcomes in girls with ADHD? *Journal of Abnormal Child Psychology*, *38*, 315–326. https://doi.org/10.1007/s10802-009-9369-2

Miller, M. R., Nevado-Montenegro, A. J., & Hinshaw, S. P. (2012). Childhood executive function continues to predict outcomes in young adult females with and without childhood-diagnosed ADHD. *Journal of Abnormal Child Psychology*, *40*, 657–668. https://doi.org/10.1007/s10802-011-9599-y

Millner, A. J., Jaroszewski, A. C., Chamarthi, H., & Pizzagalli, D. A. (2012). Behavioral and electrophysiological correlates of training-induced cognitive control improvements. *NeuroImage*, *63*, 742–753. https://doi.org/10.1016/j.neuroimage.2012.07.032

Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology. General*, *130*, 163–168.

Miyake, A., Emerson, M. J., & Friedman, N. P. (2000). Assessment of executive functions in clinical settings: Problems and recommendations. *Seminars in Speech and Language*, *21*, 169–183. https://doi.org/10.1055/s-2000-7563

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*, 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex 'frontal lobe' tasks: A latent variable analysis. *Cognitive Psychology*, *41*, 49–100. https://doi.org/10.1006/cogp.1999.0734

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, P. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*, 621–640. https://doi.org/10.1037/0096-3445.130.4.621

Miyake, A., & Shah, P. (1999). Toward unified theories of working memory: Emerging general consensus, unresolved theoretical issues, and future research directions. In Miyake & P. Shah (Eds.), & Miyake (Trans.), *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., … Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(7), 2693–2698. https://doi.org/10.1073/pnas.1010076108

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., … Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* , *340*, c869. https://doi.org/10.1136/bmj.c869

Montoya, M. F., Susperreguy, M. I., Dinarte, L., Morrison, F. J., San Martin, E., Rojas-Barahona, C. A., & Forster, C. E. (2019). Executive function in Chilean preschool children: Do short-term memory, working memory, and response inhibition contribute differentially to early academic skills? *Early Childhood Research Quarterly, 46*, 187-200. http://dx.doi.org/10.1016/j.ecresq.2018.02.009

Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, *37*, 327–328. https://doi.org/10.1016/j.intell.2009.04.005

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review*, *18*(1), 46–60. https://doi.org/10.3758/s13423-010-0034-0

Mowbray, C. T., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*, 315–340.

Mueller, S. T. (2007). The PEBL Perceptual Vigilance Task (Version 0.3). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2011a). PEBL's Berg Card Sorting Test - 64 (PBCST-64) (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2011b). The PEBL Flanker Task (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2011c). The PEBL Go/No-Go test (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2011d). The PEBL Numerical Stroop Task (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2012a). The PEBL Connections Test (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2012b). The PEBL Switcher Task (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T. (2012c). The Psychology Experiment Building Language (Version 0.13). Computer software retrieved from http://pebl.sf.net/battery.html.

Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods*, *222*, 250–259. https://doi.org/10.1016/j.jneumeth.2013.10.024

Müller, U., & Kerns, K. (2015). The development of executive function. In L. Liben & U. Müller (Eds.), *Handbook of child psychology and developmental science, Vol. 2: Cognitive processes* (7th ed., pp. 571–623). Retrieved from http://ebookcentral.proquest.com.ezproxy1.library.usyd.edu.au/lib/usyd/reader.action?docID=1895801

Naglieri, J. A. (2008). *Naglieri Nonverbal Ability Test (2nd edition)*. U.S.A.: NCS Pearson.

Naglieri, J. A. (2011). *Naglieri Nonverbal Ability Test (2nd edition) manual: Technical information and normative data*. U.S.A.: NCS Pearson.

Naglieri, J. A., & Goldstein, S. (2013). *Comprehensive Executive Function Inventory: Product overview*. Retrieved from Multi-Health Systems Inc. website: http://downloads.mhs.com/cefi/CEFI-Overview.pdf

Narayanan, N. S., Prabhakaran, V., Bunge, S. A., Christoff, K., Fine, E. M., & Gabrieli, J. D. E. (2005). The role of the prefrontal cortex in the maintenance of verbal working memory: An event-related fMRI analysis. *Neuropsychology*, *19*, 223–232. https://doi.org/10.1037/0894-4105.19.2.223

Nayfeld, I., Fuccillo, J., & Greenfield, D. B. (2013). Executive functions in early learning: Extending the relationship between executive functions and school readiness to science. *Learning and Individual Differences*, *26*, 81–88. https://doi.org/10.1016/j.lindif.2013.04.011

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2012). A Meta-analysis of executive components of working memory. *Cerebral Cortex* . https://doi.org/10.1093/cercor/bhs007

Nelwan, M., & Kroesbergen, E. H. (2016). Limited Near and Far Transfer Effects of Jungle Memory Working Memory Training on Learning Mathematics in Children with Attentional and Mathematical Difficulties. *Frontiers in Psychology*, *7*, 1384. https://doi.org/10.3389/fpsyg.2016.01384

Nettelbeck, T., & Burns, N. R. (2010). Processing speed, working memory and reasoning ability from childhood to old age. *Personality and Individual Differences*, *48*, 379–384. https://doi.org/10.1016/j.paid.2009.10.032

Neubauer, A., Gawrilow, C., & Hasselhorn, M. (2012). The Watch-and-Wait Task: On the reliability and validity of a new method of assessing self-control in preschool children. *Learning and Individual Differences*, *22*, 770–777. https://doi.org/10.1016/j.lindif.2012.05.006

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence new findings and theoretical developments. *The American Psychologist*, *67*, 130–159. https://doi.org/10.1037/a0026699

O'Connell, R. G., & Robertson, I. H. (2012). Training the brain: Nonpharmacological approaches to stimulating cognitive plasticity. In M. I. Posner (Ed.), *Cogntive neuroscience of attention* (2nd ed.). New York, NY: The Guildford Press.

Olesen, P. J., Westerberg, H., & Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature Neuroscience*, *7*, 75–79. https://doi.org/10.1038/nn1165

Owen, A. M., Hampshire, A., Grahn, J. A., Stenton, R., Dajani, S., Burns, A. S., … Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*, 1–13. https://doi.org/10.1038/nature09042

Paananen, M., Aro, T., Närhi, V., & Aro, M. (2018). Group-based intervention on attention and executive functions in the school context. *Educational Psychology Review*, *38*(7), 859–876. https://doi.org/10.1080/01443410.2017.1407407

Pardo-Vazquez, J. L., & Fernandez-Rey, J. (2012). Working memory capacity and mental rotation: evidence for a domain-general view. *The Spanish Journal of Psychology*, *15*, 881–890. https://doi.org/10.5209/rev_sjop.2012.v15.n3.39381

Parong, J., Mayer, R. E., Fiorella, L., MacNamara, A., Homer, B. D., & Plass, J. L. (2017). Learning executive function skills by playing focused video games. *Contemporary Educational Psychology*, *51*(Supplement C), 141–151. https://doi.org/10.1016/j.cedpsych.2017.07.002

Peng, J., Mo, L., Huang, P., & Zhou, Y. (2017). The effects of working memory training on improving fluid intelligence of children during early childhood. *Cognitive Development*, *43*, 224–234. https://doi.org/10.1016/j.cogdev.2017.05.006

Pepi, A., Alesi, M., & Rappo, G. (2008). Personal conceptions of intelligence affect outcome in a multimedia reading training program. *Perceptual and Motor Skills*, *107*, 963–973. https://doi.org/10.2466/pms.107.3.963-973

Pepinsky, T. B. (2018). A note on listwise deletion versus multiple imputation. *Political Analysis,* 26(4), 480-488. https://doi.org/ 10.1017/pan.2018.18

Pickering, S., Gathercole, S., & Peaker, S. (1998). Verbal and visuospatial short-term memory in children: Evidence for common and distinct mechanisms. *Memory & Cognition*, *26*, 1117–1130. https://doi.org/10.3758/bf03201189

Piper, B., Li, V., Eiwaz, M., Kobel, Y., Benice, T., Chu, A., … Mueller, S. (2012). Executive function on the Psychology Experiment Building Language tests. *Behavior Research Methods*, *44*, 110–123. https://doi.org/10.3758/s13428-011-0096-6

Preßler, A.-L., Krajewski, K., & Hasselhorn, M. (2013). Working memory capacity in preschool children contributes to the acquisition of school relevant precursor skills. *Learning and Individual Differences*, *23*, 138–144. https://doi.org/10.1016/j.lindif.2012.10.005

Protzko, J. (2017). Effects of cognitive training on the structure of intelligence. *Psychonomic Bulletin & Review*, *24*(4), 1022–1031. https://doi.org/10.3758/s13423-016-1196-1

Rabipour, S., & Raz, A. (2012). Training the brain: Fact and fad in cognitive and behavioral remediation. *Brain and Cognition*, *79*, 159–179. https://doi.org/10.1016/j.bandc.2012.02.006

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164. Retrieved from http://www.vcharite.univ-mrs.fr/PP/lubrano/BMC/Raftery-Socmeth1995.pdf

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences*, *20*, 110–122. https://doi.org/10.1016/j.lindif.2009.10.005

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.0). Retrieved from https://www.R-project.org/

Read, K.L., Kendall, P.C., Carper, M.M., & Rausch, J.R. (2013). Statistical methods for use in the analysis of randomized clinical trials utilizing a pretreatment, posttreatment, follow-up (PPF) paradigm. In Comer, J., & Kendall, P. (Eds.) *The Oxford Handbook of Research Strategies for Clinical Psychology*. New York: Oxford University Press.

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: Official Organ of the European Association of Psychological Assessment*, *28*, 164–171. https://doi.org/10.1027/1015-5759/a000123

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2012). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology. General*, *142 (2)*, 359–379. https://dx.doi.org/10.1037/a0029082

Redick, T. S., Shipstead, Z., Meier, M. E., Montroy, J. J., Hicks, K. L., Unsworth, N., Kane, M. J., Hambrick, D. Z., & Engle, R. W. (2016). Cognitive predictors of a common multitasking ability: Contributions from working memory, attention control, and fluid intelligence. *Journal of Experimental Psychology: General. 145 (11),* 1473-1492. http://dx.doi.org/10.1037/xge0000219

Redick, T. S., Shipstead, Z., Wiemers, E. A., Melby-Lervåg, M., & Hulme, C. (2015). What's Working in Working Memory Training? An Educational Perspective.

Educational Psychology Review, *27*(4), 617–633. https://doi.org/10.1007/s10648-015-9314-6

Reed, J. L., Gallagher, N. M., Sullivan, M., Callicott, J. H., & Green, A. E. (2017). Sex differences in verbal working memory performance emerge at very high loads of common neuroimaging tasks. *Brain and Cognition, 113,* 56-64. http://dx.doi.org/10.1016/j.bandc.2017.01.001

Reitan, R. M. (1955). The relation of the Trail Making Test to organic brain damage. *Journal of Consulting Psychology*, *19*, 393–394. https://doi.org/10.1037/h0044509

Richland, L. E., & Burchinal, M. R. (2013). Early executive function predicts reasoning development. *Psychological Science*, *24*, 87–92. https://doi.org/10.1177/0956797612450883

Roberts, G., Quach, J., Spencer-Smith, M.,; Anderson, P. J,, Gathercole, S., Gold, L., Sia, K., Mensah, F., Rickards, F., Ainley, J., & Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomised clinical trial. *JAMA Pediatrics, 170(5)*. https://doi.org/10.1001/jamapediatrics.2015.4568

Roebers, C. M., Cimeli, P., Rothlisberger, M., & Neuenschwander, R. (2012). Executive functioning, metacognition, and self-perceived competence in elementary school children: an explorative study on their interrelations and their role for school achievement. *Metacognition and Learning*, *7*, 151–173. https://doi.org/10.1007/s11409-012-9089-9

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2011). Modeling a cascade of effects: The role of speed and executive functioning in preterm/full-term differences in academic achievement. *Developmental Science*, *14*, 1161–1175. https://doi.org/10.1111/j.1467-7687.2011.01068.x

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Van Rossem, R. (2012). Information processing from infancy to 11 years: Continuities and prediction of IQ. *Intelligence*, *40*, 445–457. https://doi.org/10.1016/j.intell.2012.05.007

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, Vol. 48, pp. 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rothlisberger, M., Neuenschwander, R., Cimeli, P., Michel, E., & Roebers, C. M. (2012). Improving executive functions in 5- and 6-year-olds: Evaluation of a small group intervention in prekindergarten and kindergarten children. *Infant and Child Development*, *21*, 411–429. https://doi.org/10.1002/icd.752

Roughan, L., & Hadwin, J. A. (2011). The impact of working memory training in young people with social, emotional and behavioural difficulties. *Learning and Individual Differences*, *21*, 759–764. https://doi.org/10.1016/j.lindif.2011.07.011

RStudio Team (2018). RStudio: Integrated development environment for R. RStudio Inc., Boston, MA. URL: http://www.rstudio.com/

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. H. (2012). A Potential Spatial Working Memory Training Task to Improve Both Episodic Memory and Fluid Intelligence. *PloS One*, *7*(11), e50431. https://doi.org/10.1371/journal.pone.0050431

Rueda, M. R., Rothbart, M. K., McCandliss, B. D., Saccomanno, L., & Posner, M. I. (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*, 14931–14936. https://doi.org/10.1073/pnas.0506897102

Rushworth, M. F. S., Passingham, R. E., & Nobre, A. C. (2002). Components of switching intentional set. *Journal of Cognitive Neuroscience*, *14*, 1139–1150. https://doi.org/10.1162/089892902760807159

Sala, G., Aksayli, N., Tatlidil, K., Tatsumi, T., Gondo, Y., & Gobet, F. (2018). *Near and Far Transfer in Cognitive Training: A Second-Order Meta- Analysis*. https://doi.org/10.17605/OSF.IO/9EFQD

Sala, G. & Gobet, F. (2017). Does Far Transfer Exist? Negative Evidence From Chess, Music, and Working Memory Training. *Current Directions in Psychological Science*, 0963721417712760. https://doi.org/10.1177/0963721417712760

Sala, G. & Gobet, F. (2020). Working memory training in typically developing children: A multilevel meta-analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-019-01681-y

Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. *Frontiers in Human Neuroscience*, *6*, 1–14. https://doi.org/10.3389/fnhum.2012.00166

Salthouse, T. A. (2011). What cognitive abilities are involved in trail-making performance? *Intelligence*, *39*, 222–232. https://doi.org/10.1016/j.intell.2011.03.001

Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology. General*, 132 (4), 566–594. https://doi.org/10.1037/0096-3445.132.4.566

Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the Trail Making Test. *Neuropsychology*, *14*, 102–111. https://doi.org/10.1037//0894-4105.14.1.102

Sanchez-Cubillo, I., Perianez, J. A., Adrover-Roig, D., Rodriguez-Sanchez, J. M., Rios-Lago, M., Tirapu, J., & Barcelo, F. (2009). Construct validity of the Trail Making Test: Role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *Journal of the International Neuropsychological Society: JINS*, *15*, 438–450. https://doi.org/10.1017/s1355617709090626

Sánchez-Pérez, N., Castillo, A., López-López, J. A., Pina, V., Puga, J. L., Campoy, G., … Fuentes, L. J. (2018). Computer-Based Training in Math and Working Memory Improves Cognitive Skills and Academic Achievement in Primary School Children: Behavioral Results. *Frontiers in Psychology*, *8*, 2327. https://doi.org/10.3389/fpsyg.2017.02327

Santa-Cruz, C., & Rosas, R. (2017). Mapping of Executive Functions / Cartografía de las Funciones Ejecutivas. *Estudios de Psicología*, *38*(2), 284–310. https://doi.org/10.1080/02109395.2017.1311459

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 1089–1096. https://doi.org/10.1037/a0015730

Schmiedek, F., Lovden, 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. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2914582/

Schneider, G., Chicken, E., & Becvarik, R. (2018). *NSM3: Functions and Datasets to Accompany Hollander, Wolfe, and Chicken - Nonparametric Statistical Methods, Third Edition*. Retrieved from https://CRAN.R-project.org/package=NSM3

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. Flanagan & P. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (4th ed., pp. 73-163). New York, NY: Guilford Press.

Schoemaker, K., Mulder, H., Deković, M., & Matthys, W. (2013). Executive functions in preschool children with externalizing behavior problems: A meta-analysis. *Journal of Abnormal Child Psychology*, *41*, 457–471. https://doi.org/10.1007/s10802-012-9684-x

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does Working Memory Training Transfer? A Meta-Analysis Including Training Conditions as Moderators. *Educational Psychologist*, *50*(2), 138–166. https://doi.org/10.1080/00461520.2015.1036274

Schweizer, K., Moosbrugger, H., & Goldhammer, F. (2005). The structure of the relationship between attention and intelligence. *Intelligence*, *33*, 589–611. https://doi.org/10.1016/j.intell.2005.07.001

Senn, T. E., Espy, K. A., & Kaufmann, P. M. (2004). Using path analysis to understand executive function organization in preschool children. *Developmental Neuropsychology*, *26*, 445–464. https://doi.org/10.1207/s15326942dn2601_5

Shah, P., Buschkuehl, M., Jaeggi, S., & Jonides, J. (2012). Cognitive training for ADHD: The importance of individual differences. *Journal of Applied Research in Memory and Cognition*, *1*, 204–205. https://doi.org/10.1016/j.jarmac.2012.07.001

Shah, P., & Miyake, A. (1999). Models of working memory: An introduction. In Miyake & P. Shah (Eds.), & Miyake (Trans.), *Models of working memory: Mechanisms of active maintenance and executive control*. New York: Cambridge University Press.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition*, *1*, 185–193. https://doi.org/10.1016/j.jarmac.2012.06.003

Shipstead, Z., Redick, T. S., & Engle, R. W. (2010). Does working memory training generalize? *Psychologica Belgica*, *50*, 245–276. (WOS:000282537700007).

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, *138*, 628–654. https://doi.org/10.1037/a0027473

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18839484

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do 'Brain-Training' Programs Work? *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *17*(3), 103–186. https://doi.org/10.1177/1529100616661983

Singh, V. (2016). Sex-differences, handedness, and lateralization in the Iowa Gambling Task. *Frontiers in Psychology, 7*, 1-15. http://dx.doi.org/10.3389/fpsyg.2016.00708

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*, 417–453. https://doi.org/10.3102/00346543075003417

Soderqvist, 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*, 1–8. https://doi.org/10.3389/fnhum.2012.00271

Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review*, *24*(4), 1077–1096. https://doi.org/10.3758/s13423-016-1217-0

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology*, *59*, 745–759. https://doi.org/10.1080/17470210500162854

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

Stins, J. F., Polderman, J. C. T., Boomsma, D. I., & de Geus, E. J. C. (2007). Conditional accuracy in response interference tasks: Evidence from the Eriksen flanker task and the spatial conflict task. *Advances in Cognitive Psychology / University of Finance and Management in Warsaw*, *3*, 409–417. https://doi.org/10.2478/v10053-008-0005-4

Streiner, D. L. (2006). Building a better model: an introduction to structural equation modelling. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *51*(5), 317–324. https://doi.org/10.1177/070674370605100507

Strobach, T., & Karbach, J. (Eds.). (2016). *Cognitive Training: An Overview of Features and Applications*. https://doi.org/10.1007/978-3-319-42662-4

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662. https://doi.org/10.1037/h0054651

Studer-Luethi, B., Jaeggi, S. M., Buschkuehl, M., & Perrig, W. J. (2012). Influence of neuroticism and conscientiousness on working memory training outcome. *Personality and Individual Differences*, *53*, 44–49. https://doi.org/10.1016/j.paid.2012.02.012

Swanson, H. L. (2011). Working memory, attention, and mathematical problem solving: A longitudinal study of elementary school children. *Journal of Educational Psychology*, *103*, 821–837. https://doi.org/10.1037/a0025114

Swanson, H. L., & O'Connor, R. (2009). The role of working memory and fluency practice on the reading comprehension of students who are dysfluent readers. *Journal of Learning Disabilities*, *42*, 548–575. https://doi.org/10.1177/0022219409338742

Takacs, Z. K. & Kassai, R. (2019). The efficacy of different interventions to foster children's executive function skills: A series of meta-analyses. *Psychological Bulletin*, 145(7), 653-697. https://doi.org/10.1037/bul0000195

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., … Gabrieli, J. D. E. (2013). Failure of working memory training to enhance cognition or intelligence. *PloS One*, *8*(5), 1–15. https://doi.org/10.1371/journal.pone.0063614

Thorell, L. B., Lindqvist, S., Nutley, S. B., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*, 106–113. https://doi.org/10.1111/j.1467-7687.2008.00745.x

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic Bulletin & Review, 21,* 620-628. http://dx.doi.org/10.3758/s13423-013-0560-7

Titz, C., & Karbach, J. (2014). Working memory and executive functions: effects of training on academic achievement. *Psychological Research*, *78*(6), 852–868. https://doi.org/10.1007/s00426-013-0537-1

Toll, S. W. M., Van der Ven, S. H. G., Kroesbergen, E. H., & Van Luit, J. E. H. (2011). Executive functions as predictors of math learning disabilities. *Journal of Learning Disabilities*, *44*, 521–532. https://doi.org/10.1177/0022219410387302

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Practitioner Review: Do performance-based measures and ratings of executive function assess the same construct? *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *54*, 131–143. https://doi.org/10.1111/jcpp.12001

Traverso, L., Viterbori, P., & Usai, M. C. (2015). Improving executive function in childhood: evaluation of a training intervention for 5-year-old children. *Frontiers in Psychology*, *6*, 525. https://doi.org/10.3389/fpsyg.2015.00525

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language*, *49*, 446–468. https://doi.org/10.1016/s0749-596x(03)00095-0

Uka, F., Gunzenhauser, C., Larsen, R. A., & von Suchodoletz, A. (2019). Exploring a bidirectional model of executive functions and fluid intelligence. *Intelligence, 75*, 111-121. http://dx.doi.org/10.1016/j.intell.2019.05.002

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16405146

Unsworth, N., Miller, J. D., Lakey, C. E., Young, D. L., Meeks, J. T., Campbell, W. K., & Goodie, A. S. (2009). Exploring the Relations Among Executive Functions, Fluid Intelligence, and Personality. *Journal of Individual Differences*, *30*(4), 194–200. https://doi.org/10.1027/1614-0001.30.4.194

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory* , *17*, 635–654. https://doi.org/10.1080/09658210902998047

Unsworth, N., & Spillers, G. J. (2010). Working memory capacity: Attention control, secondary memory, or both? A direct test of the dual-component model. *Journal of Memory and Language*, *62*, 392–406. https://doi.org/10.1016/j.jml.2010.02.001

van den Bos, R., Homberg, J., & de Visser, L. (2013). A critical review of sex differences in decision-making tasks: Focus on the Iowa Gambling Task. *Behavioural Brain Research, 238*, 95-108. http://dx.doi.org/10.1016/j.bbr.2012.10.002

van der Donk, M. L. A., van Viersen, S., Hiemstra-Beernink, A.-C., Tjeenk-Kalff, A. C., van der Leij, A., & Lindauer, R. J. L. (2017). Individual Differences in Training Gains and Transfer Measures: An Investigation of Training Curves in Children with

Attention-Deficit/Hyperactivity Disorder. *Applied Cognitive Psychology*, *31*(3), 302–314. https://doi.org/10.1002/acp.3327

van der Molen, M. J., Van Luit, J. E. H., Van der Molen, M. W., Klugkist, I., & Jongmans, M. J. (2010). Effectiveness of a computerised working memory training in adolescents with mild to borderline intellectual disabilities. *Journal of Intellectual Disability Research: JIDR*, *54*, 433–447. https://doi.org/10.1111/j.1365-2788.2010.01285.x

van der Sluis, S., de Jong, P. F., & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence*, *35*, 427–449. https://doi.org/10.1016/j.intell.2006.09.001

van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2012). The development of executive functions and early mathematics: A dynamic relationship. *The British Journal of Educational Psychology*, *82*, 100–119. https://doi.org/10.1111/j.2044-8279.2011.02035.x

van der Ven, S. H. G., Kroesbergen, E. H., Boom, J., & Leseman, P. P. M. (2013). The structure of executive functions in children: A closer examination of inhibition, shifting, and updating. *The British Journal of Developmental Psychology*, *31*, 70–87. https://doi.org/10.1111/j.2044-835X.2012.02079.x

van Muijden, J., Band, G. P. H., & Hommel, B. (2012). Online games training aging brains: Limited transfer to cognitive control functions. *Frontiers in Human Neuroscience*, *6*(221), 1–13. https://doi.org/10.3389/fnhum.2012.00221

Vaughan, L., & Giovanello, K. (2010). Executive function in daily life: Age-related influences of executive processes on instrumental activities of daily living. *Psychology and Aging*, *25*, 343–355. https://doi.org/10.1037/a0017729

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). Fourth. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Visu-Petra, L., Cheie, L., Benga, O., & Miclea, M. (2012). The structure of executive functions in preschoolers: An investigation using the NEPSY battery. *Procedia - Social and Behavioral Sciences*, *33*, 627–631.

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychological Bulletin & Review, 24*, 307-334. http://dx.doi.org/10.3758/s13423-016-1085-7

Walton, G. M., & Dweck, C. S. (2009). Solving social problems like a psychologist. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *4*, 101–102. https://doi.org/10.1111/j.1745-6924.2009.01098.x

Wass, S. V., Scerif, G., & Johnson, M. H. (2012). Training attentional control and working memory: Is younger, better? *Developmental Review: DR*, *32*, 360–387. https://doi.org/10.1016/j.dr.2012.07.001

Watson, P. (2019). Rules of thumb on magnitudes of effect sizes. *Cognition and Brain Sciences Unit: University of Cambridge.* Retrieved February 10, 2020 from http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize.

Weber, H. S., Lu, L., Shi, J., & Spinath, F. M. (2013). The roles of cognitive and motivational predictors in explaining school achievement in elementary school. *Learning and Individual Differences*, *25*, 85–92. https://doi.org/10.1016/j.lindif.2013.03.008

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., … Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Supplement 3), S54–S64. https://doi.org/10.1212/WNL.0b013e3182872ded

Weissberg, R., Ruff, H., & Lawson, K. (1990). The usefulness of reaction-time tasks in studying attention and organization of behavior in young-children. *Journal of Developmental and Behavioral Pediatrics: JDBP*, *11*(2), 59–64.

Westerberg, H., Jacobaeus, H., Hirvikoski, T., Clevberger, P., Ostensson, M. L., Bartfai, A., & Klingberg, T. (2007). Computerized working memory training after stroke: A pilot study. *Brain Injury*, *21*, 21–29. https://doi.org/10.1080/02699050601148726

Wexler, B.E., Vitulano, L. A., Moore, C., Katsovich, L., Smith, S. D., Rush, C., Grantz, H., Dong, J., & Leckman, J. F. (2020). An integrated program of computer-presented and physical cognitive training exercises for children with attention-deficit/hyperactivity disorder. *Psychological Medicine,* 1-12. https://doi.org/10.1017/S0033291720000288

Wheldall, K., Wheldall, R., Madelaine, A., Reynolds, M., & Arakelian, S. (2017). Further evidence for the efficacy of an evidence-based, small group, literacy intervention program for young struggling readers. *Australian Journal of Learning Difficulties*, 22 (1), 3-13. https://doi.org/10.1080/19404158.2017.1287102

White, L. K., McDermott, J. M., Degnan, K. A., Henderson, H. A., & Fox, N. A. (2011). Behavioral inhibition and anxiety: The moderating roles of inhibitory control and attention shifting. *Journal of Abnormal Child Psychology*, *39*, 735–747. https://doi.org/10.1007/s10802-011-9490-x

Whittaker, T. A. (2012). Using the Modification Index and Standardized Expected Parameter Change for Model Modification. *Journal of Experimental Education*, *80*(1), 26–44. https://doi.org/10.1080/00220973.2010.531299

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. A. C., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-old children. *Journal of Experimental Child Psychology*, *108*, 436–452. https://doi.org/10.1016/j.jecp.2010.08.008

Wiemers, E. A., Redick, T. S., & Morrison, A. B. (2019). The influence of individual differences in cognitive ability on working memory training gains. *The Journal of Cognitive Enhancement, 3*, 174-185. https://doi.org/10.1007/s41465-018-0111-2

Wilbourn, M. P., Kurtz, L. E., & Kalia, V. (2012). The Lexical Stroop Sort (LSS) picture-word task: A computerized task for assessing the relationship between language and executive functioning in school-aged children. *Behavior Research Methods*, *44*, 270–286. https://doi.org/10.3758/s13428-011-0142-4

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing - 4th Edition*. Retrieved from https://www.elsevier.com/books/introduction-to-robust-estimation-and-hypothesis-testing/wilcox/978-0-12-804733-0

Willoughby, M. T., Wirth, R. J., Blair, C. B., & Family Life Project Investigators. (2012). Executive function in early childhood: longitudinal measurement invariance and developmental change. *Psychological Assessment*, *24*(2), 418–431. https://doi.org/10.1037/a0025779

Willoughby, M. T., Wirth, R. J., Blair, C. B., & Greenberg, M. (2010). The measurement of executive function at age 3 years: Psychometric properties and criterion validity of a new battery of tasks. *Psychological Assessment*, *22*, 306–317. https://doi.org/10.1037/a0018708

Witt, M. (2011). School based working memory training: Preliminary finding of improvement in children's mathematical performance. *Advances in Cognitive Psychology / University of Finance and Management in Warsaw*, *7*, 7–15. https://doi.org/10.2478/v10053-008-0083-3

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H.-J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition*, *81*, 82–94. https://doi.org/10.1016/j.bandc.2012.09.010

Wu, K. K., Chan, S. K., Leung, P. W. L., Liu, W.-S., Leung, F. L. T., & Ng, R. (2011). Components and developmental differences of executive functioning for school-aged children. *Developmental Neuropsychology*, *36*, 319–337. https://doi.org/10.1080/87565641.2010.549979

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*, 302–314. https://doi.org/10.1080/00461520.2012.722805

Yeniad, N., Malda, M., Mesman, J., van Ijzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences*, *23*, 1–9. https://doi.org/10.1016/j.lindif.2012.10.004

Zelazo, P. D., Blair, C., & Willoughby, M. T. (2017). *Executive function: Implications for education*. Retrieved from http://ies.ed.gov/ncer/pubs/20172000/pdf/20172000.pdf

Zenner, C., Herrnleben-Kurz, S., & Walach, H. (2014). Mindfulness-based interventions in schools – a systematic review and meta-analysis. *Frontiers in Psychology, 5 (603),* 1-20. https://doi.org/ 10.3389/fpsyg.2014.00603

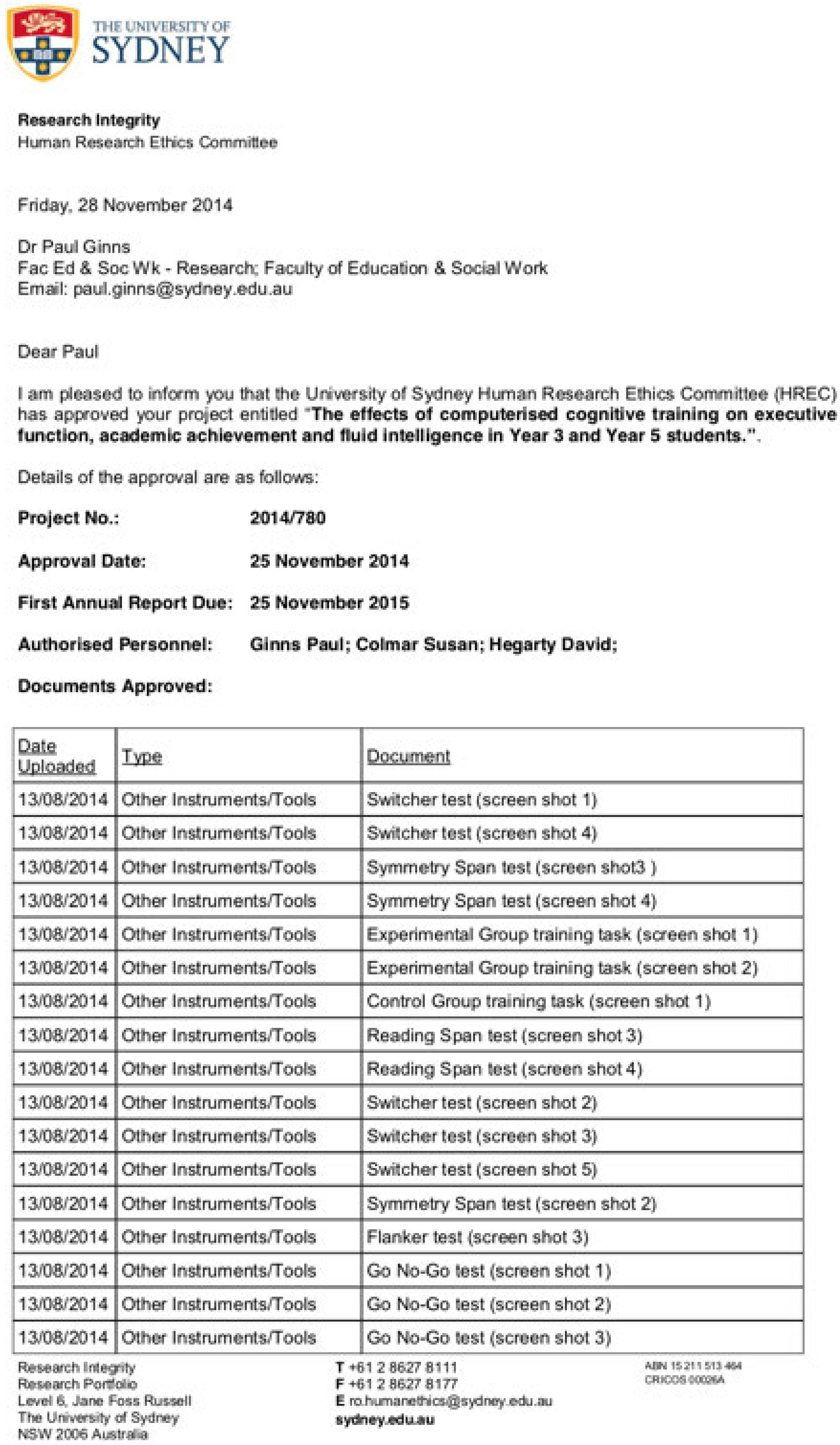
Zhao, X., Xu, Y., Fu, J., & Maes, J. H. R. (2018). Are training and transfer effects of working memory updating training modulated by achievement motivation? *Memory & Cognition*, *46*(3), 398–409. https://doi.org/10.3758/s13421-017-0773-5

Zhou, Q., Chen, S. H., & Main, A. (2012). Commonalities and differences in the research on children's effortful control and executive function: A call for an integrated model of self-regulation. *Child Development Perspectives*, *6*, 112–121. https://doi.org/10.1111/j.1750-8606.2011.00176.x

Zinke, K., Einert, M., Pfennig, L., & Kliegel, M. (2012). Plasticity of executive control through task switching training in adolescents. *Frontiers in Human Neuroscience*, *6*, 1–15. https://doi.org/10.3389/fnhum.2012.00041

9 APPENDICES

9.1 Appendix A: Letters of Ethical Approval



THE UNIVERSITY OF SYDNEY

Research Integrity
Human Research Ethics Committee

Friday, 28 November 2014

Dr Paul Ginns
Fac Ed & Soc Wk - Research; Faculty of Education & Social Work
Email: paul.ginns@sydney.edu.au

Dear Paul

I am pleased to inform you that the University of Sydney Human Research Ethics Committee (HREC) has approved your project entitled "**The effects of computerised cognitive training on executive function, academic achievement and fluid intelligence in Year 3 and Year 5 students.**".

Details of the approval are as follows:

Project No.: **2014/780**

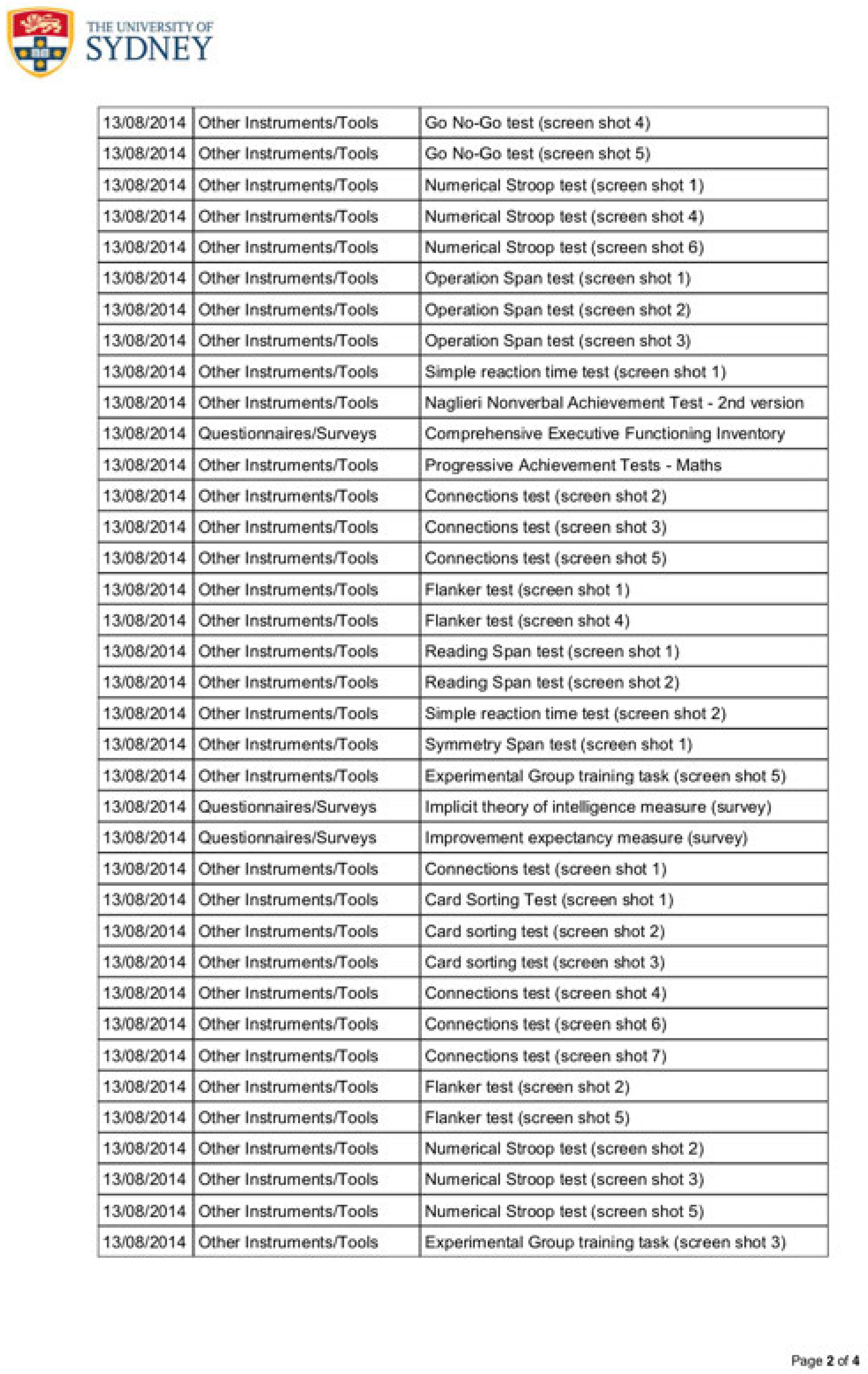
Approval Date: **25 November 2014**

First Annual Report Due: **25 November 2015**

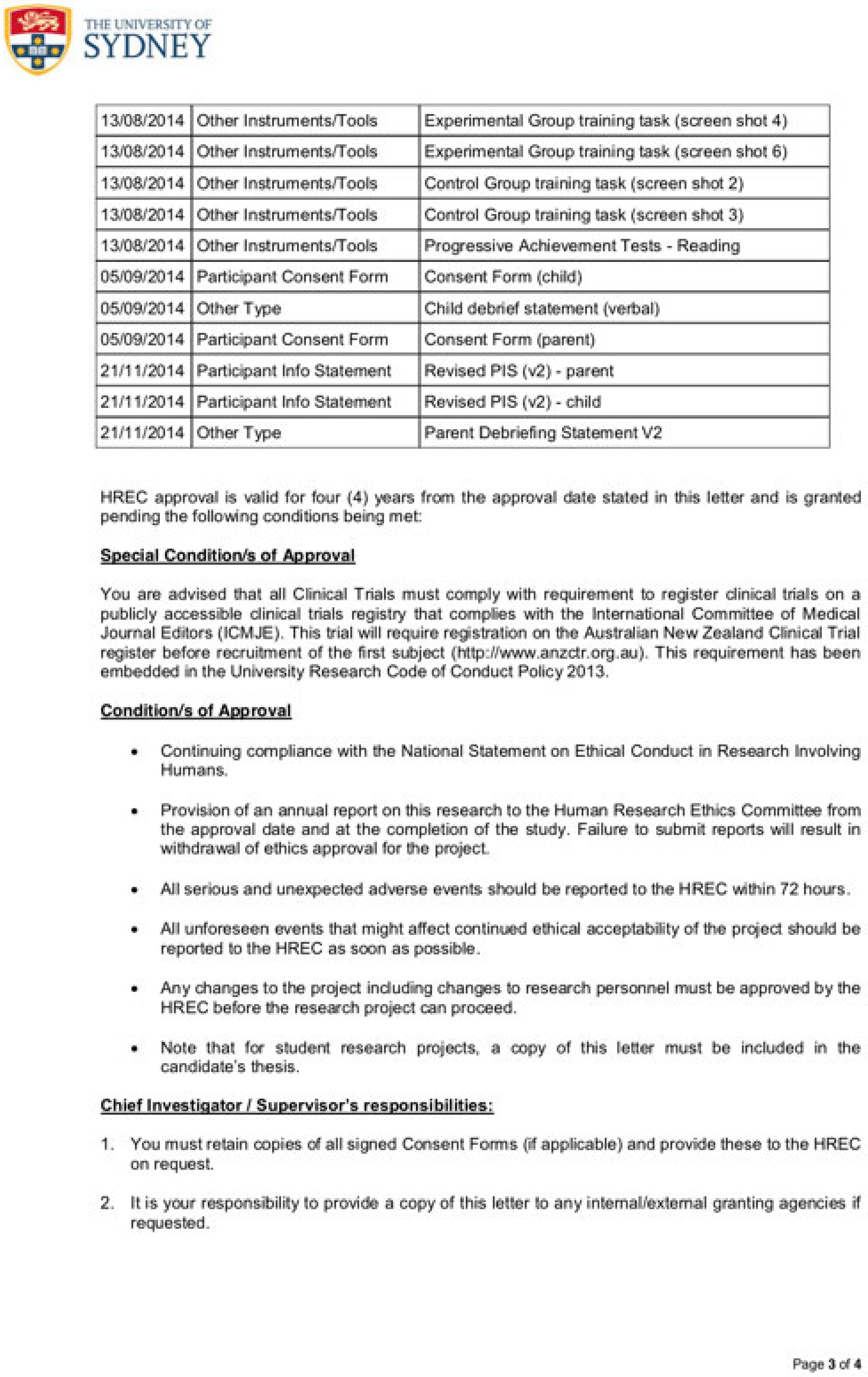
Authorised Personnel: **Ginns Paul; Colmar Susan; Hegarty David;**

Documents Approved:

Date Uploaded	Type	Document
13/08/2014	Other Instruments/Tools	Switcher test (screen shot 1)
13/08/2014	Other Instruments/Tools	Switcher test (screen shot 4)
13/08/2014	Other Instruments/Tools	Symmetry Span test (screen shot3)
13/08/2014	Other Instruments/Tools	Symmetry Span test (screen shot 4)
13/08/2014	Other Instruments/Tools	Experimental Group training task (screen shot 1)
13/08/2014	Other Instruments/Tools	Experimental Group training task (screen shot 2)
13/08/2014	Other Instruments/Tools	Control Group training task (screen shot 1)
13/08/2014	Other Instruments/Tools	Reading Span test (screen shot 3)
13/08/2014	Other Instruments/Tools	Reading Span test (screen shot 4)
13/08/2014	Other Instruments/Tools	Switcher test (screen shot 2)
13/08/2014	Other Instruments/Tools	Switcher test (screen shot 3)
13/08/2014	Other Instruments/Tools	Switcher test (screen shot 5)
13/08/2014	Other Instruments/Tools	Symmetry Span test (screen shot 2)
13/08/2014	Other Instruments/Tools	Flanker test (screen shot 3)
13/08/2014	Other Instruments/Tools	Go No-Go test (screen shot 1)
13/08/2014	Other Instruments/Tools	Go No-Go test (screen shot 2)
13/08/2014	Other Instruments/Tools	Go No-Go test (screen shot 3)



13/08/2014	Other Instruments/Tools	Go No-Go test (screen shot 4)
13/08/2014	Other Instruments/Tools	Go No-Go test (screen shot 5)
13/08/2014	Other Instruments/Tools	Numerical Stroop test (screen shot 1)
13/08/2014	Other Instruments/Tools	Numerical Stroop test (screen shot 4)
13/08/2014	Other Instruments/Tools	Numerical Stroop test (screen shot 6)
13/08/2014	Other Instruments/Tools	Operation Span test (screen shot 1)
13/08/2014	Other Instruments/Tools	Operation Span test (screen shot 2)
13/08/2014	Other Instruments/Tools	Operation Span test (screen shot 3)
13/08/2014	Other Instruments/Tools	Simple reaction time test (screen shot 1)
13/08/2014	Other Instruments/Tools	Naglieri Nonverbal Achievement Test - 2nd version
13/08/2014	Questionnaires/Surveys	Comprehensive Executive Functioning Inventory
13/08/2014	Other Instruments/Tools	Progressive Achievement Tests - Maths
13/08/2014	Other Instruments/Tools	Connections test (screen shot 2)
13/08/2014	Other Instruments/Tools	Connections test (screen shot 3)
13/08/2014	Other Instruments/Tools	Connections test (screen shot 5)
13/08/2014	Other Instruments/Tools	Flanker test (screen shot 1)
13/08/2014	Other Instruments/Tools	Flanker test (screen shot 4)
13/08/2014	Other Instruments/Tools	Reading Span test (screen shot 1)
13/08/2014	Other Instruments/Tools	Reading Span test (screen shot 2)
13/08/2014	Other Instruments/Tools	Simple reaction time test (screen shot 2)
13/08/2014	Other Instruments/Tools	Symmetry Span test (screen shot 1)
13/08/2014	Other Instruments/Tools	Experimental Group training task (screen shot 5)
13/08/2014	Questionnaires/Surveys	Implicit theory of intelligence measure (survey)
13/08/2014	Questionnaires/Surveys	Improvement expectancy measure (survey)
13/08/2014	Other Instruments/Tools	Connections test (screen shot 1)
13/08/2014	Other Instruments/Tools	Card Sorting Test (screen shot 1)
13/08/2014	Other Instruments/Tools	Card sorting test (screen shot 2)
13/08/2014	Other Instruments/Tools	Card sorting test (screen shot 3)
13/08/2014	Other Instruments/Tools	Connections test (screen shot 4)
13/08/2014	Other Instruments/Tools	Connections test (screen shot 6)
13/08/2014	Other Instruments/Tools	Connections test (screen shot 7)
13/08/2014	Other Instruments/Tools	Flanker test (screen shot 2)
13/08/2014	Other Instruments/Tools	Flanker test (screen shot 5)
13/08/2014	Other Instruments/Tools	Numerical Stroop test (screen shot 2)
13/08/2014	Other Instruments/Tools	Numerical Stroop test (screen shot 3)
13/08/2014	Other Instruments/Tools	Numerical Stroop test (screen shot 5)
13/08/2014	Other Instruments/Tools	Experimental Group training task (screen shot 3)



13/08/2014	Other Instruments/Tools	Experimental Group training task (screen shot 4)
13/08/2014	Other Instruments/Tools	Experimental Group training task (screen shot 6)
13/08/2014	Other Instruments/Tools	Control Group training task (screen shot 2)
13/08/2014	Other Instruments/Tools	Control Group training task (screen shot 3)
13/08/2014	Other Instruments/Tools	Progressive Achievement Tests - Reading
05/09/2014	Participant Consent Form	Consent Form (child)
05/09/2014	Other Type	Child debrief statement (verbal)
05/09/2014	Participant Consent Form	Consent Form (parent)
21/11/2014	Participant Info Statement	Revised PIS (v2) - parent
21/11/2014	Participant Info Statement	Revised PIS (v2) - child
21/11/2014	Other Type	Parent Debriefing Statement V2

HREC approval is valid for four (4) years from the approval date stated in this letter and is granted pending the following conditions being met:

Special Condition/s of Approval

You are advised that all Clinical Trials must comply with requirement to register clinical trials on a publicly accessible clinical trials registry that complies with the International Committee of Medical Journal Editors (ICMJE). This trial will require registration on the Australian New Zealand Clinical Trial register before recruitment of the first subject (http://www.anzctr.org.au). This requirement has been embedded in the University Research Code of Conduct Policy 2013.

Condition/s of Approval

- Continuing compliance with the National Statement on Ethical Conduct in Research Involving Humans.
- Provision of an annual report on this research to the Human Research Ethics Committee from the approval date and at the completion of the study. Failure to submit reports will result in withdrawal of ethics approval for the project.
- All serious and unexpected adverse events should be reported to the HREC within 72 hours.
- All unforeseen events that might affect continued ethical acceptability of the project should be reported to the HREC as soon as possible.
- Any changes to the project including changes to research personnel must be approved by the HREC before the research project can proceed.
- Note that for student research projects, a copy of this letter must be included in the candidate's thesis.

Chief Investigator / Supervisor's responsibilities:

1. You must retain copies of all signed Consent Forms (if applicable) and provide these to the HREC on request.
2. It is your responsibility to provide a copy of this letter to any internal/external granting agencies if requested.

THE UNIVERSITY OF
SYDNEY

Please do not hesitate to contact Research Integrity (Human Ethics) should you require further information or clarification.

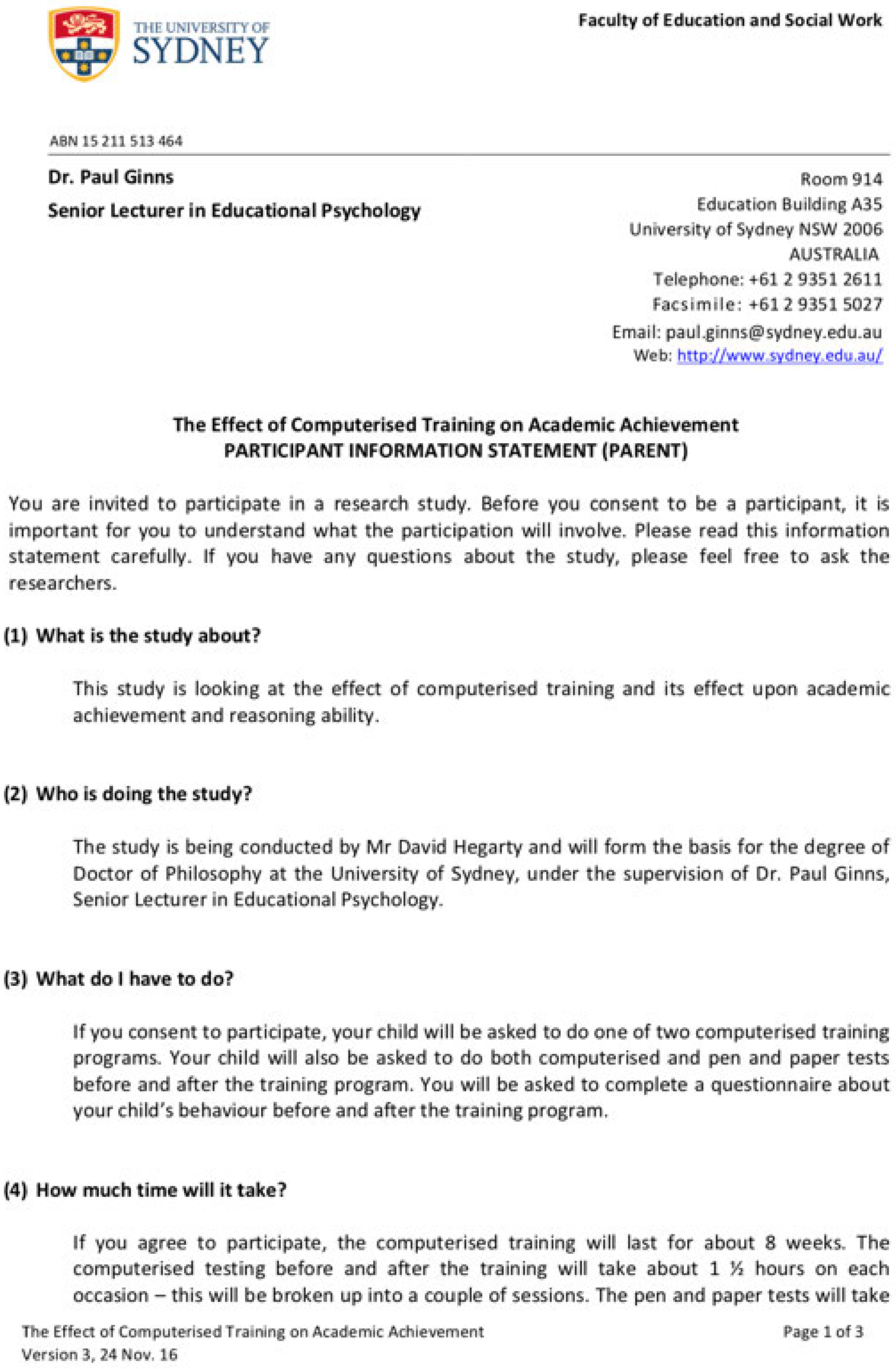
Yours sincerely

[redacted]

Dr Stephen Assinder
Chair
Human Research Ethics Committee

This HREC is constituted and operates in accordance with the National Health and Medical Research Council's (NHMRC) National Statement on Ethical Conduct in Human Research (2007), NHMRC and Universities Australia Australian Code for the Responsible Conduct of Research (2007) and the CPMP/ICH Note for Guidance on Good Clinical Practice.

9.2 Appendix B: Participant Information Sheet



THE UNIVERSITY OF SYDNEY

Faculty of Education and Social Work

ABN 15 211 513 464

Dr. Paul Ginns
Senior Lecturer in Educational Psychology

Room 914
Education Building A35
University of Sydney NSW 2006
AUSTRALIA
Telephone: +61 2 9351 2611
Facsimile: +61 2 9351 5027
Email: paul.ginns@sydney.edu.au
Web: http://www.sydney.edu.au/

The Effect of Computerised Training on Academic Achievement
PARTICIPANT INFORMATION STATEMENT (PARENT)

You are invited to participate in a research study. Before you consent to be a participant, it is important for you to understand what the participation will involve. Please read this information statement carefully. If you have any questions about the study, please feel free to ask the researchers.

(1) What is the study about?

This study is looking at the effect of computerised training and its effect upon academic achievement and reasoning ability.

(2) Who is doing the study?

The study is being conducted by Mr David Hegarty and will form the basis for the degree of Doctor of Philosophy at the University of Sydney, under the supervision of Dr. Paul Ginns, Senior Lecturer in Educational Psychology.

(3) What do I have to do?

If you consent to participate, your child will be asked to do one of two computerised training programs. Your child will also be asked to do both computerised and pen and paper tests before and after the training program. You will be asked to complete a questionnaire about your child's behaviour before and after the training program.

(4) How much time will it take?

If you agree to participate, the computerised training will last for about 8 weeks. The computerised testing before and after the training will take about 1 ½ hours on each occasion – this will be broken up into a couple of sessions. The pen and paper tests will take about 1 ½ hours on each occasion – these will also be broken up into a couple of sessions. One year after completing the first training session, this will all be repeated once again (training and testing). The questionnaire that you are asked to fill in should take about 20 minutes to complete. You will be asked to fill in this questionnaire on four separate occasions over the period of just over a year.

The computerised training sessions will take place at approximately 8:00am in the school's computer lab. It is expected that your child *might* miss 5 -10 minutes of their class administration time at the beginning of the day. Teachers will be informed if your child is involved in the training program so that they don't miss out on anything.

(5) Do I have to do the study?

Being in this study is completely voluntary. It is your choice to take part or not to take part in the study. If you do decide to take part, you can still choose to pull out if you wish at any time without any consequences. If you choose not to participate or if you participate but then decide to withdraw, this will not affect your relationship with either the University of Sydney or the school. If you choose not to participate, your son will simply go about their normal school day.

(6) Will anyone else know?

The school's Learning Support department will get access to the results from two academic achievement tests that are used as part of the research (a Reading comprehension and Mathematics test) for the purposes of supporting your child. If your child does not participate in the research, the school will administer the academic achievement results anyway as part of their normal process.

All other aspects of the study, including the study results, will be strictly confidential. All tests are completed anonymously and only the researchers will know your and your son's answers. A report of the study may be submitted for publication or be presented in conferences, but you or your son will not be named in the report.

(7) Do I get anything for being part of the study?

Your son will get the opportunity to participate in a computerised training program. After the study is complete, you will receive the findings via email if you have provided your email address on the consent form. The findings of the research study will be whether or not the computerised training program has helped the group improve their academic achievement and reasoning abilities.

(8) Can I tell other people about the study?

Yes, you can tell people that you are one of the participants in this study. However, before you receive the findings from the researchers, the researchers ask you kindly not to discuss the details of the study with others.

(9) What if I have any questions?

If you have any questions, you are very welcome to contact David Hegarty or Dr. Paul Ginns. They will be happy to answer your questions.

David Hegarty
Email: dheg5471@uni.sydney.edu.au

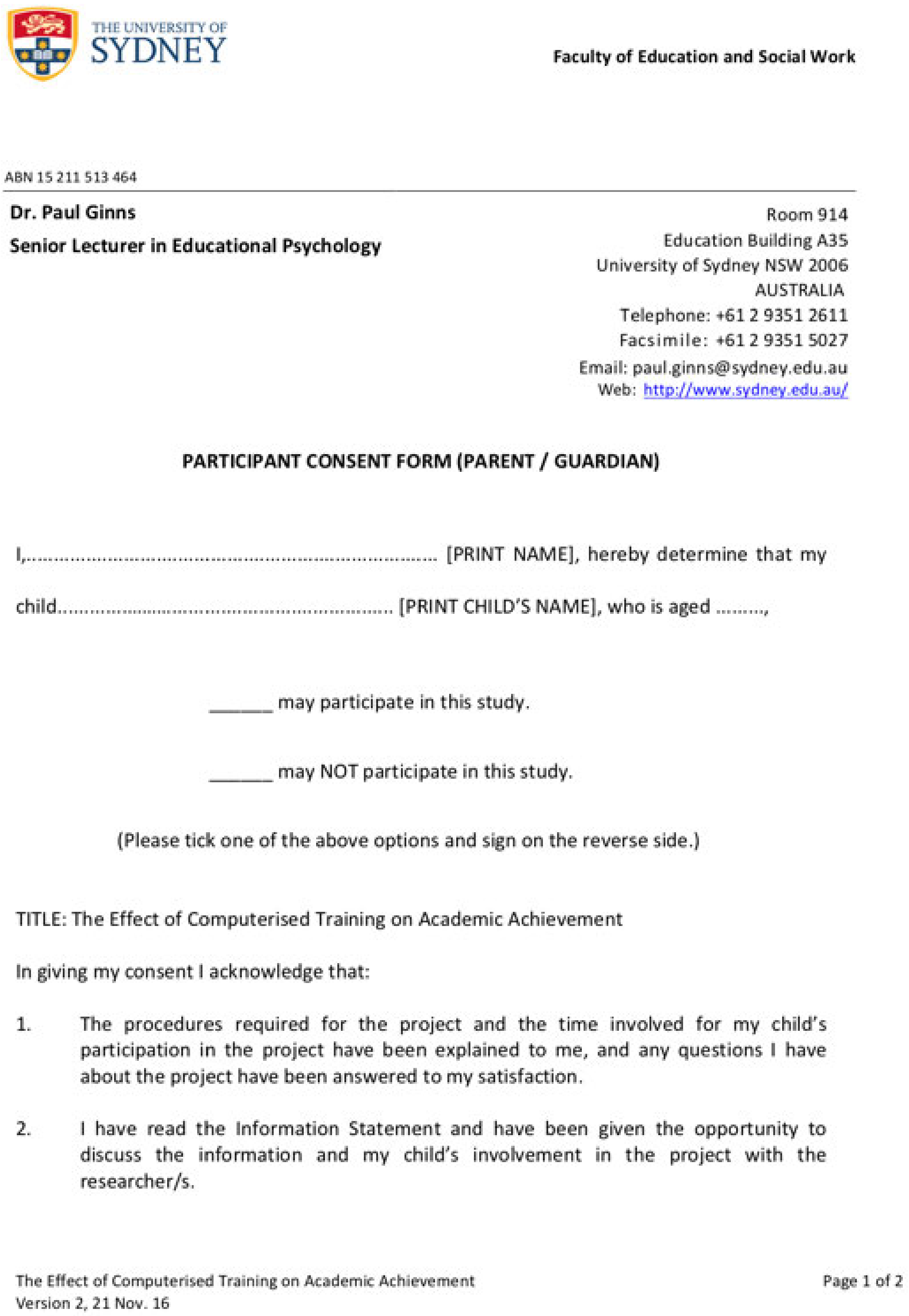
Dr. Paul Ginns
Email: paul.ginns@sydney.edu.au

(10) What if I am not happy with the study?

If you have any concerns or complaints you can contact The University of Sydney on +61 2 8627 8176 (Telephone); +61 2 8627 8177 (Facsimile) or ro.humanethics@sydney.edu.au (Email).

This information sheet is for you to keep

9.3 Appendix C: Participant Consent Form



THE UNIVERSITY OF SYDNEY

Faculty of Education and Social Work

ABN 15 211 513 464

Dr. Paul Ginns
Senior Lecturer in Educational Psychology

Room 914
Education Building A35
University of Sydney NSW 2006
AUSTRALIA
Telephone: +61 2 9351 2611
Facsimile: +61 2 9351 5027
Email: paul.ginns@sydney.edu.au
Web: http://www.sydney.edu.au/

PARTICIPANT CONSENT FORM (PARENT / GUARDIAN)

I,... [PRINT NAME], hereby determine that my

child.. [PRINT CHILD'S NAME], who is aged ....,

______ may participate in this study.

______ may NOT participate in this study.

(Please tick one of the above options and sign on the reverse side.)

TITLE: The Effect of Computerised Training on Academic Achievement

In giving my consent I acknowledge that:

1. The procedures required for the project and the time involved for my child's participation in the project have been explained to me, and any questions I have about the project have been answered to my satisfaction.

2. I have read the Information Statement and have been given the opportunity to discuss the information and my child's involvement in the project with the researcher/s.

The Effect of Computerised Training on Academic Achievement
Version 2, 21 Nov. 16
Page 1 of 2

3. I understand that being in this study is completely voluntary – I am not under any obligation to consent to my child's participation.

4. I understand that the school's Learning Support department will get access to the results from the academic achievement tests and they may use this information to support my child.

5. I understand that my child's involvement is strictly confidential. I understand that research data gathered from the results of the study may be published however no information about my child nor I will be used in any way that is identifiable.

6. I understand that I can withdraw my child from the study at any time without prejudice to my or my child's relationship with the researcher/s, St. Aloysius' College, or the University of Sydney now or in the future.

7. I understand that I can stop my participation in the study at any time if I do not wish to continue and the information provided will not be included in the study.

8. I consent to:
   - Receiving Feedback YES ☐ NO ☐

   If you answered YES to the "Receiving Feedback" question, please provide your details i.e., email address.

   Feedback Option

   Email: ____

....
Signature of Parent/Caregiver

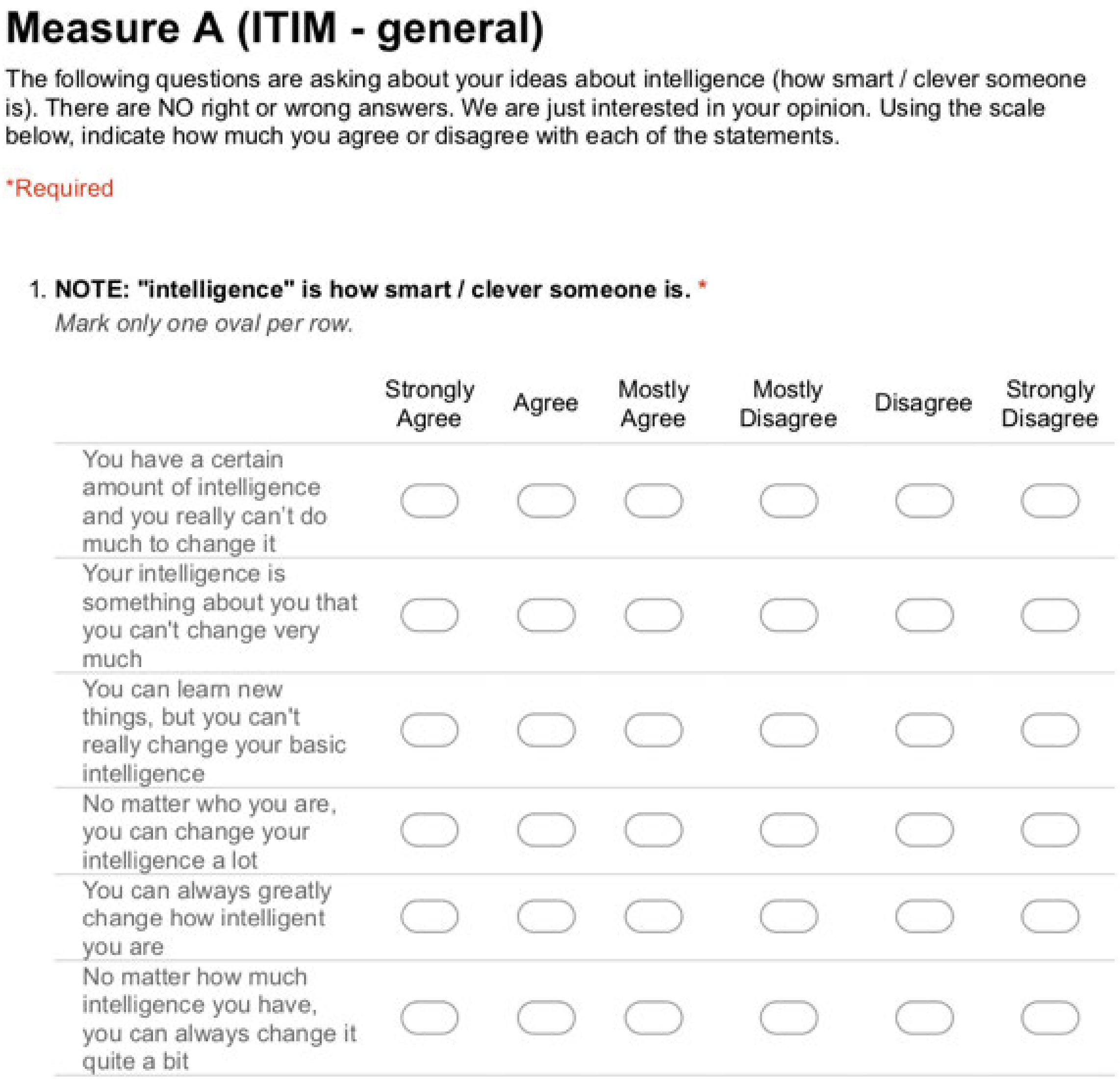
....
Please PRINT name

....
Date

9.4 Appendix D: Copies of Questionnaires

- Implicit Theory of Intelligence Measure (ITIM) as outlined by Blackwell, Trzesniewki, and Dweck (2007) and Dweck (2000)
- Improvement Expectancy Measure (IEM)

9.4.1 Implicit Theory of Intelligence Measure (ITIM)



Measure A (ITIM - general)

The following questions are asking about your ideas about intelligence (how smart / clever someone is). There are NO right or wrong answers. We are just interested in your opinion. Using the scale below, indicate how much you agree or disagree with each of the statements.

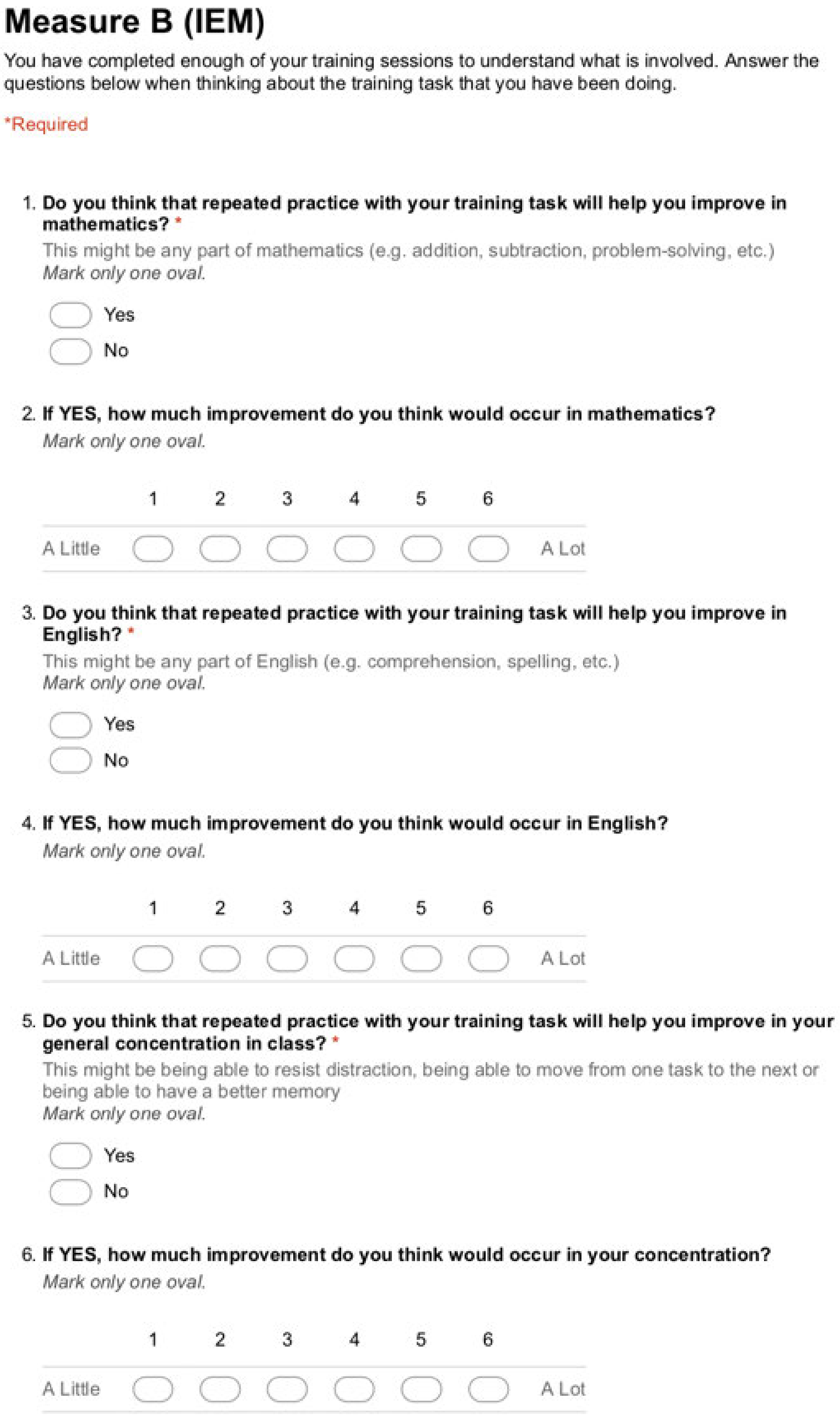
*Required

1. **NOTE: "intelligence" is how smart / clever someone is.** *

Mark only one oval per row.

	Strongly Agree	Agree	Mostly Agree	Mostly Disagree	Disagree	Strongly Disagree
You have a certain amount of intelligence and you really can't do much to change it	○	○	○	○	○	○
Your intelligence is something about you that you can't change very much	○	○	○	○	○	○
You can learn new things, but you can't really change your basic intelligence	○	○	○	○	○	○
No matter who you are, you can change your intelligence a lot	○	○	○	○	○	○
You can always greatly change how intelligent you are	○	○	○	○	○	○
No matter how much intelligence you have, you can always change it quite a bit	○	○	○	○	○	○

9.4.2 Improvement Expectancy Measure (IEM)



Measure B (IEM)

You have completed enough of your training sessions to understand what is involved. Answer the questions below when thinking about the training task that you have been doing.

*Required

1. **Do you think that repeated practice with your training task will help you improve in mathematics? ***
   This might be any part of mathematics (e.g. addition, subtraction, problem-solving, etc.)
   Mark only one oval.
   ○ Yes
   ○ No

2. **If YES, how much improvement do you think would occur in mathematics?**
   Mark only one oval.

	1	2	3	4	5	6	
A Little	○	○	○	○	○	○	A Lot

3. **Do you think that repeated practice with your training task will help you improve in English? ***
   This might be any part of English (e.g. comprehension, spelling, etc.)
   Mark only one oval.
   ○ Yes
   ○ No

4. **If YES, how much improvement do you think would occur in English?**
   Mark only one oval.

	1	2	3	4	5	6	
A Little	○	○	○	○	○	○	A Lot

5. **Do you think that repeated practice with your training task will help you improve in your general concentration in class? ***
   This might be being able to resist distraction, being able to move from one task to the next or being able to have a better memory
   Mark only one oval.
   ○ Yes
   ○ No

6. **If YES, how much improvement do you think would occur in your concentration?**
   Mark only one oval.

	1	2	3	4	5	6	
A Little	○	○	○	○	○	○	A Lot

9.5 Appendix E: Data and Data Analysis Script

Data is available at:

https://raw.githubusercontent.com/dlhegarty/PhD_Data/master/data.csv

Data analysis script is available at:

https://raw.githubusercontent.com/dlhegarty/PhD_Data_Script/master/PhD%20Data%20Analysis.R

Note. Data analysis script automatically loads data from GitHub repository.

9.6 Appendix F: Complex Span Data Analysis

In Section 9.6.2 below, an analysis is completed using results from individual CSTs to determine if there is any difference in the results between groups when compared to results provided by factor scores as a result of the potentially biased imputation process using lavaan (e.g., that the control group predominantly scores higher when T1 scores are controlled for). It is recognised that there is still removal of data listwise in this process and this does introduce potential biases in data analysis (Pepinsky, 2018). However, Section 9.6.3 provides an analysis where no data is removed at all. Although, as outlined in Sections 6.3 and 7.9.6, not removing data potentially changes what the CSTs are actually measuring, it is a useful analysis to perform so that results can be compared to the methodology used for calculation of factor scores and when data was removed.

9.6.1 Differences between groups.

Prior to analysis of individual CST data, a comparison was performed between those who had valid CST scores and those who did not. Participants were classified into two groups: those who achieved the required accuracy for the distractor task and therefore had a valid score for each CST and those who did not achieve the required accuracy and their data was removed. When comparing these two groups, there was a significant difference between the age of the participants in each group, Welch's $t(102.32) = 4.30$, $p < .001$, $d = 0.81$, $\text{CI}_{95\%}[0.40, 1.22]$, with those who had a valid CST score older than those who did not have a valid score. There was also a significant difference between the fluid intelligence scores in each group, Welch's $t(91.61) = 2.75$, $p = .007$, $d = 0.55$, $\text{CI}_{95\%}[0.15, 0.95]$, with the CST valid group scoring higher on the NAI than the CST problem group.

9.6.2 Missing data removed listwise

Operation Span. Figure F1 shows the OSPAN partial scores across each of the three measurement periods for both intervention groups. For data analysis, parametric ANCOVAs were attempted in R (R Core Team, 2019) and all models satisfied the assumption of homogeneity of slopes and all standardised residuals appeared to be distributed normally as demonstrated by Q-Q plots. However, T2 data for the OSPAN score ($F(1,87) = 9.91$, $p = .002$) violated the assumptions of homogeneity of variance. Consequently, robust ANCOVAs using WRS2 (Mair & Wilcox, 2018) were used to determine if there was a significant difference between intervention groups using T2 OSPAN partial scores while controlling for T1 partial scores.

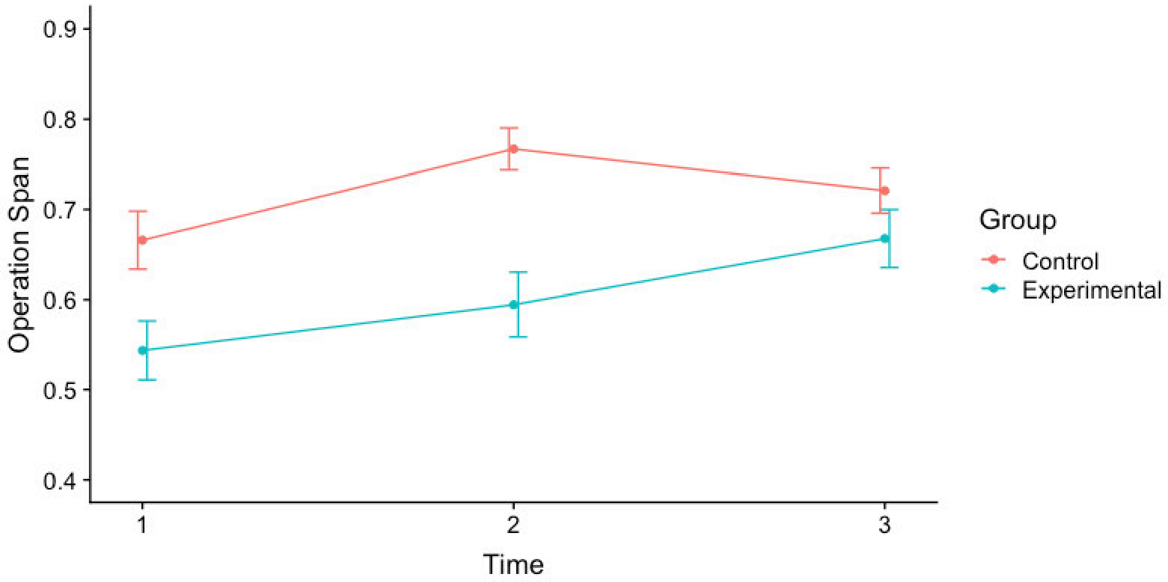


Figure F1. Operation span (OSPAN) partial score plots (missing data removed listwise). Error bars represent standard errors.

Using robust ANCOVA there was no significant effect of the intervention group on T2 OSPAN partial scores after controlling for T1 OSPAN partial scores, as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure F2 shows a scatterplot for the OSPAN partial scores for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points.

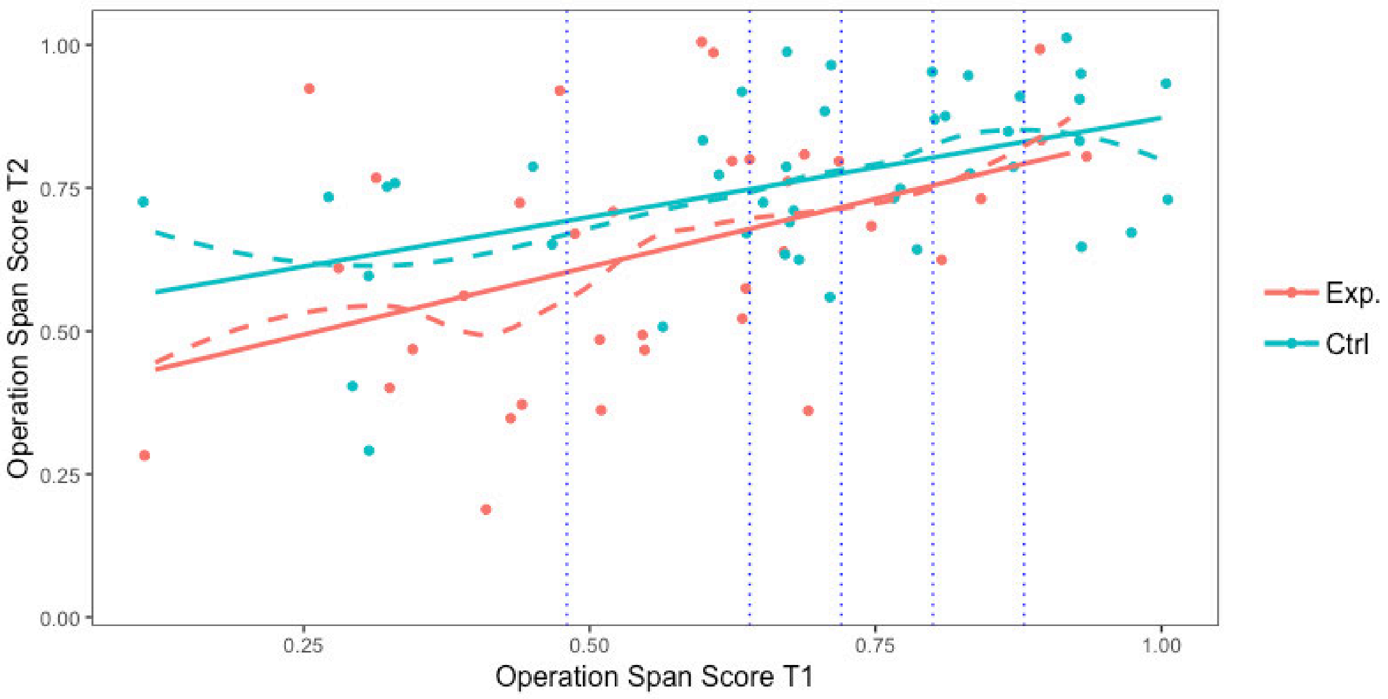


Figure F2. Operation span (OSPAN) partial scores scatterplot (missing data removed listwise) demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

T3 OSPAN partial scores satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 OSPAN partial scores after controlling for the T1 partial scores on OSPAN, $F(1, 78) = 0.001$, $p = .974$, ($\eta_p^2 = 0.00$, $CI_{90\%}[0.00, 1.00]$) with the covariate, T1 partial scores, significantly related to the T3 partial scores, $F(1, 78) = 5.85$, $p = .018$, ($\eta_p^2 = 0.08$, $CI_{90\%}[0.01, 0.18]$).

Reading span. A plot showing RSPAN partial scores for T1-T3 is shown in Figure F3. T2 data satisfied all assumptions for running a parametric ANCOVA and there was a significant effect of group on T2 RSPAN partial scores after controlling for the T1 partial scores, $F(1, 77) = -9.66$, $p = .002$, ($\eta_p^2 = 0.11$, $CI_{90\%}[0.02, 0.23]$) with the covariate, T1 partial scores, significantly related to the T2 partial scores, $F(1, 77) = 60.98$, $p < .001$, ($\eta_p^2 = 0.49$, $CI_{90\%}[0.36, 0.59]$). However, this difference is the opposite of what was expected: that is, the control group scored higher at T2 after controlling for T1 partial scores (difference = -0.10, $t = -3.11$, $p = .003$, $d = -0.70$, $CI_{95\%}[-1.15, -0.24]$).

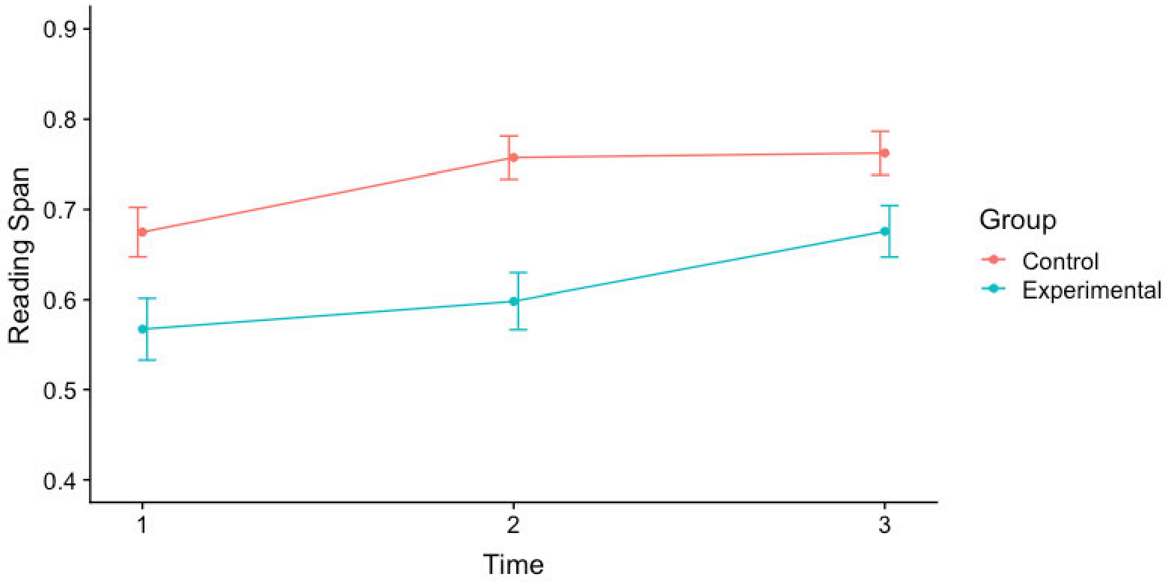


Figure F3. Reading span (RSPAN) partial score plots (missing data removed listwise). Error bars represent standard errors.

T3 data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 RSPAN partial scores after controlling for the T1 partial scores, $F(1, 79) = 0.79$, $p = .376$, ($\eta_p^2 = 0.01$, $CI_{90\%}[0.00, 0.07]$) with the covariate, T1 partial scores, significantly related to the T3 partial scores, $F(1, 79) = 28.57$, $p < .001$, ($\eta_p^2 = 0.30$, $CI_{90\%}[0.16, 0.41]$).

Symmetry span. Figure F4 shows the T1-T3 data for the SSPAN task. T2 data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 SSPAN partial scores after controlling for the T1 partial scores, F(1, 81) = -3.33, *p* = .072, (η_p^2= 0.04, $CI_{90\%}$[0.00, 0.13]) with the covariate, T1 partial scores, significantly related to the T3 partial scores, *F*(1, 81) = 17.22, *p* < .001, (η_p^2= 0.21, $CI_{90\%}$[0.09, 0.33]).

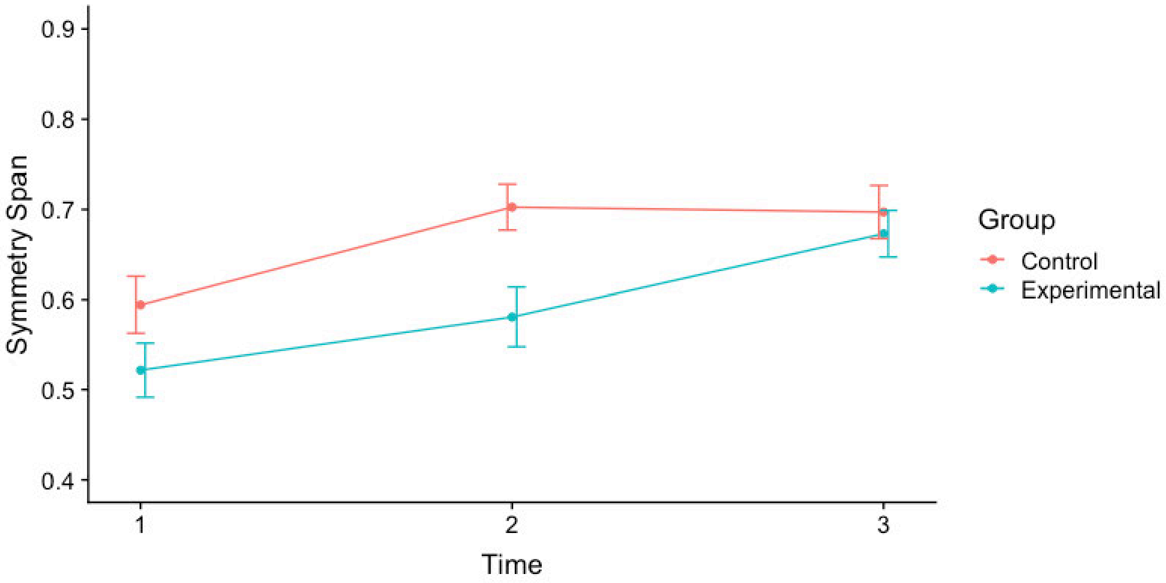


Figure F4. Symmetry span (SSPAN) partial score plots (missing data removed listwise). Error bars represent standard errors.

The SSPAN T3 partial score did not meet assumptions for running a parametric ANCOVA due to violating the assumption of homogeneity of slopes (*F*(1, 76) = 4.94, *p* = .029). Using robust ANCOVA there was no significant effect of the intervention group on T3 SSPAN partial scores after controlling for T1 SSPAN partial scores, as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure F5 shows a scatterplot for the SSPAN partial scores for T1 vs T3 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points.

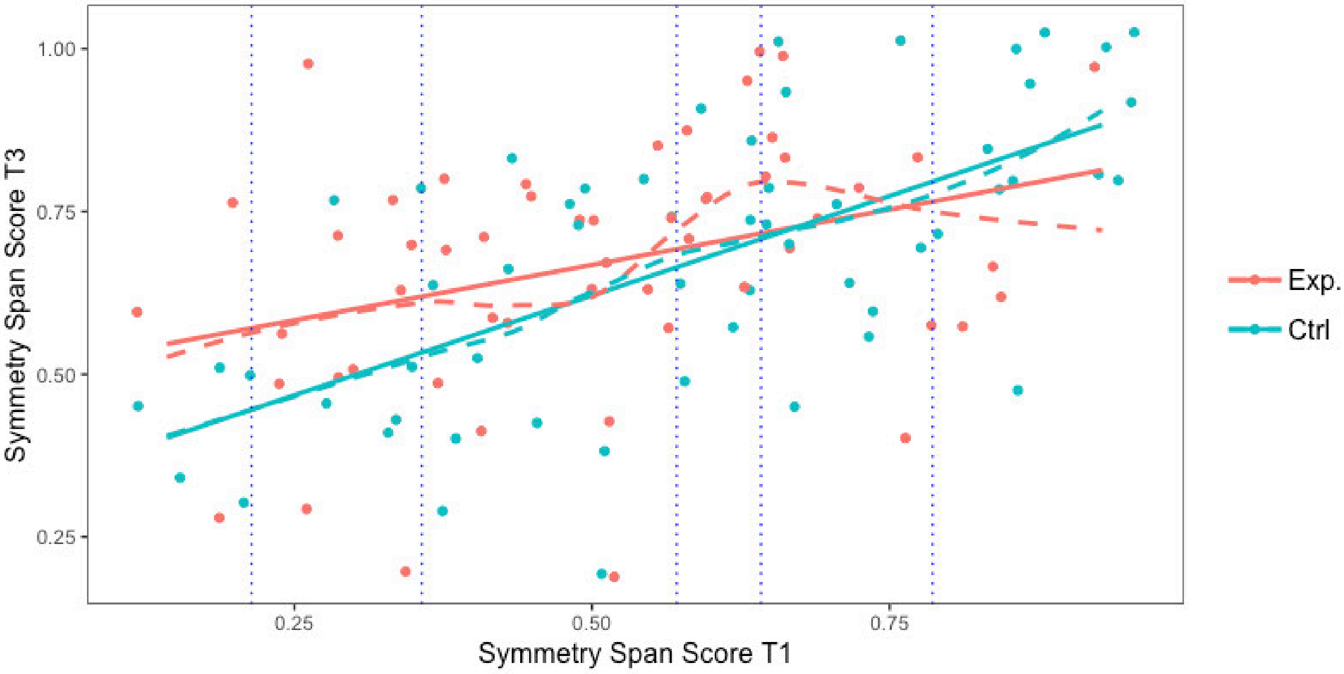


Figure F5. Symmetry span (SSPAN) partial scores scatterplot (missing data removed listwise) demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

9.6.3 Data not removed

To determine if there was any difference in results if the data were not removed due to distractor accuracy, analysis was run with complete data.

It is important to note that there was a similar pattern of differences at baseline for each of the CST measures when no data were removed, with significant differences between the control and experimental groups on T1 OSPAN partial scores, Welch's $t(103) = -2.84$, $p = .005$, $d = -0.54$, $CI_{95\%}[-0.94, -0.15]$ and T1 RSPAN partial scores, Welch's $t(93.78) = -2.43$, $p = .017$, $d = -0.46$, $CI_{95\%}[-0.86, -0.07]$, with the control group scoring higher on both measures.

Operation span. Figure F6 shows the OSPAN partial scores across each of the three measurement periods for both intervention groups. For data analysis, parametric ANCOVAs were attempted in R (R Core Team, 2019) and all models satisfied the assumption of homogeneity of slopes and all standardised residuals appeared to be distributed normally as demonstrated by Q-Q plots. However, T2 data for the OSPAN score ($F(1,103) = 5.98$, $p = .016$) violated the assumptions of homogeneity of variance. Consequently, robust ANCOVAs using WRS2 (Mair & Wilcox, 2018) were used to determine if there was a significant difference between intervention groups using T2 OSPAN partial scores while controlling for T1 partial scores.

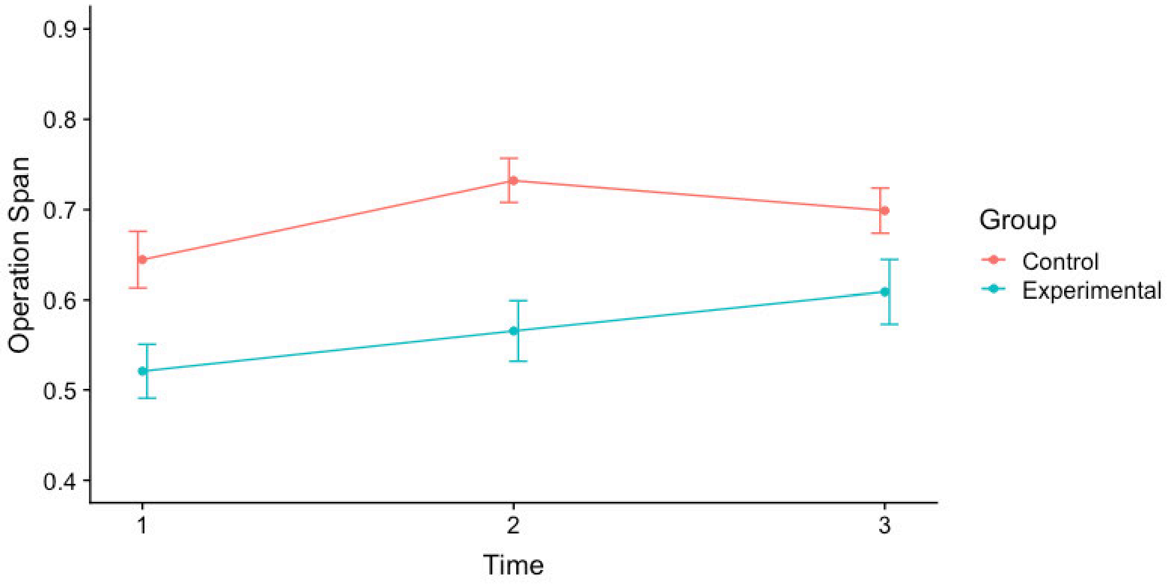


Figure F6. Operation span (OSPAN) partial score plots (no data removed). Error bars represent standard errors.

Using robust ANCOVA there was no significant effect of the intervention group on T2 OSPAN partial scores after controlling for T1 OSPAN partial scores, as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure F7 shows a scatterplot for the OSPAN partial scores for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points.

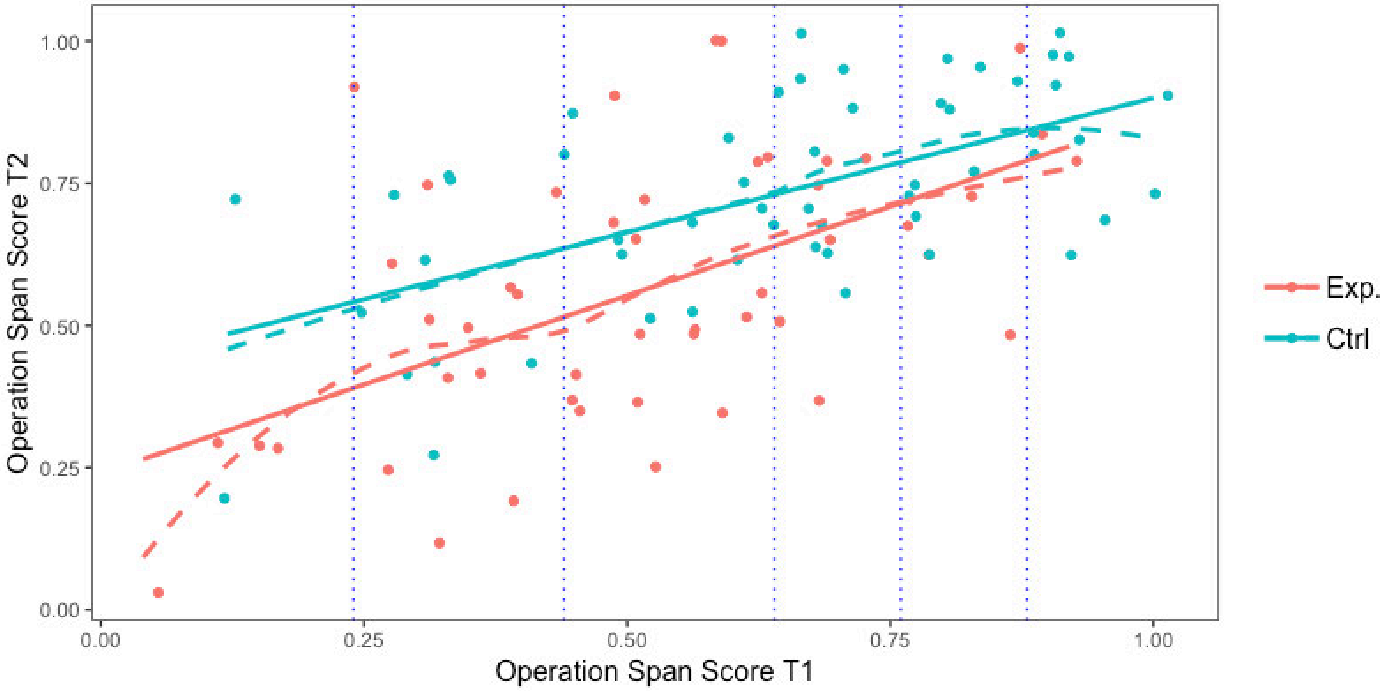


Figure F7. Operation span (OSPAN) partial scores scatterplot (no data removed) demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

T3 OSPAN partial scores satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 OSPAN partial scores after controlling for the T1 partial scores on OSPAN, $F(1, 102) = 1.46$, $p = .230$, ($\eta_p^2 = 0.14$, $CI_{90\%}[0.00, 0.07]$) with the covariate, T1 partial scores, significantly related to the T3 partial scores, $F(1, 102) = 11.27$, $p = .001$, ($\eta_p^2 = 0.13$, $CI_{90\%}[0.04, 0.22]$).

Reading span. A plot showing RSPAN partial scores for T1-T3 is shown in Figure F8. T2 data satisfied all assumptions for running a parametric ANCOVA and there was a significant effect of group on T2 RSPAN partial scores after controlling for the T1 partial scores, $F(1, 100) = -5.40$, $p = .022$, ($\eta_p^2 = 0.05$, $CI_{90\%}[0.004, 0.14]$) with the covariate, T1 partial scores, significantly related to the T2 partial scores, $F(1, 100) = 65.99$, $p < .001$, ($\eta_p^2 = 0.44$, $CI_{90\%}[0.32, 0.53]$). However, this difference is the opposite of what was expected: that is, the control group scored higher at T2 after controlling for T1 partial scores (difference = -0.07, $t = -2.33$, $p = .022$, $d = -0.45$, $CI_{95\%}[-0.84, -0.05]$).

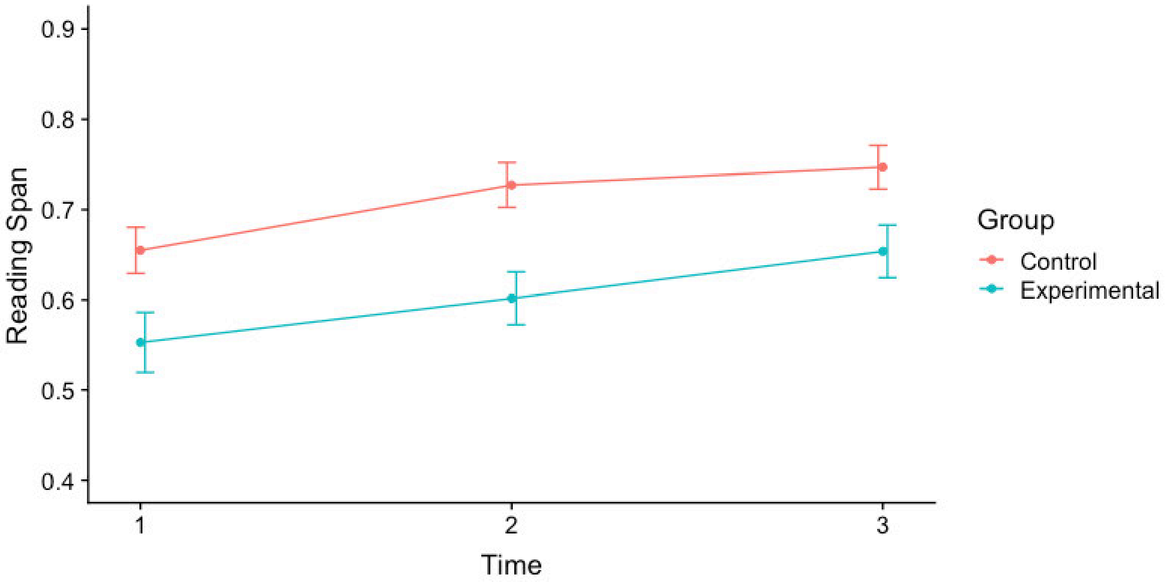


Figure F8. Reading span (RSPAN) partial score plots (no data removed). Error bars represent standard errors.

T3 data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 RSPAN partial scores after controlling for the T1 partial scores, $F(1, 101) = 1.96$, $p = .164$, ($\eta_p^2 = 0.02$, $CI_{90\%}[0.00, 0.08]$) with the covariate, T1 partial scores, significantly related to the T3 partial scores, $F(1, 101) = 35.11$, $p < .001$, ($\eta_p^2 = 0.29$, $CI_{90\%}[0.17, 0.40]$)..

Symmetry span. Figure F9 shows the T1-T3 data for the SSPAN task. T2 data satisfied all assumptions for running a parametric ANCOVA and there was a significant effect of group on T2 SSPAN partial scores after controlling for the T1 partial scores, $F(1, 102) = -5.28$, $p = .024$, ($\eta_p^2 = 0.05$, $CI_{90\%}[0.004, 0.13]$) with the covariate, T1 partial scores, significantly related to the T2 partial scores, $F(1, 102) = 22.30$, $p < .001$, ($\eta_p^2 = 0.21$, $CI_{90\%}[0.10, 0.32]$). However, this difference is the opposite of what was expected: that is, the control group scored higher at T2 after controlling for T1 partial scores (difference = -0.08, $t = -2.30$, $p = .024$, $d = -0.45$, $CI_{95\%}[-0.84, -0.06]$).

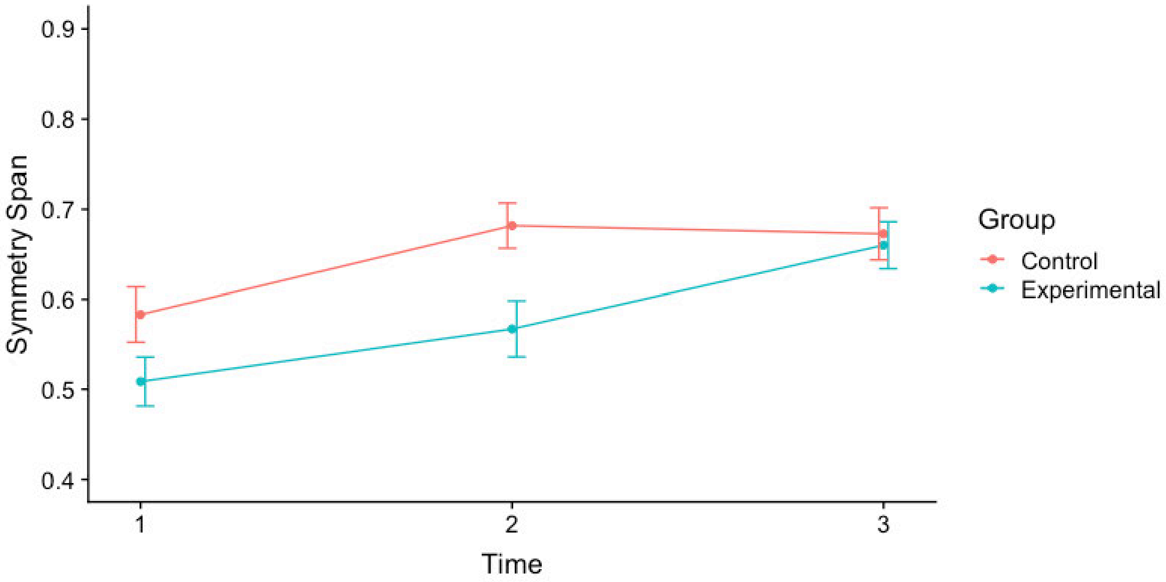


Figure F9. Symmetry span (SSPAN) partial score plots (no data removed). Error bars represent standard errors.

T3 data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 SSPAN partial scores after controlling for the T1 partial scores, $F(1, 102) = 0.48$, $p = .49$, ($\eta_p^2 = 0.005$, $CI_{90\%}[0.00, 0.05]$) with the covariate, T1 partial scores, significantly related to the T3 partial scores, $F(1, 102) = 36.97$, $p < .001$, ($\eta_p^2 = 0.26$, $CI_{90\%}[0.15, 0.37]$).

9.7 Appendix G: Near-Transfer at the Individual Task Level

As there were some difficulties in getting the CFA to provide a solution with a decent fit without removing some tasks, it was decided to test near-transfer at the individual task level. Individual WM data (CSTs) have been analysed separately in Appendix F, so only shifting and inhibition tasks are reported here.

9.7.1 *Shifting Tasks*

BCST task. BCST perseverative errors data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 BCST perseverative errors after controlling for the T1 perseverative errors on BCST, *F*(1, 102) = 0.87, *p* = .353 (η_p^2= 0.01, $CI_{90\%}$[0.00, 0.06]) with the covariate, T1 perseverative errors, significantly related to the T2 perseverative errors, *F*(1, 102) = 13.70, *p* = .0004, (η_p^2= 0.12, $CI_{90\%}$[0.04, 0.22]). There was no significant effect of group on T3 BCST perseverative errors after controlling for the T1 perseverative errors on BCST, *F*(1, 102) = 0.40, *p* = .529 (η_p^2= 0.004, $CI_{90\%}$[0.00, 0.05]) with the covariate, T1 perseverative errors, significantly related to the T3 perseverative errors, *F*(1, 102) = 11.08, *p* = .0012, (η_p^2= 0.096, $CI_{90\%}$[0.02, 0.19]).

BCST CLR data did not satisfy the assumption of homogeneity of variance. Using robust ANCOVA there was no significant effect of the intervention group on T2 BCST CLR after controlling for T1 BCST CLR, as all confidence intervals for trimmed mean differences included zero (all *p*’s > .05). Figure G*1* shows a scatterplot for the BCST CLR for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points.

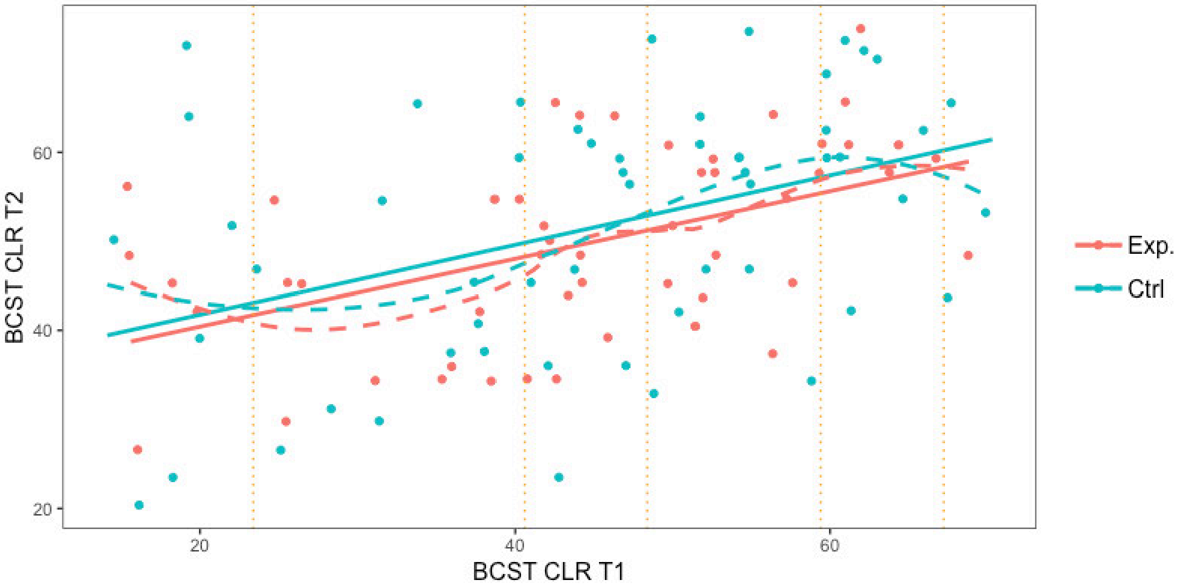


Figure G*1*. BCST CLR T1 v T2 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

Using robust ANCOVA there was no significant effect of the intervention group on T3 BCST CLR after controlling for T1 BCST CLR, as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure G2 shows a scatterplot for the BCST CLR for T1 vs T3 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points.

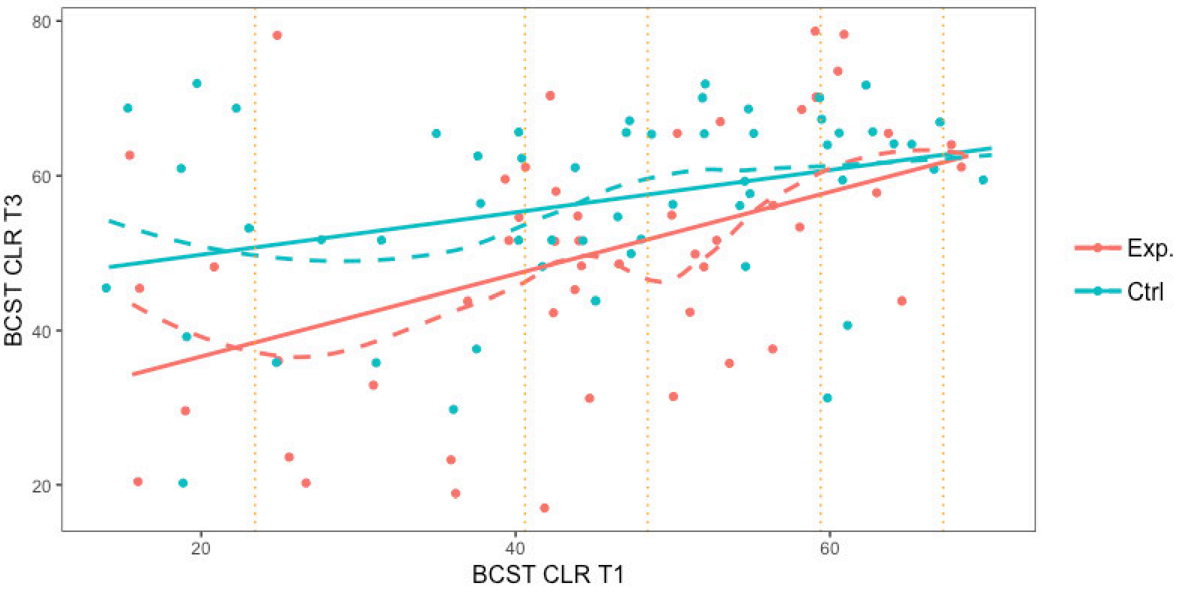


Figure G2. BCST CLR T1 v T3 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

Connections Task. Switching completion data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 switching completion after controlling for the T1 switching completion on the Connections Task, $F(1, 102) = 0.40$, $p = .529$ ($\eta_p^2 = 0.004$, $CI_{90\%}[0.00, 0.05]$) with the covariate, T1 switching completion, significantly related to the T2 switching completion, $F(1, 102) = 89.40$, $p < .001$, ($\eta_p^2 = 0.48$, $CI_{90\%}[0.36, 0.57]$). There was no significant effect of group on T3 switching completion after controlling for the T1 switching completion on the Connections Task, $F(1, 102) = 1.46$, $p = .23$ ($\eta_p^2 = 0.01$, $CI_{90\%}[0.00, 0.07]$) with the covariate, T1 switching completion, significantly related to the T3 switching completion, $F(1, 102) = 145.22$, $p < .001$, ($\eta_p^2 = 0.60$, $CI_{90\%}[0.50, 0.67]$).

Switcher Task. Switcher errors data did not satisfy the assumption of homogeneity of regression slopes. Using robust ANCOVA there was no significant effect of the intervention group on T2 errors after controlling for T1 errors as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure G3 shows a scatterplot for the switcher errors for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the four T1 design points.

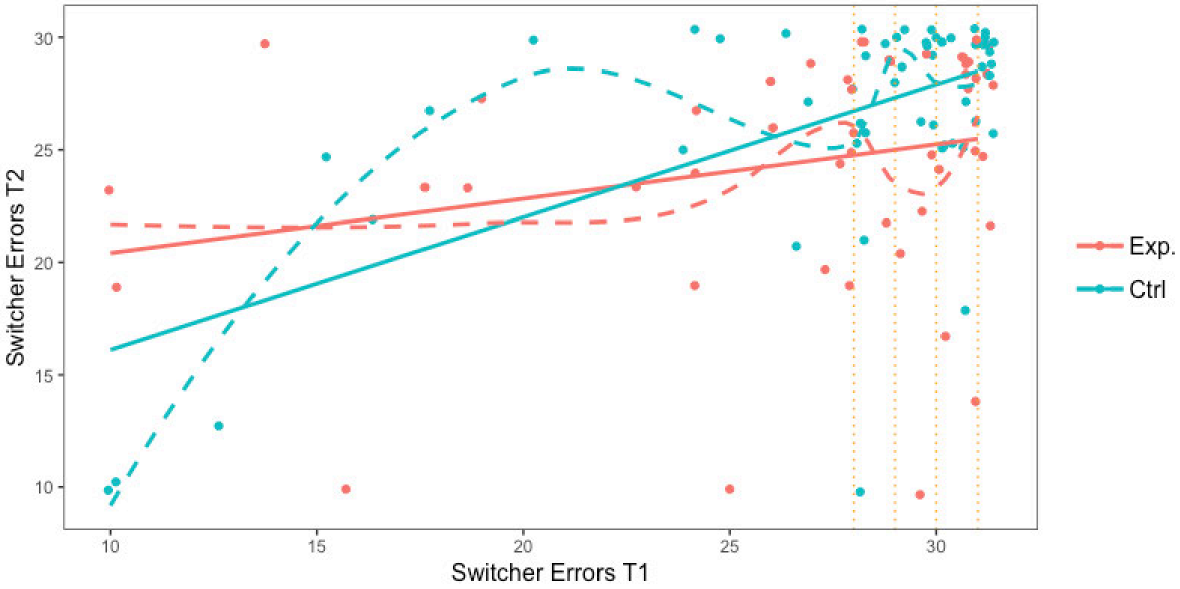


Figure G3. Switcher errors T1 v T2 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

Using robust ANCOVA there was no significant effect of the intervention group on T3 errors after controlling for T1 errors as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure G4shows a scatterplot for the error scores for T1 vs T3 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the four T1 design points.

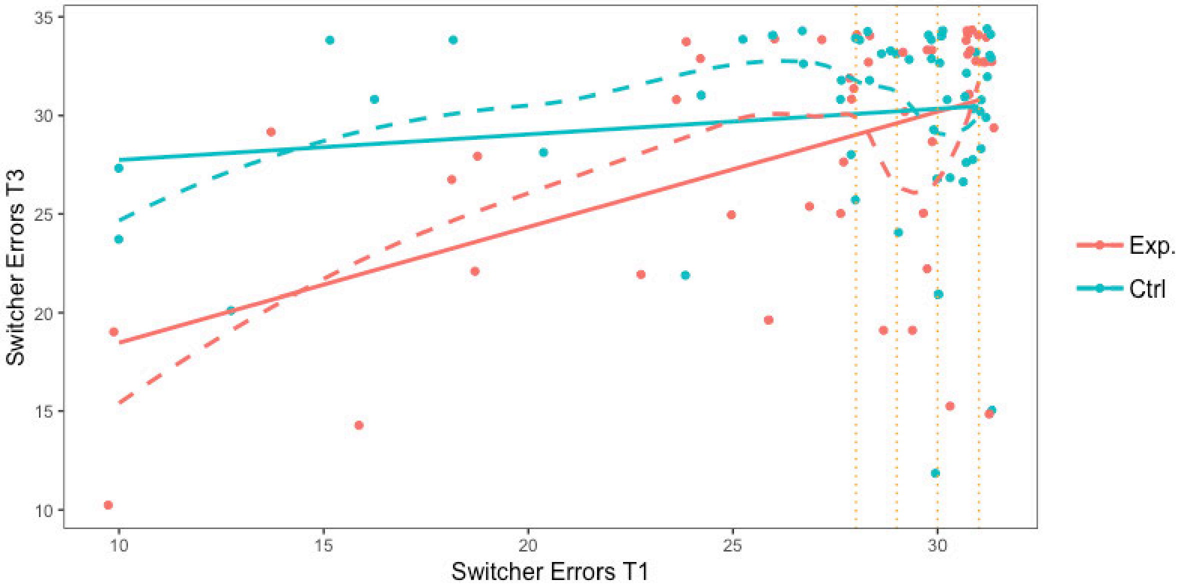


Figure G4. Switcher errors T1 v T3 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

Switcher error difference score data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 error difference scores after controlling for the T1 error difference scores on the Switcher Task, *F*(1, 101) = 0.82, *p* = .367 (η_p^2= 0.01, CI$_{90\%}$[0.00, 0.06]) with no significant relationship between the T2 error difference scores and the covariate, T1 error difference scores, *F*(1, 101) = 1.65 *p* = .202, (η_p^2= 0.02, CI$_{90\%}$[0.00, 0.08]). There was no significant effect of group on T3 error difference scores after controlling for the T1 error difference scores on the Switcher Task, *F*(1, 101) = 0.15, *p* = .702 (η_p^2= 0.00, CI$_{90\%}$[0.00, 0.04]) with no significant relationship between the T3 error difference scores and the covariate, T1 error difference scores, *F*(1, 101) = 0.04 *p* = .846, (η_p^2= 0.00, CI$_{90\%}$[0.00, 0.02]).

9.7.2 Inhibition Tasks

Flanker Task. Conflict cost data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 conflict cost after controlling for the T1 conflict cost on the Flanker Task, $F(1, 98) = 3.34$, $p = .071$ ($\eta_p^2 = 0.03$, CI$_{90\%}$[0.00, 0.11]) with no significant relationship between the T2 conflict cost and the covariate, T1 conflict cost, $F(1, 98) = 1.78$ $p = .185$, ($\eta_p^2 = 0.02$, CI$_{90\%}$[0.00, 0.08]). There was no significant effect of group on T3 conflict cost after controlling for the T1 conflict cost on the Flanker Task, $F(1, 98) = 0.93$, $p = .337$ ($\eta_p^2 = 0.01$, CI$_{90\%}$[0.00, 0.06]) with the covariate, T1 conflict cost, significantly related to the T3 conflict cost, $F(1, 98) = 5.00$, $p = .028$, ($\eta_p^2 = 0.05$, CI$_{90\%}$[0.003, 0.13]).

Flanker incongruent accuracy data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 incongruent accuracy after controlling for the T1 incongruent accuracy on the Flanker Task, $F(1, 98) = 1.16$, $p = .284$ ($\eta_p^2 = 0.01$, CI$_{90\%}$[0.00, 0.07]) with the covariate, T1 incongruent accuracy, significantly related to the T2 incongruent accuracy, $F(1, 98) = 20.26$, $p < .001$, ($\eta_p^2 = 0.17$, CI$_{90\%}$[0.07, 0.28]). There was no significant effect of group on T3 conflict cost after controlling for the T1 incongruent accuracy on the Flanker Task, $F(1, 98) = 0.04$, $p = .841$ ($\eta_p^2 = 0.00$, CI$_{90\%}$[0.00, 0.02]) with the covariate, T1 incongruent accuracy, significantly related to the T3 incongruent accuracy, $F(1, 98) = 4.77$, $p = .031$, ($\eta_p^2 = 0.05$, CI$_{90\%}$[0.002, 0.13]).

Go/No-Go Task. No-Go accuracy data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 no-go accuracy after controlling for the T1 no-go accuracy on the Go/No-Go Task, $F(1, 101) = 2.00$, $p = .161$ ($\eta_p^2 = 0.02$, CI$_{90\%}$[0.00, 0.08]) with the covariate, T1 no-go accuracy, significantly related to the T2 no-go accuracy, $F(1, 101) = 13.86$, $p < .001$, ($\eta_p^2 = 0.11$, CI$_{90\%}$[0.03, 0.21]). T3 v T1 data did not satisfy the assumption of homogeneity of regression slopes. Using robust ANCOVA there was no significant effect of the intervention group on T2 no-go accuracy after controlling for T1 no-go accuracy as all confidence intervals for trimmed mean differences included zero (all *p*'s > .05). Figure G5 shows a scatterplot for the no-go accuracy for T1 vs T3 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points.

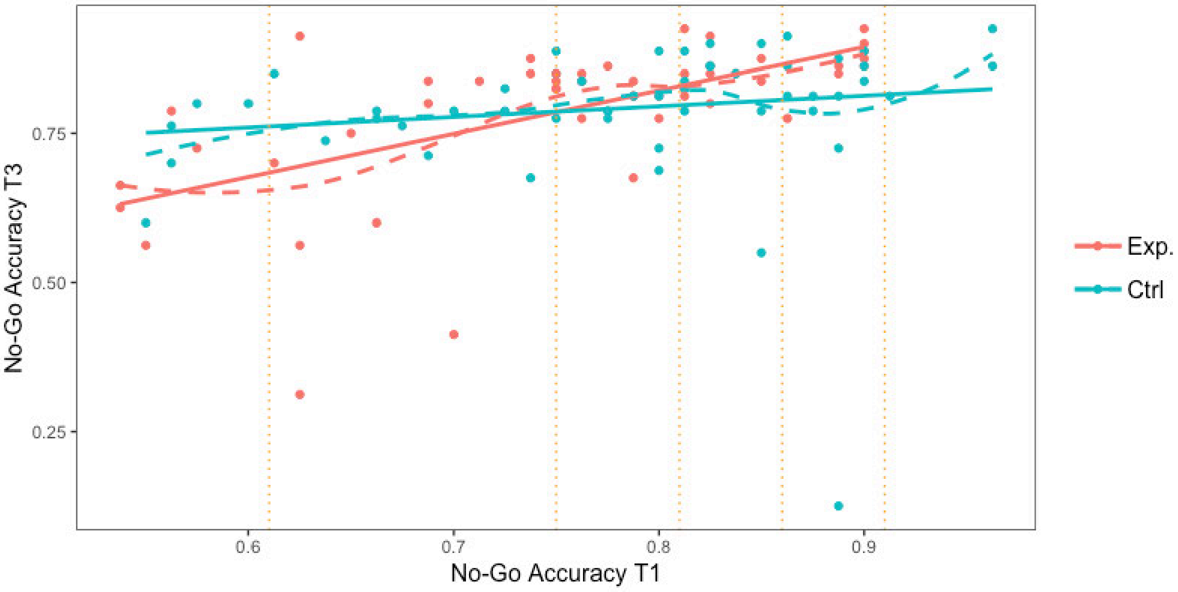


Figure G5. No-Go accuracy T1 v T3 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

T2 v T1 data for Go RT average did not satisfy the assumption of homogeneity of variance. A robust ANCOVA showed that there was a significant effect of the intervention group on T2 Go RT averages after controlling for T1 Go RT averages, at T1 scores of 71.2 (difference = 63.8, *p* = .01, $CI_{95\%}$[14.03, 113.66], *n* = 27), 119.4 (difference = 42.09, *p* = .01, $CI_{95\%}$[7.95, 76.24], *n* = 61), 159.4 (difference = 35.91, *p* = .01, $CI_{95\%}$[4.70, 67.11], *n* = 70), and 180.1 (difference = 31.58, *p* = .028, $CI_{95\%}$[-1.55, 64.70], *n* = 64). Figure G6 shows a scatterplot for the Go RT averages for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points. At the T1 scores where there was a significant difference between groups, the effect of the intervention group was the opposite of what was expected. That is, the control group scored higher on T2 Go RT averages after controlling for T1 averages.

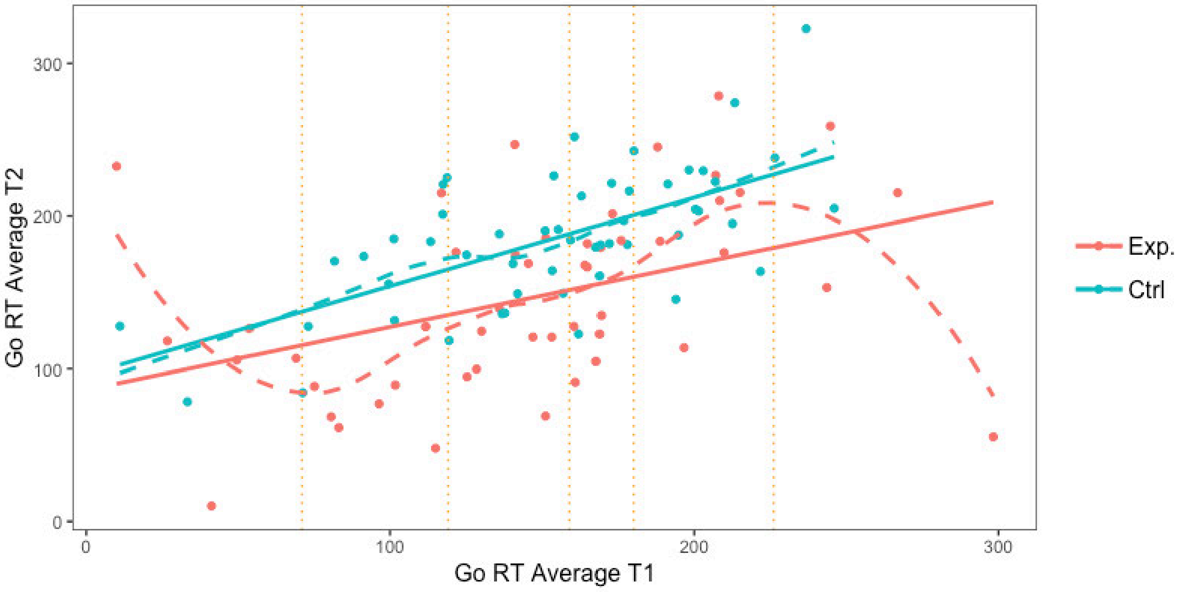


Figure G6. Go RT averages T1 v T2 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

T3 v T1 data for Go RT average satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 Go RT average after controlling for the T1 Go RT average on the Go/No-Go Task, *F*(1, 98) = 0.32, *p* = .576 (η_p^2= 0.003, $CI_{90\%}$[0.00, 0.05]) with the covariate, T1 Go RT average, significantly related to the T3 Go RT average, *F*(1, 98) = 76.26, *p* < .001, (η_p^2= 0.44, $CI_{90\%}$[0.31, 0.53]).

Numerical Stroop Task. T2 v T1 data for numerical Stroop incongruent RT did not satisfy the assumption of homogeneity of variance. A robust ANCOVA showed that there was a significant effect of the intervention group on T2 incongruent RT after controlling for T1 incongruent RT, at the T1 score of 427.16 (difference = 74.27, *p* = .047, $CI_{95\%}$[1.12, 147.43], *n* = 75). Figure G7 shows a scatterplot for the incongruent RT for T1 vs T2 across the experimental and control groups where the nonparametric regression lines (dashed lines) for both groups are shown as well as the linear regression fit and each of the five T1 design points. At the T1 scores where there was a significant difference between groups, the effect of the intervention group was the opposite of what was expected. That is, the control group scored higher on T2 incongruent RT after controlling for T1 RT.

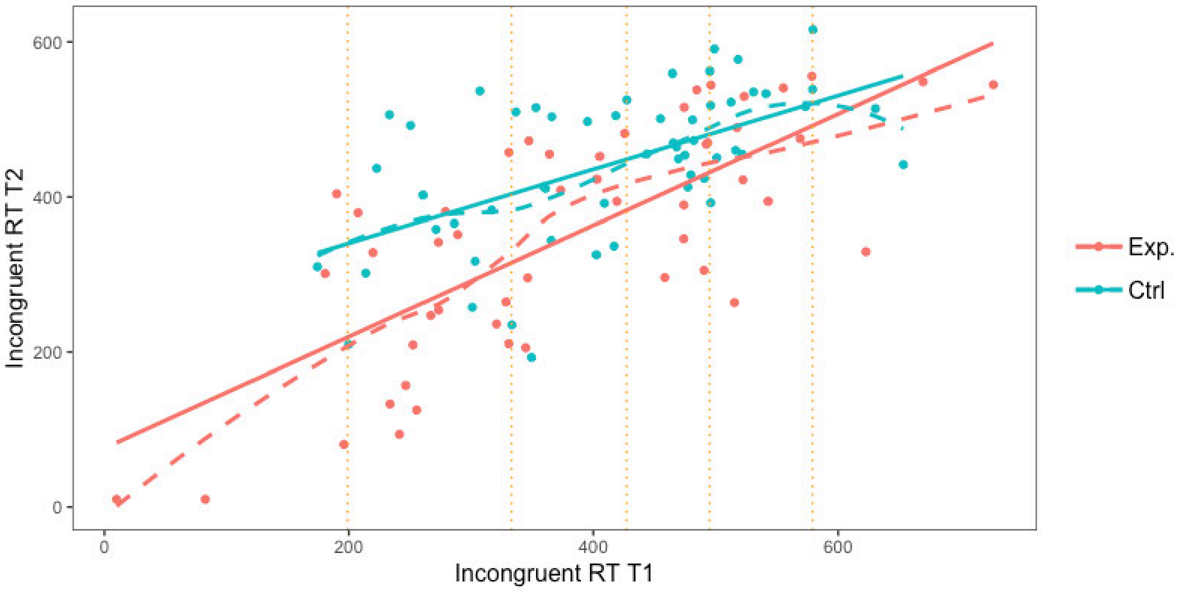


Figure G7. Numerical Stroop Incongruent RT T1 v T2 scatterplot demonstrating the comparison points for the robust analysis of covariance (ANCOVA).

T3 v T1 data for Numerical Stroop incongruent RT satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T3 incongruent RT after controlling for the T1 incongruent RT on the Numerical Stroop task, $F(1, 102) = 6.41$, $p = .0591$ ($\eta_p^2 = 0.06$, $\mathrm{CI}_{90\%}[0.01, 0.15]$) with the covariate, T1 incongruent RT, significantly related to the T3 incongruent RT, $F(1, 102) = 61.841$, $p < .001$, ($\eta_p^2 = 0.40$, $\mathrm{CI}_{90\%}[0.28, 0.50]$).

Numerical Stroop incongruent accuracy data satisfied all assumptions for running a parametric ANCOVA and there was no significant effect of group on T2 incongruent accuracy after controlling for the T1 incongruent accuracy on the Numerical Stroop

Task, $F(1, 102) = 0.14$, $p = .712$ ($\eta_p^2 = 0.001$, $CI_{90\%}[0.00, 0.03]$) with the covariate, T1 incongruent accuracy, significantly related to the T2 incongruent accuracy, $F(1, 102) = 13.00$, $p < .001$, ($\eta_p^2 = 0.11$, $CI_{90\%}[0.03, 0.21]$). There was no significant effect of group on T3 conflict cost after controlling for the T1 incongruent accuracy on the Numerical Stroop Task, $F(1, 102) = 0.0001$, $p = .991$ ($\eta_p^2 = 0.00$, $CI_{90\%}[0.00, 1.00]$) with the covariate, T1 incongruent accuracy, significantly related to the T3 incongruent accuracy, $F(1, 102) = 18.12$, $p < .001$, ($\eta_p^2 = 0.16$, $CI_{90\%}[0.06, 0.26]$).

9.8 Appendix H: Test-retest reliability of near- and far-transfer tasks

9.8.1 Near-transfer

Test-retest reliability was determined using intraclass correlation coefficients (ICC). Since the research design assumed that participants were considered as random effects for the EF measures, a one-way random effects ANOVA model to derive ICC was used (Aldridge, Dovey, & Wade, 2017; Shrout & Fleiss, 1979). The irr package (Gamer, Lemon, Fellows, & Singh, 2019) in R (R Core Team, 2019) was used for calculation of all ICC. As seen in Table H1, the test-retest reliability for all individual EF measures across all three measurement points were very low.

Table H1

Test-Retest Reliability for Executive Function (EF) Measures.

EF Domain	Task	Test-Retest Reliability	$CI_{95\%}$
Inhibition	Flanker RT Conflict Cost	-0.05	-0.15 – 0.07
	Flanker Incongruent Accuracy	0.29^{***}	0.16 – 0.42
	No-Go Accuracy	0.31^{***}	0.18 – 0.43
	Go RT Average	0.40^{***}	0.28 – 0.52
	Stroop RT Incongruent	0.55^{***}	0.44 – 0.65
	Stroop Incongruent Accuracy	0.40^{***}	0.28 – 0.52
Shifting	Switching Completion	0.55^{***}	0.45 – 0.65
	Card Sorting Perseverative Errors	0.19^{***}	0.07 – 0.32
	Card Sorting Conceptual Level Responses	0.34^{***}	0.22 – 0.47
	Switcher Errors	0.30^{***}	0.18 – 0.43
	Switcher Error Difference	-0.02	-0.12 – 0.10
WM	OSPAN Partial Score	0.32^{***}	0.18 – 0.46
	RSPAN Partial Score	0.57^{***}	0.44 – 0.68
	SSPAN Partial Score	0.38^{***}	0.24 – 0.53
Other	Mean Reaction Time (PS)	0.48^{***}	0.37 – 0.59

Note. EF = executive function; RT = reaction time; WM = working memory; OSPAN = operation span; RSPAN = reading span; SSPAN = symmetry span; PS = processing speed; $CI_{95\%}$ = 95 % confidence interval. All test-retest reliability coefficients are ICC determined by a one-way random effects ANOVA model (Aldridge, Dovey, & Wade, 2017; Shrout & Fleiss, 1979).
*** $p < .001$.

As the whole purpose of using latent variables and creating factor scores is to increase the reliability of EF measures (Aksayli et al., 2019), ICC was used to determine the test-retest reliability of the WM and combined EF factor scores. Similar to the ICC calculation for the individual EF measures, the research design assumed that participants were considered as random effects for the factor scores and a one-way random effects ANOVA model was used (Aldridge et al., 2017; Shrout & Fleiss, 1979). The WM factor score test-retest reliability across all three measurement points was 0.76 with a 95% confidence interval from 0.69 to 0.82 (F(104,210) = 10.60, *p* < .001). The combined EF factor score test-retest reliability across T1, T2 and T3 was 0.82 with a 95% confidence interval from 0.76 to 0.87 (F(104,210) = 14.50, *p* < .001). As would be expected, when compared to the test-retest reliability of the individual EF measures (Table H1), the creation of factor scores appeared to significantly improve the test-retest reliability of EF measurement.

9.8.2 Far-transfer

Test-retest reliability for the far-transfer measures was determined by ICC. Since the research design assumed that only participants were considered as random effects for the PAT Maths, PAT Reading and gF measure (NNAT2), a one-way random effects ANOVA model to derive ICC was used (Aldridge et al., 2017; Shrout & Fleiss, 1979). For behavioural reports, although a one-way random effects ANOVA model was used for parents and T1-T2 teacher observations, a two-way random effects ANOVA was used for teacher observations across T2-T3 as the teacher had changed between these time periods (Aldridge et al., 2017). As seen in Table H2, the ICC ranged from 0.39 to 0.76, indicating that there was some variability in the scores across all three time periods.

Table H2

Test-Retest Reliability Data for the Far-Transfer Measures.

	Test	Test-Retest Reliability***	$CI_{95\%}$
Academic Achievement	PAT Maths	0.67	0.58 – 0.75
	PAT Reading	0.61	0.50 – 0.70
Fluid Intelligence (gF)	NNAT2 (NAI)	0.63	0.53 – 0.72
CEFI - Parent	Inhibition	0.46	0.34 – 0.58
	Shifting	0.39	0.27 – 0.51
	WM	0.56	0.45 – 0.66
CEFI -Teacher T1 v T2	Inhibition	0.76	0.62 – 0.85
	Shifting	0.67	0.49 – 0.79
	WM	0.74	0.60 – 0.84
CEFI- Teacher T2 v T3	Inhibition	0.62	0.39 – 0.78
	Shifting	0.51	0.20 – 0.71
	WM	0.46	0.22 – 0.66

Note. PAT = Progressive Achievement Test; NNAT2 = Naglieri Nonverbal Achievement Test – Second Edition; NAI = Naglieri Ability Index; CEFI = Comprehensive Executive Function Inventory; WM = working memory. All test-retest reliability coefficients are ICC determined by a one-way random effects ANOVA model apart from T2 v T3 teacher behavioural reports which are determined by a two-way random effects ANOVA (Aldridge, Dovey, & Wade, 2017).

*** All test-retest ICC are significant at $p < .001$.