This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Studies Concerning the Application of Psychological Science to Education

Stuart J. Ritchie, MSc.

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology to The University of Edinburgh September 2013
Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Signed,

Stuart J. Ritchie
Acknowledgements

Avoidance of cliché is always a virtue, but is not always possible. For that reason, I have no qualms in saying that the experience of completing my PhD was so enjoyable mainly due to the people with whom I’ve lived and worked over the past three years.

I had the extraordinary luck of having two marvelous PhD supervisors: Sergio Della Sala imparted a zero-tolerance policy on flim-flam that is, I hope, evident throughout this thesis, and Rob McIntosh inspired with his unutterable passion for one particular band. Rob and Sergio’s help and advice on designing studies, approaching schools, analyzing data, writing papers, and responding to reviews was beyond the supervisor’s call of duty, and for that I’m beyond grateful. I was also very fortunate to work on several projects with Tim Bates, who has a remarkable, near-constant stream of fascinating thoughts, ideas, and arguments. There isn’t a single person I’ve learned more from (or quarreled with more) in my time at university.

My friends, colleagues, officemates, and fellow students made the atmosphere in our department—and our office in particular—a joy. Of particular note were Gary Lewis, Sebastià Sandoval Similà, Iva Ćukić, Max Greene, and Ian Scott, but there were many others; it’s been a pleasure to work, live, chat, and argue with you all.

Much of the research described herein would have been impossible without the participation of the children and the help of the staff at Newark Primary School, Port Glasgow High School, Towerbank Primary School, Bruntsfield Primary School, and Bruntsfield Kidzcare. Particular thanks go to Christine McNeill, Patricia Morrison, Joyce Gilmour, Jenny Dobie, Harriet Averill-Spence, Carol-Anne Kyle and Karen McInnes for their invaluable organization and patience when I appeared day after day, asking to test seemingly endless numbers of their pupils. Alex Gray’s voice-acting skills were also invaluable for the recordings described in Chapters 3 and 4.

In some other projects conducted during the past three years, it was an honour to collaborate with a number of other researchers, all of whom were enormously helpful and kind; thanks go—in no particular order—to Richard Wiseman, Chris French, Ian Deary, Geoff Der, Robert Plomin, Michelle Luciano, and Eric Chudler. Thanks also to Caroline Watt and Keith Laws for encouragement and opportunities to give talks and review manuscripts.

For endless support in matters large and small I thank my parents, but could never begin to thank them enough. I hope this thesis proves to them that I haven’t completely wasted the past three years.

Finally, Katharine deserves some form of award (or perhaps a sainthood) for putting up with my repetitive jokes and neurotic nonsense for three solid years. Sadly there is no such award, so she’ll have to make do with this: Thank you, Katharine.

----

To acknowledge the many contributions of my coauthors, beyond the abstract of this thesis “we” is always used instead of “I”.
Thesis Abstract

The purpose of this thesis is to apply an evidence-based perspective to educational interventions and techniques, and more specifically, to examine areas in which techniques ostensibly derived from an understanding of the psychological literature are applied to children’s learning. Broadly, the thesis moves from a review of ‘alternative’ techniques in education, toward empirical studies in areas where techniques informed by psychological science may, or may not, inform educational practice, toward a final empirical study examining the outcomes of the educational process.

In an introductory review (Chapter One), I assess the evidence base of five ‘alternative’—but still commonly-used—educational techniques, providing a sketch of the sometimes seductive claims made by their proponents, and the reasons teachers may decide to use them in their classrooms.

Chapter Two describes a study of one such ‘alternative’ technique: coloured filters for alleviating reading difficulties. I followed-up a sample of eighteen children who had used these filters—plastic overlays or tinted spectacle lenses—while reading for one year, and showed that, similar to the analysis at initial diagnosis, the lenses did not appear to improve reading on either a rate-of-reading test or a reading comprehension measure.

In Chapter Three, I describe an investigation of a second technique: brief relaxation and exercise periods designed to improve children’s attention and
concentration in the classroom. In two experiments, one involving two hundred and twelve children and the second involving two hundred and seventy children, I found inconclusive results: small detrimental effects of exercise on attention, but small positive effects on memory.

Chapter Four addresses a technique related to those in the previous chapter: wakeful resting. Shown to be effective for learning in amnesic patients and older individuals, and theoretically important for our understanding of memory consolidation and forgetting, this technique had not yet been applied to children learning in the classroom. Here, I provide evidence from two experiments—one large-scale, in which the technique was carried out by two hundred and eighty-four children in the classroom, and one small-scale, where the measures were administered to fourteen children in a controlled setting—that both show the technique does not appear to improve memory in young children.

In Chapter Five, I describe a simultaneous study of two educational techniques, one popular but poorly-evidenced, and one unpopular but with a strong evidential basis: mind-mapping and retrieval practice, respectively. In two samples of one hundred and nine and two hundred and nine children, I showed that retrieval practice, with or without mind mapping, improved fact learning in primary school children.

In Chapter Six, I focus on the effects of education, examined in a large, longitudinal, publicly available birth cohort dataset ($n > 18,000$). Using structural equation modeling, I show that reading and mathematics skills measured in childhood
predict socioeconomic status in midlife, even after controlling for socioeconomic status of origin, general intelligence, motivation, and educational duration.

Finally, in Chapter Seven, I summarize the findings of the thesis, give recommendations for future research, and discuss potential contributions to education from three other fields of psychology: neuroscience, social psychology, and differential psychology.
## Contents

Declaration ......................................................................................................................... 2
Acknowledgements ........................................................................................................... 3
Thesis Abstract ................................................................................................................... 4
Contents .............................................................................................................................. 7
List of Figures ..................................................................................................................... 10
List of Tables ....................................................................................................................... 12

Chapter One: A Review of Selected ‘Alternative’ Educational Techniques .................. 14
  Abstract .......................................................................................................................... 15
  Introduction .................................................................................................................... 16
  Brain gym ....................................................................................................................... 17
  Drinking water ............................................................................................................... 20
  Fish oils ......................................................................................................................... 22
  Brain training ................................................................................................................. 27
  Chewing gum ................................................................................................................. 33
  The attraction of ‘alternative’ educational techniques .................................................. 36
  The present thesis: Research questions and aims ....................................................... 40

Chapter Two: Irlen Coloured Filters in the Classroom: A 1-Year Follow-Up ............... 43
  Abstract .......................................................................................................................... 44
  Introduction .................................................................................................................... 45
  Method ............................................................................................................................. 48
  Results .............................................................................................................................. 53
  Discussion ....................................................................................................................... 58

Chapter Three: Effects of In-Class Exercise and Relaxation Sessions on
Children’s Preparedness for Learning ........................................................................... 63
  Abstract .......................................................................................................................... 64
  Introduction .................................................................................................................... 65
  Effects of exercise on cognitive performance ............................................................ 65
  Effects of relaxation on cognitive performance ......................................................... 69
Chapter Four: Does wakeful resting boost recall of newly-learned information in primary school children? ................................................................. 106

Abstract ......................................................................................... 107
Introduction .................................................................................... 108
Experiment 4.1 ............................................................................. 113
  Method ....................................................................................... 114
  Results ...................................................................................... 118
  Discussion .................................................................................. 123
Experiment 4.2 ............................................................................. 125
  Method ....................................................................................... 126
  Results ...................................................................................... 129
  Discussion .................................................................................. 133
General Discussion ......................................................................... 135

Chapter Five: Retrieval practice, with or without mind mapping, boosts fact learning in primary school children ........................................... 139

Abstract ......................................................................................... 140
Introduction .................................................................................... 141
Experiment 5.1 ............................................................................. 145
  Method ....................................................................................... 145
  Results ...................................................................................... 149
  Discussion .................................................................................. 153
Experiment 5.2 ............................................................................. 154
  Method ....................................................................................... 154
  Results ...................................................................................... 155
  Discussion .................................................................................. 160
General Discussion ......................................................................... 161
## Chapter Six: Enduring Links from Childhood Mathematics and Reading Achievement to Adult Socioeconomic Status

### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>167</td>
</tr>
<tr>
<td>Introduction</td>
<td>168</td>
</tr>
<tr>
<td>Method</td>
<td>170</td>
</tr>
<tr>
<td>Results</td>
<td>174</td>
</tr>
<tr>
<td>Discussion</td>
<td>179</td>
</tr>
</tbody>
</table>

## Chapter Seven: Summary and Discussion of the Application of Psychological Science to Education

### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>184</td>
</tr>
<tr>
<td>Summary of the present thesis and implications for future research</td>
<td>185</td>
</tr>
<tr>
<td>Beyond cognitive psychology in education</td>
<td>190</td>
</tr>
<tr>
<td>Neuroscience</td>
<td>191</td>
</tr>
<tr>
<td>Social psychology</td>
<td>194</td>
</tr>
<tr>
<td>Differential psychology</td>
<td>196</td>
</tr>
<tr>
<td>Conclusion</td>
<td>200</td>
</tr>
</tbody>
</table>

## References

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
</tr>
</tbody>
</table>

## Appendix A: Supplementary Material for Chapter Three

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>231</td>
</tr>
</tbody>
</table>

## Appendix B: Supplementary Material for Chapter Five

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>240</td>
</tr>
</tbody>
</table>

## Appendix C: Supplementary Material for Chapter Six

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>248</td>
</tr>
</tbody>
</table>

## Appendix D: Publication List

<table>
<thead>
<tr>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
</tr>
</tbody>
</table>
List of Figures

Figure 2.1. Scatterplot of prescribed filter benefit on WRRT for original study and follow-up .......................................................................................................................... 55

Figure 2.2. Optimal-condition a) WRRT scores and b) GORT scores, for original study and follow-up ............................................................................................................. 56

Figure 3.1. Not-to-scale example of one form of the Lollipops Test .............................................................. 77

Figure 3.2. Example of one form of the Square Completion Task ...................................................................... 79

Figure 3.3. Mean results by condition for each test in Experiment 3.1 ............................................................. 82

Figure 3.4. Design of Experiment 3.2 ............................................................................................................. 93

Figure 4.1. Design of Experiment 4.1 ............................................................................................................. 116

Figure 4.2. Means for recall in each condition in Experiment 4.1 across time .............................................. 119

Figure 4.3. Design of Experiment 4.2 ............................................................................................................. 127

Figure 4.4. Performance on the initial and one-week measures in Experiment 4.2 ............................ 131

Figure 5.1. Example factsheets for Primary 4 and Primary 7 in Experiment 5.2; example notes from one Primary 4 child and an example mind map from one Primary 7 child in Experiment 5.2 ............................................................................. 146

Figure 5.2. Percentage of facts recalled at the one- and five-week tests in Experiment 5.2 for the retrieval practice and non-retrieval conditions, and the mind maps and no mind maps conditions ................................................................................................. 157

Figure 6.1. Associations of age-7 mathematics and reading achievement with age-42 attained SES for male and female cohort members .............................................................................. 175

Figure 6.2. Associations of mathematics ability at age 7 and reading ability at age 7 with attained socioeconomic status at age 42, including all control variables .......... 177

Figure A1. Remaining two alternate forms of the Lollipops Test ................................................................. 232

Figure A2. Example of the 5×5 grid Visual Patterns Test administered to Primary 4 and Primary 6 classes in Experiment 3.1 ................................................................................................................ 233

Figure A3. Quiz sheets for the younger and older children for both animals ............................................ 238

Figure C1. Associations of age-7 mathematics and reading achievement with attained socioeconomic status, prior to the sex-equation process described in Chapter 6 ............................................................ 251

Figure C2. Associations of age-7 mathematics achievement with attained socioeconomic status, excluding reading ........................................................................................................ 252
Figure C3. Associations of age-7 reading achievement with attained socioeconomic status, excluding mathematics .......................................................... 253

Figure C4. Associations of age-7 mathematics and reading achievement with attained SES, with the inclusion of age-7 oral language ability ............................................. 254

Figure C5. Associations of age-7 mathematics and reading achievement with attained SES, with the inclusion of age-16 aggression ...................................................... 255
List of Tables

Table 2.1. WRRT scores for each overlay condition by group ........................................54

Table 2.2. Mean GORT scores for each group for both years ..........................................57

Table 3.1. Number of children in each class, with mean age and time of test administration ........................................................................................................................................75

Table 3.2. Testing orders across the three experimental weeks in Experiment 3.1 ...........80

Table 3.3. Mean scores by year group and condition for the Lollipops Test, the VPT, and the SCT ........................................................................................................................................81

Table 3.4. Regression coefficients for each fixed and random effect for the Lollipops Test ........................................................................................................................................85

Table 3.5. Number of children in each year group involved in Experiment 3.2, with mean ages ........................................................................................................................................89

Table 3.6. Mean Lollipops Test and Memory Test scores by year group, condition, and time for Experiment 3.2 ........................................................................................................................................93

Table 3.7. Regression coefficients for each fixed and random effect for the Lollipops Test in Experiment 2 ........................................................................................................................................95

Table 3.8. Results of binomial generalized linear mixed-effects model for the Memory Test in Experiment 3.2 ........................................................................................................................................97

Table 4.1. Descriptive statistics for story recall, theme recall, and quiz performance at each measurement ..................................................................................................................................118

Table 4.2. Results of the generalized linear mixed model for Experiment 4.1 .................121

Table 4.3. Results of the generalized linear mixed model for Experiment 4.2 ...............133

Table 5.1. Numbers and ages of children per year group in the two experiments .........145

Table 5.2. Distributions of each factsheet across the four cells of the experiment in Experiment 5.1 ........................................................................................................................................149

Table 5.3. Performance in the initial learning phase for Experiment 5.1, broken down by condition ........................................................................................................................................150

Table 5.4. Mean percentage scores and sample sizes for Experiment 5.1 .................150

Table 5.5. Results of the generalized linear mixed model for Experiment 5.1 ...............153

Table 5.6. Distributions of each factsheet across the four cells of the experiment in Experiment 5.2 ........................................................................................................................................156
Table 5.7. Performance in the initial learning phase for Experiment 5.2, broken down by condition .................................................................156

Table 5.8. Mean percentage scores and sample sizes for the 1-week and 5-week recall tests .................................................................157

Table 5.9. Results of the generalized linear mixed model for Experiment 5.2........160

Table 6.1. Heterogeneous correlation matrix and descriptive statistics for all variables, for female and male cohort members.................................................................171

Table 6.2. Changes in model fit and degrees of freedom, with associated p-values, after dropping paths in the sex-equated model .......................................................176
Chapter One

A review of selected ‘alternative’ educational techniques

* A version of this chapter was published as Ritchie, S. J., Chudler, E. H., & Della Sala, A. (2012). Don’t try this at school: The attraction of ‘alternative’ educational techniques. In S. Della Sala & M. Anderson (Eds.), Neuroscience and Education: The Good, the Bad, and the Ugly (pp. 244-264). Oxford, UK: Oxford University Press.
Abstract

A variety of ‘alternative’ educational techniques are being used in classrooms worldwide, and it is often unclear whether or not they are supported by rigorous evidence. In this chapter, the evidence for a selection of these techniques—the classroom uses of Brain Gym, drinking water, brain training games, fish oil supplements, and chewing gum—is reviewed. The evidence is mainly found wanting. We provide a brief discussion of why such techniques may nevertheless be attractive to some educators. The chapter concludes with a summary of the questions asked in the remaining chapters of the present thesis.
Introduction

This introductory chapter reviews the evidence base for a variety of ‘alternative’ educational techniques, used in classrooms worldwide. Proponents claim that each of these techniques will improve children’s cognitive function or learning abilities, and the claims are based on appeals to psychological or neuroscientific studies and principles. An examination of the evidence underpinning these techniques, then, sets the scene for the remainder of the present thesis, which reports tests of a variety of classroom learning techniques of varying plausibility, but often high popularity. At the close of this chapter, a brief summary of the questions asked in the remainder of the thesis is provided.

The techniques examined in this chapter are ‘Brain Gym’, a set of in-class movements and exercises that are claimed to boost a variety of cognitive functions; drinking water, supposed by some to be a constant requirement for peak cognitive function; fish oil supplements, shown to be effective for some medical conditions, but also claimed to be effective in improving learning abilities; brain training video games, phenomenally popular and now being introduced to classrooms in an attempt to enhance cognition; and chewing gum, shown in some experiments to improve functioning in adults, but also advocated by some for the school classroom. Often, very little evidence whatsoever exists for the efficacy of these interventions; this chapter attempts to review the most important data related to each technique, but makes no claims to be systematic.

Rather like an ‘alternative’ medicine, there is no clear-cut definition of an ‘alternative’ educational technique. Indeed, ‘complementary’ - another word used by
advocates of herbalism, homeopathy, and the like - may seem more appropriate for
the techniques discussed in this chapter, as they are not seriously entertained as an
alternative to mainstream education. However, in this chapter we use ‘alternative’ to
mean ‘unconventional’ or ‘non-traditional’.

**Brain Gym**

It would be of incredible significance if a few brief, simple exercises and
movements, performed several times during a school day, could significantly improve
children’s academic performance. It would be of even greater significance if there
were a variety of specific actions that improved particular skills, so each child could
have a set of movements tailored to their learning needs. This may sound too good to
be true, but proponents of ‘Brain Gym’, a programme of movements which
apparently promote “whole-brain learning” (Dennison & Dennison, 2010), say they
have found just such a thing.

Many educators rate the Brain Gym activities extremely favourably - indeed,
after denigrating the ‘data-driven’ approach to education as unhelpful, McChesney (in
a foreword to Dennison & Dennison 2010) calls Brain Gym “the triumphant uplifting
and transforming of the human condition” (p. xiii). Cecilia Koester, a proponent,
states on her website: “I have seen miraculous improvement in both children and
adults who have used Brain Gym. In fact, three children with whom I’ve worked have
gone from blindness to sight. One child began to walk independently at age five and,
now seven and a half, has never returned to his wheelchair” (Koester, 2002). This
confidence is unrelenting for the entirety of the *Brain Gym: Teacher’s Edition* book (Dennison & Dennison, 2010).

Twenty-six separate movements and activities are described by Dennison and Dennison (2010), and all are claimed to have impressive effects on a variety of an individual’s physical, cognitive, and social abilities. Examples include the ‘Thinking Cap’, where the ears are pulled and massaged (this activity apparently enhances listening comprehension and public speaking), ‘Brain Buttons’, where children rub areas under their collarbones (to improve ‘ease of eye movements’ and hand-eye coordination (p. 57)), ‘Think of an X’, wherein children imagine an ‘x’-shape linking their hips and shoulders (to improve ‘spatial awareness and visual discrimination’ (p. 51)), and ‘Sipping Water’, of which more below.

A host of assertions accompany the descriptions of the movements. For example, Dennison and Dennison (2010) claim that children need to be taught to ‘listen with both ears’ (p. 38), while ‘excessive exposure to electronic sounds… will “switch off” the ears’ (p. 66); that riding in a car or bus can ‘adversely affect depth perception and binocular vision’ (p. 96); that liquids other than water ‘are said to be processed in the body as food’ (p. 54); that yawning can aid creative writing skills (p. 64); and that Brain Gym can be used effectively for children with learning disabilities, individuals with depression, survivors of natural disasters, and attendees of drug rehabilitation clinics (pp. 97-98). They cite no scientific evidence for any of these claims.
Clearly some logical leaps are in effect, here. Whereas it is plausible (and as we will see below, evidenced) that incorporating periodic gaps in lessons for fun activities will reduce boredom in children, potentially enhancing learning (see Chapter 3 of this thesis), this should not be taken as proof that the very specific activities described by Brain Gym proponents are effective. In addition, whereas the brain is split into two hemispheres, there is no evidence that there is a ‘midline’ between them which children must be taught to utilise, as Dennison and Dennison (2010, p. 10) claim.

The lack of a theoretical basis for Brain Gym would not necessarily be of concern if there were well-conducted peer-reviewed studies showing the programme’s effectiveness at improving learning. There are, however—to our knowledge—no such studies. Indeed, Dennison and Dennison (2010) advise readers that the effectiveness of the activities “can be personally validated by anyone who takes a few minutes to do [them]… Brain Gym International doesn’t conduct research on its own methodologies” (p. xiv). A small number of studies have, nevertheless, been performed on Brain Gym, and are reviewed by Hyatt (2007). Unfortunately, all the studies have been of extremely poor quality. For example, Wolfsont’s (2002) study had only four participants, one of whom was the lead experimenter, while Sifft and Khalsa (1991) did not measure the pre-experiment baseline performance of their participants, leaving them unable to show an improvement over time. Hyatt (2007) also notes that some statistical techniques are used that make the findings difficult to interpret. The review concludes that there is no sound research evidence for the positive effects of Brain Gym (one more study that specifically examined Brain Gym, alongside some other interventions, is described below under ‘Brain Training’).
There is, however, a small set of higher-quality studies investigating the effects of exercise breaks – not specific Brain Gym activities - on children’s learning. These are reviewed in more detail in Chapter 3 of this thesis, and are accompanied by reports of two new experiments carried out to assess the effectiveness of in-class exercise techniques. However, such studies are far from support for Brain Gym, and much less an endorsement of their programme. Coupling fun breaks – which may reduce boredom and improve concentration – with demonstrably false claims about ‘Brain Buttons’, ‘midlines’, is unfortunate, since it has the potential to mislead children about the workings of the brain and the body.

There have been a number of re-writings and backtrackings in recent years; as noted above, the Brain Gym manual has been revised and updated to remove some of the more conspicuous errors (see Dennison & Dennison, 1994; Sense About Science, 2008). However, until all such errors are removed, Brain Gym remains an exercise in pseudoscience.

**Drinking water**

The claim that regularly drinking water can aid children’s classroom performance can be found in Brain Gym literature, but is also a fairly widespread belief amongst the public (Grandjean, Reimers, Haven, & Curtis, 2003). Whereas the popular idea that eight glasses of water are required per day is a myth (Valtin, 2002), and studies of the effects of dehydration on cognition in adults are inconsistent (Cian,
Observational studies indicate that many schoolchildren are under-hydrated (e.g. Kaushik, Mullee, Bryant, & Hill, 2007), and dehydration can certainly result in a variety of negative behaviours such as irritability and restlessness in children (D’Anci, Constant, & Rosenberg, 2006). Addressing hydration needs is certainly important for nutritional reasons, but a more interesting question is whether providing children with water in the classroom will immediately improve cognitive function.

This has been directly addressed by—to our knowledge—only two studies to date. The first (Benton & Burgess, 2009) was a repeated-measures experiment where forty children were tested on memory and attention tests on two consecutive days, on one day having drunk 300ml of water just prior to testing, and on the other having had no extra water. Scores on memory, but not attention, tests were significantly higher on the day where water was drunk. The second study (Edmonds & Jeffes, 2009) used a between-subjects design, splitting twenty-three children into two groups who were tested twice in one day on memory, visual attention, visual search and visuomotor tasks. One group drank a glass of water in between the testing sessions, and the authors reported a significant effect of group on an increase in scores on visual attention and visual search, but not memory or visuomotor tasks.

Thus, the two small-scale studies that have been carried out so far are to some extent contradictory in their results, and certainly do not represent a generalisable literature, not least because the studies were not double-blinded (although, short of
using an IV drip, as suggested for adult participants by Lieberman (2007), it is difficult to imagine how such blinding could be achieved). Additionally, in neither study was each child’s dietary or hydration status at the time of testing taken into account. In any case, as Edmonds and Jeffes (2009) note, even if it is conclusively shown that drinking water facilitates classroom learning, a host of questions remain: how much water is necessary, and is it dependent on body mass? How long before starting schoolwork should a child drink? Lastly, and importantly, by what mechanism does drinking water—beyond satisfaction of basic hydration needs—improve cognitive function? A great deal more research is required before these findings can be utilised in regular classroom situations.

**Fish oils**

Some authors have advocated the use of fish oils, specifically the two ‘omega-3’ polyunsaturated fatty acids, EPA (eicosapentanoic acid) and DHA (docosahexaenoic acid) to improve educational outcomes. These are ‘essential’ fatty acids - they cannot be synthesized by the body, and must be obtained through diet - which are involved in various metabolic processes, including many of those in the brain (e.g. Das, 2003), and appear to be particularly important during early brain development in utero (e.g. Dunstan, Simmer, Dixon, & Prescott, 2008).

A marked reduction in consumption of oily fish in modern, as compared to prehistoric, diets has seen the quantity of omega-3 fatty acids decrease, while omega-6 fatty acids—found in poultry, eggs, and vegetable oils—have become far more common. Omega-3 and omega-6 fatty acids compete for the same enzyme in
metabolic processes, and thus an excess of one type inhibits the body’s ability to use the other (Kirby, Woodward, & Jackson, 2009). The underlying hypothesis here is that many individuals, especially those not eating a balanced diet, are deficient in omega-3s, and therefore supplementation could have measurable benefits; since omega-3s are important in brain processes, these benefits could be cognitive or neurological.

Unfortunately, media reports (e.g. Cassidy, 2002) publicised fish oils as a potential treatment for dyslexia, dyspraxia, Attention Deficit Hyperactivity Disorder (ADHD) and autism before any data on their effectiveness for these conditions had been published. The leap thus was made from the essential nature of the nutrients, and a few anecdotal accounts, to claims that they have effects on a wide variety of serious conditions - even in a newspaper article informing readers that a study of those effects had just begun.

The study in question—the ‘Oxford-Durham Study’—was reported by Richardson and Montgomery (2005). A total of one hundred and seventeen children with a diagnosis of Developmental Coordination Disorder (DCD, or dyspraxia) were split into two groups, one receiving capsules containing omega-3 and omega-6 fatty acids, and one receiving an olive oil placebo for three months. Pre- and post-study testing failed to show any change in the main outcome measure—coordination improvements in the treatment group were no different from those in the placebo group—but the treatment group’s reading and spelling abilities were significantly better.
A number of other studies on groups of children with various developmental disorders such as autism, dyslexia and ADHD have since been performed. Studies on fish oil supplementation and autism spectrum disorders are reviewed by Bent, Bertoglio, and Hendren (2009), while Kirby et al. (2009) provide an excellent review of the totality of the evidence. Ultimately, due to the wide variation in supplement compositions, sample characteristics, experimental methodologies, time periods, and outcome measures (not to mention the inconsistent results reported) reviewers find it impossible to draw any firm conclusions about the evidence for the use of the supplements. There is simply no adequate body of evidence upon which to base the claim that fish oil supplementation can improve any cognitive or behavioural outcome for children with special educational needs (incidentally, this is not to say that fish oils do not have efficacy for treating some medical conditions, such as rheumatoid arthritis – see Miles & Calder, 2012).

A great many claims have nonetheless been made on the basis of this inconsistent evidence. Most notably, it has become widely assumed that even typically-developing children could benefit from fish oil supplements, regardless of the poverty of the evidence for this belief. The media reported heavily on the Durham Education Authority’s announcement that they would be administering regular fish oil supplements to a sample of up to five thousand children who were sitting their GCSE exams in 2007/8, in an effort to improve their performance. Unfortunately, this ‘study’ had no control group, was not double blind, and was not ultimately published in a scientific journal. Indeed, it was so poor that Goldacre (2008) uses it as an example to instruct a popular audience how not to conduct research. Nevertheless, the impression members of the public would likely have taken from the media
coverage—published before the study had even begun—was that fish oils are an intervention almost guaranteed to improve children’s school performance.

Thankfully, the Durham GCSE trial is not the only evaluation of fish oils’ ability to enhance the learning of typically-developing children. Osendarp et al. (2007) and Kirby, Woodward, Jackson, Wang, and Crawford (2010) have investigated this alleged effect in large samples with high-quality study designs, and both reported discouraging results. In the Osendarp et al. (2007) study, while supplementing children with a variety of micronutrients (iron, zinc, and several vitamins) did improve learning and memory performance, simply using omega-3 fatty acids did not. This study used two samples, one from a developed country with good nutrition (Australia) and one from a developing country with a poor nutrition (Indonesia). Perhaps counter-intuitively, the Indonesian children, despite their inferior diet, did not benefit substantially more from the micronutrient supplements than the Australian children. Of a range of over thirty-five cognitive, physical, and emotional outcomes measured, Kirby et al. (2010) found three significant differences between their omega-3 supplemented group and a placebo group after sixteen weeks of supplementation. Interestingly, one of the differences was that the behaviour of the placebo group, according to one teacher-rated measure, had become significantly better than that of the supplemented group. Small and inconsistent outcomes like these indicate that the results may have been simply due to chance. It is worth noting that as one measures more and more outcomes, the chances of finding false positive (Type I) results—in any direction—are multiplied; with thirty-five measures, it is not surprising that a few reached statistical significance, and the authors note this (Kirby et al., 2010, p. 728).
Results from the Osendarp et al. (2007) study, along with others (e.g. Gesch, Hammond, Hampson, Eves, & Crowder, 2002) point towards the fact that omega-3 fatty acids alone may not be enough to produce significant cognitive or behavioural benefits. These studies only showed a difference in ability and behaviour if nutrients other than omega-3 fatty acids were provided. Kirby et al. (2009) note that human nutritional biochemistry is deeply complex, and involves an enormous variety of contingent, interacting reactions. Several other micronutrients are required for the body to properly utilise fatty acids, and if a child is also deficient in these, it is unsurprising if omega-3-only studies are disappointing.

In a perfect world, all parents would have the knowledge and the means to provide their children with a balanced diet, including oily fish. In reality, supplementation may turn out to be an important method of providing children with essential nutrients like omega-3 fatty acids. However, at this time, it is impossible to come to firm conclusions about whether supplementation might reliably improve children’s academic performance, or especially about which supplements might have which effects. While the best research currently points towards fish oils having no educational benefits, especially in typically-developing children, a great deal more high-quality research is required before anyone can make conclusive statements about the efficacy of the technique.
Brain training

Given the huge—and growing—popularity of video games amongst children (e.g. Lenhart et al., 2008), it is not surprising that educators have begun to incorporate similar technologies in their classrooms. Numerous recent studies and reviews have looked at the effects of video games on children, in particular focusing on the potentially negative consequences of playing violent games, an area that has become an ideological and scientific battleground (Ferguson, 2010). However, as well as being entertaining, could video games be used to aid learning in school? If so, which types of game are best? Recently, it has been suggested that ‘brain training’ games could be of particular educational use, potentially improving children’s learning abilities and self-esteem (e.g. Miller & Robertson, 2010). Is this really the case?

Nintendo’s *Brain Training* (or, in the US, *Brain Age*) games are a phenomenally popular series for the hand-held Nintendo DS console. They are designed to be played for a short time each day, as gamers train themselves to complete various tasks such as basic arithmetical problems, calculations of the difference in time between two clocks, and counting the numbers of syllables in words. Then, the player can progress to the ‘brain age check’, which involves more arithmetic, Stroop tests (the console has an in-built microphone to record the player’s voice), and memory tests. Poor performance will result in the player being assigned a ‘brain age’ far higher than their chronological age. With practice, this brain age can be reduced.
Nintendo is quick to point out that the concept of brain age is not a scientific one (it appears to be merely a marketing strategy; Lorant-Royer, Munch, Mesclé & Lieury, 2010). However, the company uses Ryuta Kawashima, a Japanese neuroscientist, as the public face of the *Brain Training* games, which base their content on a book he authored (Kawashima, 2005). The brain age concept has obvious connotations – Kawashima has published gerontological research (e.g. Kawashima et al., 2005), and various other studies have assessed the efficacy of *Brain Training* for slowing the effects of dementia (e.g. Smith et al., 2009; Willis et al., 2006). Butcher (2008) notes that concerns have been raised about Nintendo’s marketing of *Brain Training*, especially in advertisements in which actors are seen failing to remember names, then using *Brain Training*, after which their memory swiftly improves.

Exaggerated claims like these are incongruous with the experimental data into the efficacy of brain training, which are equivocal (Butcher, 2008; Nacke, Nacke, & Lindley, 2009). A recent large study in adults (Owen et al., 2010) recruited 11,430 participants online, and after assessing them on a battery of cognitive tests, organised them into three groups: those who were to receive training in reasoning and problem-solving, those who were to receive *Brain Training*-like exercises in arithmetic, memory, and attention, and those who were to have no training and instead were simply to answer trivia questions on a range of subjects. The groups performed these tasks for six weeks, at which point they completed the test battery once more. The test scores of all three groups improved to a broadly similar degree – that is, not only did the type of training make no difference, but even those undergoing no specific training performed just as well.
The Owen et al. (2010) study illustrated a very important point about brain training. While gamers do show improvements on the game itself – in the study, the brain training group showed significantly enhanced performance on the brain training tasks – this does not imply that those benefits will be apparent outside of the circumscribed gaming context. Anyone claiming - on the basis of their gaming score - that the benefits can translate into daily life, or that the benefits are peculiar to brain training, is basing their ideas on mere analogy. Indeed, Fuyuno (2007) quotes neuroscientist Michael Marsiske: ‘users may get the illusion of huge gains when starting with Brain Age, but these have more to do with learning the device than actual mental improvements.’ (p.20).

A recent set of experiments (e.g. Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Jaeggi, Buschkuehl, Jonides, & Shah, 2011) have shown that training on specific working memory tasks known as ‘n-backs’ appears to improve fluid intelligence, a component of general intelligence (g) which involves reasoning and problem-solving abilities (the evidential basis for this technique is extremely controversial: see Melby-Lervåg & Hulme, 2013, Morrison & Chein, 2011, and Shipstead, Redick, & Engle, 2012 for a series of critical reviews; Boot, Simons, Stothart, and Stutts, 2013, Clifford and Getz, 2010, and Moody, 2009, offer substantial methodological criticisms of the research; and Redick et al., 2013 report a failure to replicate the effect using appropriate controls). The studies received a great deal of media attention, much of it making the leap from ‘a specific task has been shown to work for a specific measure’, to in effect stating ‘Nintendo’s Brain Training has been validated by scientists’ (e.g. Derbyshire, 2011). However, the studies did not in fact test Nintendo’s Brain Training, so are of limited relevance to our discussion in this chapter.
In any case, before claiming that brain training can aid children’s learning of specific classroom subjects, ecologically valid studies in the classroom should be carried out, testing outcomes involving learning new information, not simply executive functioning. To our knowledge, three studies have been performed on the effects of Nintendo’s *Brain Training* in the classroom: two in France (Lorant-Royer, Spiess, Goncalves, & Lieury, 2008; Lorant-Royer et al., 2010) and one in Scotland (Miller & Robertson, 2010). They provide conflicting data, but are illustrative of the issues one might encounter in carrying out and reporting such classroom studies.

Lorant-Royer et al. (2008) split forty-nine ten-year-old children, who had been tested on two school subjects (science and geography), self-esteem questionnaires, and on measures from an intelligence test, into four groups: a group playing *Dr. Kawashima’s Brain Training*, a group playing a similar Nintendo game, *Brain Academy*, a group completing paper-and-pencil puzzles, and a control group who had normal school lessons. When tested after a period of eleven weeks, there were some increases in performance in the *Brain Training* groups, but also in the paper-and-pencil and control groups. No between-group differences were found on any of the measures. *Brain Training*, then, was found to be no better than paper-and-pencil games, or indeed normal school activities, at improving academic achievement, or self-esteem. The authors concluded that *Brain Training* is a game, and nothing more.

Miller and Robertson (2010) reported a study with a somewhat similar design, which tested seventy-one ten to eleven-year-old children from three separate classes at three separate schools. In the first class, twenty-one children used *Dr. Kawashima’s*
Brain Training game for twenty minutes each school day for ten weeks. In the second, thirty-one children performed a set of Brain Gym exercises each day for the same time period. In the third, which was the control group, nineteen children had normal school lessons. Their mathematics performance, as well as their self-esteem, was assessed at both the start and the end of the ten-week period.

The maths scores of all groups increased; the Brain Training group by 10%, the Brain Gym group by 2%, and the control group by 5% (for self-esteem, no significant differences were found). Surprisingly, as noted by Logie & Della Sala (2010) in a critique of the study, the authors did not perform a between-groups test on the results. That is, they did not test whether the increase of 10% in the Brain Training group was statistically significantly greater than the increase of 5% in the control group. Subsequent analyses showed that the difference was not, in fact, significant, meaning that the researchers had essentially reported null results.

If one uses Miller and Robertson’s (2010) logic of simply comparing the raw size of the gain in score, it would appear that Brain Gym has a detrimental effect on maths performance, as the no-treatment control group achieved a greater score gain than the Brain Gym group. Puzzlingly, Miller and Robertson (2010) suggest – albeit cautiously – that Brain Gym proponents may “find encouragement” in the results (p. 252). However, a simple increase in a score, which may appear impressive on its own, should not be considered outside of the context of other results; in this case, the surrounding context makes the Brain Gym group’s result look far from remarkable.
This was not the only flaw in Miller and Robertson’s (2010) results. Even if the results of their study hold, they do not show the effectiveness of *Brain Training per se* – a control group who played a different game on the same console should have been included, to test the specific effects of of the tasks included in the *Brain Training* game (Logie & Della Sala 2010). We could then see whether the results could be caused by any computer game – perhaps an effect of gaming in general, or the nonspecific placebo and Hawthorne effects of being given a desirable console and being part of a study (for an example of this latter effect in education, see Rosas et al., 2003). More recently, Lorant-Royer et al. (2010) performed a study with controls of this type, and found no particularly noteworthy cognitive test score differences between no-treatment controls and groups playing *Dr. Kawashima’s Brain Training*, another Nintendo game (*Super Mario Bros.*), or pencil-and-paper games.

The evidence for the effectiveness of brain training in the classroom is at best questionable. There is, however, considerable interest in other forms of video games as learning aids; for teachers to neglect investigating the educational potential of such a popular form of entertainment would be foolish. Annetta (2010) provides an interesting theoretical discussion of the aspects of video gaming that are likely to enhance learning. It could be argued that brain training does not quite fit the bill; due to the fairly trivial nature of the puzzles, important factors such as ‘immersion’ and ‘instruction’ are likely to be low (see also Howard-Jones, 2010, for more discussion of educational games). Spence and Feng (2010), in a review of the effects of video games on spatial cognition, note that several studies show that ‘first-person shooter’ (FPS) action gamers have improved visuo-spatial abilities which are both persistent through time and generalise to outside of the game’s particular context. Of brain
training, they note that the simplicity of the tasks – especially when compared to complex modern action games – would not likely lead to any general cognitive benefits: “games modelled on the FPS genre have much more promise for remedial training than the kinds of simple puzzle games promoted by most brain-training enterprises” (p. 102).

We are some way yet from designing the perfect educational video game, either for typically-developing children or for children with developmental disorders (Durkin, 2010). But once again, an ‘alternative’ educational technique provides us with a peculiar irony: it may be that brain training is, despite its promises, the type of game least likely to translate to more general increases in cognitive abilities because it does not share the immersive properties of more mainstream video games. Naturally, if individuals enjoy the puzzles and challenges presented by brain training, we would not attempt to discourage them. However, as we have seen before, enjoyment does not necessarily translate into enhanced learning.

**Chewing gum**

In December 2010, the *Guardian* ran a story under the headline ‘German pupils told to keep chewing as scientists extol virtues of gum’ (Connoly, 2010), which described a Bavarian study of chewing gum’s effects on concentration in the classroom. Could it be the case that chewing gum—that notorious classroom irritation—can aid important learning skills like memory and attention (James, 2007; Jensen, 2005)? Despite the enthusiasm of the German scientists and teachers quoted in the *Guardian* report, we should adopt a healthy scepticism towards such claims,
especially since—as with the claims about fish oils, discussed above—the study had not been published at the time of the media report.

Claims about the effects of chewing gum on learning are made on the basis of laboratory studies such as that of Wilkinson, Scholey and Wesnes (2002), who showed that adult volunteers who chewed sugar-free gum performed better on memory tests— but not attention tests—than others who did not chew. It should be noted that another group, participants who mimicked chewing movements but without any gum, performed only slightly worse than those with the gum. The authors suggested that the jaw activity involved in chewing might increase cortical blood flow, thereby enhancing cognition. It is certainly true that the action of mastication (chewing) is accompanied by neurophysiological changes (e.g. Onozuka et al., 2003; Sakamoto, Nakata, Hondam, & Kakigi, 2009). However, the research into the effects themselves is controversial: a study published in the same journal, by Tucha, Mecklinger, Maier, Hammerl, and Lange (2004), failed to replicate Wilkinson et al. (2002)’s results.

More recent results are similarly inconclusive. Two studies by Smith (2009, 2010), showed that, whereas gum-chewing did appear to improve to a modest degree alertness, reaction time, and intelligence scores, it had no effect on memory tasks compared to a ‘no chewing’ control condition. Allen, Norman, and Katz (2008) also found no effect of chewing gum on memory. Studies of context-dependent learning—those that test whether chewing whilst learning facts enhances recall of those facts whilst chewing later, not due to the chewing per se, but simply due to the similar experience at learning and recall—have generally been negative (e.g. Johnson &
Miles, 2007; Overman, Sun, Golding, & Prevost, 2009), though see Rickman, Johnson, and Miles (2013) for a more encouraging result relating to chewing ‘resistance’ (that is, the number of pieces of gum in the mouth).

Perhaps chewing gum does not directly influence learning, but instead reduces stress and anxiety – rather like using a ‘stress ball’ toy – and allows learning to go ahead unimpeded? Zibell and Madansky (2009) found that many people self-report using chewing gum to lower stress levels. The two studies that investigate the effects of chewing gum on stress during difficult tasks (Scholey et al., 2009; Torney, Johnson, & Miles, 2009) are directly contradictory; different assessment tasks were used, however, which may explain the inconsistency. Unfortunately, no studies have directly investigated the effects that reduced stress due to chewing might have on cognitive ability or, for that matter, classroom performance.

The studies discussed here, then, in no way represent a conclusive evidence base for the beneficial effects of chewing gum on cognition (and this is not to mention the plausibility of the highly speculative mechanism—regarding cortical blood flow—that is invoked for these effects). Even if they did, however, anyone recommending the use of chewing gum in the classroom would be taking the results out of the laboratory context. No published studies, to our knowledge, have yet directly assessed the effects of chewing gum on learning either in a sample of children or in a school classroom setting - indeed, the one study that mimicked a classroom setting (though with adults; Allen et al., 2008) reported no effect. To make matters even more complicated, the brain activation caused by chewing in young adults is significantly
different from that in children (Onozuka et al., 2003); one cannot, then, expect data from studies using adult participants to generalise to children.

The practice of using chewing gum in the classroom to indirectly improve learning is not utterly implausible, but it is currently unproven. However, any benefits future studies may discover will in all likelihood be rather small. Teachers may have to weigh these against the practical, aesthetic, and hygiene aspects of encouraging children to chew more gum in the classroom.

The attraction of ‘alternative’ educational techniques

In this chapter, we have reviewed a variety of ‘alternative’ educational techniques, some with somewhat plausible theoretical rationales (e.g. Brain Training), and others that appear extremely unlikely to improve classroom learning, at least for the reasons their proponents assert (e.g. Brain Gym). To conclude, we will examine some proposed explanations for the attraction of these techniques to educators.

Unfortunately, the enthusiasm of proponents too frequently gives the impression that unproven or disproven educational techniques are effective, and sales pitches are often highly irresponsible. Unregulated websites and reckless reports in the media—frequently written before any proper trials have been conducted—only add to consumers’ misconceptions. The ‘alternative’ techniques discussed in this chapter, and those like them, seem particularly convincing and easy to sell. This seems to be due to a variety of factors. First, a ‘quick fix’ is extremely attractive. We would all like to be able to rapidly improve the memory and intelligence of our
children by getting them to perform a set of simple exercises, giving them fish oil pills, chewing gum, or glasses of water, or purchasing the latest computer game. The sad reality is that genuine learning takes long hours of hard work, and quick fixes are very rare indeed. Tempering our fervent desires with scientific evidence is not always easy.

Second, there is a common misunderstanding of the scientific method, and where the burden of proof lies. One particular line of argument—that the claims are not disproved and are ‘working theories’, so there is no problem in recommending their use—is made regularly to support controversial educational treatments (Hyatt, Stephenson & Carter, 2009). However, following this reasoning, one should be prepared to uncritically accept for classroom use literally any technique, whether pseudoscientific or otherwise, as long as there is no evidence either way of its efficacy. We believe the public, and especially children, deserve better than this illogical reasoning.

Third, it has been suggested that there is something specific about the beguiling mystery of the brain that seems to encourage credulity. The authors of two recent studies have argued that the public is particularly receptive to arguments that include ‘scientific’-seeming information about the brain (McCabe & Castel, 2008; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008). In these two studies, participants were more likely to rate scientific descriptions as more plausible when they were accompanied with a picture of the brain (McCabe & Castel, 2008) or with meaningless information about neuroscience (Weisberg et al., 2008). It is easy to see how the conclusions of these studies can apply to some of the ‘alternative’ techniques
discussed here, perhaps especially Brain Gym. It is certainly the case that teachers have been shown to have a fragile understanding of the neuroscience— and psychology— of learning (Dekker, Lee, Howard-Jones, & Jolles, 2012).

However, the findings of the two 2008 studies have been questioned by Farah and Hook (2013), who note that, in McCabe and Castel’s (2008) study, it was often the case that the descriptions accompanied by neuroscience pictures did provide more information to readers, and Weisberg et al.’s findings do not only apply to neuroscience per se. Farah and Hook (2013) also point to a variety of failed attempts to directly replicate the original results. It may be, then, that the reasons behind the popularity of techniques marketed with pseudo-neuroscientific rationales lie elsewhere.

It is certain that more of these techniques will appear in future. Thus, critical responsibilities for researchers are both to perform rigorous studies of the techniques (as is attempted in several chapters of this thesis), and then communicate the evidence, along with the skills to properly evaluate new claims, to the public. We echo the suggestion of Hyatt et al. (2009), that a well-designed, easily-accessible online resource - similar to the Cochrane Collaboration website, which provides reviews of medical research - should be set up to collect high-quality reviews of the evidence for controversial educational techniques. Such a website would be of great assistance to parents, teachers and other educators who are inundated by the claims of various proponents, and simply do not have the time to research the evidence for each individual technique.
In the context of special education, Mostert and Crockett (2000) provide a strong argument for educators being informed of the history of their discipline, and the many methods that had been adopted before being discredited by subsequent scientific investigation. Regrettably, a great many illustrative examples of the use of disproven techniques can be found in the annals of special education – one particularly tragic example being the use of ‘facilitated communication’ (Jacobson, Foxx, & Mulick, 2005; Offit, 2008). Whereas it is the case that children and families with special educational needs are particularly vulnerable to overblown and irresponsible claims about interventions, we would argue that Mostert and Crockett’s (2000) argument should be extended to all educators. The important contribution made by education to the lives of children must not be marred by the use of techniques that may raise hopes, but are at best unproven or at worst disproved. In the words of Mostert and Crockett (2000):

We [educators] cannot inspire confidence in our university colleagues, preservice teachers, or students and parents by hawking the baubles of every latest fad that washes through the popular media, latest in-service workshop, or educational journal. We must be able to discern the effective wheat from the useless chaff – discernment enhanced by attending to the history of our practices over time, what we have achieved, and perhaps most importantly, by closely attending to the historical mistakes we have made (p. 138).
The present thesis – research questions and aims

This final section provides a brief overview of the questions asked in the subsequent chapters of this thesis.

We begin, in Chapter 2, by testing an ‘alternative’ technique that—similarly to those reviewed in this chapter—has an extremely controversial evidence base: The use of coloured filters, which are proposed as a treatment for a disorder that also has controversial status as a diagnostic entity. This disorder, Irlen Syndrome, which causes reading difficulties, is said to be alleviated by a person-specific coloured tint, used either in spectacle lenses or as a plastic overlay which covers text to be read. We follow up on some participants—primary and high school-age children—from a previously-published study to assess whether the filters have an effect on reading speed or comprehension after one year of use.

Next, in Chapter 3, we test a technique related to Brain Gym, the first technique assessed in this Chapter: Brief exercise sessions at the start of a class, which are believed by many teachers to enhance attention and memory in primary school children. Previous results on the effectiveness of brief exercise techniques for raising cognition in the classroom are equivocal. This technique was compared, in two experiments, to relaxation sessions at the start of class, a technique also commonly used in the primary schools we visited, though thus far very poorly researched.
Having investigated the effect of relaxation before a lesson, in Chapter 4 we next tested the effect of relaxation sessions after learning some new information. This question relates to a current debate in memory research regarding memory consolidation, and follows some previous research demonstrating striking improvements in memory in adults after a very brief period of wakeful resting. This is tested in two experiments with primary school children, one in-class and one in more controlled conditions.

We then, in Chapter 5, report a simultaneous test of two educational techniques. The first is retrieval practice, which employs the ‘testing effect’, the well-replicated finding from the cognitive psychological literature that recall is improved more by being tested on previously-learned information than by restudying it. The second is mind mapping, a very popular, but—again, as with the techniques discussed in this chapter—poorly-evidenced technique. Two studies, both in highly ecologically valid classroom settings, tested these two techniques in concert with one another, and investigated whether a combination of the techniques would provide more benefit to learning than either one alone.

A final empirical chapter, Chapter 6, describes a different kind of analysis, this time examining the outcomes, rather than the process, of education. In a large, nationally-representative, publicly-available dataset, we test whether skills emphasized in early education—mathematics and reading—predict socioeconomic status later in life, even after controlling for a wide variety of common confounds, such as general intelligence and social class of origin. Such a finding would
emphasize the importance of understanding better ways to teach basic academic skills.

The thesis ends with a summary of some possible future directions for the union of psychology and education, considering the potential contributions of neuroscience, social psychology, and differential psychology to educational outcomes.
Irlen coloured filters in the classroom: A one-year follow-up†

Abstract

Coloured filters are used to treat Irlen Syndrome, a controversial disorder posited to be the cause of a substantial proportion of reading difficulties. In a previous study, we found that Irlen coloured filters do not produce any short-term alleviation of reading difficulties in schoolchildren aged 7-12. Here, we tested whether coloured filters show benefits with longer-term use, in a subset of the original sample. We measured reading rate with and without filters in 18 children diagnosed with Irlen Syndrome, who had been using the filters for one year, and compared the progression of their reading ability across the year against 10 poor-reader control children. The Irlen-treatment group did not read any faster when using their coloured filter, and showed no disproportionate gain in reading progress across the year compared to controls. We conclude that Irlen filters do not benefit reading, even after one year of use.
CHAPTER 2: IRLEN COLOURED FILTERS IN THE CLASSROOM

Introduction

Irlen Syndrome (IS), also known as Meares-Irlen Syndrome, Scotopic Sensitivity Syndrome or visual stress, is a controversial diagnostic entity which purportedly causes visual distortions and illusions when the person views text or other high-contrast patterns (Irlen, 1991; Wilkins, 2003). The Irlen Institute posits that these symptoms are often the cause of reading difficulties in up to 46% of individuals with ‘reading problems, dyslexia, and learning difficulties’ (Perceptual Development Corporation, 1998), but also that they can be alleviated by the use of individually-prescribed coloured filters (Irlen, 2010). The coloured filter treatment is not marketed as a cure for reading difficulties, but is believed to remove a barrier to reading development (Irlen, 2010). These filters, either in the form of tinted lenses or coloured plastic overlays, are used worldwide, often in school classrooms (Hyatt, Stephenson & Carter, 2009, p. 321), and regularly receive mass media coverage.

Considerable controversy surrounds the efficacy of this treatment; three recent narrative reviews (American Academy of Pediatrics, 2009; Hyatt et al., 2009; Royal College of Ophthalmologists, n.d.) and one systematic review (Albon et al., 2008) have concluded that it should not be recommended for individuals with reading difficulties until more rigorous research shows positive effects.

Our recent study (Ritchie, Della Sala, & McIntosh, 2011) drew similar conclusions. We administered reading tests with and without coloured overlays to 61 primary school children aged 7-12 years, 77% of whom had been diagnosed with IS by an Irlen Institute Diagnostician. Importantly, and differently to all previous work in this area, the children diagnosed with Irlen Syndrome were not informed of the
colour of their prescribed overlay, or even their diagnostic status, before testing. The experimenter was also not informed of the child’s prescribed colour or diagnosis before testing. This design was implemented in an attempt to reduce placebo effects to the greatest extent possible. Under these masked conditions, the overlays failed to produce any significant increase in reading rate, as measured by the Wilkins Rate of Reading Test (WRRT), or global reading ability, as measured on the Gray Oral Reading Test (GORT). We concluded that Irlen coloured filters do not alleviate reading difficulties.

Our previous study, like most previous work in this area, focused on the immediate effects of coloured filters (though see Noble, Orton, Irlen, & Robinson, 2004; Robinson & Foreman, 1999a, 1999b). We compared reading with and without filters at a single time point in poor-reader children who had been prescribed the filters very recently. It is clear that the coloured filter theory predicts an immediate benefit at this stage, but a further key prediction regards the longer-term benefits. Since the filters are intended as a long-term aid, to be used on a continuing basis, they should facilitate reading even after extended periods of use. More crucially, since the filters purportedly remove a prior barrier to reading development, we should expect the children using them to begin to catch-up with their peers (those not diagnosed with Irlen Syndrome) in terms of reading development; they should therefore show disproportionate gains in reading fluency and comprehension within the first year of treatment.

This study aimed to assess reading outcomes one year after a school-level intervention by the Irlen Institute, and to provide practical data on what educators
might expect after such an intervention. Only two previous experiments have examined Irlen filters over the longer-term. Noble et al. (2004) reported large gains in reading ability over three months for two groups of 31 and 40 children using coloured filters (see 'Statistical Power' section of Results, below, for more details). More mixed results (improved accuracy and comprehension, but not rate of reading or reading strategy) were reported for a sample of 38 children using coloured filters over 20 months (Robinson & Foreman, 1999a, 1999b). However, this study has been criticised for various methodological problems such as failure to include an eye examination as part of the study and inappropriate statistical analyses (Albon et al., 2008; Hyatt et al., 2009). In addition, the 95% confidence interval for the reading accuracy results reported by Robinson and Foreman (1999b) crosses zero, potentially indicating a lack of effect (Albon et al., 2008). There is therefore a clear need for more evidence on this issue.

In the present chapter, we followed up as many children as possible from our previous sample (Ritchie et al., 2011) one year after their IS diagnosis and coloured filter prescription. Little evidence exists on the continuation rate of Irlen filters; this is in contrast the information available for Intuitive Filters (Wilkins, 1994), a separate, competing system that involves diagnosis using an “Intuitive Colorimeter” (Wilkins, Nimmo-Smith, & Jansons, 1992), an instrument that allows patients to select an exact coloured tint that may not be catered-for with the limited options available during the Irlen diagnosis session. Wilkins, Lewis, Smith, & Rowland (2001) showed that over half of the children diagnosed using this instrument and prescribed filters from this system continued using their Intuitive overlays for up to three months; it is unclear,
given the differences between the diagnostic processes, whether children will show differential continuation rates with Irlen filters.

Here, we first ascertained the number of children who were still using their prescribed overlay after one year. Next, we evaluated reading rate with and without coloured filters, and global reading ability, comparing poor-reader children who had been diagnosed with IS and were still using their coloured filter with poor readers who had not been diagnosed with IS and had never used coloured filters.

If coloured filters benefit reading in children with a diagnosis of IS, then these children should read more fluently when using their filter than when using a filter of a different colour, or when using no filter. Finally, we compared the progression of reading abilities of children diagnosed with IS across one year. If the coloured filters have removed a prior perceptual “barrier” (Irlen, 2010) to reading development, the children that continued to use them should show a greater improvement in reading development across the year than poor-reader children without IS, for whom no such barrier has been removed.

**Method**

**Participants**

*Original sample.* The recruitment sample was the group of 61 children who had taken part in our previous study (Ritchie et al., 2011). All had been identified by their teachers as having below-average reading ability. Children with a diagnosis of autism were excluded from the original sample, and thus were not included in the
Treatment continuation. For the children diagnosed with IS, we obtained information from the school and from parents/guardians about whether the child was still using a coloured filter. 22 children with IS were still using a form of coloured filter treatment after one year. In 14 cases, this was the coloured overlay prescribed originally, whilst 8 children had progressed to coloured-lensed spectacles. An analysis of the records from the original study showed that, of the 47 children who had been diagnosed with IS, those that had discontinued treatment did not differ significantly from those that had continued in terms of mean age (9.60 vs. 9.68 years; \( t(45) = .79, p = .81 \)), mean Mini Mental State Examination score (23.64 vs. 24.27; \( t(45) = .54, p = .59 \)), WRRT performance with a colourless overlay (81.28 vs. 77.83; \( t(38.97) = .43, p = .67 \)), GORT oral reading quotient (ORQ; 78.50 vs. 76.41; \( t(44) = .52, p = .61 \)), percentage of orthoptic tests failed at examination in the original study (13.71% vs. 17.43%; \( t(44) = .91, p = .37 \)), or overlay benefit at the original study (1.00 vs. .90 extra words read correctly with the prescribed overlay than the colourless overlay; \( t(45) = .03, p = .98 \)). That is, in terms of the measures available, there was no obvious pattern to distinguish the IS children that continued treatment from those that did not.

During the study period, we were able to ask 13 of these latter children about why they had discontinued treatment. Three reported that they had found that the overlay made reading more difficult; two had visited an optician and had their vision corrected, which they saw as a replacement for the overlays; one had lost the overlays and not sought replacements; one could not remember having received an overlay;
and six did not give a specific reason. Due to the variation in responses, it is difficult to know whether these children should be regarded as children for whom the filter did not work, children for whom the filter worked only initially, children who lacked motivation to persist with the treatment, or (most probably) some mixture of these and other descriptors. Given the uncertain status, and likely heterogeneity of this group, we decided not to include ‘treatment-discontinued’ children in the present study’s behavioral analyses.

Follow-up groups. Of the 22 IS children who had continued treatment, 18 were available for and completed follow-up testing (5 using lenses, 13 using overlays). Of the 13 children not diagnosed with IS, 10 were available for and completed testing. The follow-up testing was therefore completed by two groups, an Irlen-treatment group (n=18) and a Non-Irlen group (n=10). An analysis of the records from the original study showed that the Irlen-treatment group was older than the Non-Irlen group (9.78 vs. 8.70 years; t(26) = 2.28, p = .03), but the groups did not differ on MMSE score (23.61 vs. 24.20; t(26) = .59, p = .70), or colourless WRRT performance (87.90 vs. 79.39; t(26) = .24, p = .24). The Irlen-treatment group did have a lower GORT ORQ than the Non-Irlen group (74.83 vs. 91.00; t(26) = 3.63, p = .001).

Statistical Power

As far as we are aware, the most recent long-term study of the effects of Irlen filters is that of Noble et al. (2004), who reported that in their two groups of children, the filters were associated with an increase in reading grade equivalents of an average
of 1 year, 5 months in a period of 3 months. After this period of time, the authors suggested that ‘any further rate of development may [have been] at normal grade expectations rather than at accelerated levels’ (p. 21). While insufficient information was included by Noble et al. (2004) to calculate a precise effect size, on the basis of their results we would expect the effect of the overlays on reading to be moderate-to-large. Power analysis using G*Power 3.0 (Faul, Erdfeler, Lang, & Buchner, 2007) indicated that, with our 18 Irlen-diagnosed participants, we had 80% power to detect any cross-year effect above approximately $d = .61$.

**Reading measures**

The *Wilkins Rate of Reading Test* (WRRT; Wilkins, Jeanes, Pumfrey, & Laskier, 1996) consists of four lists of 15 familiar words arranged in 10 lines, each with a different random word order, in closely-spaced type and a small font size. Each 150-word test is read for one minute, and any deviations from the text are recorded. As in our previous study, the larger-font size form of the WRRT was used (the font used was Times New Roman, 12 point). Two extra word lists were created to allow for the procedure described below. A ‘practice’ sheet, read for 30 seconds before first administration of the task, was also created.

The *Gray Oral Reading Test* (GORT; Wiederholt & Bryant, 2001) is a global reading measure in which participants read stories aloud while being assessed for accuracy and fluency, then answer comprehension questions. This measure was included because the WRRT, with its random word list, does not assess many important aspects of reading ability. An overall age-standardised ‘Oral Reading
Quotient’ is then calculated. Form B of the GORT was used in our original experiment, so the parallel Form A was used for this follow-up. During GORT testing, Irlen-treatment children used their usual coloured filter (overlay or lenses), and Non-Irlen children used no filter.

Procedure

The children were tested individually in a quiet room at their school (Newark Primary School, Port Glasgow, or Port Glasgow High School).

For the WRRT, an ABCCBA design identical to that of our previous study (Ritchie et al., 2011) was used, with each participant reading twice in each of three conditions: prescribed filter, non-prescribed filter, and colourless filter. For the Non-Irlen group we used the mock ‘prescribed’ colour from the original study (these had been chosen to match approximately the frequency of colours prescribed to children in the Irlen group). The ‘non-prescribed’ filters were determined by the same fixed pairings of filter colours as used in the original study - chosen, where possible, from the complementary end of the spectrum. The Irlen-treatment children who had progressed to coloured lenses used these as their prescribed filter; all other filters used were plastic overlays. The order of the three conditions within the ABCCBA design was cycled from participant to participant within each group.

For the GORT, each participant read the stories in their favored filter condition. That is, all Irlen-treatment children used their usual filter (overlay or lenses), while Non-Irlen children used no filter.
After completing both reading tests, the children were debriefed – the Irlen-treatment group were asked what difference they felt their filter made to their reading, while the Non-Irlen group were simply thanked for their participation. After the study was complete, class teachers were surveyed about their pupils' coloured filter use; for each child, they were asked whether, in the period after their initial diagnosis, the child used their filter ‘always’, ‘regularly’, ‘sometimes’, or ‘not at all’.

Results

Usage

Information from teachers on overlay/filter usage was available for 17 of the 18 Irlen Syndrome children. This indicated that, after first being diagnosed with Irlen Syndrome and prescribed a filter, 4 of the children 'always' used their filter in class (three of these were the children who used coloured lenses). 8 used it 'regularly' (these included the remaining two coloured lens-users), while the remaining 5 used their filter 'sometimes'. This indicates a good general level of compliance with the treatment.

WRRT results

Table 2.1 shows the scores for the Irlen-treatment and Non-Irlen groups in each filter condition (prescribed filter, non-prescribed overlay, colourless overlay) in this follow-up study. Given that the children in the Irlen-treatment group had been
using the filters for a full year, suggesting that they felt that the treatment improved
their reading, we predicted that they would read significantly faster in the prescribed
filter condition than in the other conditions. A 3 x 2 ANOVA tested for differences
between the three WRRT conditions for the two groups (Irlen-treatment and Non-
Irlen). There was no main effect of condition ($F(2, 52) = .34, p = .71$), or of group
($F(1, 26) = 1.49, p = .23$), and no significant condition $\times$ group interaction ($F(2, 52) =
2.08, p = .14$). The filters made no measurable difference to reading rate in either
group.

\textit{Table 2.1} Mean 2011 WRRT scores (words per minute) for each overlay condition by
group, with SDs in parentheses.

\begin{tabular}{|l|c|c|c|}
\hline
Group & Prescribed & Non- & Colourless \\
& & Prescribed & \\
\hline
Irlen- & 89.33 (20.28) & 86.19 (19.93) & 82.83 (20.04) \\
treatment & & & \\
Non-Irlen & 93.45 (16.78) & 92.90 (13.45) & 96.65 (14.89) \\
Total & 90.80 (18.89) & 88.59 (17.92) & 87.77 (19.30) \\
\hline
\end{tabular}

This overall lack of filter benefit on reading rate replicates, in this sub-sample,
the findings made in our original study at the time of filter prescription. Figure 2.1
shows the filter benefit (the number of words per minute more that were read with the
prescribed filter than with the colourless filter) for each child both for the original
study and at follow-up. The filter benefit scores are scattered evenly around zero for
the original study, and also for the follow-up, with the exception of two Irlen-
treatment children who showed a large positive filter benefit at follow-up. These two
children are responsible for driving the numerical trend toward higher prescribed filter
scores apparent in Table 2.1. However, the overall group effect at follow-up remains
non-significant, and almost as many children read more slowly with the filters as read
more quickly, indicating that the clinical significance of the filters is low. Moreover, there is no significant correlation between the filter effect in 2010 and in 2011 ($\rho(26) = .16, p = .40$), indicating a very low test-retest reliability of any filter effect.

*Figure 2.1.* Scatterplot of prescribed filter benefit on WRRT for original study and follow-up.

To compare the progression of reading rate between the two groups, across the year, we analysed WRRT scores from the Irlen-treatment group reading in the prescribed filter condition and the Non-Irlen groups reading in the colourless overlay condition (we would expect the groups to be most comfortable in these conditions). Using a 2 x 2 ANOVA to test for effects of group (Irlen-treatment, Non-Irlen) and time (first study, follow-up), we found only a significant main effect of year ($F(1,26)$
= 12.68, \( p = .001, \eta_p^2 = .33 \), confirming that reading rate increased between study and follow-up (null results were also obtained using similar analyses between all filter conditions). There was no significant main effect of group \( (F(1,26) = 1.45, p = .24) \), and no significant year \( \times \) group interaction \( (F(1,26) = .12, p = .74) \), indicating that the improvement in reading rate across the year was similar between the two groups. This is not consistent with a prior obstacle to reading progress having been removed for the Irlen-treatment group. These results are illustrated in Figure 2.2a.

Figure 2.2. Optimal (chosen)-condition a) WRRT scores and b) GORT scores, for original study and follow-up.

GORT results

The mean GORT Fluency, Comprehension, and overall Oral Reading Quotient (ORQ) scores, are shown for each group at each time in Table 2.2. The overall ORQ scores are also plotted in Figure 2.2b. A 2 x 2 ANOVA showed a significant main
effect of group \((F(1,26) = 14.93, p = .001, \eta_p^2 = .37)\), and of time \((F(1,26) = 7.23, p = .01, \eta_p^2 = .22)\), indicating that the Irlen-treatment group ORQ was below that of the Non-Irlen group, and that ORQ decreased in both groups during the year. Crucially, no significant group by time interaction was found \((F(1,26) = .14, p = .71)\), indicating that the groups’ reading performances changed comparably across the year. This is not consistent with a prior obstacle to reading progress having been removed for the Irlen-treatment group. Finally, a correlation between the ORQ scores in the first study and at follow up was strong and highly significant, confirming the test-retest reliability of this reading measure \((r(26) = .87, p < .001)\).

Table 2.2. Mean GORT scores for each group for both years (Flu. = Fluency; Comp. = Comprehension, ORQ = Oral Reading Quotient), with SDs in parentheses.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Irlen-treatment</td>
<td>3.94 (2.41)</td>
<td>7.56 (2.43)</td>
<td>7.44 (1.72)</td>
<td>74.83 (11.87)</td>
<td>71.67 (10.09)</td>
<td></td>
</tr>
<tr>
<td>Non-Irlen</td>
<td>7.70 (2.71)</td>
<td>9.30 (1.57)</td>
<td>8.40 (1.90)</td>
<td>91.00 (10.10)</td>
<td>86.80 (10.88)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.29 (3.08)</td>
<td>8.18 (2.29)</td>
<td>7.79 (1.81)</td>
<td>80.61 (13.60)</td>
<td>77.07 (12.56)</td>
<td></td>
</tr>
</tbody>
</table>

One potential complication in comparing the GORT scores across time is that, in the first study, half of the children were administered the GORT with a coloured overlay, and half without, as dictated by the between-groups design of that study. For the present Irlen-treatment group, 8 had used their prescribed overlay at the first assessment, and 10 had not. It should be noted that any bias caused by this factor would have been to the advantage of the Irlen-treatment group at follow-up, as all were using their prescribed filter at follow-up, whilst fewer than half had used their prescribed filter in the original study. Given that the Irlen-treatment group did not perform differentially better at follow-up, this potential source of bias seems to have
had no relevant influence. Nonetheless, in order to double-check, we ran an additional repeated-measures ANOVA, confined to the Irlen-treatment group, to compare the effects of time across the subgroups who had and had not used their prescribed filter at the first GORT assessment. This confirmed the main effect of time \( (F(1,16) = 5.30, \ p < .05) \), reflecting the reduction in GORT ORQ across the year, but found no main effect of subgroup \( (F(1,16) = .31, \ p = .59) \), nor time by subgroup interaction \( (F(1,16) = 3.64, \ p = .08) \). This provides further reassurance that the variation in the original GORT testing condition did not distort the present results.

**Debrief**

At debrief, all 18 of the Irlen-treatment group children reported positive subjective effects of the overlays. 14 children reported of a reduction of specific purported IS symptoms – for example, ‘the words stopped moving around on the page’, ‘the page is less blurry’, ‘the overlay stops the white page from hurting my eyes’, and ‘the page feels zoomed in’ – while 4 reported general statements of preference – for example, ‘I prefer reading with this overlay’, and ‘this overlay makes my reading much better’.

**Discussion**

This follow-up study examined the effects of Irlen coloured filters after one year of use in the same sample tested in our original study (Ritchie et al., 2011). Using the same reading tests - the WRRT and the GORT – we reassessed the children who were still using their coloured overlay or coloured lenses, and compared them with those who were not diagnosed with IS. To our knowledge, this is the first one-
year follow-up study on the Irlen treatment. On the WRRT, no significant differences were found between coloured filters and colourless overlays, and no between-group differences in reading progress across the year were found. On the GORT, both groups significantly decreased in skill across the year, and again, no differences in the trajectory between the groups were found.

From a practical perspective, it is useful to note that, according to our results, around 53% of children diagnosed with IS and prescribed a coloured filter will discontinue the treatment within one year. Continuation rates for Irlen overlays have not been previously reported, but as noted above, research using Intuitive overlays—a comparable treatment—in a sample with normally-distributed reading ability has found very similar continuation rates (Wilkins et al, 2001).

The most immediate and parsimonious explanation of the WRRT results is that the Irlen overlays have made no difference; both groups improved their reading rate to a similar degree. This was despite all of the Irlen-treatment group children indicating they felt a subjective benefit when reading with overlays. The placebo effects that might be expected to accompany such beliefs were evidently not strong enough to improve reading scores on either of the tests while using the preferred overlay.

Only two children had large ‘overlay difference’ scores at follow-up, as shown in Figure 2.1; the outlying individual with the largest score stopped reading in the colourless overlay condition after an average of 26 seconds, stating that her eyes were tired. This was not the case in the original study, when she was able to complete
reading for a full minute in both iterations of the colourless overlay condition. In addition, as can be seen in Figure 2.1, this individual read substantially faster in the colourless overlay condition in the original study (i.e. had a negative overlay effect score), suggesting her performance was highly variable.

Since the GORT ORQ results are age-standardised, the ORQ score should remain constant across time, provided that the children are improving in line with the standardised norms. In fact, while the original and follow-up ORQ scores were highly correlated, they significantly decreased in both groups across the year. This may reflect a genuine loss of global reading ability. As noted above, given that the Irlen treatment is intended to remove a “barrier” to reading ability (Irlen, 2010), one would expect that the individuals with IS who had been using the Irlen treatment for a full year would no longer be disrupted by the visual symptoms of the disorder, and would make concomitant gains in their scores disproportionate to those of the Non-Irlen group. That this has not occurred, and the two groups changed parallel to one another (see Figure 2.2b), suggests that the overlays have made no measurable difference to the global reading ability of the IS-diagnosed children in our sample.

Irlen (2010) has claimed that coloured filters are often life-changing, resulting in greatly improved reading ability and thereby improved educational attainment. Such claims predict large effect sizes for the treatment. While the present study had a relatively small sample size, it had adequate power to detect medium-to-large effects of the order required for practical or educational significance; these effects should have been particularly large in the within-subject filter comparison. Smaller effects
might have remained undetected by our study, but we do not believe these would have had important practical significance (see Ferguson, 2009).

Our study has a number of limitations beyond its small sample size. Firstly, the control group of poor readers not diagnosed with Irlen Syndrome is not the ideal comparison, which would be a group of children diagnosed with Irlen Syndrome who did not receive Irlen treatment. Unfortunately, no such sample was available. Comparing the groups of children with and without Irlen Syndrome as we did here introduces an extra confound to the analyses, since these groups may differ in the developmental trajectory of their reading ability. Children suffering from the syndrome who received treatment may have improved more in their reading across time (or, in the case of the ORQ results, decreased less) than similar individuals who did not receive treatment, and our study was unable to test this possibility.

Secondly, the small number of controls limits the generalizability of the between-groups comparison. However, the similarities in reading trajectories between the Irlen-treatment and control groups (see Figure 2.2) are nevertheless striking. Thirdly, Irlen (2010) claims that coloured filters can have effects other than improving reading ability, such as a reduction in headaches or visual discomfort. While the children in the present study were included on the basis of their poor reading ability, it is possible that the filters had other effects across the year that the study did not measure. Finally, it should be noted that conclusions drawn from our results are limited to the Irlen Institute’s coloured filter method; other methods, such as the Intuitive system (Wilkins, 1994) exist, and utilize different diagnostic techniques. On the other hand, our study has real-world relevance and value, being a
direct assessment of a school-level intervention for reading, as conducted by the Irlen Institute.

Conclusion

Data from this study suggest that the filters provided no significant benefit to reading ability, at least in the majority of cases in our modest sample, even after one year of continued use. These results, combined with those from our previous study (Ritchie et al., 2011), imply that the Irlen coloured filter treatment does not have a statistically or clinically significant effect on reading fluency or comprehension, either in the short or longer term.
Effects of In-Class Exercise and Relaxation Sessions on Children’s Preparedness for Learning‡

‡ Both experiments reported in this chapter were designed, performed, and analyzed with the assistance of my supervisors. The first experiment was presented as: Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2012). The effect of in-class exercise and relaxation sessions on children’s preparedness for learning. Poster presented at the Meeting of the European Association for Research on Learning and Instruction: Special Interest Group 22 – Neuroscience and Education, Institute of Education, London.
Abstract

Two experiments examined the effects of two different in-class activities on preparedness for learning in primary school children. We compared the effects of exercise and relaxation, both in short, under-five-minute sessions performed prior to tests of attention and memory. In the first experiment, we unexpectedly found a detrimental effect of exercise on an attention test thirty minutes later. In the second experiment, this effect was not replicated, but a possible beneficial effect of exercise on memory was found. No effects of relaxation were found in either experiment. Significant effects in both experiments were very small in size (in Experiment 3.1, exercise reduced attention by around one sixth of a standard deviation; in Experiment 3.2, exercise improved memory by around one eighth of a standard deviation). We conclude that, whereas exercise and relaxation may be generally beneficial for physical health and mental wellbeing, teachers using these techniques in class should not expect to see practically significant benefits to their students’ preparedness for learning.
Introduction

Effects of Exercise on Cognitive Performance

A large number of studies attest to the important health benefits of regular physical exercise undertaken by children (Janssen & LeBlanc, 2010; Strong et al., 2005). In addition, incorporating ‘recess’ sessions involving play and physical activity into the school day has been shown to reduce behaviours that have an indirect negative effect on learning, such as misbehaviour in the classroom (Barros, Silver, & Stein, 2009) and fidgeting and listlessness (Jarrett et al., 1998; Pellegrini & Davis, 1993). Other indirect benefits may come from exercise’s positive effect on self-esteem (Tremblay, Inman, & Willms, 2000). However, Hillman, Erikson, and Kramer (2008) suggest that aerobic exercise may not just benefit health and learning-conducive behavior, but may directly improve children’s academic achievement via its effects on cognitive processes such as executive functions, which are known to be associated with academic performance (e.g. Mazzocco & Kover, 2007).

Hillman et al.’s (2008) suggestion is supported by studies such as that of Coe, Pivarnik, Womack, Reeves and Malina (2006), who found a positive correlation between amount of vigorous aerobic exercise performed and grade level in various academic subjects in a group of 214 eleven year-old children. In addition, Chomitz et al. (2009) showed a positive relationship between physical fitness level and academic achievement. Reviews show that the majority of studies have found a positive relationship between both in-class and separate-session exercise activities and various measures of cognitive performance such as perceptual-motor, memory, verbal and mathematical reasoning, and attention tasks in children (Sibley & Etnier, 2003; St-
Louis-Deschêne & Ellemberg, 2010; Tomporowski, Davis, Miller, & Naglieri, 2008; Tomporowski, Lambourne, & Okumura, 2011; Voss, Nagamatsu, Liu-Ambrose, & Kramer, 2011), but it should be noted that some studies have found no relationship, or a negative relationship, between these variables (e.g. Daley & Ryan, 2000; Tremblay, Inman, & Willms, 2000). As with all such correlational research, however, questions of causality cannot be resolved by these studies; experimental research is therefore required to clarify whether exercise leads to improved or reduced academic achievement.

In Chapter 1 of this thesis, we discussed a number of in-class exercise programmes that have appeared, most notably perceptual-motor programmes such as Brain Gym (Dennison & Dennison, 2010). It is unlikely that such programmes induce levels of aerobic exercise to confer any cognitive benefits, since physical activity is not their major focus (Howard-Jones, 2010). In any case, as noted in Chapter 1, very little research has been performed on these techniques, despite their popularity. For this reason, the following literature review will focus only on interventions that have been adequately experimentally tested, and that do not, like Brain Gym, have deeply implausible theoretical rationales.

Mahar et al. (2006) studied an in-class exercise programme, ‘Energizers’, in which teachers of third- and fourth-grade pupils (aged 8-10) ran a 10-minute session that combined aerobic exercise with simple games. ‘On-task behaviour’ (where the child focuses on their work rather than distractors for a 5-minute period) was measured by raters in the classroom at baseline in the 30 minutes before, and then again in the 30 minutes after the Energizers session, and general physical activity
throughout the day was measured by having the children wear pedometers. On-task behaviour increased after the Energizer session, and physical activity was higher on days that included Energizer sessions; the authors concluded that Energizers activities “are recommended for teachers who may want to increase physical activity and/or on-task behavior in their students” (Mahar et al., 2006, p. 2093). In a sample of older children (13-14 years), Kubesch et al. (2009) showed that a 30-minute physical education session, but not a 5-minute in-class ‘movement break’, improved performance on a computerized executive function task, though unfortunately they did not report the length of time between exercise and testing.

Two more recent studies have assessed the effects of in-class exercise sessions on children’s attention using a battery of neuropsychological tests that, in a novel method, were administered to classes en masse. In a counterbalanced crossover design, Hill et al. (2010) administered the tests—consisting of mental tracking tasks such as serial paced addition—to 1224 Primary School children (8-12 years) over two weeks. On each of the five days of week 1, the children were administered a different neuropsychological test one hour after either a 10-15 minute exercise session or a control condition with no exercise. In week 2, the groups crossed over to the other condition.

Results showed an interaction: There were no significant differences in the overall performance of the groups in week 1, but a small effect of exercise in week 2 (we calculate Cohen’s $d = .11$), over and above the practice effects found in the group in the non-exercise condition. This result is somewhat counter-intuitive, because a main effect of exercise (i.e. a full crossover interaction of group by week) would be
expected on the theory that exercise enhances cognitive processes. The authors propose that the observed pattern is explained by the novelty of the experiment wearing off in the second week. They suggest that the children in the non-exercise condition of week 1 would have had added arousal due to their inclusion in a scientific study, and would thus have behaved more attentively than normal during the cognitive tasks, masking the effects of exercise in week 1. Only once these novelty effects were extinguished, by week 2, would the benefits of exercise have become apparent.

These findings were replicated in a sample of 552 children by Hill et al. (2011); they were also shown not to be moderated by body mass index or level of ADHD symptoms. Hill et al. (2011) add to the explanation given by Hill et al. (2010) for the lack of effect of exercise in week 1: In this week, they suggest, the children were able to learn how to perform the tasks, and the expression of this learning was enhanced by exercise in the second week. However, it is theoretically possible to interpret their results as reflecting a detrimental effect of exercise on task learning in week 1. On this interpretation, exercise would actually have impeded task learning in week 1, causing a performance decrement for this group in week 2, when tested in the no-exercise condition, thus creating the apparent benefit of exercise in week 2 as an artefact. Unfortunately, neither Hill et al. (2010) nor Hill et al. (2011), included a control group who performed no exercise in either week, in order to exclude this alternative explanation of their results, which they did not consider.

Nevertheless, the results from studies of in-class exercise sessions are overall suggestive of small positive effects of exercise on cognitive performance in children
Potential mechanistic explanations of this effect come from studies of adults as well as non-human animals, where molecules such as brain-derived neurotrophic factor (BDNF), which is required for long-term potentiation and thus learning, are shown to increase during exercise (Hillman et al., 2008; Vaynman, Ying, & Gomez-Pinilla, 2004).

**Effects of Relaxation on Cognitive Performance**

Relaxation sessions are a second strategy often used by teachers to ‘set the scene’ for learning. While the exercise sessions described above are intended to increase arousal and thus allow children to deploy more attentional resources, the aim of relaxation sessions is to reduce arousal at times of the day where children may be easily distractible.

School-based relaxation activities are sometimes implemented in an attempt to reduce stress-related problems such as test anxiety or headache, and not to prepare all children for learning per se (e.g. King, Ollendick, Murphy, & Molloy, 1998; Rasid & Parish, 1998). In general, programmes teaching children relaxation techniques using long-term (i.e. over weeks or even full school years) interventions appear to have some success in reducing stress and improving coping (Kraaga, Zeegers, Koka, Hosmand, & Abu-Saad, 2006), and in some cases preventing a fall in academic achievement (Thomson, Greibstein, & Kulenschmidt, 1980). However, the direct effects of an untrained in-class relaxation session on children’s preparedness for learning in the subsequent lesson have not, to our knowledge, been directly tested.
The effects of relaxation on cognition have sometimes been assessed only indirectly, that is, as control conditions in experiments focusing on other techniques. Rauscher et al. (1993), for example, found no effect of relaxation (induced by ‘relaxing’ music) on students’ cognitive performance while using it as a control condition in an experiment on the ‘Mozart Effect’ (see Gray & Della Sala, 2007), where the music is played during the cognitive task. Only one experiment with children, to our knowledge, has directly assessed the effects of in-class relaxation on cognitive outcomes: Hallam, Price, and Katsarou (2002) found positive effects of relaxing music—again, played during the task—compared to no music, on the speed, but not accuracy, of the mathematical performance of 30 children aged 10-12 \((d = .98)\). In a second experiment with 31 children, accuracy on a memory task was greater in a group who listened to relaxing music compared to those with no music \((d = 1.38)\), and performance while listening to ‘aggressive’, atonal jazz was poorer than in the no music condition. This study had a number of methodological problems, however – no baseline data were collected for the music-listening groups, and it is thus unclear whether there was initial group equivalence. A stronger methodology would have been a crossover design, with each group being assessed in each condition.

Theories relating to arousal, mood, and cognitive performance appear at first to be inconsistent: Hallam et al. (2002) showed that lowering children’s arousal level with relaxing music improves performance, while some ‘Mozart Effect’ studies suggest that increasing arousal level with complex music will result in the same effect (Thompson, Schellenberg, & Husain, 2001). However, the particular timing of the intervention may be of great importance. Hallam et al. (2002) suggest, for example,
that calming music could be played to children at times of predominantly high
arousal, such as their return from lunch break, to reduce arousal and improve
concentration. Such a practice is common in the schools we visited during the present
study.

The suggestion from Hallam et al. (2002) suggests that, conversely, relaxation
periods at times of low arousal, such as early in the morning, would be detrimental to
cognitive performance. In addition to investigating the effects of exercise, then, the
present study investigated this possibility by comparing between classes who had
undergone relaxation sessions in the morning and following lunch break.

**Experiment 3.1**

In Experiment 3.1, we examined the effects of short in-class exercise and
relaxation activities on visual attention, memory, and mental rotation tasks in a
Primary School at which teachers routinely include such activities to prepare pupils
for learning. The teachers at the school in the present experiment (Towerbank Primary
School, Edinburgh) often included an exercise session (‘Energizers’) at the start of the
school day, and Relaxation when classes returned from their lunch break. However, to
assess the differential effects of time and condition, the classes in the present study
were split into an AM group and a PM group, and randomly assigned to perform
Exercise, Relaxation, or no activity at a set time in each of three consecutive weeks.

We included a practice/pilot session in Experiment 3.1. The hypothesis of Hill
et al. (2011) would suggest that, since the children in our experiment had seen the test
materials prior to the experimental sessions, their knowledge of the tasks should be activated particularly strongly by the exercise session, and so we should see particularly high scores in the exercise condition. On the other hand, if exercise has any detrimental effect on task learning (another possible interpretation of Hill et al.’s data), then this preliminary practice session should eliminate this negative effect. In either case, the inclusion of preliminary practice should maximize our ability to detect positive benefits of exercise, in addition to reducing the influence of initial practice effects. As regards time of day, Hill et al. (2010) suggest that exercise should have most benefits when children are most tired. It is unclear whether the children would be most tired early in the morning or after lunch break; we thus did not make a directional prediction for the effects of time.

If relaxation sessions have an effect on preparedness for learning, we would expect it to be greatest for the PM-group classes. After children have returned from an active lunch break, reducing their level of arousal should improve their performance (Hallam et al., 2002). It is not clear whether performing an exercise session in the afternoon or a relaxation session in the morning will be detrimental to performance.

To assess preparedness for learning, we chose to test three abilities: Visual attention (assessed using a novel “Lollipops Test”), working memory, (assessed using the Visual Patterns Test; Della Sala, Gray, Baddeley, & Wilson, 1997) and mental rotation (assessed using a novel “Square Completion Test”). Some justification of these choices is required. Visual attention is of critical importance for reading, a modality used routinely in the teaching of new information in the classroom; poor performance on visual search tasks is associated with lower score on reading tests
(e.g. Franceschini, Gori, Ruffino, Pedrolli, & Faccetti, 2012; Mayer, Mayringer, & Landerl, 1998). As noted above, measures of working memory have previously been shown to correlate positively with later mathematics attainment (Mazzocco & Kover, 2007), and the Visual Patterns Test has been used in previous studies assessing educational performance. For instance, Gathercole, Alloway, Willis, and Adams (2006) found that visuospatial short-term memory, partly measured using the Visual Patterns Test, was positively associated with performance on reading and mathematics tests. Finally, whereas mental rotation is not a commonly used skill in educational contexts, it is a task that requires concentration, a skill that itself may be influenced by having recently exercised or relaxed. Importantly, this argument could be made for all three tasks: In a repeated-measures design, better performance on these tasks in some conditions and not others would indicate higher levels of concentration and focus on the task at hand.

Nonetheless, it could be argued that there are two problems with these measures. First, they do not cover the full range of cognitive functions required for optimal classroom performance. Whereas a full neuropsychological battery administered to each child (or a range of tasks all assessing one broad ability, as in Hill et al., 2010) would have been optimal, it was not practical in our experiments. Second, the tasks do not directly measure the learning of any educationally-relevant material, such as facts or skills. These possibilities are discussed below, and are in part rectified by the use of a more educationally-relevant task in the second experiment in this chapter.
The ecologically-valid nature of the experiment, with the full procedure taking
place in the classroom and to a large extent implemented by the classroom teacher, is
an advantage of our design, but it also introduces noise to the data. Since we did not
have the opportunity to train the teachers to perform the procedure in a consistent
manner, there may have been idiosyncrasies between classes that we could not
predict. For this reason (along with some statistical points that are discussed below),
in addition to a more “standard” ANOVA-based analysis of the data, we also
implemented mixed-effects models that allowed the inclusion of random effects for
classroom. This same analysis strategy was also used in Chapters 4 and 5 of this
thesis.

Method

Participants

Participants were two hundred and twelve children from eight classes at
Towerbank Primary School, Edinburgh, Scotland. These were three Primary 2 classes
(aged approximately 6 years), two Primary 4 classes (aged approximately 8 years),
and three Primary 6 classes (aged approximately 10 years). Written permission was
received from all parents/guardians prior to the experiment, and the study was
approved by the Psychology Research Ethics Committee of The University of
Edinburgh.

The number of children in each class is shown in Table 3.1, along with their
ages and the time of day at which they carried out their activities. Times of testing
were balanced such that each year group had at least one AM group and one PM group.

Table 3.1. Number of children in each class, with mean age and time of test administration

<table>
<thead>
<tr>
<th>Class</th>
<th>Year group</th>
<th>n (male, female)</th>
<th>Mean age in years (SD)</th>
<th>Time of testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>29 (13, 16)</td>
<td>6.36 (0.30)</td>
<td>AM</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>23 (10, 13)</td>
<td>6.37 (0.33)</td>
<td>PM</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>29 (13, 16)</td>
<td>6.31 (0.30)</td>
<td>PM</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>28 (14, 14)</td>
<td>8.35 (0.34)</td>
<td>PM</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>29 (17, 12)</td>
<td>8.28 (0.27)</td>
<td>AM</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>23 (9, 14)</td>
<td>10.33 (0.27)</td>
<td>AM</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
<td>27 (8, 19)</td>
<td>10.29 (0.27)</td>
<td>AM</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>24 (10, 14)</td>
<td>10.17 (0.34)</td>
<td>PM</td>
</tr>
<tr>
<td>Total</td>
<td>212 (94, 118)</td>
<td>8.25 (1.70)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: AM = start of school day; PM = after lunch break.

Materials

Each class teacher was provided with a sealed envelope containing their assigned trial order, and two CDs on which music was recorded for the activity sessions (all teachers had access to the audio equipment required to play the CDs). For Exercise sessions, teachers were provided with a selection of four one-minute-long pieces of up-tempo music, as consecutive tracks on the CD. Teachers were asked to encourage children to perform four exercises (star jumps, jogging on the spot, hopping with either leg, touching toes then jumping) for at least thirty seconds each to the four pieces of music respectively. These exercises were chosen to be similar to the sorts of activities normally engaged in by the children at the school during their ‘Engergizers’ sessions.
For Relaxation sessions, teachers were again provided with music on CD – the track ‘Lichen’ (the duration of which was 4:15) from the Aphex Twin’s *Selected Ambient Works Volume II*. Children were asked to sit at their desks quietly, possibly placing their heads on their desk, closing their eyes, and listening to the music.

For the No Activity conditions, the teachers were asked to lead the class in their normal lessons, and not to perform any activities similar to the exercise and relaxation conditions described above.

The effects of the three conditions (Exercise, Relaxation, no activity) on preparedness for learning were assessed using a short battery of tests. At the start of each testing session, A4-sized answer booklets were distributed to all children, with a cover page, and 4 pages corresponding to the following three tests:

*Visual attention* was tested using a novel measure we named the ‘Lollipops Test’. This test is similar to the Balloons Test (Edgeworth, Robinson, & MacMillan, 1998), but made more challenging by incorporating different visual designs on each item, and a denser test array. The array—a full A4 page in the answer booklet; see Figure 3.1—contained 120 ‘lollipops’, 20 of which had no stick (a straight line emanating from the bottom of the circle). Children were asked to scan the array, find the lollipops with no stick and, using a pencil, add a stick to those lollipops.
Figure 3.1. Not-to-scale example of one form of the Lollipops Test. Actual sheets printed on full A4. Participants added ‘sticks’ to those lollipops without them.

Pilot testing during the introductory session (see below), where each group completed the test under the same time limit (60 seconds), indicated the following test durations would allow children of each age group to find a mean of ten lollipops (i.e. 50% of the total): Primary 2 class: 48 seconds; Primary 4 class: 33 seconds; Primary 6 class: 25 seconds. After this time period, children were asked to put their pencils down and turn the page to the next test. The classroom teacher and the experimenter monitored compliance with this instruction. Three versions of this test were created for each of the three weeks of the experiment, with the same number of ‘stickless’ lollipops (targets) at different positions in the array (see Appendix A, Figure A1 for the two forms different to that in Figure 3.1). Answers were recorded as targets (stickless lollipops) hit per minute. Because the non-target lollipops already had their sticks in place, false alarms (where the children added a stick to a lollipop that already had one) were extremely rare, and were thus ignored in the scoring process.
Short-term memory was tested using an adapted version of the Visual Patterns Test (VPT; Della Sala et al., 1997). The answer booklet contained an empty square grid (4×4 squares for the P2 classes, 5×5 squares for the P4 and P6 classes; see Appendix A, Figure A2 for an example), and the children were shown a similar grid with crosses in eight of the boxes on A2-sized posters held up by the experimenter and the class teacher. The patterns were shown for fifteen seconds in the P2 and P4 classes, and ten seconds in the P6 classes. At this point, the children were instructed to recreate in their answer booklet the pattern they had seen, then turn the page to the next test. VPT scores were calculated by summing the correct answers and subtracting the ‘false alarms’.

Mental Rotation was tested using a novel Square Completion Task (SCT) similar to that used by Rauscher et al (1993). This involved two shapes, one of which had to be mentally rotated by 90° to assess whether together the shapes made a perfect square (see Figure 3.2). Children were asked to tick YES or NO boxes provided on their sheet. Each iteration of the test included four such puzzles, of increasing difficulty. Children were given as much time as they required to complete this test, and the test was scored simply as a mark out of four.
Figure 3.2. Example of one form of the Square Completion Task. This image took up half of one sheet of A4 paper, with a second form in the other half.

Procedure

Introductory session. Over a week before the experimental sessions began, the experimenters visited each class for an approximately ten-minute period and informed them that they would be taking part in a scientific experiment. They had the Lollipops Test and the VPT explained to them, and were shown an example of the SCT. The children then completed practice versions of the three tests, which served as a pilot (pilot data were used to adjust task difficulty and time-limits for the experimental versions of these tests; the).

Experimental sessions. The following procedure was identical for AM and PM classes, who performed it at their assigned time.

On Week 1 of the experiment, during the first ten minutes of class (either at the start of the school day for the AM classes or just after lunch break for the PM classes), the teacher opened the envelope provided to them, and performed their assigned activity for that day (Energizers, Calm Time, or no activity). This activity
was then performed (for ‘no activity’, teachers were asked to continue with their usual classroom activities). Exercise and Relaxation sessions were performed entirely by the teachers, and the experimenters were not present at the sessions. Teachers were asked not to make any link explicit to the children between the session and the experimenter’s subsequent visit, nor to inform the experimenter of the activity that had been performed on that day.

Forty minutes after the start of class, the experimenter arrived and administered the Lollipops Test, the VPT, and the SCT in that order. After test administration, the experimenter left the classroom and the class continued with their regular school day.

On Weeks 2 and 3 of the experiment, the same procedure was followed, so that all eight classes were administered tests in all three experimental conditions. The timetable was arranged such that, for each class, the experimental sessions were exactly one week apart. Table 3.2 shows the testing orders.

<table>
<thead>
<tr>
<th>Class</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>E</td>
<td>R</td>
<td>E</td>
<td>E</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>R</td>
</tr>
<tr>
<td>Week 2</td>
<td>R</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>R</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>Week 3</td>
<td>N</td>
<td>E</td>
<td>R</td>
<td>R</td>
<td>E</td>
<td>E</td>
<td>R</td>
<td>N</td>
</tr>
</tbody>
</table>

Note: E = Exercise, R = Relaxation, N = Neutral
Results

Answer booklets were scored by the experimenter while still blind to the order of activity conditions in each class. In Week Three of the study, Classes 1 and 2 were unable to complete their assigned activity, so their scores from that week (52 children) were removed from the analysis. In addition, a further thirty-four children were absent on at least one of the test days. Thus, we were left with full data from all three weeks for one hundred and twenty six children, with two hundred and four children contributing data on at least two test weeks. Mean results for each test in each condition are shown in Table 3.3 split by school year, and are shown in Figure 3.3.

Table 3.3. Mean scores (SD) by year group and condition for the Lollipops Test (targets per minute), the VPT (hits minus false alarms), and the SCT (correct responses).

<table>
<thead>
<tr>
<th>Test</th>
<th>Condition</th>
<th>Primary 2</th>
<th>Primary 4</th>
<th>Primary 6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lollipops</td>
<td>Neutral</td>
<td>16.15 (3.70)</td>
<td>24.72 (4.91)</td>
<td>32.59 (6.37)</td>
<td>25.38 (8.53)</td>
</tr>
<tr>
<td></td>
<td>Exercise</td>
<td>14.78 (3.92)</td>
<td>22.27 (5.58)</td>
<td>31.20 (5.66)</td>
<td>22.94 (8.48)</td>
</tr>
<tr>
<td></td>
<td>Relaxation</td>
<td>15.58 (4.14)</td>
<td>24.29 (5.50)</td>
<td>32.48 (5.97)</td>
<td>23.57 (8.85)</td>
</tr>
<tr>
<td>VPT</td>
<td>Neutral</td>
<td>4.22 (2.34)</td>
<td>3.56 (3.46)</td>
<td>3.74 (2.95)</td>
<td>3.82 (2.96)</td>
</tr>
<tr>
<td></td>
<td>Exercise</td>
<td>4.37 (2.40)</td>
<td>2.96 (2.82)</td>
<td>4.55 (2.43)</td>
<td>3.98 (2.64)</td>
</tr>
<tr>
<td></td>
<td>Relaxation</td>
<td>4.22 (2.23)</td>
<td>4.86 (2.69)</td>
<td>3.73 (2.82)</td>
<td>4.24 (2.59)</td>
</tr>
<tr>
<td>SCT</td>
<td>Neutral</td>
<td>2.96 (.95)</td>
<td>3.51 (.74)</td>
<td>3.38 (.82)</td>
<td>3.30 (.86)</td>
</tr>
<tr>
<td></td>
<td>Exercise</td>
<td>2.98 (1.03)</td>
<td>3.46 (.78)</td>
<td>3.61 (.74)</td>
<td>3.35 (.89)</td>
</tr>
<tr>
<td></td>
<td>Relaxation</td>
<td>3.04 (.98)</td>
<td>3.49 (.89)</td>
<td>3.63 (.70)</td>
<td>3.36 (.91)</td>
</tr>
</tbody>
</table>

Note: VPT = Visual Patterns Test; SCT = Squares Completion Test.
Since the tests were novel, we first assessed their reliability using Cronbach’s Alpha. The Lollipops Test showed high reliability ($\alpha = .92$ across the three experimental administrations). However, both the VPT and the SCT showed poor reliability ($\alpha = .34$ and .44 respectively). Inspection of these latter tests showed that they exhibited substantial ceiling effects, indicating that even after an increase in difficulty since the pilot testing, they were not of sufficient difficulty for any of the age groups involved.

Kolmogorov-Smirnoff tests on all three conditions for the VPT and the SCT were significant (all $p$-values < .05), showing deviation from normality. For the
Lollipops test, the Kolmogorov-Smirnoff test was significant only for the Exercise condition \((p = .003)\); although on visual inspection the data from this condition closely approximated a (somewhat flattened) normal distribution. Nevertheless, we first used Friedman’s ANOVA to test between the three conditions for each of the tests. No significant results were found for the VPT \((\chi^2(2) = 5.31, p = .07)\), or the SCT \((\chi^2(2) = 1.10, p = .58)\), and thus no further analyses were carried out; there was no effect of condition on these tests. However, for the Lollipops Test, a significant result was found \((\chi^2(2) = 23.54, p < .001)\), and a Wilcoxon signed ranks test confirmed the significant differences were due to the lower score in the Exercise condition compared to the Neutral condition \((Z = 4.26, p < .001, r = .26)\) or the Relaxation condition \((Z = 2.97, p = .003, r = .17)\). There was no difference between the Neutral and Relaxation conditions \((Z = .39, p = .70)\). This initial analysis, then, showed that Exercise appeared to have a detrimental effect on attention, as measured by the Lollipops Test.

As noted above, an alternative method with which to analyze these results is linear mixed-effects modeling. Whereas the Friedman ANOVA is simple to understand and may be more familiar to researchers in this area, linear mixed-effects modeling has the following advantages over it: first, it allows us to use maximum-likelihood estimation, a superior method of fitting the model to the listwise deletion employed in the ANOVA, especially with the missing data as described above (see Schafer & Graham, 2002); second, it allows us to add a between-subjects effect of time of day (AM or PM) and covariates of year group and sex, which we were unable to account for using Friedman’s ANOVA; third, it allows us to test for random effects in the data, particularly of class (teachers may have performed the exercise and
relaxation sessions in idiosyncratic ways); fourth, it is more robust to small deviations from normality, as in the Exercise condition.

Using the ‘lme’ function in the nlme package for R (v3.1-113; Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2013), we re-analysed the Lollipops Test by estimating a model using restricted maximum likelihood estimation (REML). The model included the within-subjects effect of condition (Neutral, Exercise, Relaxation), the between-subjects effect of time of day (AM or PM), the covariates of year group (2, 4, or 6), and sex, and the random intercept effect of class (1-8). The covariate of age was available, but not included in the analysis here due to its strong correlation with year.

For the Lollipops test, interaction terms added both separately and together significantly worsened model fit (ΔAIC after inclusion of all three interaction terms > +11.00; they thus did not have significant effects on test scores), so they were discarded in favour of the basic model. Results from this model are shown in Table 3.4.
### Table 3.4. Regression coefficients for each fixed and random effect for the Lollipops Test.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b$</th>
<th>$SE$</th>
<th>$t$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed Effects:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>15.45</td>
<td>0.98</td>
<td>15.79</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Condition: Neutral</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition: Exercise</td>
<td>-1.51</td>
<td>0.56</td>
<td>-2.68</td>
<td>.008</td>
</tr>
<tr>
<td>Condition: Relaxation</td>
<td>-0.26</td>
<td>0.54</td>
<td>-0.48</td>
<td>.630</td>
</tr>
<tr>
<td>Time: AM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time: PM</td>
<td>1.32</td>
<td>0.82</td>
<td>1.60</td>
<td>.184</td>
</tr>
<tr>
<td>Year Group: 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year Group: 4</td>
<td>8.51</td>
<td>1.01</td>
<td>8.42</td>
<td>.001</td>
</tr>
<tr>
<td>Year Group: 6</td>
<td>17.07</td>
<td>0.96</td>
<td>17.79</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex: Male</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex: Female</td>
<td>-0.52</td>
<td>0.45</td>
<td>-1.15</td>
<td>.25</td>
</tr>
<tr>
<td><strong>Random Effect:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class (1-8)</td>
<td>0.83</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>26.00</td>
<td>5.10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Dashes = Reference levels; all $p$-values uncorrected.

Significant main effects of condition were found for the Lollipops Test; pairwise comparisons using the Tukey HSD correction showed that the difference between the targets-per-minute scores was significant between the Exercise and Neutral conditions (mean difference ($SE$) = -1.51 lollipops per minute (.56), $p = .02$) but not between the Neutral and Relaxation conditions (mean diff. ($SE$) = -0.26 (.54), $p = .88$) or between the Relaxation and Exercise conditions (mean diff. ($SE$) = 1.25 (.53), $p = .053$). Thus, the Exercise condition significantly decreased performance on the Lollipops Test, whereas the Relaxation condition had no effect. This analysis showed no effect of time of day: Lollipops Test scores were not significantly higher in the PM classes than the AM classes (mean diff. ($SE$) = 1.32 (.82), $p = .18$). Despite the titration of the task for each age group, older children’s performance was significantly superior to that of younger children (mean diff. between Primaries 6 and 2 ($SE$) = 17.07 (.96), $p < .001$). No significant effect was found for sex. Adding a random effect for each individual, instead of each class, did not substantially alter the results reported here.
Discussion

Using a repeated measures design in eight classes of Primary School children, we tested the effects of short Exercise and Relaxation sessions on the attention, memory, and mental rotation ability of Primary School children. The study represents, to our knowledge, the first investigation of short in-class relaxation techniques on children's executive functions, and the first study of in-class exercise techniques that included children under the age of eight years.

The results showed that a short Exercise session, far from improving attention ability as has been found in previous studies (e.g. St-Louis-Deschêne & Ellemberg, 2011), significantly decreased scores on a novel, reliable test of visual attention, the Lollipops Test, 30 minutes later, compared to a neutral control condition. This finding is superficially inconsistent with those of Hill et al., (2010, 2011) and Mahar et al. (2006), who reported positive effects of in-class exercise activities, and those of Kubesh et al. (2009) who found no effect of such an activity.

A potential explanation of our result lies in the timing of the exercise and testing sessions. Hill et al. (2010, 2011) measured attentional performance one hour after exercise, and it is conceivable that the beneficial effects of exercise for cognitive function appear only after a period of time longer than the 30-minute gap between exercise and testing in the present study. During this gap, over-arousal may lead to slightly reduced attentional resources, as reflected in our results. Such an effect would be consistent with results meta-analysed by Lambourne and Tomporowski (2010), who concluded that in the first twenty minutes of exercise, cognitive performance was
significantly impaired; only after twenty minutes of exercise did benefits appear. It should, however, be noted that these results were for experiments where the cognitive tasks were completed while exercising, and tended to be for older individuals than those studied here. Nevertheless, it may be the case that exercise causes distraction in the short-term: Experiment 3.2, which manipulated the time between the activity and testing, was designed in part to assess this possibility.

The results from our other two tests (memory and mental rotation) are far more ambiguous than those of the Lollipops Test. For both the VPT and SCT, even after our raising of the difficulty after the pilot testing, near-ceiling effects were observed. In addition, the reliabilities of both the measures were low. Null results were found across the three experimental conditions for both of these measures, but it may be the case that different, more reliable and valid measures would show the same decrement in scores after exercise that we observed for the attentional measure. Alternatively, the negative effects of exercise may be confined to attention. In Experiment 3.2, we investigated these alternatives in a somewhat larger sample in a second school.

Finally, an obvious limitation of Experiment 3.1, and those preceding it in the classroom exercise literature, is the lack of direct examination of educationally-relevant outcomes. This is discussed above in the introductory section for Experiment 3.1. Whereas executive functions such as attention and working memory are clearly of great importance for preparedness for classroom performance (e.g. Mazzocco & Kover, 2007), and concentration on any task should be related to the preparedness to learn in the classroom, they do not necessarily relate directly to outcomes: It could be
the case, for instance, that new facts are learned more effectively after exercise without a concomitant improvement in attention, working memory, or mental rotation performance. The new memory test in Experiment 3.2 involved recall of facts, an outcome variable more closely related to education.

**Experiment 3.2**

Experiment 3.2 was carried out in a different school to Experiment 3.1 (Bruntsfield Primary School, Edinburgh), and used the same measure of attention but a different memory measure (in place of the VPT). Due to the strong ceiling effects observed in the SCT, we did not use it in this second experiment. In addition, in Experiment 3.2 we were able to test children not just from years 2, 4, and 6, but from all years between Primary 2 and Primary 6.

We manipulated the time between activity and testing (either thirty minutes or one hour), and hypothesized, on the basis of the literature surveyed above, that exercise would improve children’s cognition only at a longer time after the activity. On the basis of Experiment 3.1, we would predict that exercise would be detrimental to performance (at least attentional performance) with only a short time (around 30 minutes) between the activity and testing. Since Experiment 3.1 showed no effect of relaxation on cognition, we used a relaxation condition as a control in Experiment 3.2, comparing it to an exercise session identical to that used in Experiment 3.1.

Also, as noted above, Experiment 3.2 tested more educationally-relevant outcomes: The memory test used involved learning facts about two animals and
recalling them on a later written quiz. It also had the advantage of being a more long-
term measure, assessing memory one day later rather than immediately (that is, it was
not a working memory measure like the VPT in Experiment 3.1). We thus went
further than the previous experiment—and previous published studies in the
literature—which focused solely on executive functioning.

Method

Participants

Two hundred and seventy pupils at Bruntsfield Primary School, Edinburgh,
aged from 6-12 years, took part in Experiment 3.2. Two classes from each school year
(Primary 2 to Primary 6) were included. Table 3.6 shows the breakdown of age and
sample size by year groups. Written consent was received from the parents or
guardians of each child before their participation, and ethical approval was given by
the University of Edinburgh Psychology Research Ethics Committee.

Table 3.5. Number of children in each year group involved in Experiment 3.2, with
mean ages.

<table>
<thead>
<tr>
<th>Year group</th>
<th>n (male, female)</th>
<th>Mean age in years (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>55 (27, 28)</td>
<td>6.55 (0.36)</td>
</tr>
<tr>
<td>3</td>
<td>49 (32, 17)</td>
<td>7.70 (0.30)</td>
</tr>
<tr>
<td>4</td>
<td>59 (30, 29)</td>
<td>8.82 (0.28)</td>
</tr>
<tr>
<td>5</td>
<td>50 (27, 23)</td>
<td>9.70 (0.28)</td>
</tr>
<tr>
<td>6</td>
<td>57 (30, 27)</td>
<td>10.72 (0.32)</td>
</tr>
<tr>
<td>Total</td>
<td>270 (146, 124)</td>
<td>8.73 (1.51)</td>
</tr>
</tbody>
</table>
Measures

Attention was measured using the same test as in Experiment 3.1: the Lollipops Test. The same time limits were used for years 2, 4, and 6 (48, 33, and 25 seconds, respectively). Times for intervening years were calculated as the average of the years above and below; thus, Primary 3 were given 41 seconds and Primary 5 29 seconds to complete the task. A different form of the test, with a different arrangement of the same number of targets and distractors, was used on each of the testing days (the same as in Experiment 3.1; see Appendix A, Figure A1).

Memory was measured using a novel task (henceforth Memory Test) involving two sets of facts about rare animals (that the children were unlikely to have heard of previously): the pangolin and the slow loris. Shorter fact sets were used for children in Primary 2 and 3 than for those in Primary 4, 5, and 6. Fact sets were recorded by a voice actor and this recording was given to all teachers on compact disc to ensure consistent administration. The older children’s version of the pangolin fact set is quoted below; the version for younger children, and the slow loris forms, can be found in Appendix A:

“The Pangolin is an animal that is found in Africa. It is covered in scales, like a suit of armour, and it is yellow and brown in colour. Pangolins are usually around 88 centimetres long. Pangolins mainly eat ants. They dig the ants out of their nests with their sharp claws, and pick them up with their very long, thin, sticky tongue. The pangolin is nocturnal, and spends most of the day sleeping. If a predator, for example a tiger, tries to eat a pangolin, the
pangolin will curl up into a ball, and spray the predator with a horrible smell.

There are eight different species of pangolin, and these include the giant pangolin and the tree pangolin.”

To measure memory for the fact sets, quizzes were created, again for the different year groups (Primary 2-3 and 4-6). The questions for the quizzes—see Appendix A for each question—were recorded by a voice actor and, again, given to teachers on compact disc. For years 2-3, the quiz consisted of five questions, all of which were answered by circling the correct answer from multiple choices on an A4 sheet of paper: for three questions, the choices were all pictures (of the animal, its habitat, etc.); for two questions, the choices were numbers. Thus, the quiz for the younger groups assessed recognition memory. For years 4-6, the quiz included nine questions, one of which was answered by circling a picture (of three choices), and the remainder of which were answered by writing the answer in a space provided. Some questions required multiple answers (for example, for the pangolin quiz: “can you name two of the different species?”), meaning that the total score on the quiz was fifteen points. The quiz for the older children, then, mainly assessed recall memory, with only one recognition item. The differences between the test types reflected our expected difference in the ability levels of the children – older children should be able to recall facts with no visual ‘prompt’ (arguably a more educationally-relevant ability) whereas the younger children would have found this task more difficult. All Memory Test quiz sheets are reproduced in Appendix A.
CHAPTER 3: IN-CLASS EXERCISE AND RELAXATION

Procedure

The experiment took place on the same day of two consecutive weeks. All sessions took place in the morning, just after the start of the school day, which began at 8:50AM (this time was most optimal for teachers, and we decided against examining effects of time of day since no such effects were found in Experiment 3.1). The teachers were asked to have their children complete an Exercise or Relaxation condition—both performed identically to those in Experiment 3.1, with the same musical accompaniment—in the first ten minutes of the school day. Conditions were counterbalanced so that each class from each year performed a different activity on each of the two test days. After the activity, the children continued with their regular classroom activities.

The experimenter visited the classroom at either at 9:30AM or at 10:00AM (that is, either thirty minutes or one hour after the Exercise or Relaxation activity was completed), and administered the Lollipops test in the same fashion as described in Experiment 3.1, above. Next, the experimenter played the recording of the fact set—which of the two fact sets was used first in each class was counterbalanced—twice as the children listened.

Teachers were asked to run the quiz for the memory test at approximately the same time in the morning of the next day; the experimenter was thus not present during the quiz administration. The full design of the experiment is shown in Figure 3.4.
Figure 3.6 Design of Experiment 3.2, showing each session across the two experimental weeks.

Results

Scores for the attention and memory tests are shown in Table 3.7, below, for each condition and year group.

Table 3.6. Mean Lollipops Test (targets per minute) and Memory Test (correct answers) scores (SDs) by year group, condition, and time for Experiment 3.2. For the Memory Test, total scores are separated by younger and older groups; see below.

<table>
<thead>
<tr>
<th>Test</th>
<th>Condition</th>
<th>Time</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
<th>Total (all years)</th>
<th>Total younger</th>
<th>Total older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lollipops Test</td>
<td>Exercise</td>
<td>&lt;30mins</td>
<td>16.42</td>
<td>18.27</td>
<td>20.88</td>
<td>25.43</td>
<td>27.56</td>
<td>21.66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;30mins</td>
<td>(2.91)</td>
<td>(3.20)</td>
<td>(4.88)</td>
<td>(6.09)</td>
<td>(6.11)</td>
<td>(6.34)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Relaxation</td>
<td>&lt;30mins</td>
<td>12.98</td>
<td>18.84</td>
<td>21.26</td>
<td>25.70</td>
<td>30.51</td>
<td>22.04</td>
<td>(7.65)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;30mins</td>
<td>(3.48)</td>
<td>(3.57)</td>
<td>(5.41)</td>
<td>(5.21)</td>
<td>(5.38)</td>
<td>(7.03)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Memory Test</td>
<td>Exercise</td>
<td>&lt;30mins</td>
<td>14.53</td>
<td>19.67</td>
<td>22.70</td>
<td>25.26</td>
<td>30.12</td>
<td>22.44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;30mins</td>
<td>(3.86)</td>
<td>(3.46)</td>
<td>(4.36)</td>
<td>(5.72)</td>
<td>(5.61)</td>
<td>(7.03)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Relaxation</td>
<td>&lt;30mins</td>
<td>14.90</td>
<td>18.17</td>
<td>19.93</td>
<td>25.38</td>
<td>32.23</td>
<td>22.34</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;30mins</td>
<td>(3.92)</td>
<td>(2.44)</td>
<td>(4.19)</td>
<td>(5.56)</td>
<td>(6.00)</td>
<td>(7.72)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: For the Memory Test, learning took place at the times noted in the table, with the testing session the day after. Maximum Memory Test scores: 5 in Primary 2 and 3 (younger); 15 in Primary 4-6 (older). P = Primary year.

Unfortunately, due to other classroom activities taking precedence over the experiment, the teachers were unable to follow the instructions precisely. None of the classes, then, were tested at precisely thirty and sixty minutes after their assigned
activity, as planned. Nevertheless, we were able to code each class as having performed the activity less than, or more than, thirty minutes before the experimenter started the testing/learning session. All analyses below use this variable where ‘time’ is mentioned. Mean test scores are shown split for time in Table 3.7.

Across the two weeks, the Lollipops Test again showed acceptable reliability ($\alpha = .72$). The results from the two conditions approximated a normal distribution, and so were first analyzed using a repeated-measures ANCOVA with the within-subjects factor of condition (Exercise versus Relaxation), the between subjects factor of time (<30 mins before versus >30 mins before), and the covariates of school year and sex (again, age was excluded as it was strongly collinear with year; in an analyses including both as covariates, age was never significant).

This analysis showed no main effect of condition ($F(1, 266) = .70, p = .40$) or time ($F(1, 266) = .002, p = .97$), and importantly, no time×condition interaction ($F(1, 266) = .70, p = .40$). The covariate of school year had a significant influence, with older children scoring higher ($F(1, 266) = 458.35, p < .001, \eta^2_p = .63$); sex had no effect ($F(1, 266) = 2.04, p = .16$) Our hypotheses, then, were not supported for the Lollipops Test data: there was no influence of exercise on attentional performance, whatever the time between the activity and testing.

Again, an alternative analysis of the Lollipops Test data was carried out using a generalized linear mixed-effects model, which included the same factors and covariates as the ANOVA but also added a random effect of class again to account for any teacher-specific differences in the administration of the Exercise and Relaxation
conditions. Experiment 3.2 did not have the same extent of missing data as Experiment 3.1, but nonetheless REML was used. The results of this analysis are shown in Table 3.8; similar to the results of the ANCOVA, condition had no effect on the number of targets successfully found within the time limit, and there was no condition × time interaction. This analysis includes year group as a factor, but the results of interest are not substantially changed if this is included as a continuous variable. Thus, the finding from Experiment 3.1, that exercise had a detrimental effect on attention as measured by the Lollipop Test, was not replicated here, even where the time between exercise and testing was less than 30 minutes.

Table 3.7. Regression coefficients for each fixed and random effect for the Lollipops Test in Experiment 3.2.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient $b$</th>
<th>SE</th>
<th>$t$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Effects:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>13.96</td>
<td>0.96</td>
<td>14.54</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Condition: Exercise</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Condition: Relaxation</td>
<td>0.79</td>
<td>0.57</td>
<td>1.39</td>
<td>.16</td>
</tr>
<tr>
<td>Time: &lt;30mins</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time: &gt;30mins</td>
<td>0.19</td>
<td>0.84</td>
<td>0.23</td>
<td>.75</td>
</tr>
<tr>
<td>Year Group: 2</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year Group: 3</td>
<td>4.14</td>
<td>1.17</td>
<td>3.55</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Year Group: 4</td>
<td>6.51</td>
<td>1.15</td>
<td>5.67</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Year Group: 5</td>
<td>10.76</td>
<td>1.16</td>
<td>9.26</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Year Group: 6</td>
<td>15.41</td>
<td>1.15</td>
<td>13.39</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Sex: Male</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex: Female</td>
<td>0.77</td>
<td>0.41</td>
<td>1.86</td>
<td>.06</td>
</tr>
<tr>
<td>Time × Condition</td>
<td>−0.49</td>
<td>0.82</td>
<td>−0.60</td>
<td>.55</td>
</tr>
<tr>
<td>Random Effect:</td>
<td>Variance</td>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class (1-8)</td>
<td>0.92</td>
<td>0.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>22.39</td>
<td>4.73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Dashes = Reference levels; all $p$-values uncorrected.

Turning to the Memory Test, the decision was made to analyze the scores from years 2-3 and years 4-6 separately, given the wide differences between the test types (entirely recognition memory in the younger children versus mainly recall in the older children; combining these results into one test would thus, arguably, lead to
invalid results). The test had poor reliability for both the younger ($\alpha = .55$) and older ($\alpha = .56$) children. An ANCOVA similar to that used for the Lollipops test could not be applied due to failure of normality, though no clear ceiling effects were observed in the Memory Test scores. A Wilcoxon signed ranks test was thus carried out between the Exercise and Relaxation conditions; this was not significant for younger participants ($Z = .48, p = .64$), but was significant in older participants ($Z = 2.04, p = .041$). In this case, where no other variables (such as year group, sex, or time) were controlled, exercise had a positive effect on memory, though only for older children.

To follow up these results with a more detailed analysis, generalized linear mixed-effects models were estimated for the Memory Test data; the factor of time, covariates (sex and primary school year), and the random class effect were included. These models were different from that used above; since the data could be considered as binomial counts, in that there were numbers of correct and incorrect answers for the quiz (Jaeger, 2008), binomial-family ‘lmer’ models were estimated using the ‘lme4’ package for R (Bates, Maechler, Bolker, & Walker, 2014). These models had the same structure as that for the Lollipops data, above, but had the outcome variable arranged as ‘correct’ versus ‘incorrect’ as opposed to a single measure. The results of these models are shown in Table 3.9 for younger and older participants.
Table 3.8. Results of binomial generalized linear mixed-effects model for the Memory Test in Experiment 3.2.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Younger participants (P2-3)</th>
<th>Older participants (P4-6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.98 ± 0.17, z = 5.84, p &lt; .001</td>
<td>0.10 ± 0.10, z = -3.55, p &lt; .001</td>
</tr>
<tr>
<td>Condition</td>
<td>-0.11 ± 0.19, z = -0.56, p = .58</td>
<td>-0.36 ± 0.10, z = -3.55, p &lt; .001</td>
</tr>
<tr>
<td>Time</td>
<td>-0.39 ± 0.19, z = -2.05, p = .04</td>
<td>-0.08 ± 0.11, z = -0.76, p = .45</td>
</tr>
<tr>
<td>Sex</td>
<td>0.26 ± 0.14, z = 1.84, p = .07</td>
<td>-0.11 ± 0.07, z = -1.50, p = .13</td>
</tr>
<tr>
<td>Year: Primary 3</td>
<td>-0.15 ± 0.14, z = -1.12, p = .26</td>
<td>-0.06 ± 0.10, z = -0.55, p = .58</td>
</tr>
<tr>
<td>Year: Primary 5</td>
<td>- - - -</td>
<td>0.02 ± 0.21, z = .83</td>
</tr>
<tr>
<td>Year: Primary 6</td>
<td>- - - -</td>
<td>0.02 ± 0.21, z = .83</td>
</tr>
<tr>
<td>Time x Condition</td>
<td>0.18 ± 0.27, z = 0.66, p = .51</td>
<td>0.15 ± 1.06, z = 2.9</td>
</tr>
</tbody>
</table>

Random effects | Var. | SD | Var. | SD |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Class (Intercept)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.003</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Note: Younger participants’ quiz maximum score = 5; older participant’s quiz maximum score = 15. For Condition, negative scores favour Exercise over Relaxation; for Time, negative scores favour <30mins over >30mins. Obs. = observations.

As shown in the table, whereas there was no effect of condition on memory in the younger children, a significant effect was found for the older year groups: Exercise improved memory performance compared to relaxation, albeit with a very small effect size. In the younger group only, scores were significantly higher when the condition was performed fewer than 30 minutes before the memory materials were learned: that is, there was a significant main effect of time. No significant interaction was found in either group; that is, the conditions did not have pronounced effects whether they were performed earlier or later in the morning. None of the covariates for either group reached significance.
Discussion

The main result of Experiment 3.1, that exercise had a detrimental effect on attention, was not replicated in Experiment 3.2, which included a somewhat larger sample and fewer missing observations. Furthermore, our hypothesis that exercise would be detrimental to attention if performed temporally close to measurement was not supported; no interactions between time and condition were found. However, we did find that exercise had a positive effect on memory in this second experiment, albeit measured using a different task than in Experiment 3.1, only in children above primary 4 (aged approximately 8-9 years), and with a small effect size. Potential explanations for the disparity in results between experiments are discussed in the next section.

General Discussion

Taken together, the two experiments reported in this chapter—both testing the effects of brief exercise and relaxation sessions on preparedness for learning in large samples of schoolchildren in naturalistic classroom settings—present a somewhat contradictory picture. In the first experiment, performance on our novel measure of attention, the Lollipops Test, was reduced when exercise had been performed thirty minutes prior. However, this was not found in the second experiment, where Lollipops Test scores were no different in the relaxation or exercise conditions, and were the same regardless of time. No effects of either exercise or relaxation on memory were found in Experiment 3.1, but a beneficial effect of exercise on memory
was found in Experiment 3.2 (though it should be borne in mind that the memory tests were assessing very different kinds of memory).

Our hypothesized mechanism for the negative effect of exercise on attention in Experiment 3.1 was that, consistent with the conclusion of the Lambourne and Tomporowski (2010) meta-analysis, exercise might only produce such performance decrements in the very short term, possibly by causing distraction due to over-stimulation. We had thus expected that tests at times up to an hour after the exercise activity, a time interval used in previous studies where positive effects have been found (e.g. Hill et al., 2010), would show improvements in performance. However, Experiment 3.2 showed that the effect was not present; exercise had no differential effects on cognitive abilities depending on the time it was performed relative to the testing (that is, there was no time×condition interaction). This may be due to the effect in Experiment 3.1 being very small; in that experiment, exercise reduced attentional performance by around one sixth of a standard deviation. This rather trivial difference may have been a spurious result. In any case, small effects such as this one are unlikely to be of practical significance to teachers in the classroom.

Previous studies have, however, found positive effects of exercise on attentional abilities. It may be that the previous positive effects were due to the longer exercise durations used in published studies; Kubesch et al. (2009) explicitly stated that “a 5-min aerobic endurance exercise is too short to lead to an increase of free fatty acids and therefore to an increased serotonin biosynthesis” (p. 237) which would hypothetically increase executive function, while Coe et al. (2006) suggest that a “a threshold of activity intensity may be needed to bring about changes in the child that
contribute to increased academic achievement” (p. 1518). Our Exercise conditions may not have been adequate to evoke substantial performance improvements, and a similar argument could be made that our Relaxation conditions did not sufficiently reduce arousal. Unfortunately, longer sessions were not practical, and in any case we aimed to test short activities similar to those that were already being performed by the teachers in the school.

Furthermore, our results are also inconsistent with those of Mahar et al. (2006), who found positive effects in the thirty-minute period immediately following their ‘Energizer’ session, though it should be noted they did not measure cognitive performance, but “on-task behaviour”. The fact that our measure of time in Experiment 3.2 was inexact—the teachers were unable to complete our original thirty-versus-sixty-minutes design—reduces our confidence on the issue of the timing of activities, however, and future in-class experiments should be run where a more exact timetable can be followed.

Whereas our results were discouraging for attentional performance, there may have been a significant effect of exercise on memory performance, in the predicted positive direction. This effect was found only in Experiment 3.2, on the fact-based Memory Test, and the most obvious reason for the differences in results between the experiments is the different measures used. Not only was the measure in Experiment 3.1 an immediate working memory (WM) task and that used in Experiment 3.2 a more complex test of recall and recognition of facts, but the tasks had differential difficulty: ceiling effects were observed for the Visual Patterns Task in Experiment
3.1, but—whereas the data were still not normally distributed—no such effects appeared for the fact-learning-based Memory Test in Experiment 3.2.

We are only aware of one previous study assessing the effects of exercise on the memory performance of young people, that of Budde et al. (2010). That study showed that score on a letter-digit-span task, recruiting WM, improved after a 12-minute intensive exercise session compared to no exercise, but only for those participants who performed generally poorly on the task. This is contradictory to our finding of no benefit for a WM task; however, the Budde et al. (2010) experiment was performed in participants substantially older than those in the present study (15-16 years old). In addition, our novel version of the VPT, administered to an entire class at once, did not prove to have high reliability, and thus the conditions did not necessarily have equal difficulty.

It may be the case, then, that exercise sessions prior to learning are beneficial for learning of new facts. Again though, these effects are very small, and they only appeared in our group of older children (above 8 years old). Unfortunately, we cannot distinguish whether this effect is due to the children’s later developmental stage, or their completion of a heavily recall-based test (in comparison to the largely recognition-based test given to the younger children). This memory result does, however, come with two important caveats, the first of which is the lack of a ‘neutral’ condition in Experiment 3.2. The condition was not included for reasons of time; unfortunately, without it we cannot ascertain whether the relaxation session was detrimental to this particular memory task, and thus the positive effect of the exercise was inflated for this reason. The second caveat is the poor reliability of the Memory
Test in Experiment 3.2; again, the conditions were not necessarily similar in difficulty.

Contrary to the suggestion of Hallam et al. (2002), who found effects of relaxing music on cognitive performance, relaxation had no effect on either attention or memory at either time of day in Experiment 3.1 (there were no condition×time interactions), and was either worse than, or no different to, the Exercise condition in Experiment 3.2. However, as mentioned above, Hallam et al. (2002) found positive results when playing music to children as they completed the outcome measures (cf. “Mozart Effect” studies such as Rauscher et al., 1993), not as a separate session beforehand. If relaxing music were required to be played simultaneously with the measure to improve performance, we would not have found its benefits in our study. Nevertheless, our study represents a rare investigation of the effects of relaxation on children’s neuropsychological functioning in the classroom, and we would encourage further research in this area, especially given the popularity of relaxation techniques in primary schools.

It should be noted that a weakness of Experiment 3.2 was the lack of a neutral, control condition – to maximize sample size, and due to time constraints, we used relaxation as the control in this experiment since it was no different to the neutral condition in Experiment 3.1. This does, however, leave open the possibility that positive or negative effects of relaxation compared to no activity would have been found in Experiment 3.2; future experiments should always include a neutral control condition if at all possible.
A further limitation of our study is that we did not verify the activity levels of the children, for example with a pedometer (as in Mahar et al., 2006) or a heart rate monitor (as in Budde et al., 2008). This would have allowed us to ensure that the children engaged in more activity in the Exercise conditions than in the other conditions. However, such measurement would have been impractical given our attempts at a naturalistic experiment in the classroom setting, and we did observe effects of exercise, indicating that the activity levels were sufficient to evoke at least some psychological changes. Nevertheless, studies with more exact measures of physical arousal should be performed to assess more precisely the activity levels required to evoke cognitive changes.

Best (2010) has suggested that cognitively-involving exercise sessions—those which combine physical exercise with some form of engagement with complex rules or strategies (for example, team sports)—may be more effective at improving executive functions than the simple kinds of activity engaged in in the present chapter. Future research should try to identify the precise exercise activities that are most effective at increasing preparedness to learn; these may involve in-class team games or some other more structured activity. A similar argument could be made for relaxation: particular relaxation techniques, such as meditation or specific breathing exercises, may be more effective than the simple techniques employed currently in classrooms and, by extension, in this chapter.

Finally, it should be noted that we only assessed the short-term effects of exercise and relaxation sessions, testing outcomes up to one day after the session. It may be the case that implementing such sessions across, for example, a school term,
may have positive effects on academic outcomes, as suggested by correlational studies such as Chomitz et al. (2009). Longer-term experimental studies will be required to answer this question. Another method to address the potential causal relationship between exercise and academic performance would be the monozygotic twin differences design (e.g. Burt, McGue, Iacono, & Krueger, 2006). This design would allow investigators to control for genetic and shared-environmental effects to assess, in this case, whether the twin who happens to engage in more exercise also has superior academic performance.

**Conclusion**

The results of the two experiments in the present chapter indicate only very small effects—both positive and negative—of exercise on neuropsychological tests. The initial finding of a decrement in attentional performance after exercise in Experiment 3.1 did not replicate in Experiment 3.2, and may therefore have been a Type I error; this is especially plausible given its small effect size (around one sixth of a standard deviation on our novel Lollipops Test). The improvement in scores on the more educationally-relevant recall memory test observed in Experiment 3.2 was also small, but is nonetheless of potential interest for future studies; it should be followed up using a more reliable measure of recall memory, though if similar-sized effects are found, they are unlikely to be important for educators in a practical sense. Finally, no effects of relaxation before learning were found in either study; the next chapter of the present thesis (Chapter 4) examines the potential for effects of relaxation *after* learning, as part of a memory consolidation process.
Even faced with the lack of substantive cognitive benefits (and it should be noted that our study is unusual in not finding positive effects, though this may to some extent reflect publication bias in the literature), we would not recommend on this basis that in-class exercise and relaxation activities be discontinued. Whatever the effects on cognitive performance, teachers may find that pupils enjoy such sessions and, as described above (e.g. Janssen & LeBlanc, 2010), the effects of exercise on health are well-documented and of great importance, and this argument can also be made, to a lesser extent in our opinion, for the utility of relaxation sessions. Teachers should not, however, necessarily expect practically significant benefits from engaging their classes in such activities.
Does wakeful resting boost recall of newly-learned information in primary school children?§

§ The experiments described in this chapter were designed, performed, and analyzed with the assistance of Robert D. McIntosh, Michaela Dewar, and Sergio Della Sala.
Abstract

Previous research in amnesic patients and older adults has shown that a period of wakeful resting with minimal sensory or cognitive interference after learning new information can substantially improve recall of the information up to one week later. In two experiments, we tested whether this technique is effective for children’s learning. In Experiment 4.1, two hundred and eighty-four children participated in an in-class test with teachers leading relaxation sessions after a story was learned. At testing one and five weeks later, no benefits of post-learning wakeful resting were observed. In Experiment 4.2, fourteen children participated in a controlled, individually-administered test. Again, one week later, no benefits of wakeful resting were observed compared to learning in a distractor condition. The limitations of these experiments and their possible implications for future research on memory consolidation in children are discussed.
CHAPTER 4: WAKEFUL RESTING AND MEMORY IN CHILDREN

Introduction

A number of findings from cognitive psychology and neuropsychology indicate that a brief period of wakeful resting, involving minimal sensory and cognitive interference, after learning some new information leads to better later recall than if the learning is followed by another task, or other forms of interference (for reviews, see Wixted, 2004, 2010). Such evidence may support a ‘consolidation’ theory of forgetting (Dewar, Cowan, and Della Sala, 2010), where new memories are vulnerable to interference for a time after their creation, and is of relevance to educators since it raises the possibility that a brief intervention—an in-class wakeful resting period—could enhance students’ learning. In this chapter, we follow the experiments of the previous chapter, which investigated pre-learning relaxation, with two experiments investigating the possibility that post-learning relaxation sessions improve the learning of new information.

The effects of a period of minimal retroactive interference on later memory were first studied in the early 20th Century in healthy individuals by Müller and Pilzecker (1900; see also Dewar, Cowan, and Della Sala, 2007 for an historical review), who showed that if a list of syllables to be remembered was followed immediately by a second list, later recall was poorer than if the second list was shown some minutes after the first. They took this as evidence for a consolidation process in memory (indeed, this term in relation to memory was coined by Müller and Pilzecker, 1900) wherein a period of time post-learning is required to ‘stabilize’ the memory trace (Dudai, 2004), and which it is possible to interrupt with a secondary task.
After a long period in which consolidation tended not to play a crucial part in cognitive psychological theories of memory and forgetting (Brown & Lewandowsky, 2010), some recent work has rediscovered the effect in the context of patients with amnesic deficits. In two studies of six amnesic patients, Cowan, Beschin, and Della Sala (2004) found that a period of minimal interference, where the participants simply sat in a quiet, darkened room for either 10 minutes or 1 hour (for the two experiments respectively), immediately after learning word lists or stories led to greatly improved delayed recall compared to a condition in which further cognitive tasks were performed after learning. For instance, the authors reported that in three patients, “after 1 h long intervals… the retention of what had been recalled before the interval was 0% when the interval included interference but 63, 78, and 85% of that material after no interference” (Cowan et al., 2004, p. 831). This impressive effect was not, the authors argue, due to conscious repetition of the information to hold it in memory – several of the patients who displayed improvements due to minimal interference were found to be asleep during at least some of the minimal interference period, ruling out this explanation, at least for those particular participants.

Such results (and those conceptually replicating them; see, e.g. Dewar, Fernandez Garcia, Cowan, & Della Sala, 2009), suggest that in at least some amnesics—potentially depending on the specific location of the lesion causing their amnesia (Cowan et al., 2004, p. 833)—a process of memory consolidation has been weakened, and is thus particularly vulnerable to retroactive interference. On the basis of these results, Dewar et al. (2010) propose a model of forgetting drawing on Dudai (2004), who differentiated two types of consolidation. The first, ‘synaptic consolidation’, is related to the well-studied process of long-term potentiation at the
cellular level, though not necessarily only at the synapse; Dudai (2004) notes that, although morphological changes have been found in the synapse after new memory formation, the synapse and other neuronal regions, such as the cell nucleus, must act in concert for the process to occur. Synaptic consolidation is present in all species capable of forming long-term memories, and is hypothesized to involve a number of cellular mechanisms including modification of synaptic gene expression (see Dudai, 2004, Figure 2, for a model), and to last up to around 90 minutes after learning takes place.

The second form of consolidation, ‘systems consolidation’, is hypothesized to involve hippocampal networks that temporarily store a memory trace before it is passed to higher neocortical systems, after which the hippocampal trace decays (Dudai, 2004). Systems consolidation occurs over a longer time period (days or weeks) than synaptic consolidation (minutes or hours), and the two processes may occur either in parallel, or as a result of one another (Dewar et al., 2010).

The theoretical model of Dewar et al. (2010) states that that, after learning new information, an automatic (non-conscious) neural process occurs in which the information is encoded by a process of ‘replay’ of the information. Such a replay mechanism is not merely inferred from behavioural data and theory. For example, Peigneux et al. (2006) trained fifteen healthy participants on a spatial or procedural memory task between two functional magnetic resonance imaging (fMRI) sessions. At the second fMRI measurement, the authors observed persisting training-related activity, which lasted into a third fMRI measurement performed after the participants had a 30-minute (wakeful) rest. This study did not test whether retroactive
interference during this 30-minute session would interrupt the task-associated brain activity; given the evidence for the effects of retroactive interference discussed here, such an investigation would be a worthwhile avenue for future imaging research.

The consolidation view states that interruption of synaptic and systems consolidation processes by intervening tasks was at the root of the recall differences found by Müller and Pilzecker (1900), Cowan et al. (2004), and others. However, the consolidation theory has been questioned by Brown and Lewandowsky (2010), who note that the finding from several studies of retroactive interference (see Wixted, 2004), that retroactive interference both immediately after learning and immediately before the retrieval task impairs recall (that is, a ‘U’-shaped distribution of forgetting) cannot be explained by a simple view of consolidation. They propose a ‘temporal distinctiveness’ view, in which memory traces that are “temporally crowded” (p. 64) are more difficult to discriminate, and thus interfere with one another. Brown and Lewandowsky (2010) argue that this explains both ends of the ‘U’-shaped distribution without invoking consolidation, since the learning and retrieval at these times has been contaminated with the temporally close interference. It should be noted that the current research does not attempt to discriminate between these two theories of the retroactive interference effect, instead focusing on the practical applications of retroactive interference to the classroom.

Recent studies involving consolidation theory have moved beyond samples of neurological patients, and extended the investigation of minimal retroactive interference to healthy participants (that is, they have begun to follow up upon the work of Müller and Pilzecker, 1900, using modern experimental and statistical
methods). In their theoretical discussion, Dewar et al. (2010) state that improved memory following minimal retroactive interference should be found in healthy individuals, albeit to a lesser extent than in amnesics; this idea was tested by Dewar, Alber, Butler, Cowan, and Della Sala (2012) in two experiments with healthy older participants (aged around 70 years). Importantly, this experiment tested whether such effects can be measured over one week, a longer term than had previously been studied in this area (for example, as described above, the final test in Cowan et al., 2004 was after one hour). In the first experiment, fourteen participants learned two stories from a commonly-used memory scale followed either by wakeful resting with minimal interference—as with the studied described above, in a quiet, darkened room, for ten minutes—or by a distractor ‘spot-the-difference’ task. After a general recall test 15-30 minutes after learning, the participants left the laboratory and returned one week later for a surprise recall test on both stories.

Results showed large effects (all $\eta^2_p > .5$) of wakeful resting on recall one week later, compared to the distractor condition; the finding was replicated in a second experiment without a 15-30 minute test (it was possible that this test mediated the effects of learning in the original experiment, but this did not appear to be the case). As in the study of Cowan et al. (2004), the authors argue that the effect was not due to conscious practice: only a small fraction of the participants reported thinking about the stories during the relaxation period or between the two laboratory visits, and removing those participants from the analysis did not substantively alter the results. The authors conclude that the minimal interference paradigm applies not just beyond the boundaries of patient samples, but also across periods of at least one week.
The above results have clear implications not just for everyday life, but for education. If the wakeful-resting paradigm has effects on learning, would it be possible for teachers to instantiate it in their classrooms, leading to improved recall of new information taught to their pupils? Here, in two experiments involving primary school children, we tested this possibility. In the first, larger-scale experiment, we compared the effectiveness of teachers using in-class relaxation techniques before and after learning new material. In the second experiment, we tested a smaller number of children in a one-to-one setting, under more controlled conditions. Positive effects of minimal retroactive interference in either of these situations would provide compelling evidence that the consolidation process, discussed above, is also present in young children, and that this can be successfully utilized in the educational context.

**Experiment 4.1**

The first experiment reported here attempted to apply the results from previous wakeful-resting paradigms to classroom learning in primary school children. As noted in Chapter 3 of the present thesis, primary school teachers commonly use relaxation sessions (‘calm time’) to start lessons. Here, we designed a study in which teachers would simply change the timing of the relaxation session to either before or after the children learned some new information, and tested the children’s recall at one and five weeks after learning (a longer time interval than has been investigated in any previous retroactive interference study).

Given the evidence from the previous chapter that relaxation sessions performed prior to learning had no effect on memory, we used the ‘relax-before’
condition as a control condition, comparing it to a ‘relax-after’ condition involving wakeful resting after learning. If wakeful resting after learning has positive effects on consolidation of memories, we would expect a time × condition interaction, such that the recall in the ‘relax-after’ condition at later testing points should be higher than that in the ‘relax-before’ (control) condition.

**Method**

**Participants**

Two hundred and eighty-four primary school-aged children attending Towerbank Primary School, Edinburgh, participated in the experiment. Precise age data were not available, as the experimenter only collected testing data from the school and was not present in the school during the testing sessions, but the sample included eighty children from Primary 4 (three classes), sixty-six from Primary 5 (three classes), fifty-nine from Primary 6 (two classes), and seventy-nine from Primary 7 (three classes). Sex data were also unavailable. Consent forms were received from the parents or guardians of each child before their participation, and permission was received for the experiment from the University of Edinburgh Psychology Research Ethics Committee.

**Materials**

*Memory.* Memory was tested using the immediate and delayed recall stories from the Children’s Memory Scale (CMS; Cohen, 1997). There were two stories, one
for each experimental condition, that were scored for a potential forty-one ‘story units’ (for recall of each individual story detail) and six ‘theme units’ (for recall of overarching story themes). The stories were recorded by a voice actor to allow en masse administration to each class of children. Fifteen test questions (all yes/no questions) were available for each story.

Music. To accompany the ‘calm time’ relaxation sessions, two four-minute pieces of relaxing ambient music—Articulate Silences, Pt. 1 and Pt. 2, by minimalist electronic group Stars of the Lid—were provided to teachers. Music was played during the relaxation sessions in order to improve the ecological validity of the experiment, since teachers reported commonly using music during ‘calm time’ sessions.

Procedure

The experiment, as illustrated in Figure 4.1, involved activities on four separate days: two ‘learning phase’ days in which the relaxation session was performed either before or after learning new facts; a one-week test which included a free recall test and a quiz, and a five-week test that included only a quiz. All activities were completed in the participants’ normal classrooms, and were administered and supervised by their regular classroom teacher. The experimenter was not present at any of the activities described below.
Learning phase. On both consecutive learning phase days, the classroom teacher opened an envelope provided to them by the experimenter. Inside was an instruction run a ‘calm time’ relaxation session either before or after the children had heard the recording of the day’s assigned CMS story. The conditions were counterbalanced such that the initial condition, either ‘relax-after’ or ‘relax-before’, alternated from class to class.

For the relaxation session, the teachers were asked to instruct the children to relax for four minutes, to not talk to their classmates, and to try not to think about anything. During these four minutes, the teachers played the accompanying relaxing music, supplied to them in their envelopes on marked compact discs.

The recordings of the stories were played to the children twice, and after hearing the story twice, the children were asked to write down as much as they could remember from it on a blank sheet of paper. They were given five minutes to complete this task, before—in the relax-before condition—returning to their
classroom activities or—in the relax-after condition—relaxing in the manner described above.

The initial recall sheets for each participant were collected by the teachers and scored by the experimenter, blind to the condition under which the task was performed.

Testing phase 1. On the first testing day, one week after the second learning phase day, the children were reminded by the teacher of the general subject of each of the stories (the teachers told them the story was about “a robbery” or “a lifeguard”), and were again asked to note down as much as they could remember from each on a blank sheet of paper. They were given five minutes to complete this task.

Next, the classroom teachers read aloud seven of the questions from the CMS for each story, and the children responded by circling ‘yes’ or ‘no’ on printed quiz sheets they had been provided. On the back of these sheets, the children also answered ‘yes’ or ‘no’ to three questions: “Did you think about the stories at all during the last week?”; “Did you talk about the stories to your friends or family during the last week?”; and “Did you expect to do this quiz today?”

Testing phase 2. Finally, five weeks after the initial learning sessions, the teachers administered a test including seven different questions on each story from the CMS. Again this test was read aloud by the classroom teacher, and answered by circling ‘yes’ or ‘no’ on the quiz sheets provided.
CHAPTER 4: WAKEFUL RESTING AND MEMORY IN CHILDREN

Results

Not all teachers were able to administer the five-week test at the exact time requested: one class (in Primary 6) had their test administered seven weeks after the learning sessions, and one class (in Primary 7) had theirs eight weeks after the learning sessions. These differences were taken into account in the analysis described below. Unfortunately, the free-recall sheets from some classes were lost: the one-week recall test sheets were lost from two Primary 4 classes while the initial recall test notes were lost from one Primary 6 class and one Primary 7 class. All other free recall test notes were available, and quiz sheets were available from all classes from both quizzes. Table 4.1 provides descriptive statistics for all the tests and testing points in Experiment 4.1.

Table 4.1. Descriptive statistics for story recall, theme recall, and quiz performance at each measurement.

<table>
<thead>
<tr>
<th>Time</th>
<th>Measure (max. score)</th>
<th>Relaxation-After (Experimental)</th>
<th>Condition</th>
<th>Relaxation-Before (Control)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean (SD)</td>
<td>n</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Initial</td>
<td>Story units recalled (41)</td>
<td>18.89 (7.78)</td>
<td>219</td>
<td>19.25 (8.65)</td>
</tr>
<tr>
<td></td>
<td>Theme units recalled (6)</td>
<td>3.35 (1.73)</td>
<td>219</td>
<td>2.97 (1.87)</td>
</tr>
<tr>
<td>1 week</td>
<td>Story units recalled (41)</td>
<td>16.17 (7.93)</td>
<td>179</td>
<td>17.04 (8.87)</td>
</tr>
<tr>
<td></td>
<td>Theme units recalled (6)</td>
<td>3.64 (1.93)</td>
<td>179</td>
<td>3.06 (1.91)</td>
</tr>
<tr>
<td>5 week</td>
<td>Quiz (7)</td>
<td>5.53 (1.31)</td>
<td>260</td>
<td>5.69 (1.24)</td>
</tr>
<tr>
<td></td>
<td>Quiz (7)</td>
<td>5.31 (1.10)</td>
<td>257</td>
<td>5.43 (1.04)</td>
</tr>
</tbody>
</table>

The distributions of the story unit variables from the initial recall and the 1-week test were found not to be different from normal by visual inspection, whereas the remaining variables (theme unit recall and quiz scores) were all non-normally
distributed, with ceiling effects evident in the quiz score distributions, which were all negatively skewed. We thus carried out an separate analysis involving only the normally-distributed variables (this also allowed us to test the effects of relaxation on only recall memory, and not include both recall and recognition in the same model – see below for discussion of this issue). The means of these variables are shown in Figure 4.2 for each condition.

*Figure 4.2.* Means for story recall for each condition in Experiment 4.1 across time. Error bars are standard errors of the mean.

We used a 2x2 ANCOVA with the factors of time (initial versus 1 week) and condition (relaxation versus control), and the covariates of primary school year and weeks between the initial and final tests, to test whether relaxation slowed forgetting of the story material across one week. This showed that there was no main effect of
time \((F(1,141) = .00, p = .96)\), and a main effect of condition \((F(1,141) = 8.89, p = .003, \eta^2_p = .06)\). This latter effect indicated that performance in the relax-before (control) condition was higher at both testing points, compared to that in the relax-after condition. It is not clear why this should be so; contrary to the results of Chapter 3 of the present thesis, it may be that pre-relaxation produced better initial learning. Crucially for the present experiment, however, there was no time×condition interaction \((F(1,141) = .50, p = .48)\); that is, across one week, relaxation did not lead to better recall of story elements. The covariate of year had a significant effect, such that older children tended to have better recall scores \((F(1,141) = 5.43, p = .02, \eta^2_p = .04)\), and weeks between tests was also a significant influence \((F(1,141) = 9.92, p = .002, \eta^2_p = .07)\), possibly due to the longer times between testing tending (by chance) to be in classes of older children. Analysis without these covariates (i.e. a 2×2 ANOVA) also showed no time×condition interaction, suggesting a similar conclusion to the ANCOVA analysis.

Next, we estimated a fuller model of the data, this time using generalized linear mixed-effects modeling. This analysis was justified by the following characteristics of our data: first, some observations were missing, and thus the analyses were fit with maximum-likelihood estimation. Second, the data could be considered binomial: numbers of correct and incorrect answers at each measurement (Jaeger, 2008). Using an analysis that accounts for the binomial structure of the responses also allows us to avoid potential problems caused by the non-normality of some of the score distributions. The scores for each measure were therefore recoded as ‘correct’ and ‘incorrect’, and a binomial-family `lmer` from the ‘lme4’ package in R (Bates et al., 2014) was used to test the fixed effects of time (with two levels: initial
recall, and one week recall) and condition (relaxation versus control), again with the covariate of primary school year. Random effects were included to control for the differential abilities of individual children (intercept) to recall facts across time (slope); including instead a random effect of class, to control for differential administration of the tests by different teachers, made no substantial difference to the results reported below.

As shown in Table 4.2, significant main effects were found for time and for weeks between testing, but not (as in the ANCOVA analysis above) for condition. However, there was no time×condition interaction, again suggesting no slowing of forgetting across the week associated with the relaxation-after condition.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>Coefficient</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−2.62</td>
<td>0.29</td>
<td>−9.12</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time (week 1)</td>
<td>−0.53</td>
<td>0.05</td>
<td>−9.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Condition (relax-after)</td>
<td>−0.03</td>
<td>0.03</td>
<td>−.82</td>
<td>.42</td>
</tr>
<tr>
<td>Year</td>
<td>0.28</td>
<td>0.04</td>
<td>6.40</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Weeks between tests</td>
<td>0.19</td>
<td>0.06</td>
<td>3.38</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Time×Condition interaction</td>
<td>−0.05</td>
<td>0.05</td>
<td>−1.01</td>
<td>.31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance</th>
<th>SD</th>
<th>No. obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant (Intercept)</td>
<td>0.50</td>
<td>0.71</td>
<td>805</td>
</tr>
<tr>
<td>Time (Slope)</td>
<td>0.29</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

As can be seen from Table 4.2—and as would be expected—recall was significantly poorer with time. A similar result was found in a separate analysis including only the quiz results (that is, only recognition memory) from the week 1 quiz to the week 5 quiz: there was a significant effect of time ($b = −0.24$, $SE = .08$, $z =$
−2.96, \( p = .003 \)) but not of condition \((b = −0.21, \ SE = .18, \ z = −1.13, \ p = .26)\), year \((b = −0.01, \ SE = .03, \ z = −.27, \ p = .78)\), or weeks between testing \((b = −0.21, \ SE = 0.18, \ z = −1.13, \ p = .26)\). There was no significant time×condition interaction \((b = 0.06, \ SE = 0.11, \ z = 0.54, \ p = .59)\). So far, then, the analyses have shown that neither the story recall across one week or the quiz recognition across five weeks showed a positive influence of minimal retroactive interference.

Next, theme unit recall was examined as an outcome—for the first two measurement occasions, with the quiz as the outcome on the third occasion—using a similar mixed-effects model. This time, significant main effects were found for both time \((b = 0.66, \ SE = 0.06, \ z = 11.91, \ p < .001)\) and condition \((b = 0.44, \ SE = 0.08, \ z = 5.40, \ p < .001)\), and a significant time×condition interaction was found \((b = −0.22, \ SE = 0.06, \ z = −3.76, \ p < .001)\). The covariates were not significant influences: Non-significant effects were found for year \((b = 0.004, \ SE = 0.04, \ z = 0.11, \ p = .91)\) and for weeks between tests \((b = −0.003, \ SE = 0.05, \ z = −0.09, \ p = .93)\). The significant interaction here appeared to show that the slope of forgetting for theme units in the relax-after group was not as steep as that for the relax-before (control) group; that is, in line with our experimental hypothesis, wakeful resting after learning appeared to significantly slow forgetting. We discuss some caveats to this interpretation in the Discussion and General Discussion sections below.

It should be noted that running the analyses above including only those children who answered ‘no’ to the three questions on the reverse of the quiz sheets—that is, those children who did not report thinking or talking about the stories, and did
not expect to complete the quiz that day, leaving one hundred and forty-seven children for analysis—did not substantively change the results described above.

Discussion

In an experiment involving over two hundred primary school-age children, we attempted to replicate (conceptually) the effects of wakeful resting on memory, previously documented only in adults (e.g. Dewar et al., 2012). In our experimental design, such effects would be shown by a significant interaction between condition (relaxing before versus relaxing after the learning session) and time (initial recall versus subsequent measurements), signifying that the relaxation-after condition slowed forgetting. In the present experiment, only one such significant interaction was found: a mixed-effects model suggested that theme units—a measure of gist recall of a story—were more accurately remembered in the relax-after condition. However, in a similar analysis, there was no similar effect for story units, a far more detailed measure of precisely how many units of the story the participants recalled. For this reason, we would urge caution in the interpretation of the significant result found here; due to the number of alternative analyses performed, we cannot rule out the possibility that a spurious, false-positive result was found in this case.

Interestingly, one unexpected finding from Experiment 3.1 contradicts the results from our experiments in Chapter 3: the ‘relax-before’ condition, in which participants relaxed for a four-minute period before the learning session, appeared to cause a small improvement in recall across both time points. One possible reason for this contradiction could be that the memory measures used in the two experiments in
Chapter 3, both of which were novel, had poor reliability and were not standardized measures such as the CMS stories used here. The size of the effect of relaxation before learning was modest ($\eta^2_p = .06$), but was still above Ferguson’s (2009) critical minimal value for a practically significant effect.

Why were the strong effects of relaxation after learning, found in previous studies (Dewar et al., 2012), not found under the present circumstances? It may be that the memory consolidation effect found in previous studies of adults does not extend to young children. However, a number of experiment-specific factors may have led to our failure to replicate the effect. First, the relaxation time used here was shorter than that in previous studies, which have tended to use a ten-minute period of wakeful resting after learning. Here, only four minutes was used due to the difficulty of having a class of children relax quietly for ten full minutes. It may be the case that a longer period is needed to boost memory consolidation. Second, previous experiments have used one-to-one testing, and have not had participants relax in a classroom with around thirty fellow participants. The potential distractions inherent in this environment may have reduced the effectiveness of the wakeful resting period.

Third, as noted above, music was played in the relaxation session; this was not the case in any previous wakeful resting experiment. Fourth, to avoid repetitious testing, we split the available quiz questions into seven at each measurement point; there may not have been enough variance in these results to uncover any effects of wakeful resting, and near-ceiling effects were observed, despite our use of a standardized, age-normed test instead of the novel measures used in Chapter 3 of the present thesis. Fifth, since the experimenter was not present in the classrooms, we
cannot be certain that the teachers complied completely with the instructions. However, our analysis with a random effect of class should have to some extent controlled for this latter limitation.

Thus, the wakeful resting paradigm may not be appropriate for in-class use. However, it remains to be seen whether the paradigm works for children’s memory in the same manner it appears to for adults (Dewar et al., 2012). The tradeoff for the high ecological validity of Experiment 4.1 was low experimental control; a close replication under more controlled conditions is therefore necessary to test whether children benefit from minimal retroactive interference in the same conditions as used in previous experiments. Experiment 4.2 describes a closer replication of a previous experimental design (that of Dewar et al., 2012), again with children as participants.

**Experiment 4.2**

Given the constraints on our study design due to the classroom setting, Experiment 4.1 was not a direct attempt to replicate of previous wakeful-resting experiments (e.g. Dewar et al., 2012). For this reason, we could not tell whether the null results we obtained were due to the technique being ineffective in children, or the ineffectiveness of the particular design we used, which was not under controlled conditions, and had the potential for distractions of many kinds in the relaxation period. Thus, to test the theoretical validity of the minimal retroactive interference paradigm in children, regardless of its practical applicability, in Experiment 4.2, we carried out a more exact replication of Dewar et al.’s (2012) experimental design,
with the only substantial change being the much younger participant age (primary school-age children as opposed to older adults).

Method

Participants

Fourteen primary school-aged children, nine boys and five girls, aged an average of 8.43 years (SD = 1.40) attending an after-school club (Bruntsfield Kidzcare, Edinburgh) participated in the experiment. The University of Edinburgh Psychology Research Ethics Committee gave permission for the study, and signed informed consent notes were returned from each child’s parent or guardian before they took part (twenty forms were returned, but time constraints—the experiment was conducted one month from the end of term before the summer holiday period began—only allowed us to test fourteen children). The after-school club was gifted £5 in book tokens for each child’s participation (£70 in total).

Materials

*Recall task.* As in Experiment 4.1, memory was tested using the immediate and delayed recall stories from the CMS.

*Distractor task (puzzles).* A series of paper and pencil ‘odd-one-out’ and ‘match-the-pair’ puzzles were taken from a children’s puzzle book. These tasks involved spotting and circling very small differences in one of nine similar cartoon
pictures on A4 sheets of paper. One puzzle of high difficulty was used at the end of the ‘relax’ condition (see below), while a booklet of fourteen items was administered in the ‘puzzle’ condition.

**Procedure**

The procedure of Experiment 4.2 closely replicates that of Dewar et al. (2012).

*Learning phase.* Children sat comfortably in a quiet room in the school to which Bruntsfield Kidzcare is attached (Bruntsfield Primary School, Edinburgh). The only other people in the room were the experimenter and a representative from Bruntsfield Kidzcare. The experimental procedure is shown in Figure 4.3.

*Figure 4.3. Design of Experiment 4.2 (adapted from Figure 1 in Dewar et al., 2012).* For illustration, this shows the design for half of the participants; the design was counterbalanced as per the description below.

First, the participants were informed that they were going to hear two stories, complete some puzzles, and relax for a time. The experimenter then read the first story aloud twice, and the participants were asked to repeat back as many story details as they could remember. This recall was audio recorded and later scored as each participant’s initial recall measure. The participant was then asked to relax for the next ten minutes, in any way they chose (sitting back in their chair, putting their head
on the table in front of them, etc.), trying not to think about anything at all, and also trying to stay awake. The experimenter and the Kidzcare representative monitored participants’ compliance with the instructions, reminding them to continue relaxing if they showed signs of distraction such as looking around the room or talking. A stopwatch running to ten minutes was displayed in full view of the participants throughout the silent relaxation session. After this period, the participant was given two minutes to complete the single high-difficulty puzzle mentioned above. They were also asked whether they had thought about the content of the stories during the ten minutes of relaxation.

Next, the participant was read the second story aloud twice, and again audio recorded while they recalled story details for the initial recall measure. The experimenter then gave the participant the puzzle booklet, and allowed them twelve minutes to complete it, instructing them that they could skip any puzzles they found too difficult, and could return to any skipped puzzles after coming to the end of the booklet. After this twelve-minute period, the learning phase was complete and the children returned to their after-school club activities.

The order of the stories was counterbalanced in a repeating ABAB order throughout the participants, while the order of the conditions was counterbalanced in a repeating AABB order, such that the first participant had Story 1-Relaxation-Story 2-Puzzle, the second had Story 2-Relaxation-Story 1-Puzzle, the third Story 1-Puzzle-Story 2-Relaxation, and so on.
Testing phase. One week later, the experimenter returned to the after-school club and administered a surprise test consisting of the fifteen questions related to each story from the CMS. The experimenter read each question aloud, and participants responded by circling the correct answer on a sheet with a printed ‘YES’ and ‘NO’ for each question. Most testing sessions involved multiple participants simultaneously, but efforts were always made to ensure each participant could not see another participant’s answer sheet. After an answer had been provided for each of the 30 questions (15 for each story), the experiment was complete. The order of tests was counterbalanced in a repeating AABB order, such that the first two participants had tests for Story 1, then Story 2, the second two participants the reverse, and so on.

Results

All participants stayed awake for the full duration of the relaxation session, and all participants completed at least some of the puzzles in the booklet. Eleven participants were tested exactly one week after the learning phase; the remaining three were tested five, six, and fourteen days afterwards due to differences in the times they were present at the after-school club. Only three participants reported having thought about the stories during the relaxation session; removing these three participants from the analysis did not substantially alter the results presented below.

Initial recall. In the relaxation condition, the participants initially recalled an average of 25.64 story details (SD = 8.78), and in the puzzle condition, the participants recalled on average 28.29 details (SD = 6.21), both from a maximum of 41 units. Shapiro-Wilk tests for normality in small samples indicated that neither
distribution was significantly different from normal. The higher free recall scores found here compared to Experiment 4.1 may reflect the differences in test administration – unlike in Experiment 3.1 where the free recall test was written, in Experiment 4.2, the children recalled the story units orally, so could more easily convey a larger number of responses before forgetting occurred.

 Tested memory. On the 1-week follow-up quiz, the participants correctly answered 11.50 questions on average (SD = 1.45) in the relaxation condition, and 11.29 questions on average (SD = 1.94) in the puzzles condition, both from a maximum score of 15. Again, normality testing indicated that the distributions were not significantly different from normal. A related-samples t-test on the quiz performance across the two conditions indicated no significant difference between them, t(13) = .29, p = .78; relaxation after learning the information did not improve performance compared to completing a distractor task immediately afterwards.

 ANCOVA analysis. The data, illustrated in Figure 4.4, were first analyzed using a 2×2 ANCOVA with the within-subjects factors of time (initial versus tested memory) and condition (relaxation versus puzzle), and the covariates of age, primary school year, sex, and days between the learning and testing phases. For analysis, scores were converted to percentages of the maximum scores (41 and 15 for initial recall and tested recognition, respectively). It should be noted that this analysis merges recall data (initial) and recognition data (quiz), a weakness discussed in the General Discussion section, below.
Figure 4.4. Memory performance (z-scores) on the initial and 1-week (quiz) measures in Experiment 4.2. Error bars are standard errors of the mean.

The ANCOVA analysis found no significant main effect of time; participants recalled no greater proportion of the maximum available test score at initial recall than in the testing phase \(F(1,10) = .01, p = .93\). There was no significant main effect of condition \(F(1,10) = .89, p = .37\); nor was there a significant time×condition interaction \(F(1,10) = .23, p = .68\). Thus, the ANCOVA found no effects of relaxation on story recall across the week. None of the covariates had effects; non-significant results were found for age \(F(1,10) = .01, p = .94\), year \(F(1,10) = .01, p = .91\), sex \(F(1,10) = .07, p = .80\) and days between phases \(F(1,10) = .00, p = .97\).

A second approach taken was to take into account each child’s initial performance by scaling the quiz answers by the number of correct responses at the initial recall test. We used these scores in a similar ANCOVA to that described above. This approach showed no effect of time \(F(1,10) = 1.76, p = .21\) or of condition
(F(1,10) = .18, p = .68); the effects of the covariates were similar to those reported above.

Mixed-effects model analysis. The ANOVA approach above has several weaknesses. First, the initial test was one of free recall, and the second was a written yes/no quiz; aside from the format differences, the tests had different maxima, and are thus not directly comparable. Second, the test phase results are binary choices, and should thus be analysed using a binomial test. Third, it does not allow the estimation of random effects in the dataset. For these reasons, as in the previous chapter, we re-analysed the data using a generalized linear mixed-effects model ‘lmer’ in the ‘lme4’ package for R (Bates et al., 2014). We first arranged the data into numbers of correct and incorrect responses (the latter including lack of responding), then entered the data, along with a random effect of each participant’s ability to recall story elements across time. As in Experiment 4.1, the model was fit using maximum likelihood estimation.

The results of the mixed-effects model are shown in Table 4.3. Unlike the previous analysis, time had a significant main effect, indicating that a greater proportion of correct answers were given in the testing phase than at the initial recall measurement. There was also a main effect of condition: once the random-effects structure of the data was taken into account, participants remembered more facts in the puzzle condition than in the relaxation condition. However, importantly, there was no time×condition interaction, indicating that neither of the conditions slowed forgetting across time. As in the previous analysis, none of the covariates had significant effects.
Table 4.3. Results of the generalized linear mixed model for Experiment 4.2.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>b</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.61</td>
<td>1.85</td>
<td>0.33</td>
<td>.74</td>
</tr>
<tr>
<td>Time – Testing Phase</td>
<td>0.63</td>
<td>0.29</td>
<td>2.20</td>
<td>.03</td>
</tr>
<tr>
<td>Condition – Puzzle</td>
<td>0.32</td>
<td>0.13</td>
<td>2.43</td>
<td>.02</td>
</tr>
<tr>
<td>Age</td>
<td>0.06</td>
<td>0.30</td>
<td>0.20</td>
<td>.84</td>
</tr>
<tr>
<td>Primary school year</td>
<td>−0.06</td>
<td>0.28</td>
<td>−0.22</td>
<td>.83</td>
</tr>
<tr>
<td>Sex</td>
<td>0.11</td>
<td>0.30</td>
<td>0.37</td>
<td>.71</td>
</tr>
<tr>
<td>Days between phases</td>
<td>−0.05</td>
<td>0.06</td>
<td>−0.82</td>
<td>.41</td>
</tr>
<tr>
<td>Time×Condition</td>
<td>−0.39</td>
<td>0.26</td>
<td>−1.50</td>
<td>.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effects</th>
<th>Variance</th>
<th>SD</th>
<th>No. obs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant (Intercept)</td>
<td>0.53</td>
<td>0.73</td>
<td>56</td>
</tr>
<tr>
<td>Time (Slope)</td>
<td>0.66</td>
<td>0.81</td>
<td></td>
</tr>
</tbody>
</table>

Discussion

In this experiment, a close replication of that of Dewar et al. (2012) but with children instead of older adults as participants, we failed to find an effect of wakeful resting on recall one week later. Children performed just as well on the final test for stories after which they rested as for stories after which they completed the distractor task.

One reason this experiment may not have replicated previous results concerns experimental power. Although the effects of wakeful resting found by Dewar et al. (2012) were large, and our sample size \((n = 14)\) matched that of their first experiment, it is likely that our power to detect any effects was low: the power of a related \(t\)-test to detect even large effects \((d = .50)\) between two conditions with fourteen participants
is around 40%, and our power was lower given our more complex $2 \times 2$ ANCOVA analysis. As noted above, due to time constraints, we were unable to test all of the 20 children for whom we had received permission before the end of term, reducing our statistical power to low levels; power of 80% to detect a large effect would have required a sample of around 35 participants, which was beyond even the number who handed in permission forms.

Whereas Experiment 4.2 was a very close replication of Dewar et al. (2012), it was not exact. For instance, we did not have participants rest in a darkened room, as in Dewar et al.’s cubicle-based testing. In addition, testing was done on paper, with the experimenter verbally leading the child through each test, rather than the computerized testing in Dewar et al. experiment. These differences may have been enough to attenuate the effects of wakeful resting, though it is unlikely they would have reduced the large effects found in Dewar et al. so much as to render them non-significant.

In addition, Experiment 4.2 was a replication of the second experiment in Dewar et al. (2012), which did not include a test 15 minutes after the initial learning session. It may have been the case that the benefits of minimal retroactive interference manifest in the short-, but not the longer-term; we were not able to test this possibility. However, even if such a short-term effect exists, it is unlikely that it would be practically significant in an educational context.
General Discussion

Two experiments involving primary school children, in two quite different settings, tested the hypothesis that a brief session of wakeful resting—involving minimal cognitive or sensory interference—immediately after learning some new to-be-recalled information should reduce forgetting over time. Experiment 4.1, in a sample of over two hundred children who learned new information in a classroom setting, showed that wakeful resting did not facilitate story recall across one or five weeks. Experiment 4.2, in a sample of fourteen children in a one-to-one setting under tighter experimental conditions, showed that relaxation immediately after hearing a story did not significantly slow forgetting of the story elements across one week. In neither experiment was the time × condition interaction necessary to demonstrate a slowing of forgetting by wakeful resting present.

These findings are, at least at first glance, incompatible with both ‘consolidation’ (Dewar et al., 2010) and ‘temporal distinctiveness’ (Brown & Lewandowsky, 2010) theories of forgetting, since they do not show the previously well-replicated pattern of results. Since this experiment represents, to our knowledge, the first investigation of this effect in young children, our main finding may be that the effect of minimal retroactive interference on memory, previously found in adults, does not apply to young children. There are two potential (and mutually contradictory) reasons for this finding. First, it could be that children’s memories are more robust to retroactive interference than those of adults; this seems unlikely given the higher general distractability and lower self-control of children in comparison to adults. Second, children may engage in more distracting rumination and inner
thoughts than adults when instructed to relax; this would have lessened the effectiveness of the relaxation session, and raises the possibility that children who are trained in a relaxation technique would benefit more than untrained children from minimal retroactive interference. However, before drawing either of these conclusions, we should enumerate a number of methodological issues that may have led to our failure to find the effect.

Specific reasons behind the failure of each experiment to conclusively demonstrate the wakeful resting effect are described in the Discussion sections for each, above. For instance, idiosyncratic factors relating to the manner in which teachers carried out the experiment, or the shorter relaxation time, may have affected the results of Experiment 4.1; our lack of statistical power may have caused us to make a Type II error in Experiment 4.2. Here, we outline some more general explanations for the failure to find the results in schoolchildren.

First, a major limitation in both experiments concerns the use of recognition tests, in which participants were given a statement and asked whether they agreed with it (by circling ‘yes’ or ‘no’), rather than free recall tests at the final testing occasion. Previous studies have demonstrated effects of minimal retroactive interference on recall – for instance, Cowan et al. (2004) used a test of free recall of word lists, and whereas Dewar et al. (2012) used a test involving stories, as in the present experiment, they did not use the standardized tests at follow-up testing, instead repeating the free recall test. Recall tests, although related to recognition, use distinct neural subsystems: Staresina and Devachi, (2006), for example, found evidence for free-recall-specific brain networks, not activated during a recognition
memory task, in an fMRI study of eighteen participants. No studies of the retroactive interference effect so far have systematically compared recall and recognition memory, or investigated whether these are differentially affected by interference during the purported consolidation period, although the separate analyses we ran in Experiment 4.1 would indicate that neither recall nor recognition memory was affected by the relaxation-after condition. Future studies should investigate whether this is the case, and more fully identify the types of memory performance we should expect minimal retroactive interference to facilitate.

Time constraints in Experiment 4.2, and at the five-week mark in Experiment 4.1, obliged us to use a brief quiz at the final measurement. However, in Experiment 4.1 a recall test was used at immediate testing and at the one-week test; we were thus able to assess scores on recall-only tests across time. No effect was found over this time, suggesting that the effects are not present in children for recall tests. Nevertheless, the experiment-specific effects described above may be the explanation for the lack of a finding in this analysis.

Second, the precise setting of the wakeful resting was different from previous studies. This was most obvious in Experiment 4.1, where the participants relaxed in a room full of their peers (as opposed to alone). Even in the less-distracting environment of the relaxation session in Experiment 4.2, the room was not darkened as in previous experiments, and the participant was not alone in the room (the experimenter and an after-school-club representative were both present). It may be that a more sensory-depriving atmosphere, with both low light levels and low noise levels, is required for more effective consolidation of memories.
Given the wide-ranging methodological issues with these experiments, we cannot draw any firm theoretical conclusions from our failure to replicate previous results. Future research should attempt to overcome the methodological issues and should, in the initial stages, focus on one-to-one experiments with children, to establish whether the minimal retroactive interference phenomenon applies to young children in more controlled settings. Recall tests should be used, and large samples should be tested to ensure adequate power. Finally, the recall conditions outlined in previous studies (quiet and darkness) should be followed more closely to ensure direct replications of the wakeful resting technique under which effects have been found.

**Conclusion**

Our two studies produced initial, though flawed, evidence that the boosting of recall by minimal retroactive interference found previously in older adults and patient samples does not apply in children in a classroom or a one-on-one context. Future studies should be performed to rule out the methodological issues that could have given rise to these results and identify under which circumstances, if any, children’s learning can be boosted using this technique.
Retrieval practice, with or without mind mapping, boosts fact learning in primary school children.

** The experiments reported in this chapter are published as: Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. *PLoS ONE, 8*, e78976.
Abstract

Retrieval practice is a method of study in which testing is incorporated into the learning process. This method is known to facilitate recall for facts in adults and in secondary-school-age children, but existing studies in younger children are somewhat limited in their practical applicability. In two studies of primary-school-age children of 8-12 years, we tested retrieval practice along with another study technique, mind mapping, which is more widely-used, but less well-evidenced. Children studied novel geographical facts, with or without retrieval practice and with or without mind mapping, in a crossed-factorial between-subjects design. In Experiment 5.1, children in the retrieval practice condition recalled significantly more facts four days later. In Experiment 5.2, this benefit was replicated at one and five weeks in a novel, larger sample of schoolchildren. No consistent effects of mind mapping were observed. These results underline the effectiveness of retrieval practice for fact learning in young children.
Introduction

Students tend to believe that the best way to learn new facts is by prolonged or repeated exposure (Karpicke, Butler, & Roediger, 2009), but a body of evidence in cognitive psychology attests to the value of retrieval practice for boosting learning (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Karpicke, 2012; Roediger, Finn, & Weinstein, 2012; Roediger & Karpicke, 2006a). This is sometimes known as the ‘testing effect’, because the critical feature is that studying should include test periods, during which the student tries to recall the facts without checking the source material.

Whereas most experimental evidence for retrieval practice comes from adult participants tested in laboratory settings, a number of studies have applied the technique in school classrooms, mostly with children aged 11 years and above (Carpenter, Pashler, & Cepeda, 2009; McDaniel, Agarwal, Huesler, McDermott, & Roediger, 2011; Metcalfe, Kornell, & Finn, 2009; Roediger, Agarwal, McDaniel, & McDermott, 2011; Spitzer, 1939). These studies have shown that retrieval practice facilitates learning of a variety of materials in the classroom. For instance, Roediger et al. (2011) showed that retrieval practice can be incorporated into the school curriculum, with low stakes quizzing at various points throughout the term showing learning benefits on later exams in US sixth grade children (aged 11-12).

Some experiments have also demonstrated benefits of retrieval practice in samples including younger children. An early experiment (Gates, 1917) had children aged 6-14 years learn nonsense syllables and biographical material, with varying amounts of time devoted to self-testing by silent recitation. As self-testing time
increased, so did the amount of material recalled three to four hours later. Much more recently, in two experiments (Rohrer, Taylor, & Sholar, 2010), both with 28 children aged 9-11, groups who were tested immediately after learning fictional map locations had better recall for the locations one day later than a ‘study only’ group, and the testing effect ‘transferred’ to questions more complex than those on the immediate test (for other experiments including young children, see Bouwmeester, & Verkoeijen, 2011; Fishman, Keller, & Atkinson, 1968; Rea & Modigliani, 1985).

However, these studies have a number of limitations to their practical applicability. They have tended to focus on small samples (e.g. Fishman et al., 1968; Rea & Modigliani, 1985; Rohrer et al., 2010), to use short testing intervals (e.g. Gates, 1917; Rohrer et al., 2010), or to test relatively simple materials such as word lists (e.g. Bouwmeester & Verkoeijen, 2011). To our knowledge, no one study has addressed these limitations in a sample that includes children of primary school age (normally up to 11/12 years of age in the UK). The present study tests the effect of retrieval practice in two reasonably large samples of children, across intervals of up to five weeks, on educationally-relevant geographical facts.

Despite its growing evidence base, retrieval practice is used rarely in schools (Roediger & Karpicke, 2006a), especially by comparison with some other less well-evidenced techniques. For instance, mapping techniques have become popular in classrooms worldwide (Buzan & Buzan, 2006). These include ‘mind-mapping’, the drawing of diagrams to organize facts into categories, and the more sophisticated ‘concept-mapping’, in which the diagram visually represents the inter-relations between facts. Proponents claim that individuals with a ‘visual’ learning style benefit
from these techniques (Abi-El-Mona & Adb-El-Khalick, 2008). However, a recent review concluded that there is no good evidence for the claim that a student’s preferred ‘learning style’ influences their learning outcome from different instructional techniques (Pashler, McDaniel, Rohrer, & Bjork, 2009).

In addition, the evidence regarding the effect of mind mapping on learning is sparse and mixed. One study (Farrand, Hussein, & Hennessy, 2002) reported a benefit of mind mapping on fact learning in medical students, but other studies in similar groups have found no effect (D’Antoni, Zipp, Olson, & Cahill, 2010; Wickramasinghe, Widanapathirana, Kuruppu, Liyanage, & Karunathilake, 2007). An encouraging finding in a younger population (sixty-two 13-14 year olds) was that the use of mind maps throughout a science course yielded higher scores on a later test than did standard note-taking (Abi-El-Mona & Adb-El-Khalick, 2008). We are aware of no similar studies in children of primary school age, despite the fact that mind mapping is very commonly used with younger students.

One prior study (Karpicke & Blunt, 2011a) has incorporated both mapping and retrieval practice. Undergraduate students who were given an immediate test on facts they had studied had better recall at one week than did those who studied the facts once, studied them repeatedly, or who drew a concept map. The authors concluded that retrieval practice was superior to concept mapping (for further discussion of this result, see Mintzes et al., 2011; Karpicke & Blunt, 2011b). Even so, it should be emphasized that retrieval practice does not preclude mapping techniques, and it is possible that their combination (e.g. self-testing using a mapping technique) would be more beneficial than either technique alone. This potential to combine
techniques was noted by the authors (Karpicke & Blunt, 2011a), while Roediger (2013) has discussed the need for studies on combinations of learning techniques.

This study, like that of Karpicke and Blunt (2011a), tested the effects of retrieval practice and a mapping technique on fact learning, but with three major differences. First, we focused on much younger participants, primary school children aged 8-12 years. Second, for this age group we used simple mind mapping (with which the children were already familiar) rather than more complex concept mapping. Third, we used a crossed-factorial design to test not only the effects of retrieval practice and mind mapping, but also their combination. In Experiment 3.1, we tested the effects of these techniques in a sample of 109 children, within a school week. In Experiment 5.2, we replicated our findings in a larger sample, with a longer interval between the learning and test phases, and a somewhat more challenging task.

We hypothesized, consistent with previous work, that retrieval practice should improve memory for facts across time. Given the lack of solid previous evidence, we made no directional prediction regarding mind mapping, or its combination with retrieval practice, which may variously prove to have additive positive effects on memory beyond retrieval practice by allowing the use of a visual encoding strategy, distract from the task at hand and prove detrimental, or make no appreciable learning difference.
Experiment 5.1

Method

Participants

Participants in Experiment 5.1 were 109 pupils (59 female) from Primary 5 and 7 classes (two classes from each year) at Towerbank Primary School, Edinburgh, aged 8-12 years ($M = 10.29$ years, $SD = 1.07$; numbers and ages per class are shown in Table 5.1, including those from Experiment 5.2, for comparison). The experiment was approved by the Psychology Research Ethics Committee at the University of Edinburgh, and written informed consent was obtained from the parent or guardian of each participating child before the study began.

Table 5.1. Numbers and ages of children per year group across the two experiments.

<table>
<thead>
<tr>
<th>Experiment 5.1</th>
<th>Experiment 5.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary 5</td>
<td>Primary 7</td>
</tr>
<tr>
<td>$n$</td>
<td>59</td>
</tr>
<tr>
<td>$M$ age in years ($SD$)</td>
<td>9.35 (0.31)</td>
</tr>
</tbody>
</table>

Materials

Four single-sided four-paragraph ‘factsheets’ each concerned a different country likely to be unfamiliar to young children in the UK (Senegal, South Korea, Peru, and Iran). Two example factsheets (from Experiment 5.2, which had similar materials; see below) are shown in Figs 1a and 1b; the text of all the fact sheets is provided in Appendix B. Sheets for Primary 5 children had 11 facts (~100 words), and sheets for Primary 7 children had 16 facts (~130 words). All factsheets were
placed in unmarked envelopes; those to be given to the non-retrieval group also contained a separate blank sheet of paper for note-taking.

_Figure 5.1._ Example factsheets for Primary 4 (a) and Primary 7 (b) in Experiment 5.2; example notes from one Primary 4 child (c) and an example mind map from one Primary 7 child (d) in Experiment 5.2.

![factsheets examples](image)

**Learning phase**

In the learning phase, children were randomly assigned to a retrieval practice condition within classes (half of the children in each class used retrieval practice), while mind mapping conditions were arranged between classes (one of the two classes from each year used mind mapping). The following paragraphs describe the details of this procedure.

On the Monday of the experimental week, the experimenter visited each classroom and teachers split the classes into two groups (‘retrieval practice’ and ‘non-retrieval’), by running through the class register and assigning successive children to
alternating groups (the assortment of children to groups was thus arbitrary). The two
groups were then seated on opposite sides of the classroom, in sub-groups of four.
Within each subgroup, each child was given a different one of the four factsheets. In a
few classrooms, where seating required that some subgroups were larger than four,
factsheets were handed out so that no child sat directly beside a classmate with the
same country. In these larger groups, and in some groups with fewer than four
children, factsheets were given out in the order Senegal-South Korea-Peru-Iran to
attempt an approximately equal distribution of the four sheets across classrooms and
experimental conditions.

The experimenter explained that the children were to learn some facts for a
quiz at the end of the week. They were asked to open their envelopes and to read their
fact sheets without writing anything. This initial reading period lasted five minutes.
For the next five minutes, the children made notes on the facts. Children in one group
(non-retrieval) kept the factsheet in view throughout this period and made notes on
the blank sheet of paper. Children in the other group (retrieval practice) were required
to turn the factsheet over, and to make notes on the blank side. The experimenter and
the teacher monitored compliance with instructions.

In half of the classes (one class from each year, randomly selected), the
children made notes in the form of a mind map, writing the name of the country in the
center of the page, and drawing lines outwards to facts grouped by their categories.
All classes at the school regularly used this form of mind mapping to represent facts,
such as historical knowledge, or the attributes of characters in reading books. In the
other half of the classes, the children were asked to make notes in any way they liked except for mind mapping. Examples of mind maps and notes made in this study are shown in Figures 5.1c and 5.1d, respectively.

Next, all children read the sheets again for three minutes, without writing anything. Finally, all children continued making notes, as before, for five minutes. The experimenter then collected the sheets. Note that the learning phase thus lasted for 18 minutes overall, but the retrieval practice group had the factsheets visible for only eight minutes, whilst the non-retrieval group had the factsheets visible throughout.

**Testing phase**

On the Friday of the experimental week, the children were given a written recall test, with one question for each fact on their factsheet. This test was administered by the experimenter, who read the questions out loud to the whole class. The first question asked the name of the country; the remaining questions were in a fixed, pseudo-random order. Each question was read once, and the children were given as much time as needed to write each answer before the next question. All questions are shown in Appendix B. Children were told they would still gain a mark for misspelled correct responses. The experimenter later scored the tests; since quiz sheets were identical in all four experimental conditions, this scoring was blind to the conditions in which each child participated. Half-marks were awarded for partial, but correct, answers (e.g. “Korea” for “South Korea”; “hot” for “hot and dry”).
Results

Approximately equal numbers of children learned about each of the four countries (Senegal: 31; South Korea: 28; Peru: 25; Iran: 25; see Table 5.2 for the distributions of these sheets across the conditions of the experiment). One sheet from the learning phase was lost, leaving 108 sheets. For descriptive purposes, scores are given below (and in the tables and figures) in percentage terms; however, z-scores were calculated to make the Primary 5 and Primary 7 tests comparable, and are used in all analyses below (re-running the analyses on the percentage data produced near-identical results). Overall, 82.85% of facts were recorded on the note sheets in the learning phase, and a 2×2 ANOVA (retrieval practice condition × mind maps condition) confirmed that there were no significant differences in percentage of facts recorded between conditions [mean difference = 4.95% in favour of no mind maps ($F(1, 105) = 1.58, p = .21$); mean difference = .47% in favour of non-retrieval ($F(1, 105) = .01, p = .91$)]. Descriptive statistics, with means and standard deviations for performance in the learning phase, are shown in Table 5.3.

Table 5.2. Distributions of each factsheet across the four cells of the experiment in Experiment 5.1.

<table>
<thead>
<tr>
<th>Country on Factsheet</th>
<th>Retrieval Practice</th>
<th>Non-retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mind Maps</td>
<td>No Mind Maps</td>
</tr>
<tr>
<td>Senegal</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>South Korea</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Iran</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Peru</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 5.3. Mean performance (SD) in the initial learning phase (percentage of all potential facts written on notes) for Experiment 5.1, broken down by condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mind maps</th>
<th>No mind maps</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>82.87 (17.55)</td>
<td>81.95 (19.12)</td>
<td>82.40 (18.18)</td>
</tr>
<tr>
<td>Non-retrieval</td>
<td>77.47 (24.63)</td>
<td>88.29 (19.00)</td>
<td>83.26 (22.27)</td>
</tr>
<tr>
<td>Total score</td>
<td>80.17 (21.35)</td>
<td>85.34 (19.15)</td>
<td>82.85 (20.31)</td>
</tr>
</tbody>
</table>

The mean percentage scores on the recall test are shown in Table 5.4 for each condition. To test the effects of retrieval practice and mind mapping on fact recall, we ran a 2×2 ANCOVA, with the between-subjects factors of retrieval practice group (retrieval practice vs. non-retrieval) and mind mapping group (mind map vs. no mind map). A multiple regression including all potential covariates—age, sex, test/year (Primary 5 vs. Primary 7), country on the factsheets (of the four available), number of facts recorded during the learning phase—indicated that only the number of facts recorded in the learning phase was significantly related to the final test score ($p < .001$; $p$-values for other variables = .50-.94), and thus only this variable was included as a covariate in the ANCOVA. An additional analysis that, instead of using this covariate, scaled the test scores by the number of facts recorded during the initial session, did not appreciably alter the main results reported here.

Table 5.4. Mean percentage scores (SDs) and sample sizes for Experiment 5.1, for those with valid data.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mind maps</th>
<th>No mind maps</th>
<th>Total score</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>72.09 (21.58)</td>
<td>77.31 (24.22)</td>
<td>74.70 (22.86)</td>
<td>52</td>
</tr>
<tr>
<td>Non-retrieval</td>
<td>68.31 (29.80)</td>
<td>64.43 (28.26)</td>
<td>66.23 (28.79)</td>
<td>56</td>
</tr>
<tr>
<td>Total score</td>
<td>70.20 (25.83)</td>
<td>70.41 (27.02)</td>
<td>70.31 (26.33)</td>
<td>108</td>
</tr>
</tbody>
</table>
Children in the retrieval practice group recalled significantly more facts than those in the non-retrieval practice group \( (F(1, 104) = 6.33, p = .01, \eta^2_p = .06) \). There was no main effect of mind mapping \( (F(1, 104) = 1.93, p = .17) \), but an interaction was found between the conditions \( (F(1, 104) = 10.66, p = .001, \eta^2_p = .09) \): Post-hoc testing indicated that using mind maps was more effective than not when in the non-retrieval condition (mean difference = 14.93%, \( p = .001 \); this large difference was apparent only after adjusting for the covariate, and is not evident in the raw means shown in Table 5.4), but offered no learning advantage when in the retrieval practice condition (mean difference = -6.16%, \( p = .19 \)).

The covariate also had a significant influence on the outcome: Those who noted more facts tended to recall a higher percentage of facts later \( (F(1, 104) = 166.49, p < .001, \eta^2_p = .61) \). Re-running the analysis without the covariate resulted in no main effect of either retrieval practice group \( (F(1, 105) = 2.16, p = .14) \) or mind map group \( (F(1, 105) = .01, p = .91) \), and no interaction \( (F(1, 105) = .68, p = .41) \). The retrieval practice effect was thus reliant on the inclusion as a covariate of the facts recorded in the learning phase, but the inclusion of the covariate was, in our view, justified: Taking into account baseline memory ability led to a more accurate estimation of the model results.

Whereas the data used here met the assumptions for ANCOVA, it could be argued that a logit analysis is more appropriate, since the test scores are binomial counts (Jaeger, 2008). In addition, since each year group were administered tests with different numbers of answers (more difficult tests for older children), we had more information for the older children, and thus more reliable results. Collapsing the
scores across these tests into z-scores, as we did for the ANCOVA analysis above, results in a loss of this extra information.

For these reasons, we provide a secondary analysis of the data from both of our experiments, and instead of using an ANCOVA we use a generalized linear mixed-effects model that takes into account the binomial nature of the data and the different tests. This analysis was carried out using the ‘lmer’ function in the R package ‘lme4’ (Bates et al., 2014).

Instead of using the z-score of test ability, the dependent variables were the number of facts correctly written down, and the number of facts incorrectly written down or not written down. The model allowed for main effects of, and the interaction between, retrieval practice group and mind mapping group, and controlled for the number of facts recorded in the learning session. A random intercept of participant was added to indicate that participants will differ in their ability to recall facts.

The results of this analysis are shown in Table 5.5. The pattern of results was similar to that in the ANCOVA analysis: a significant effect of retrieval practice group, no significant effect of mind mapping group, and a significant retrieval group × mind mapping group interaction, along with a significant influence of the single covariate, number of facts recorded during the learning session.
Table 5.5. Results of the generalized linear mixed model for Experiment 5.1.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>b</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.82</td>
<td>0.23</td>
<td>8.00</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Group – Non-retrieval</td>
<td>-1.45</td>
<td>0.30</td>
<td>-4.75</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Group – Mind Maps</td>
<td>-0.58</td>
<td>0.31</td>
<td>-1.85</td>
<td>.06</td>
</tr>
<tr>
<td>Facts recorded in learning session</td>
<td>1.26</td>
<td>0.11</td>
<td>11.12</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Retrieval × Mind Maps</td>
<td>1.61</td>
<td>0.44</td>
<td>3.69</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

Random effects

<table>
<thead>
<tr>
<th>Variance</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Discussion

Experiment 5.1 showed that the retrieval practice effect could reliably be found in primary school children, using similar methods to those of a previous study in adults (Karpicke & Blunt, 2011). Children in the retrieval practice group had significantly higher recall scores four days later than those in the non-retrieval group. The other study technique, mind mapping, did not exert a main effect on learning, but did improve learning compared to normal note-taking in the non-retrieval practice condition.

While they did not violate the assumption of normality (Kolmogorov-Smirnov test $D(109) = .08, p = .09$), we noted that the scores from the test in Experiment 5.1 were somewhat negatively skewed, indicating that the children did not, on average, find the tasks to be particularly challenging. In addition, the mean number of facts recorded in the learning phase was over 85%, indicating that the learning phase was longer than required for many to make a note of all of the facts. For these reasons, and to test whether the main results—the significant main effect of retrieval practice and
its significant interaction with mind mapping—would replicate in a larger sample, we ran a second experiment in a different primary school.

**Experiment 5.2**

In Experiment 5.2, we increased the difficulty of the tasks by increasing the number of facts on each factsheet, reducing the duration of the learning phase, and extending the interval between the learning and testing phases to one week for a first test, and five weeks for a second. The addition of the five-week test allowed us to assess longer-term outcomes, and examine forgetting across time.

**Method**

**Participants**

Participants were 209 UK Primary School children (99 female), aged 8-12 years ($M = 10.15$ years, $SD = 1.19$), from Primaries 4, 5, 6, and 7 (two classes from each year; see Table 5.1 for numbers and ages per year) at Bruntsfield Primary School, Edinburgh. As with Experiment 5.1, the experiment was approved by the Psychology Research Ethics Committee at the University of Edinburgh, and written informed consent was obtained from the parent or guardian of each participating child.

**Materials**

Experiment 5.2 used very similar materials to Experiment 5.1, with more facts on the factsheets to increase the difficulty of the task. Sheets for Primary 4 children had 11 facts (~100 words), sheets for Primary 5 children had 18 facts (~135 words),
and sheets for Primaries 6 and 7 had 22 facts (~155 words). All materials are reproduced in full in Appendix B.

Learning Phase

The learning phase proceeded in the same manner as Experiment 5.1, with the same conditions (retrieval practice/non-retrieval, mind mapping/no mind mapping) but with a slight reduction in duration. It had the following structure: five minutes study/five minutes note-taking/three minutes study/three minutes note-taking. In this experiment, then, the learning phase lasted 16 minutes, with the factsheets visible to the retrieval practice group for eight minutes only.

Testing Phase

At one week and five weeks later, at the same time of day as the learning session had taken place, the children completed a written recall test, the same type as that in Experiment 5.1, in their regular school classroom. The one-week test was administered by the experimenter, while the classroom teacher administered the five-week test; the pseudo-random order of the questions was different at each test. The list of quiz questions is shown in Appendix B.

Results

Again, a comparable number of children learned about each country (Senegal: 52; South Korea: 54; Peru: 50; Iran: 53; see Table 5.6 for the distributions of different sheets per experimental condition). Two note sheets from the learning phase were lost, leaving 207 sheets. An average of 75.95% of the facts were recorded on the note sheets; again there were no significant differences in this between conditions [2×2 ANOVA mean difference = 5.49% in favour of no mind mapping \( F(1, 203) = 3.60, p \)
mean difference = 1.25\% in favour of non-retrieval ($F(1, 203) = .17, p = .68$). Table 5.7 details the performance in the learning phase by condition.

Table 5.6. Distributions of each factsheet across the four cells of the experiment in Experiment 5.2.

<table>
<thead>
<tr>
<th>Country on Factsheet</th>
<th>Frequency per condition in Experiment 5.2</th>
<th>Retrieval Practice</th>
<th>Non-retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mind Maps</td>
<td>No Mind Maps</td>
</tr>
<tr>
<td>Senegal</td>
<td></td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>South Korea</td>
<td></td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>Iran</td>
<td></td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Peru</td>
<td></td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 5.7. Mean performance (SD) in the initial learning phase (percentage of all potential facts written on notes) for Experiment 5.2, broken down by condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mind maps</th>
<th>No mind maps</th>
<th>Total score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>75.50 (18.85)</td>
<td>75.67 (22.25)</td>
<td>75.58 (20.42)</td>
</tr>
<tr>
<td>Non-retrieval</td>
<td>71.42 (25.70)</td>
<td>82.24 (18.28)</td>
<td>76.35 (23.16)</td>
</tr>
<tr>
<td>Total score</td>
<td>73.48 (22.50)</td>
<td>78.81 (20.60)</td>
<td>75.95 (21.75)</td>
</tr>
</tbody>
</table>

Twenty-three children were unavailable at either the one- or the five-week tests, leaving 186 children for the analysis of fact learning. Mean percentage recall scores for each condition, at each test, are shown in Table 5.8. Figure 5.2 illustrates the mean recall results across the one- and five-week tests, first for the retrieval practice and non-retrieval conditions, and second for the mind maps and no mind maps conditions.
Table 5.8. Mean percentage scores (SDs) and sample sizes for the 1-week and 5-week recall tests (including only those who provided data at both tests).

<table>
<thead>
<tr>
<th>Weeks after learning</th>
<th>Mind maps</th>
<th>No mind maps</th>
<th>Total score</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>1</td>
<td>61.43 (24.48)</td>
<td>62.24 (24.21)</td>
<td>61.82 (24.23)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>52.82 (25.37)</td>
<td>53.30 (28.08)</td>
<td>54.07 (26.55)</td>
</tr>
<tr>
<td>Non-retrieval</td>
<td>1</td>
<td>52.16 (22.83)</td>
<td>56.42 (21.77)</td>
<td>54.07 (22.33)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>42.22 (22.69)</td>
<td>49.29 (22.04)</td>
<td>45.39 (22.55)</td>
</tr>
<tr>
<td>Total score</td>
<td>1</td>
<td>56.98 (23.04)</td>
<td>59.60 (23.19)</td>
<td>58.19 (23.62)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>47.73 (24.58)</td>
<td>51.48 (25.45)</td>
<td>49.47 (24.99)</td>
</tr>
</tbody>
</table>

N

100
86
186

Figure 5.2. Percentage of facts recalled at the one- and five-week tests in Experiment 5.2 for (a) the retrieval practice and non-retrieval conditions and (b) the mind maps and no mind maps conditions. Error bars represent +/- 1 standard error.

To assess the effects of retrieval practice and mind mapping at both time-points, we ran a three-way (2×2×2) ANCOVA, including retrieval practice group (retrieval practice vs. non-retrieval) and mind mapping group (mind map vs. no mind)
map) as between-subjects factors and time of test (one or five weeks) as a within-subject factor. As in Experiment 5.1, we used multiple regression to identify related covariates: in this experiment, age, test type (Primary 4, Primary 5, or Primary 6/7) and facts recorded during the learning phase were all significantly related to the one-week score (all \( p \)-values < .001), and these variables plus sex were significantly related to the five-week score (for sex, \( p < .04 \); all other \( p \)-values < .001). Therefore, we included all four variables as covariates in the analysis.

Children in the retrieval practice group recalled significantly more facts than those in the non-retrieval practice group (\( F(1, 177) = 9.66, p = .002, \eta^2_p = .05 \)). The main effect of time was not significant (\( F(1, 177) = 1.83, p = .18 \)), and there was no interaction of time with retrieval practice group (\( F(1, 177) = .001, p = .98 \)) or with mind map group (\( F(1, 177) = .57, p = .45 \)); neither of the learning techniques altered the rate of forgetting across time. There was no main effect of mind mapping (\( F(1, 177) = .18, p = .67 \)) and, in contrast to Experiment 5.1, the interaction between retrieval practice and mind mapping was far from significance (\( F(1, 177) = .08, p = .78 \)).

Regarding covariates, there was again a large and significant effect of facts recorded in the learning phase, with those initially recording more facts tending to recall a higher percentage later (\( F(1,177) = 97.48, p < .001, \eta^2_p = .36 \)) and there were also significant influences of age (\( F(1, 177) = 20.60, p < .001, \eta^2_p = .10 \)) and test type (\( F(1, 177) = 15.63, p < .001, \eta^2_p = .08 \)), such that older participants, and those with more facts to be remembered, tended to gain higher scores. There was no effect of sex (\( F(1, 177) = 2.52, p = .11 \)).
To test whether there were any interactive effects of test type (Primary 4, 5, or 6/7; that is, whether the number of facts on the sheet was an influence on learning), in a further analysis we included it as a fixed effect, allowing it to interact with time and with the manipulated variables (retrieval practice and mind mapping group). No significant interactions were found between test type and time ($F(1, 170) = 1.63, p = .20$), retrieval practice group ($F(1, 170) = .92, p = .40$) or mind mapping group ($F(1, 170) = .27, p = .77$). All main effects and interactions between other variables remained significant or non-significant as in the original analysis. Thus, the effects of retrieval practice and mind mapping were comparable across all levels of the test.

In Experiment 5.1, the significant effect did not survive removal of the covariate. On the contrary, running the analysis with no covariates in Experiment 5.2 produced the same pattern of between-group results: significant effects of retrieval practice group ($F(1, 182) = 4.73, p = .03, \eta_p^2 = .03$) but not mind map group ($F(1, 182) = .82, p = .37$), and no significant interaction ($F(1, 182) = .27, p = .60$). Without covariates there was a main within-group effect of time ($F(1, 182) = 65.88, p < .001, \eta_p^2 = .27$), but no time $\times$ retrieval practice interaction ($F(1, 182) = .02, p = .90$), or time $\times$ mind map interaction ($F(1, 182) = .26, p = .61$).

As in Experiment 5.1, we report a secondary logit analysis of these results. A similar model to Experiment 5.1 was run, this time also including the within-subjects factor of time (one or five weeks after the learning phase) and the covariates of age, and test type (Primary 4, Primary 5, or Primary 6/7). The random effect in this model allowed for differences in individual ability to recall facts across time.
Table 5.9 shows the results of this model, again indicating that the results in this supplementary analysis did not differ strongly from those in the main ANCOVA analysis. Significant main effects were found for retrieval practice group, but not for mind mapping group, and there was no significant interaction found between the groups. Time had a significant effect, but did not interact with either of the conditions.

Table 5.9. Results of the generalized linear mixed model for Experiment 5.2.

<table>
<thead>
<tr>
<th>Fixed effects</th>
<th>b</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-4.09</td>
<td>1.13</td>
<td>-3.61</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Time – five weeks</td>
<td>-0.51</td>
<td>.10</td>
<td>-4.98</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Group – Non-retrieval</td>
<td>-0.45</td>
<td>.20</td>
<td>-2.32</td>
<td>.02</td>
</tr>
<tr>
<td>Group – Mind Maps</td>
<td>-0.18</td>
<td>.19</td>
<td>-0.98</td>
<td>.33</td>
</tr>
<tr>
<td>Age</td>
<td>0.58</td>
<td>.13</td>
<td>4.53</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Quiz – Primary 5</td>
<td>-0.81</td>
<td>.21</td>
<td>-3.85</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Quiz – Primary 6/7</td>
<td>-1.86</td>
<td>.34</td>
<td>-5.48</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Facts recorded in learning session</td>
<td>0.67</td>
<td>.07</td>
<td>10.27</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Time × retrieval interaction</td>
<td>0.08</td>
<td>.15</td>
<td>0.51</td>
<td>.60</td>
</tr>
<tr>
<td>Time × mind map interaction</td>
<td>0.11</td>
<td>.14</td>
<td>0.81</td>
<td>.42</td>
</tr>
<tr>
<td>Retrieval × mind map interaction</td>
<td>0.14</td>
<td>.26</td>
<td>0.53</td>
<td>.76</td>
</tr>
<tr>
<td>Time × retrieval × mind map interaction</td>
<td>-0.17</td>
<td>.20</td>
<td>-0.82</td>
<td>.41</td>
</tr>
</tbody>
</table>

Discussion

Experiment 5.2 was successful in increasing the difficulty of the tasks: Both the mean number of facts recorded on the sheets in the learning phase, and the mean
number of facts recalled at the first test, were lower than in Experiment 5.1. It provided a replication of the main result from Experiment 5.1 in a larger sample, across a longer time interval, and robust to the inclusion or exclusion of covariates. Children in the retrieval practice condition recalled significantly more facts at the one- and five-week tests, albeit with a smaller effect size than for the four-day test administered in Experiment 5.1. Experiment 5.2 did not replicate the interaction between retrieval practice and mind mapping discovered in Experiment 5.1; mind mapping did not affect learning outcomes in either of the retrieval conditions.

**General Discussion**

In two in-class experiments with primary school children, we compared the effects of retrieval practice and of mind mapping on later fact recall. The total time spent in the learning phase was the same in each condition, but children in the retrieval practice groups were exposed to the study materials for a far shorter time than those in the non-retrieval groups (in Experiment 5.2, exactly half as long). Despite this, children in the retrieval practice groups did not note down any fewer facts during learning, and subsequently recalled a significantly higher percentage of facts than those who did not use retrieval practice. This latter finding indicates that primary school teachers, like other educators, would benefit their children by using retrieval practice in the classroom.

Like Karpicke and Blunt (2011a), we found that retrieval practice improved fact recall, whereas a mapping technique had no main effects, even though the use of the latter is more widespread in schools. Our design additionally allowed for the evaluation of mapping *in combination* with retrieval practice, which we found did not
have any special benefits for recall. However, in Experiment 5.1, we did find that retrieval practice and mind mapping interacted significantly, such that mind mapping was superior only in the non-retrieval condition. This finding was not replicated in our larger second experiment, where the interaction was very far from significance. Since the procedures were so similar across the two experiments, this interaction in the initial experiment may have been a false-positive.

It should be noted that the mind mapping technique that our children used was necessarily simpler than the concept-mapping used by the undergraduate participants in the experiment by Karpicke and Blunt (2011a). It may also be relevant that our children did not follow any specific recommendations for optimal mind mapping, such as the use of colour and pictures (Buzan & Buzan, 2006). However, it is not clear to what extent such recommendations are evidence-based, and others advise that “...there is no necessity to retain an ideal structure or format” in mind mapping (Davies, 2010, p. 282). Our over-riding concern was to test the mapping technique that was already being used regularly in the school we visited, with which the children were comfortable. Without exception, all children in the mind mapping groups in both experiments produced maps with the country name in the centre and radiating ‘spokes’ to either individual facts or groups of facts (a mind map of the latter type is shown in Figure 5.1).

The retrieval practice effects in both experiments were statistically significant, and above the recommended effect size threshold for practical significance (Ferguson, 2009). The result from Experiment 5.1 is comparable to, though on the lower bound of, effect size estimates from previous studies in this age group: One previous paper
(Rohrer et al., 2011), for example, found effect sizes of \( d = .64 \) and \( .54 \) (corresponding approximately to \( \eta^2_p = .09 \) and \( .07 \), respectively) for retrieval practice on a test one day after learning, compared to our effect of \( \eta^2_p = .06 \) for recall 4 days after learning. The overall effect in Experiment 5.2 was slightly smaller (\( \eta^2_p = .05 \)); the effect of retrieval practice thus if anything appeared to decline across the longer gap between the learning and testing phases. However, previous experiments (e.g. Roediger et al., 2011) have found substantially larger effects with even longer learning-test intervals than in our study. This discrepancy may be explained by the younger age group involved in our study: Roediger & Karpicke (2006a) note that the retrieval practice effect may to some extent depend on age.

One potential limitation of Experiment 5.2 is that the same facts were tested at 1 and 5 weeks. The short-term test may thus have acted as retrieval practice for the longer-term test, boosting final performance. This may explain the finding that, unlike in some previous experiments (e.g. Roediger & Karpicke, 2006b, though see Spitzer, 1939), retrieval practice did not slow forgetting across the four-week gap in Experiment 5.2. However, any such influence should have raised the performance of the retrieval and non-retrieval groups equally. The difference between these groups was maintained at the longer-term test, indicating that using retrieval during initial learning is still beneficial in the longer-term, regardless of any intervening tests.

Two alternative explanations of the retrieval practice effects observed here should also be considered. First, since the participants in the retrieval group were able to look at their factsheet again after taking down some notes (on a mind map or otherwise), they potentially received feedback on their initial performance. This
feedback could have alerted them to facts that they did not recall in the first note-taking period, or they could have repaired any errors they had made during the note-taking. Thus, the retrieval practice effect observed here might not have been a direct effect of retrieval, but a ‘mediated’ testing effect (see Roediger & Karpicke, 2006a), whereby the feedback, not the ‘testing effect’, aids later recall (this effect has also been described as “test-potentiated learning”, see e.g. Arnold & McDermott, 2013). However, participants in the retrieval group did not see their notes and the factsheet at the same time, and were not permitted to write anything during the second study period, which would impede direct comparisons between their notes and the factsheet. In addition, the non-retrieval group also had a period of restudy of the facts, where they could have reflected on the factsheets and received similar feedback on their note-taking performance; the retrieval group still outperformed the non-retrieval group. Neither of these points completely rule out a mediated testing effect, however; our design could not fully tease apart direct and indirect effects of retrieval practice.

Second, since our design precluded us from recording the number of facts recorded in the first note-taking period in the learning session, it may be the case that the non-retrieval participants—those with a somewhat easier task—recorded all of their facts during this period, and did not concentrate on the task during the second note-taking period. This would mean that the effective time on-task in the retrieval group was longer, explaining the better recall on the later test. However, if this interpretation were correct, we would expect children in the non-retrieval group to have written down more facts on average than the children using retrieval practice. As can be seen in Tables 5.3 and 5.7, and in the ANOVA results reported above for both experiments, the total percentage of facts recalled in the learning phase was similar in
both conditions, with the vast majority of children in both groups failing to record all of the facts. This implies that children in both groups were still working on their notes at the end of the study period, and that time on-task is not responsible for the retrieval effect.

**Conclusion**

The two experiments reported here have substantial practical implications for primary school teachers: using simple self-testing in the classroom by asking children to make notes on their learning materials from memory will improve significantly their recall of those materials several weeks later. The retrieval practice group in our Experiment 5.2 recalled over 8.5% more facts than the non-retrieval group, five weeks after the learning session. The popular technique of mind mapping, on the other hand, may be an interesting and enjoyable way for children to visually represent their learning, but teachers should not expect it to boost fact learning—at least of the type studied here—in the short- or long-term.
Enduring Links from Childhood Mathematics and Reading Achievement to Adult Socioeconomic Status

†† The study reported in this chapter was published as Ritchie, S. J., & Bates, T. C. (2013). Enduring Links from Childhood Mathematics and Reading Achievement to Adult Socioeconomic Status. Psychological Science, 24, 1301-1308. We are grateful to the Centre for Longitudinal Studies (CLS), the Institute of Education, and the Economic and Social Data Service (ESDS) for these data. Neither the CLS nor the ESDS are responsible for the analysis or the interpretation of the data presented in this chapter.
Abstract

Understanding the determinants of socioeconomic status (SES) is an important economic and social goal. Several major influences on SES are known, yet much of the variance in SES is unexplained. In a large, population-representative U.K. sample, we tested the effects of mathematics and reading achievement at age 7 on attained SES by age 42. Mathematics and reading ability both had substantial positive associations with adult SES, above and beyond the effects of SES at birth, and other important additional measured factors, such as intelligence. Achievement in mathematics and reading was also significantly associated with intelligence scores, academic motivation, and duration of education. These findings suggest effects of improved early mathematics and reading on SES attainment across the life span.
Introduction

This final chapter of this thesis differs greatly from those preceding it, since it focuses on the outcomes of education instead of the process of learning. It thus acts as a bookend to the empirical parts of the thesis, testing the importance of various aspects of education across the lifespan.

Research aimed at understanding socioeconomic status (SES) attainment across the life course has increasingly focused on cognitive (e.g., Deary et al., 2005) and non-cognitive (e.g., Moffitt et al., 2011) differences apparent early in childhood (Reynolds, Temple, White, Ou, & Robertson, 2011). Existing analyses, however, have left around half the variance in SES unaccounted for; more specific factors may therefore play a role. In the research reported here, we tested whether early mathematics and reading skills predicted SES attainment over and above social status of origin, intelligence, academic motivation, and educational duration.

Childhood mathematics and reading skills are associated with educational achievement, translating into higher grades and greater attained qualifications (Duncan et al., 2007; McGee, Prior, Williams, Smart, & Sanson, 2002). In adulthood, these skills have also been shown to correlate with SES (e.g., Kutner et al., 2007). There are, however, at least two reasons to hypothesize that childhood mathematics and reading are causal influences on adult SES. First, these skills are associated with variables that could plausibly influence SES attainment. Individuals with greater academic skills are likely to be more successful in occupational contexts because their abilities afford them improved chances of career development. Further, numeracy is
associated with successful financial decision-making (Agarwal & Mazumder, 2013), and both poor literacy and poor numeracy are linked, through their effects on understanding of medical information, to ill health (Anker & Kaufman, 2007; Berkman, Sheridan, Donahue, Halpern, & Crotty, 2011), which limits SES (Cai & Kalb, 2006). Moreover, reading ability is positively associated with self-esteem (Kiuru et al., 2012), which predicts improved economic prospects (Trzesniewski et al., 2006), and is negatively associated with delinquency (Svensson, 2011).

Second, although mathematics ability and reading ability are cognitive skills, they are not completely subsumed by general intelligence; both mathematics (Hart, Petrill, Thomson, & Plomin, 2009) and reading (Harlaar, Hayiou-Thomas, & Plomin, 2005) show independent genetic effects, and reading and intelligence have been found to be neurally distinguishable (Tanaka et al., 2011).

We therefore hypothesized that, if mathematical and reading skills in childhood have substantial effects beyond the classroom—specifically, effects on midlife SES—we should find direct associations between these variables, even when controlling for intelligence, social class of origin, academic motivation, and educational duration. We also expected to find indirect associations between mathematics and reading ability and midlife SES, possibly via the effect of these skills on downstream variables such as education. We tested these relationships in a large, nationally representative sample, using a data set from a longitudinal study spanning from childhood to midlife.
Method

Participants

Participants were members of the National Child Development Study, which used a sample of 17,638 infants born during 1 week in 1958 in England, Scotland, and Wales, along with 920 immigrants born during the same time. The 18,558 cohort members have been followed up on in eight waves, the first of which was in 1965, when participants were about 7 years old, and the most recent of which was in 2008 and 2009, when participants were about 50 years old. The present study focuses on data collected when participants were approximately 7 years old (Wave 1; \( n = 15,425 \)), 11 years old (Wave 2; \( n = 15,337 \)), 16 years old (Wave 3; \( n = 14,647 \)) and 42 years old (Wave 6; \( n = 11,419 \)); data files are available from the Institute of Education (2008a, 2008b). The sample size for each variable is shown in Table 6.1.
**Table 6.1.** Heterogeneous correlation matrix and descriptive statistics for all variables, for female (above the diagonal) and male (below the diagonal) cohort members.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>Female N</th>
<th>Female M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>.38</td>
<td>.28</td>
<td>.20</td>
<td>.24</td>
<td>.27</td>
<td>.29</td>
<td>.31</td>
<td>.35</td>
<td>.16</td>
<td>.35</td>
<td>.26</td>
<td>.15</td>
<td>.13</td>
<td>6887</td>
<td>4.01 (1.56)</td>
</tr>
<tr>
<td>2</td>
<td>.40</td>
<td>-</td>
<td>.31</td>
<td>.17</td>
<td>.20</td>
<td>.23</td>
<td>.24</td>
<td>.21</td>
<td>.30</td>
<td>.20</td>
<td>.33</td>
<td>.24</td>
<td>.17</td>
<td>.11</td>
<td>6857</td>
<td>2.97 (.97)</td>
</tr>
<tr>
<td>3</td>
<td>.32</td>
<td>.34</td>
<td>-</td>
<td>.10</td>
<td>.09</td>
<td>.07</td>
<td>.10</td>
<td>.07</td>
<td>.15</td>
<td>.09</td>
<td>.18</td>
<td>.10</td>
<td>.07</td>
<td>.03</td>
<td>7063</td>
<td>4.76 (1.31)</td>
</tr>
<tr>
<td>4</td>
<td>.22</td>
<td>.16</td>
<td>.10</td>
<td>-</td>
<td>.64</td>
<td>.53</td>
<td>.53</td>
<td>.48</td>
<td>.52</td>
<td>.18</td>
<td>.29</td>
<td>.26</td>
<td>.19</td>
<td>.17</td>
<td>7256</td>
<td>5.00 (2.50)</td>
</tr>
<tr>
<td>5</td>
<td>.23</td>
<td>.15</td>
<td>.08</td>
<td>.67</td>
<td>-</td>
<td>.57</td>
<td>.73</td>
<td>.60</td>
<td>.60</td>
<td>.18</td>
<td>.31</td>
<td>.29</td>
<td>.19</td>
<td>.18</td>
<td>7298</td>
<td>2.8 (.84)</td>
</tr>
<tr>
<td>6</td>
<td>.27</td>
<td>.20</td>
<td>.07</td>
<td>.56</td>
<td>.59</td>
<td>-</td>
<td>.74</td>
<td>.79</td>
<td>.64</td>
<td>.17</td>
<td>.28</td>
<td>.30</td>
<td>.20</td>
<td>.19</td>
<td>7213</td>
<td>24.29 (6.69)</td>
</tr>
<tr>
<td>7</td>
<td>.28</td>
<td>.21</td>
<td>.10</td>
<td>.54</td>
<td>.72</td>
<td>.78</td>
<td>-</td>
<td>.80</td>
<td>.66</td>
<td>.20</td>
<td>.36</td>
<td>.33</td>
<td>.19</td>
<td>.21</td>
<td>7252</td>
<td>3.23 (.91)</td>
</tr>
<tr>
<td>8</td>
<td>.27</td>
<td>.16</td>
<td>.05</td>
<td>.49</td>
<td>.60</td>
<td>.78</td>
<td>.80</td>
<td>-</td>
<td>.61</td>
<td>.17</td>
<td>.36</td>
<td>.28</td>
<td>.18</td>
<td>.19</td>
<td>7303</td>
<td>2.74 (1.28)</td>
</tr>
<tr>
<td>9</td>
<td>.30</td>
<td>.25</td>
<td>.10</td>
<td>.54</td>
<td>.61</td>
<td>.65</td>
<td>.66</td>
<td>.61</td>
<td>-</td>
<td>.25</td>
<td>.40</td>
<td>.36</td>
<td>.28</td>
<td>.19</td>
<td>6878</td>
<td>44.14 (15.90)</td>
</tr>
<tr>
<td>10</td>
<td>.18</td>
<td>.18</td>
<td>.08</td>
<td>.13</td>
<td>.19</td>
<td>.20</td>
<td>.22</td>
<td>.20</td>
<td>.27</td>
<td>-</td>
<td>.33</td>
<td>.24</td>
<td>.21</td>
<td>.19</td>
<td>5612</td>
<td>19.00 (6.09)</td>
</tr>
<tr>
<td>11</td>
<td>.33</td>
<td>.27</td>
<td>.20</td>
<td>.26</td>
<td>.31</td>
<td>.29</td>
<td>.35</td>
<td>.31</td>
<td>.39</td>
<td>.31</td>
<td>-</td>
<td>.37</td>
<td>.19</td>
<td>.26</td>
<td>5647</td>
<td>17.16 (2.11)</td>
</tr>
<tr>
<td>12</td>
<td>.33</td>
<td>.25</td>
<td>.14</td>
<td>.31</td>
<td>.36</td>
<td>.37</td>
<td>.38</td>
<td>.35</td>
<td>.45</td>
<td>.34</td>
<td>.44</td>
<td>-</td>
<td>.19</td>
<td>.51</td>
<td>4552</td>
<td>3.58 (1.67)</td>
</tr>
<tr>
<td>13</td>
<td>.19</td>
<td>.20</td>
<td>.08</td>
<td>.19</td>
<td>.16</td>
<td>.21</td>
<td>.16</td>
<td>.15</td>
<td>.24</td>
<td>.15</td>
<td>.16</td>
<td>.19</td>
<td>-</td>
<td>.15</td>
<td>5708</td>
<td>2.31 (.91)</td>
</tr>
<tr>
<td>Male N</td>
<td>7227</td>
<td>7226</td>
<td>7440</td>
<td>7645</td>
<td>7706</td>
<td>7673</td>
<td>7691</td>
<td>7572</td>
<td>7254</td>
<td>5856</td>
<td>5495</td>
<td>5040</td>
<td>5486</td>
<td>3800</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male M (SD)</td>
<td>4.02 (1.56)</td>
<td>2.98 (1.32)</td>
<td>4.81 (2.50)</td>
<td>5.22 (.89)</td>
<td>2.87 (7.43)</td>
<td>2.93 (91)</td>
<td>2.35 (136)</td>
<td>41.81 (16.29)</td>
<td>20.03 (6.18)</td>
<td>17.24 (2.41)</td>
<td>3.25 (2.00)</td>
<td>2.28 (0.88)</td>
<td>10.06 (0.74)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Measures**

*SES of origin* was a formative variable calculated from three measures completed by the cohort members’ parents when the cohort member was 7 years old. The first measure was the social class of the father’s occupation, measured by the Registrar General’s Social Classes (RGSC) class scheme and scored using a 7-point scale (Office of Population Censuses and Surveys, 1980): 1 = *Class I, professional*, 2 = *Class II, managerial/technical*, 3 = *Class III* or *Class III*, skilled nonmanual or Class III*, skilled manual, 4 = *Class IV*, semiskilled nonmanual or Class IV*, semiskilled manual, 5 = *Class V, unskilled manual*. The second measure, parental housing tenure, was rated on a scale from 1 (*owner occupied*) to 4 (*rent free*). The third measure was the reported number of rooms in the parents’ home.

*Mathematics* was a latent variable constructed from two measures recorded when participants were 7 years old. The first measure was the participant’s score on the Problem Arithmetic Test (Kellmer Pringle, Butler, & Davie, 1966), which was constructed specifically for this cohort and consisted of 10 arithmetic questions (e.g., “What is half of 38?”); one point was awarded for each correct answer. The second measure was a rating of the participant’s mathematical ability provided by his or her teacher; ratings were made using a scale from 1 (*little, if any, ability in this sphere*) to 5 (*extremely good facility with number and/or other mathematical concepts*).

*Reading* was a latent variable constructed from three measures of reading recorded when participants were 7 years old. The first measure was the participant’s score on the Southgate Group Reading Test (Southgate, 1958), a 30-item test of word recognition in which participants must match words to pictures. The second measure
was the participant’s status as a reader as rated by his or her teacher, using a scale from 1 (nonreader) to 5 (avid reader). The third measure was the level of books the participant was able to read; each participant’s teacher was asked to indicate whether the participant was pre-reading (Level 1), currently at the standard of basic graded classroom reading books (Levels 2–5), or beyond basic reading books (Level 6).

*Intelligence* was measured at age 11 using a timed test involving matrices of linked words, symbols, or shapes; participants studied these and then completed an unfinished sequence from a set of alternatives. This measure contained 40 verbal items and 40 nonverbal items. The total score from this test correlates very strongly ($r = \sim .90$) with IQ-type tests and have high reliability ($\alpha = .94$; Douglas, 1964, p. 131).

*Academic motivation* was measured using a questionnaire administered to cohort members at age 16. The questionnaire assessed their level of agreement with eight statements regarding motivation, self-regulation, and planning in school and beyond. Ratings were made using 5-point Likert scales. Two example items are “I find it difficult to keep my mind on work” and “School is a waste of time.” The eight items show acceptable reliability ($\alpha = .75$).

The *educational duration* variable, reported at age 42, was the age at which the cohort member left full-time education.

*Attained SES*, measured at age 42, was a latent variable made up of three measures: the RGSC class of the cohort member’s occupation, the cohort member’s housing tenure, and the log of the cohort member’s gross income at his or her current job. For the RGSC measure, manual and nonmanual Class IV jobs were not classed
into distinct categories, but the variable was otherwise identical to that used for paternal RGSC.

Analysis

Data were analyzed using OpenMx (Boker et al., 2011) in the R environment. A saturated structural equation model containing independent submodels for male and female cohort members was constructed (because of possible sex differences in the determinants of SES trajectories; see Ceci & Williams, 2010). Covariance of all continuous and ordinal variables was computed as a heterogeneous correlation matrix using the polycor package (Fox, 2010), with pairwise deletion for missing values.

Results

The correlation matrix—with separate correlations for males and females—used as an input to the model is shown in Table 6.1. The bivariate relationships between males’ and females’ mathematics and reading achievement at age 7 and SES at age 42 are shown in Figure 6.1.
Figure 6.1. Associations of age-7 (a) mathematics and (b) reading achievement with age-42 attained SES for male and female cohort members

Model fitting proceeded as follows. We first examined the extent to which the models for males and females could be equated—that is, we tested which paths could be set equal across males and females without significant loss of fit (excluding the formative and reflective measurement-model elements). Of the paths among the variables, all but four could be equated across sex without significant loss of fit, \( \Delta \text{-log likelihood} = 18.36, \Delta \text{df} = 15, p = .24 \). Two paths were significantly stronger for male cohort members than for female cohort members: the covariance between mathematics and reading and the path from intelligence to motivation. By contrast, two paths were stronger for females than for males: the paths from both reading and education to attained SES. A diagram showing the path weights in the model prior to this sex-equation process can be found in Appendix C (Figure C1).

We then examined the significance of the hypothesized effects in the sex-equated model by setting each path to 0 and examining the resultant change in \(-2 \log \)
likelihood. All paths were significant except for the path from mathematics to academic motivation, which could be dropped without significant loss of fit (see Table 6.2 for changes in $-2 \log$ likelihood and degrees of freedom, and for the associated $p$-values, for all paths). Figure 6.2 shows the final model including path weights for both sexes, standardized within each sex.

**Table 6.2.** Changes in model fit ($-2 \log$ likelihood) and degrees of freedom, with associated $p$-values, after dropping paths in the sex-equated model.

<table>
<thead>
<tr>
<th>Path</th>
<th>$\Delta$-2LL after dropping</th>
<th>$\Delta df$</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Origin SES -&gt; Mathematics</td>
<td>927.62</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Origin SES -&gt; Reading</td>
<td>1396.50</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Origin SES -&gt; Intelligence</td>
<td>425.28</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Origin SES -&gt; Motivation</td>
<td>205.06</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Origin SES -&gt; Education</td>
<td>871.01</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Origin SES -&gt; Attained SES</td>
<td>276.12</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Mathematics &lt;-&gt; Reading</td>
<td>8069.73</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Mathematics -&gt; Intelligence</td>
<td>439.50</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Mathematics -&gt; Motivation</td>
<td>.002</td>
<td>1</td>
<td>.97</td>
</tr>
<tr>
<td>Mathematics -&gt; Education</td>
<td>30.26</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Mathematics -&gt; Attained SES</td>
<td>50.79</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Reading -&gt; Intelligence</td>
<td>638.80</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Reading -&gt; Motivation</td>
<td>7.67</td>
<td>1</td>
<td>.006</td>
</tr>
<tr>
<td>Reading -&gt; Education</td>
<td>8.23</td>
<td>1</td>
<td>.004</td>
</tr>
<tr>
<td>Reading -&gt; Attained SES</td>
<td>17.20</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Intelligence -&gt; Motivation</td>
<td>175.64</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Intelligence -&gt; Education</td>
<td>153.49</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Intelligence -&gt; Attained SES</td>
<td>138.68</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Motivation -&gt; Education</td>
<td>618.29</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Motivation -&gt; Attained SES</td>
<td>362.81</td>
<td>1</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Education -&gt; Attained SES</td>
<td>592.47</td>
<td>2</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>

*Note: For measures, see Table 6.1. For paths, see model in Figure 6.2. SES = Socioeconomic Status.*
Figure 6.2. Associations of mathematics ability at age 7 and reading ability at age 7 with attained socioeconomic status at age 42, including all control variables. Values are path coefficients (values for males are given on the left of the slash marks, and values for females are given on the right). Paths that could not be equated between the sexes (excluding formative and reflective measurement-model elements) are indicated with asterisks, and the non-significant path in the sex-equated model is indicated with a dashed line. Significance values are shown in Table 6.2. RGSC = Registrar General’s Social Classes occupational classification (Office of Population Censuses and Surveys, 1980). SES = socioeconomic status. SGRT = Southgate Group Reading Test (Southgate, 1958).
For both sexes, the latent factor of mathematics ability had a significant direct positive association with SES some 35 years later, with a model path weight of .18 for males and .10 for females (the difference between these weights was not significant). The direct path from reading ability to attained SES was significantly higher for females (.10) than for males (.04), for whom it was not significant. As can be seen from Figure 6.2, both reading and mathematics ability at age 7 had significant indirect associations with midlife SES: There were positive associations of mathematics and reading ability with both intelligence at age 11 and educational duration; reading ability was also significantly associated with academic motivation at age 16.

Two further models, each containing only one latent variable of either mathematics ability, or reading ability, were estimated, and are shown in Appendix C (Figures C2 and C3 respectively). A comparison of these models reveals that mathematics ability and reading ability have similar associations with downstream outcomes such as educational duration, which may indicate that they both reflect a general factor present at age 7. However, a test of the discriminant validity of the two variables confirmed the significant effects of reading independent of its association with mathematics. All paths to and from reading were set to 0, except for the path between mathematics and reading, which was changed to a direct path from mathematics to reading and set to 1 to model the reading measures as a reflection of a general factor defined by mathematics and reading ability at the latent level. This model fit substantially and significantly more poorly ($\Delta \chi^2 = 776.14$, $\Delta df = 17, p < .001$), which indicated that the separate inclusion of both mathematics and reading was required for the best fit.
Finally, to test the robustness of the association between mathematics ability, reading ability, and midlife SES, we ran two additional models, including a new, theoretically motivated variable (oral ability at age 7 and aggression at age 16) in each. The inclusion of these new variables did not significantly reduce the associations we found using the original model. The models are shown in Appendix C, in Figures C4 and C5, respectively.

**Discussion**

Our results—illustrated in bivariate form in Figure 6.1 and in full, modeled form in Figure 6.2—show that mathematics and reading ability at age 7 are substantially and positively associated with SES at age 42, independently of relevant confounding variables. Both direct and indirect associations were found: Mathematics and (for females) reading ability directly related to attained SES, and there were additional positive impacts of mathematics and reading ability at age 7 on subsequent intelligence, academic motivation, and educational duration.

These direct and indirect paths from mathematics and reading ability to subsequent SES may arise through a variety of mechanisms. First, it is plausible that the aforementioned links from mathematics and reading ability to intelligence, academic motivation, and, ultimately, educational duration increase one’s qualifications and, thus, open access to higher-paying occupations. Second, improved mathematics and reading skills may be of direct value in employment contexts; individuals with superior skills in these areas may, through their abilities in calculation, quantification, or extracting salient information, be more competitive in the occupational sphere. Third, the direct links from mathematics and reading ability
to SES might be explained by the associations, noted above, between high mathematics and reading ability and variables that may improve SES, such as health literacy (Berkman et al., 2011) and management of personal finances (Agarwal & Mazumder, 2013).

Four paths to attainment differed between the sexes. First, intelligence was more strongly associated with academic motivation in males than in females. The mechanisms underlying this difference are unclear; the difference seems to indicate that, given the importance of education, targeting lower-ability boys for motivational interventions might be a worthy investment. Second, reading ability and educational duration were more strongly associated with attained SES in females than in males. After finding this stronger path between education and attained SES for females in the same sample, Schoon (2008) suggested that, at the time the cohort members were entering the labor market, “education was more of a necessity for women than for men to get ahead in the occupational hierarchy, while for men there might have been more opportunities to train on a job” (p. 79). Our finding of a stronger path between reading and attained SES for females than for males, then, may be explained by differences in the types of occupations available to, or chosen by, males and females. Finally, there was a small difference in the covariance between reading and mathematics, such that the path weight for males was slightly higher than that for females. This difference may reflect sex differences in risk for reading and mathematics disorders (e.g., Hawke, Olson, Willcut, Wadsworth, & DeFries, 2009), which might alter the relationship between reading and other variables across the sexes.
Two alternative causal explanations for our findings could be suggested. First, a “generalist genes” hypothesis (Haworth, Meaburn, Harlaar, & Plomin, 2007) might predict that mathematics and reading ability are manifestations of genetic cognitive ability, which was not controlled for by the later intelligence test administered to our sample. However, our discriminant-validity analysis indicated that separate inclusion of both mathematics ability and reading ability was necessary for the best model fit. In addition, there is evidence that high early reading ability may assist the development of intelligence (Harlaar et al., 2005), and high early mathematics ability may have similar effects. Second, it remains possible that childhood mathematics and reading ability are indicators of other variables that were not included in our model. However, these variables would have to be outside the four major cognitive and noncognitive variables believed to underpin SES attainment, which were controlled for in this study.

One limitation of the present study was that the items in the measure of academic motivation assessed two related but separate constructs: motivation and self-control/self-regulation. The inclusion of cleaner measures of these two constructs in future studies would allow researchers to disentangle their separate effects on SES outcomes.

The present study did not contain a randomized experiment to test the efficacy of an intervention to improve mathematics ability, reading ability, or both. However, other work has shown positive effects of, for instance, preschool education on these variables (Tucker-Drob, 2012), and interactions between the effects of genetics and the effects of teacher quality on reading ability (Taylor, Roehrig, Soden Hensler,
Connor, & Schatschneider, 2010) indicate that levels of attainment of these skills are to some extent malleable.

**Conclusion**

The results reported in this chapter strengthen the understanding of routes to social mobility, supporting not only the roles of SES of origin, intelligence, academic motivation, and educational duration, but also the specific faculties of mathematics and reading.
Chapter Seven

Summary and discussion of the application of psychological science to education
Abstract

This final chapter first provides a summary of the findings from across the present thesis, and some broad recommendations and implications for future research that can be drawn from them. Second, the thesis having focused mainly on the possible contributions of cognitive psychology to education, the chapter briefly discusses some potential contributions of other domains of psychology, specifically cognitive neuroscience, social psychology, and differential psychology.
Summary of the present thesis and implications for future research

The opening chapter of the present thesis was, we hope, a convincing case that better evidence is needed in education. In discussing several popular classroom techniques and the claims surrounding them, we exposed some severe weaknesses of evidence. For Brain Gym (Dennison & Dennison, 2010), a theory with no evidential backing is combined with potentially enjoyable exercise activity; the few studies that have been performed suggest that the claims of proponents are however wildly overblown. Whereas beliefs about the need for large quantities of drinking water (eight glasses per day) are pervasive (Valtin, 2002), there is very little evidence underlying them. Fish oils, while effective for some medical conditions, and an important part of a balanced diet, sometimes lacking in socioeconomically-disadvantaged children, do not appear to improve academic outcomes (e.g. Kirby, Woodward, & Jackson, 2009). Commercially-available ‘Brain Training’ games have now, in our opinion, been convincingly shown to have no effects on cognition beyond placebo (Owen et al., 2010). Finally, the evidence for the effects of chewing gum on cognition is mixed (Rickman, Johnson, & Miles, 2013), but there exists to our knowledge no evidence that it can improve academic outcomes in children.

Overall, the conclusion can be drawn that teachers are in need of improved resources to aid in their evaluation of new and attractive ‘alternative’ educational techniques. Such a resource could take the form of an online collection of brief reviews of the evidence behind each condition, similar to those provided by the Macquarie University Special Education Centre (MUSEC, 2012). These briefings are aimed at special educational interventions for conditions such as Autism Spectrum
Disorder, and ‘Irlen Syndrome’ (see Chapter Two and below). However, better availability of evaluations of interventions for typically-developing children, such as those discussed in Chapters Three, Four, and Five of the present thesis, would be of great benefit for teachers.

Our discussion of poorly-evidenced techniques led us to our consideration of Irlen Syndrome in Chapter Two. A controversial disorder, not recognized by several professional bodies of paediatricians and ophthalmologists (e.g. American Academy of Pediatrics, 2009), Irlen Syndrome (Irlen, 2010) is nonetheless purported to cause a substantial proportion of reading disorders. Our follow-up testing of children who had used the Irlen Institute’s standard treatment for Irlen Syndrome, coloured filters, for a year showed that their reading is not improved by the filters compared to a plain transparent acetate sheet. We would note that no evidence of Irlen Syndrome—in many cases diagnosed only on the basis of response to the treatment—was found in our results (or those from our original experiment; Ritchie, Della Sala, & McIntosh, 2011). We would recommend that teachers focus on evidence-based treatments for, on one hand, optometric disorders, and on the other, reading difficulties (e.g. Hatcher et al., 2006), instead of investing resources in treatments that are far from well-supported. Proponents of the Irlen treatment have, in general, appeared reluctant to perform and publish high-quality, peer-reviewed research on the validity of their diagnostic instruments, the efficacy of their coloured filter treatment, and the long-term prognosis of treated and untreated Irlen Syndrome. This attitude should change before recommendations about treatments, and their associated costs, are made to teachers and parents.
Whereas it appears coloured filters are growing in popularity in UK schools, they are still relatively rare. Chapter Three tested two more commonly-used techniques: exercise and relaxation sessions at the start of lessons. Some evidence exists for potential effects of in-class exercise sessions on cognition (e.g. Hill et al., 2010, 2011), although, as we describe in the chapter, the data are in some cases open to multiple interpretations. Data on the effects of in-class relaxation sessions are very sparse indeed. In our first experiment, we found that, whereas relaxation had no effects at all, exercise had a detrimental effect on attention. A second experiment failed to replicate this effect, but instead found that exercise had a positive effect on a delayed-recall memory test. We are less confident in this latter finding due to the low reliability of the novel memory measure we designed for the study. One conclusion that can, however, be drawn is that whether the effects of these techniques are positive or negative, they are likely to be very small, at least in the conditions under which we investigated them. The effects of longer-term exercise and relaxation interventions on academic functioning should be investigated next, with researchers moving beyond correlational studies by running longitudinal experiments and, as we suggest in the chapter, using designs such as the monozygotic-differences twin design to investigate causality.

Chapter Three showed relaxation before learning to be ineffective in improving short-term outcomes. In Chapter Four, we took the lead from recent studies of minimal retroactive interference (e.g. Dewar et al., 2012) and studied the possible effects of relaxing after learning. In one large, in-class study and one small, more controlled study, we showed that the effects of wakeful resting on memory, found in several studies in adults, do not appear to be present in young children. A number of
methodological issues, from low statistical power to the form of memory test used, may explain the discrepancy between our results and the previous research; as such, no practical recommendations can be drawn about this technique until further research has been conducted. This research should involve large samples of children, tested under controlled conditions, and use tests of both recall and recognition memory to investigate possible differential effects of minimal retroactive interference on different memory subsystems. Only if effects from such studies are reliable and practically significant (e.g. Ferguson, 2009) should further attempts be made to apply minimal retroactive interference to children’s learning in the classroom.

To a considerably greater extent than minimal retroactive interference, the use of retrieval practice has been shown to be effective for improving memory in studies of adults (Roediger & Karpicke, 2006a). Some studies in children exist, but far fewer, and the studies tend to be limited in their practical applicability. In Chapter Five, we tested retrieval practice, along with a second technique, mind mapping (Buzan & Buzan, 2006), in two samples of primary school children in a classroom setting. The less popular technique, retrieval practice, facilitated memory for new facts at both one and five weeks, while the well-used technique, mind mapping, had no effects on recall. The combination of the two techniques showed no particular effects. The retrieval practice method we used in the chapter involved only pencil-and-paper tests, and produced practically significant benefits to learning in a naturalistic setting: We would thus recommend it to primary school teachers for this type of fact learning. Although mind mapping had no significant effect on learning, we would not go so far as to recommend against it: it is a technique with no costs (either financial or for learning), and children may still enjoy representing their learning visually. Future
research should follow on from our method (taken from a recommendation from Roediger, 2013) of combining potentially effective learning techniques to assess whether, in combination, they produce additive learning gains.

A concern about the experiments reported in Chapters 3, 4, and 5 is the choice of measures used. Whereas these measures covered several executive functions (e.g. attention, memory) and the learning of new facts and their retention in memory across time, they did not (with the exception of Chapter 2, which assessed an intervention targeted at reading) cover other important outcomes of education, such as skill learning. The types of intervention tested here may also influence both lower-level motor skills such as handwriting, and higher-level procedural skills such as the approach to mathematical problems or critical analysis. These factors are often more challenging to measure than, for example, recall of facts or visual attention performance, but they are nevertheless important, and for a full understanding of the psychology of education, future research should not neglect them.

The final empirical chapter of the present thesis, Chapter Six, returns to the topic of reading, but examines it from a rather different perspective to that of Chapter Two. Here, we analyzed a large, longitudinal dataset, the National Child Development Study, to test for the effects of childhood reading, and mathematics, performance on adult socioeconomic status (SES) thirty-five years later. Even after controlling for SES at origin, general intelligence, academic motivation, and educational duration (as studied by, e.g., Deary et al., 2005), the two childhood skills predicted adult SES, which was a latent variable of income, occupational status, and housing tenure. Chapter Six hinted at the impressive long-term benefits of improved education
regardless of one’s social or biological starting point. The structural equation model shown in Figure 6.2 shows only associations, not causal relationships, and thus raises many questions regarding causality. Most importantly for the hypothesis in Chapter 6, would improving reading and mathematics education lead to the benefits for social mobility implied in the (purely correlational) model? Do faculties such as mathematics and reading causally influence general intelligence (see below for discussion of one study that sheds light on this question)? Could improving intelligence have effects on academic motivation, or vice versa?

If causality is demonstrated for any or all of the relationships uncovered in Chapter Six, the next important questions regard the mechanisms of these effects. If, for instance, reading raises intelligence, how does it have this effect? What are the specific mechanisms by which childhood mathematics achievement improves adult socioeconomic status? Future research addressing these questions may lead to specific interventions aimed at promoting social mobility.

**Beyond cognitive psychology in education**

Much of the present thesis has tested ideas and techniques with their base in cognitive psychology, with varying results. In the initial chapter, and at various other points, we touched on ideas with a neuroscientific rationale, and in the final empirical chapter, we took on an ‘individual differences’ perspective. In this section, we briefly outline some further potential contributions to education that may be found in psychology, focusing on three fields in particular: neuroscience, social psychology, and differential psychology.
**Neuroscience**

A great deal has been made of the potential for ‘neuroeducation’ in recent years (Howard-Jones, 2010; Anderson & Della Sala, 2012). Here, we discuss three potential contributions of neuroscience to education: inspiration for new in-class techniques, ‘neuroprognosis’, and testing the effects of techniques from other areas.

As discussed in Chapter 1, survey research has shown that a substantial proportion of teachers harbor a number of misconceptions about the brain; for instance, Dekker, Lee, Howard-Jones, and Jolles (2012) found that 48% of UK teachers agreed with the statement “we only use 10% of our brain”, while 88% agreed that “short bouts of co-ordination exercises can improve integration of left and right hemispheric brain function”. As noted above, to combat these misunderstandings, clearer information on the aspects of the brain relevant to education could likely be provided on teacher-training courses or online.

Setting aside these misconceptions, it is the case that teachers are interested in what neuroscience could contribute to their classroom teaching, and many scientists have been optimistic about such a contribution (Howard-Jones, 2010). Unfortunately, examples of validated, evidence-based classroom techniques inspired solely by neuroscientific (and not ‘mere’ psychological) principles are at present few and far between. It has been suggested, for example, that emerging evidence relating to the ‘mental number line’ (e.g. Izard & Dehaene, 2008) indicates that educators should devote more focus to teaching basic arithmetical abilities (Seron, 2012), but it is not
clear that neuroscientific evidence, as opposed to cognitive psychological research on numerical cognition, was necessary to provide such an insight. It may be, however, that a better neural understanding of specific learning disabilities, such as dyslexia and dyscalculia, will be helpful in developing appropriate targeted interventions (e.g. Ansari, de Smedt, & Grabner, 2011; Wilson, Revkin, Cohen, & Dehaene, 2006).

However, perhaps one should not expect interventions to be directly inspired by neuroscience, which is still at a relatively early stage of investigating human learning. A different way in which (cognitive) neuroscience may contribute to education is by early prediction of children’s learning disabilities. A high-profile test of this so-called ‘neuroprognosis’ was carried out by Hoeft et al. (2010), who tested twenty-five teenagers (aged around 14 years) with dyslexia and twenty controls on a variety of reading measures and two brain measures: functional magnetic resonance imaging (fMRI) and diffusion-tensor imaging (DTI). After two years, the participants were followed up and tested again. Hoeft et al. (2010) found that, after the brain measures (specifically, from activity recorded in the right prefrontal area using fMRI and differences in white matter organization in the superior longitudinal fasciculus using DTI) were taken into account, the behavioural reading measures did not provide any better prediction of the trajectory of the dyslexic participants’ reading; predicting growth in reading directly from the brain alone, was more effective than using reading tests alone.

Naturally, strong conclusions should not be drawn from a study with a small sample size and a necessarily small range of reading tests (see also Bishop, 2010). A further limitation of such research which will affect its practical applicability is that
such brain measurements are extremely expensive, a factor that, combined with the necessary longitudinal aspect of such experiments, also greatly increases the difficulty of researching new ideas and techniques. Presumably, the goal of such research is to predict learning problems from a very early age, not fourteen, as in the Hoeft et al. (2010) study (see also Gabrieli, 2009). However, for this purpose, a simple behavioural test of the reading ability of the child’s parents may be more predictive of future problems, given the strong genetic influences on, for example, reading ability (see below for more discussion of genetics and education). Nevertheless, the study of Hoeft et al. (2010) represents a step towards a solid role for neuroscience in education, and the future of the young field of ‘neuroprognosis’ may provide methods that furnish more precise and accurate predictions.

Finally, neuroscientific techniques have been employed in various experiments as outcome measures (e.g. Moreno et al., 2011; Neville et al., 2013). The logic here is that if a particular technique causes differences in the brain, detectable using measures such as fMRI or electroencephalography (EEG), then the technique has efficacy over and above ‘mere’ psychological changes. Coltheart and McArthur (2012) provide a strong critique of such arguments, noting that the only factor relevant to educators is behavioural changes; if a new classroom technique causes measurable brain changes but has no influence on behaviour—for instance, facts memorized or attention span—then the technique is not of practical interest to teachers. Conversely, if a technique reliably changes behaviour but has no detectable neural effect, this does not speak to the failure of the technique; it could simply indicate the inadequacy of our current brain measures for reliably highlighting psychological outcomes.
Considering the mixed results from ‘neuroeducation’ so far, we would not go as far as Bruer (1997), who described the gap between neuroscience and education as “a bridge too far”, but would urge caution—at least for the present—in interpreting results that are preliminary and are not necessarily generalizable beyond the small samples involved. Teachers and researchers expecting a substantial neuroscientific influence on educational research, classroom techniques, and prediction of learning disabilities, may have some time to wait.

Social Psychology

The focus in the present thesis on interventions targeted at improving cognitive outcomes (reading, learning, memory, attention, and so on) leaves aside a major aspect of psychology in education: motivation. Our structural equation model in Chapter Six included motivational aspects and showed that these were predictive of educational duration and later socioeconomic success above and beyond cognitive aspects such as reading, mathematics, and general intelligence. It seems important, then, to investigate ways to improve motivation; several researchers from social psychology have been investigating this question. In a useful review of social-psychological interventions in education, Yeager and Walton (2011) describe two areas where motivational interventions may improve academic outcomes.

First, they describe research on students’ ‘mindsets’. A number of papers by Dweck and colleagues (e.g. Blackwell, Trzesniewski, & Dweck, 2007) have proposed that students—young and old—holding an ‘entity theory’ of intelligence (a belief that
intelligence is immutable) tend to underperform compared to those holding a ‘process theory’ (a belief that intelligence is malleable). Upon encountering an academic setback, such as a failed test, those with an entity theory are hypothesized to lose motivation, whereas those with a process theory will redouble their efforts. Blackwell et al. (2007) report an intervention designed to encourage students to change their ‘mindset’ toward a process theory. In ninety-nine students from a middle school in the US (aged around 11-14 years), they showed that this intervention—involving eight twenty-five-minute workshops—produced, as they described it, a “marginally significant” (p. 257) interaction between time and intervention group, such that children engaged in the workshops improved academically to a greater extent than a control group. They attributed the lack of significance to the low power of their experiment, but replication attempts will be necessary to test whether such gains can reliably be found.

Second, Yeager and Walton (2011) discuss ‘stereotype threat’, the phenomenon originally described by Steele and Aronson (1995), whereby a negative stereotype about a group (for instance, that “women are bad at mathematics”) causes members of that group to perform poorly on academic tests. Various ‘social-belonging’ interventions have been tested in an attempt to mitigate stereotype threat in minority students in the US. For instance, in a randomized controlled trial including 49 African-American and 43 European-American college students, Walton and Cohen (2011) had participants read material about the fact that having doubts about one’s ability in college is normal, and write an essay on that subject for future students. The authors found that, up to three years later, the minority academic achievement gap, a source of great educational and political concern in the US, could be reduced by half.
The authors argued that the intervention “untether[ed] their sense of belonging from daily hardship” (p. 1449), convincing participants that encountering everyday problems did not mean they were unsuited for college. It should be noted that the existence of stereotype threat, at least for women in the field of mathematics, has recently been questioned in a review by Stoet and Geary (2012); we may not, then, expect motivational interventions to produce such large benefits for this group.

Overall, Yeager and Walton (2011) emphasize the fact that such interventions are not ‘magic bullets’; as discussed several times in this thesis, the desire for such ‘quick fixes’ is strong, but the efficacy of all social-psychological interventions is contingent on various contextual factors, and, as noted above, a great deal more research needs to be carried out to clarify which interventions are effective and when.

Differential psychology

The last area of psychology to be considered in this section is differential psychology, an area that broadly includes intelligence, personality, and behaviour genetic research. The volume of education-specific research in this area is extremely large, and therefore this section necessarily covers only a small number of studies that highlight particularly interesting educational questions that can be answered by this kind of research.

Chapter Six of the present thesis reported a study investigating the effects of psychological variables—reading, mathematics, and intelligence—on educational duration and later socioeconomic status. The question of how cognition might affect
education can be reversed, however: does education improve cognitive ability? Just as those with higher initial cognitive ability may remain in the education system for longer, it may also be the case that schooling adds skills and abilities that improve general intellectual ability (Deary & Johnson, 2010).

Is there a way to tease apart this potentially reciprocal relationship? Brinch and Galloway (2012) found just such a way, using data from a ‘natural experiment’ that occurred in Norway in the 1960s. The government added an extra two years of compulsory education for all children, but the reform was staggered, occurring in some municipalities before others. By comparing the intelligence test scores—taken upon entry to compulsory military service some years later—of individuals from the reformed and to-be-reformed municipalities (that is, of those who had had more or fewer years of compulsory schooling), Brinch and Galloway were able to show that one extra year of schooling added on average 3.60 extra IQ points, an estimate that matches previous correlational studies (i.e. Falch and Sandgren Massih, 2011).

Above, we discussed ‘neuroprognosis’, the idea that measures of brain function could in future be used to predict learning disabilities and abilities even more effectively than behavioural measures. Could genetic testing be similarly useful to detect future problems? Whereas it is certainly the case, given substantial evidence from quantitative genetic research (twin and adoption studies) that all academic abilities—for instance, reading (e.g. Bates et al., 2007), mathematics (e.g. Kovas, Haworth, Petrill, & Plomin, 2007), and general intelligence (e.g. Deary, Spinath, and Bates, 2006)—are under substantial genetic influence, researchers have only just begun to move beyond quantitative genetics toward molecular genetics, and are
attempting to find the specific genes involved in these abilities using genome-wide association studies (GWAS), candidate-gene studies having largely failed to produce replicable results (Chabris et al., 2012).

Due to the very small effects involved—academic abilities are proposed to be highly ‘polygenic’ traits, involving small contributions from a very large number of genes (Chabris et al., 2013)—extremely large samples are required for GWAS studies in order to have the statistical power to detect them. For example, in a recent GWAS study in a sample of 126,559 individuals, Rietveld et al. (2013) were only able to find three single nucleotide polymorphisms (SNPs; DNA sequence variations) that were significantly associated with educational duration (years of schooling), and these explained around 2% of the variance. Even larger studies will be required to detect larger portions of variance; optimism about the GWAS method can at least be gained from studies of physical traits such as height, where far more variance-explaining SNPs are being found in large-sample studies (e.g. Allen, 2010).

In the era of GWAS, however, it is far from the case that twin studies have outlived their usefulness (van Dongen, Draiisma, Martin, & Boomsma, 2012). In fact, twin studies are extremely useful for the creation of controlled designs to test questions of causality in education. For instance, a long-running question in educational research concerns the effects of differences in teacher quality on the academic development of students. Observational studies are confounded by, among other factors, genetically-influenced developmental trajectories. Given that monozygotic (identical) twins share all of their genes, studying them can rule out such confounds. To address the teacher quality question in the field of reading and spelling
development, Byrne et al. (2010) studied a sample of 355 pairs of monozygotic twins, some of whom were in the same classes, and some of whom were in different classes. Larger non-genetic effects on the reading or spelling of twins in different classrooms would indicate ‘teacher effects’ (or, more precisely, ‘classroom effects’) in the development of literacy. Results showed that 8% of the variance in literacy development could be attributed to differences between teachers, a percentage that is not insubstantial, but one that is, as the authors note, “incompatible with the strident claims sometimes seen from journalists and echoed by politicians that much of the blame for poor literacy levels in some young children can be largely attributed to defective teaching” (p. 40).

Similarly, a number of studies (reviewed by Stanovich & Cunningham, 2001) suggest that reading exposure, measured using an Author Recognition Test (which involves a list of names, some of which are those of authors, some of which are lures; the participant has to select the names they think are real authors), may improve verbal intelligence. Such a transfer from a specific skill to a general ability is potentially of great importance, and would indicate that teaching reading has effects beyond literacy to improve general intellectual functioning. Again, however, the previous research is purely correlational and does not rule out potential genetic and other confounds, and so cannot answer the causal question of whether improving reading would improve intelligence (see Lee, 2012, for a superb discussion of correlation and causation issues in differential psychological research).

To avoid the issue of confounding, Ritchie, Bates, and Plomin (under review) estimated a longitudinal monozygotic-differences model in a sample of 1,890 twin
pairs, testing whether twins with better reading ability at earlier ages tended to have higher subsequent intelligence test scores, even controlling for earlier intelligence. This hypothesis was supported, and the results extended beyond verbal intelligence to nonverbal measures such as Raven’s Progressive Matrices, an indicator of reasoning ability. Methods from differential psychology, then, are powerful ways to test important educational questions where direct randomized interventions (in this case, stopping some children from reading while others continue to learn) are potentially unethical, and rule out a number of significant confounds along the way.

**Conclusion**

This chapter summarized the outcomes of the present thesis, and gave a brief outline of three broad areas with important contributions to make to education. It has emphasized the important role that psychological science has to play in informing teachers, parents, and policy-makers, and testing new ideas in education. Studies from cognitive psychology, such as those reported in the majority of chapters in the present thesis, test specific learning techniques, building them into theories of the mind that should be taught as standard in teacher-training courses. It is fair to say that the utility of neuroscientific techniques for education has been oversold, and this will remain the case until advocates of “neuroeducation” provide stronger evidence that brain scans have better predictive power than cognitive tests, and until costs associated with neuroimaging drop substantially. A growing number of studies testing techniques from social psychology that target students’ motivation to succeed may help reduce socially-influenced gaps in achievement across socioeconomic, sex, and ethnic divisions, though replication studies are required before large-scale roll-outs of the interventions. Large-scale hypotheses tested by differential psychologists may unlock
CHAPTER 7: SUMMARY AND DISCUSSION

the causes of educational differences, from genes to the brain to the mind, and assess their consequences for the individual and for society. The combination of evidence from all these fields of psychological science has the potential to overturn our current understanding of the educational process, and improve academic outcomes for all.
References


overview of an emerging field. *Neuroethics, 5*, 105-117.


Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects
models using Eigen and S4. R package version 1.0-6. Retrieved from:
http://CRAN.R-project.org/package=lme4

Bates, T. C., Castles, A., Luciano, M., Wright, M. J., Coltheart, M., & Martin, N. G.
(2007). Genetic and environmental bases of reading and spelling: A unified

spectrum disorder: A systematic review. *Journal of Autism and
Developmental Disorders, 39*, 1145-1154.

memory and attention of children. *Appetite, 53*, 143-146.

Low health literacy and health outcomes: An updated systematic review.
*Annals of Internal Medicine, 155*, 97-107.

Best, J. R. (2010). Effects of physical activity on children’s executive function:
Contributions of experimental research on aerobic exercise. *Developmental
Review, 30*, 331-351.


attainment: Influences of childhood intelligence, childhood social factors, and education. *Intelligence, 33*, 455-472.


Dewar, M., Cowan, N., & Della Sala, S. (2007). Forgetting due to retroactive interference: A fusion of Müller and Pilzecker’s (1900) early insights into forgetting and recent research on anterograde amnesia. *Cortex, 43*, 616-634.


learning than elaborative studying with concept mapping". *Science, 334*, 453- c.


automatization or a phonological deficit? *Scientific Studies of Reading*, 2, 321-
340.

of Psychology*, 55, 235-269.

Wixted, J. T. (2010). The role of retroactive interference and consolidation in
everyday forgetting. In S. Della Sala (Ed.), *Forgetting* (pp. 285-312). Sussex,
UK: Psychology Press.

Wolfsont, C. (2002). Increasing behavioural skills and level of understanding in
adults: A brief method integrating Dennison’s brain gym balance with Piaget’s

self-perception research study. *Current Medical Research and Opinion*, 25,
1491-1500.
Appendix A

Supplementary materials for Chapter Three:

The Effects of In-Class Exercise and Relaxation Sessions on Children’s Preparedness for Learning
Figure A1, Experiment 3.1 Materials I. Remaining two alternate forms of the Lollipops Test (first form shown in Figure 3.1).
Figure A2, Experiment 3.1 Materials II: Example of the 5×5 grid Visual Patterns Test administered to Primary 4 and Primary 6 classes in Experiment 3.1; other forms had the crosses arranged in different patterns. For younger children, a 4×4 grid was used.
Experiment 3.2 Materials I: Below are the three versions of the Memory Test materials not reproduced in Chapter 3.

1) Pangolin Fact Set – Primary 2 and 3

“The Pangolin is an animal that lives in Africa. It is covered in scales, like a suit of armour, and it is yellow in colour. Pangolins mainly eat ants. They dig the ants out of their nests with their sharp claws, and pick them up with their very long, thin, sticky tongue. The pangolin spends most of the day sleeping. Tigers like to eat pangolins, but if they try, the pangolin will curl up into a ball, and spray the tiger with a horrible smell. There are eight different kinds of pangolin, and these include the giant pangolin and the tree pangolin.”

2) Slow Loris Fact Set – Primary 2 and 3

“The Slow Loris is an animal that lives in forests in Asia. It has very big eyes, which help it see in the dark. Slow lorises live for about 25 years. Slow lorises mainly eat fruit and insects, but sometimes sneak into birds’ nests and steal their eggs. Slow lorises are very good tree climbers, and have a strong grip. They often hang upside-down from their feet from branches, which leaves their hands free to eat. Sometimes bears try to eat slow lorises, but the slow loris has a very painful, poisonous bite, which can scare predators away. There are five different kinds of slow loris. Sadly, the number of slow lorises in the world is getting smaller, because people take them away from the forest to keep as pets.”

3) Slow Loris Fact Set – Primary 4, 5, and 6

“The Slow Loris is an animal that is found in forests in Asia. It has very big eyes, which help it see in the dark. Slow lorises live for about 25 years. Slow lorises mainly eat fruit and insects, but sometimes sneak into birds’ nests and steal their eggs. Slow lorises are very good tree climbers, and have a strong grip. They often hang upside-down from their feet from branches, which leaves their hands free to eat. Sometimes bears and apes try to eat slow lorises, but the slow loris has a very painful, poisonous bite, which can scare predators away. There are five different species of slow loris. Sadly, the number of slow lorises in the world is getting smaller, because people take them away from the forest and keep them as pets.”
Experiment 3.2 Materials II: Quiz scripts recorded by voice actor and played to children of each age group, for each of the two animals. All questions were read twice, as shown below.

1) Pangolin Quiz, Primary 2 and 3.

“This quiz is about the pangolin, the animal you heard about yesterday. Please write your name and class at the top of the sheet.

Question 1. Which of these animals is the pangolin? Circle the picture.

[REPEAT, wait 5s]

Question 2. What colour is the pangolin? Circle the word.

[REPEAT, wait 5s]

Question 3. What does the pangolin like to eat? Circle the picture.

[REPEAT, wait 5s]

Question 4. Which of these animals likes to eat the pangolin? Circle the picture.

[REPEAT, wait 5s]

Question 5. How many different kinds of pangolin are there? Circle the number.

[REPEAT, END]”

2) Pangolin Quiz, Primary 4, 5 and 6.

“This quiz is about the animal you heard about yesterday. Please write your name and class at the top of the sheet.

Question 1. What is the name of the animal you learned about yesterday?

[REPEAT, wait 5s]

Question 2. Where is the animal found?

[REPEAT, wait 5s]

Question 3. Which of these animals do you think is the animal you heard about? Circle the picture.

[REPEAT, wait 5s]

Question 4. What colour is the animal?

[REPEAT, wait 5s]

Question 5. If you measured the animal, how long would it be?

[REPEAT, wait 5s]

Question 6. What is the main thing the animal eats?

[REPEAT, wait 5s]

Question 7. How does the animal get its food?
[REPEAT, wait 5s]

Question 8. The animal has a predator. What is it?

[REPEAT, wait 5s]

Question 9. What does the animal do if something tries to eat it?

[REPEAT, wait 5s]

Question 10. How many different species of the animal are there?

[REPEAT, wait 5s]

Question 11. Can you name two of the different species?”

[REPEAT, END]

3) Slow Loris Quiz, Primary 2.

“This quiz is about the slow loris, the animal you heard about yesterday. Please write your name and class at the top of the sheet.

Question 1. Which of these animals is the slow loris? Circle the picture.

[REPEAT, wait 5s]

Question 2. How many years can the slow loris live for? Circle the number.

[REPEAT, wait 5s]

Question 3. Where does the slow loris like to eat its food? Circle the picture.

[REPEAT, wait 5s]

Question 4. Which of these animals likes to eat the slow loris? Circle the picture.

[REPEAT, wait 5s]

Question 5. How many different kinds of slow loris are there? Circle the number.

[REPEAT, END]”

4) Slow Loris Quiz, Primary 4, 5, and 6.

“This quiz is about the animal you heard about yesterday. Please write your name and class at the top of the sheet.

Question 1. What is the name of the animal you learned about yesterday?

[REPEAT, wait 5s]

Question 2. Where is it found?

[REPEAT, wait 5s]

Question 3. Which of these animals do you think is the animal you heard about? Circle the picture.
Question 4. How long does the animal live?
[REPEAT, wait 5s]

Question 5. Can you name three things the animal likes to eat?
[REPEAT, wait 5s]

Question 6. Where does it eat its food?
[REPEAT, wait 5s]

Question 7. Can you name two of the predators of the animal?
[REPEAT, wait 5s]

Question 8. Why do predators stay away from the animal?
[REPEAT, wait 5s]

Question 9. Why is the number of those animals in the world getting smaller?
[REPEAT, wait 5s]

Question 10. How many different species of the animal are there?
[REPEAT, wait 5s]

Question 11. Can you name two of the different species?
[REPEAT, END]”
Figure A3. Experiment 3.2 Materials III: Quiz sheets for the younger and older children for both animals (not actual size; all quiz sheets printed on A4 paper).

(1) Pangolin Quiz, Primary 2 and 3.

<table>
<thead>
<tr>
<th>Week 1 Quiz</th>
<th>NAME</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>YELLOW</td>
<td>RED</td>
</tr>
<tr>
<td>3.</td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
<tr>
<td>4.</td>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

(2) Pangolin Quiz, Primary 4, 5, and 6.

<table>
<thead>
<tr>
<th>Quiz</th>
<th>NAME</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
</tbody>
</table>
(3) Slow Loris Quiz, Primary 2 and 3

<table>
<thead>
<tr>
<th>Quiz</th>
<th>NAME</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.   31   40   7   25

3.   ![Image](image1)

4.   ![Image](image2)

5.   12   5   9   2

(4) Slow Loris Quiz, Primary 4, 5, and 6.

<table>
<thead>
<tr>
<th>Quiz</th>
<th>NAME</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.   

3.   ![Image](image3)

4.   

5.   

6.   

7.   

8.   

9.   
Supplementary Materials for Chapter Five:

Retrieval practice, with or without mind mapping, boosts fact learning in primary school children
Experiment 5.1 Materials I. Text of all fact sheets for Primary 5 in Experiment 5.1.

1) Senegal
The country. Senegal is a country in Africa. In a list of the world’s biggest countries, Senegal is number 67. The capital city of Senegal is called Dakar. It takes 6 hours to fly there from Edinburgh.
The people. Most people in Senegal live in the countryside, and most of them follow the religion of Islam.
What is it like to live there? The weather in Senegal is hot all year round. Senegal is famous for its music.
What do they make there? Two things made in Senegal are peanuts and coffee, which are sold to people in other countries.

2) South Korea
The country. South Korea is a country in Asia. In a list of the world’s biggest countries, South Korea is number 109. The capital city of South Korea is called Seoul. It takes 10 hours to fly there from Edinburgh.
The people. Most people in South Korea live in cities, and most of them do not have a religion.
What is it like to live there? The weather in South Korea is hot and wet. South Korea is famous for its computers.
What do they make there? Two things made in South Korea are boats and robots, which are sold to people in other countries.

3) Peru
The country. Peru is a country in South America. In a list of the world’s biggest countries, Peru is number 20. The capital city of Peru is called Lima. It takes 12 hours to fly there from Edinburgh.
The people. Most people in Peru live in cities, and most of them follow the religion of Christianity.
What is it like to live there? The weather in Peru is hot all year round. Peru is famous for its mountains.
What do they make? Two things made in Peru are fish and gold, which are sold to people in other countries.

4) Iran
The country. Iran is a country in Asia. In a list of the world’s biggest countries, Iran is number 18. The capital city of Iran is called Tehran. It takes 6 hours to fly there from Edinburgh.
The people. Most people in Iran live in cities, and most of them follow the religion of Islam.
What is it like to live there? The weather in Iran is hot and dry. Iran is famous for its history.
What do they make? Two things made in Iran are oil and carpets, which are sold to people in other countries.
Experiment 5.1 Materials II. Text of all fact sheets for Primary 7 in Experiment 5.1.

1) Senegal
The country. Senegal is a country in Africa. Senegal is a medium-sized country, and in a list of the world’s biggest countries, Senegal comes 67th. The capital city of Senegal is called Dakar. It takes 6 hours to fly there from Edinburgh.
The people. Thirteen million people live in Senegal, and most of them live in the countryside. Most of the people in Senegal follow the religion of Islam.
What is it like to live there? Senegal has a tropical climate, meaning it is hot all year round. Senegal is famous for its music.
What do they make there? One kind of Senegalese food is yassa, which is made of fish and garlic. Two things produced in Senegal are peanuts and coffee, which are sold to people in other countries.

2) South Korea
The country. South Korea is a country in Asia. South Korea is a small country, and in a list of the world’s biggest countries, South Korea comes 109th. The capital city of South Korea is called Seoul. It takes 10 hours to fly there from Edinburgh.
The people. Forty-five million people live in South Korea, and most of them live in cities. Most of the people in South Korea do not follow any religion.
What is it like to live there? South Korea has a humid climate, meaning it is hot and wet. South Korea is famous for its technology.
What do they make there? One kind of South Korean food is kimchi, which is made of vegetables and spices. Two things produced in South Korea are ships and robots, which are sold to people in other countries.

3) Peru
The country. Peru is a country in South America. Peru is a large country, and in a list of the world’s biggest countries, Peru comes 20th. The capital city of Peru is called Lima. It takes 12 hours to fly there from Edinburgh.
The people. Thirty million people live in Peru, and most of them live in cities. Most of the people in Peru follow the religion of Christianity.
What is it like to live there? Peru has a tropical climate, meaning it is hot all year round. Peru is famous for its mountains.
What do they make? One kind of Peruvian food is pachamanca, which is made from meat and spices. Two things produced in Peru are fish and gold, which are sold to people in other countries.

4) Iran
The country. Iran is a country in Asia. Iran is a large country, and in a list of the world’s biggest countries, Iran comes 18th. The capital city of Iran is called Tehran. It takes 6 hours to fly there from Edinburgh airport.
The people. Seventy-five million people live in Iran, and most of them live in cities. Most of the people in Iran follow the religion of Islam.
What is it like to live there? Iran has an arid climate, which means it is hot and dry. Iran is famous for its history.
What do they make? One kind of Iranian food is tah-chin, which is made from rice and chicken. Two things produced in Iran are oil and carpets, which are sold to people in other countries.
Experiment 5.1 Materials III: Quiz questions used (for all countries simultaneously) in Experiment 5.1 for Primary 5.

1) What is the name of the country you learned about?
2) Name two things made in the country that are sold to other countries.
3) Where is the country?
4) Which number is the country in a list of world’s biggest countries?
5) What is the name of the country’s capital city?
6) How many hours does it take to fly there from Edinburgh?
7) What is the weather like in the country?
8) What is the country famous for?
9) Where do most people in the country live?
10) Do most people in the country have a religion? If they do, which one is it?

Experiment 5.1 Materials IV: Quiz questions used (for all countries simultaneously) in Experiment 5.1 for Primary 7.

1) What is the name of the country you learned about?
2) Name a kind of food from the country.
3) What is the food made of?
4) Name two things made in the country that are sold to people in other countries.
5) Where is the country?
6) Which number is the country in a list of the world’s biggest countries?
7) What is the name of the country’s capital city?
8) How many hours does it take to fly there from Edinburgh?
9) What is the climate of the country, and what does this mean?
10) What is the country famous for?
11) How many people live in the country, in millions?
12) Where do most of the people in the country live?
13) Do most people in the country have a religion? If they do, which one is it?
**Experiment 5.2 Materials I:** Text of all fact sheets for Primary 5 in Experiment 5.2 (nb. fact sheets for Primary 4 were identical to those used for Primary 5 in Experiment 5.1).

1) Senegal

*The country.* Senegal is a country in Africa. In a list of the world’s biggest countries, Senegal comes 67th. The capital city of Senegal is Dakar. It takes 6 hours to fly there from Edinburgh.

*The people.* Thirteen million people live in Senegal, and most of them live in the countryside. Most of the people in Senegal follow the religion of Islam.

*What is it like to live there?* Senegal has a tropical climate, meaning it is hot all year round. Senegal is famous for its music and its football. The money used in Senegal is called the Franc.

*What do they make there?* One kind of Senegalese food is yassa, which is made of fish and garlic. Two things made in Senegal are peanuts and coffee, which are sold to people in other countries.

2) South Korea

*The country.* South Korea is a country in Asia. In a list of the world’s biggest countries, South Korea comes 109th. The capital city of South Korea is called Seoul. It takes 10 hours to fly there from Edinburgh.

*The people.* Forty-five million people live in South Korea, and most of them live in cities. Most of the people in South Korea do not follow any religion.

*What is it like to live there?* South Korea has a humid climate, meaning it is hot and wet. South Korea is famous for its technology and its martial arts. The money used in South Korea is called the Won.

*What do they make there?* One kind of South Korean food is kimchi, which is made of vegetables and spices. Two things made in South Korea are ships and robots, which are sold to people in other countries.

3) Peru

*The country.* Peru is a country in South America. In a list of the world’s biggest countries, Peru comes 20th. The capital city of Peru is called Lima. It takes 12 hours to fly there from Edinburgh.

*The people.* Thirty million people live in Peru, and most of them live in cities. Most of the people in Peru follow the religion of Christianity.

*What is it like to live there?* Peru has a tropical climate, meaning it is hot all year round. Peru is famous for its mountains and its animals. The money used in Peru is called the Sol.

*What do they make?* One kind of Peruvian food is pachamanca, which is made from meat and spices. Two things made in Peru are fish and gold, which are sold to people in other countries.

4) Iran

*The country.* Iran is a country in the Middle East. In a list of the world’s biggest countries, Iran comes 18th. The capital city of Iran is called Tehran. It takes 6 hours to fly there from Edinburgh airport.

*The people.* Seventy-five million people live in Iran, and most of them live in cities. Most of the people in Iran follow the religion of Islam.

*What is it like to live there?* Iran has an arid climate, which means it is hot and dry. Iran is famous for its history and its culture. The money used in Iran is called the Rial.

*What do they make?* One kind of Iranian food is tah-chin, which is made from rice and chicken. Two things made in Iran are oil and carpets, which are sold to people in other countries.


Experiment 5.2 Materials II: Text of all fact sheets for Primary 6 and 7 in Experiment 5.2.

1) Senegal

The country. Senegal is a country in Africa. In a list of the world’s biggest countries, Senegal comes 67th. The capital city of Senegal is Dakar. It takes 6 hours to fly there from Edinburgh.

The people. Thirteen million people live in Senegal, and most of them live in the countryside. Most of the people in Senegal follow the religion of Islam, and the language most of them speak is French. The leader of Senegal is called Macky Sall.

What is it like to live there? Senegal has a tropical climate, meaning it is hot all year round. Senegal is famous for its music and its football. The money used in Senegal is called the Franc.

What do they make there? One kind of Senegalese food is yassa, which is made of fish and garlic. Three things produced in Senegal are cotton, peanuts, and coffee, which are sold to people in other countries.

2) South Korea

The country. South Korea is a country in Asia. In a list of the world’s biggest countries, South Korea comes 109th. The capital city of South Korea is called Seoul. It takes 10 hours to fly there from Edinburgh.

The people. Forty-five million people live in South Korea, and most of them live in cities. Most of the people in South Korea do not follow any religion, and the language most of them speak is Korean. The leader of South Korea is called Lee Myung-Bak.

What is it like to live there? South Korea has a humid climate, meaning it is hot and wet. South Korea is famous for its technology and its martial arts. The money used in South Korea is called the Won.

What do they make there? One kind of South Korean food is kimchi, which is made of vegetables and spices. Three things produced in South Korea are ships, cars, and robots, which are sold to people in other countries.

3) Peru

The country. Peru is a country in South America. In a list of the world’s biggest countries, Peru comes 20th. The capital city of Peru is called Lima. It takes 12 hours to fly there from Edinburgh.

The people. Thirty million people live in Peru, and most of them live in cities. Most of the people in Peru follow the religion of Christianity, and the language most of them speak is Spanish. The leader of Peru is called Ollanta Humala.

What is it like to live there? Peru has a tropical climate, meaning it is hot all year round. Peru is famous for its mountains and its animals. The money used in Peru is called the Sol.

What do they make? One kind of Peruvian food is pachamanca, which is made from meat and spices. Three things produced in Peru are wood, fish, and gold, which are sold to people in other countries.

4) Iran

The country. Iran is a country in the Middle East. In a list of the world’s biggest countries, Iran comes 18th. The capital city of Iran is called Tehran. It takes 6 hours to fly there from Edinburgh airport.

The people. Seventy-five million people live in Iran, and most of them live in cities. Most of the people in Iran follow the religion of Islam and the language most of them speak is Persian. The leader of Iran is called Ali Khamenei.

What is it like to live there? Iran has an arid climate, which means it is hot and dry. Iran is famous for its history and its culture. The money used in Iran is called the Rial.

What do they make? One kind of Iranian food is tah-chin, which is made from rice and...
APPENDIX B

chicken. Three things produced in Iran are oil, nuts, and carpets, which are sold to people in other countries.

Experiment 5.2 Materials III: Quiz questions used (for all countries simultaneously) at one week in Experiment 5.2 for Primary 5 (n.b. questions for Primary 4 were identical to those for Primary 5 in Experiment 5.1), followed by the order of the questions at the five-week test.

1) What is the name of the country you learned about last week?
2) Where is the country?
3) Where does the country come in a list of the world’s biggest countries?
4) How many people live in the country?
5) Where do most of them live?
6) Name two things the country is famous for.
7) What is the name of the money used in the country?
8) Which religion do most of the people follow in the country?
9) What kind of climate does the country have, and what does this mean?
10) Name two things that are made in the country.
11) What is the name of the country’s capital city?
12) How long would it take to fly there from Edinburgh?
13) Name a kind of food made in the country.
14) What two things is the food made of?

Order of the above questions at the five-week test: 1, 2, 5, 8, 3, 11, 10, 13, 14, 5, 6, 7, 4, 9.

Experiment 5.2 Materials IV: Quiz questions used (for all countries simultaneously) at one week in Experiment 5.2 for Primary 6 and 7 followed by the order of the questions at the five-week test.

1) What is the name of the country you learned about last week?
2) Where is the country?
3) Where does the country come in a list of the world’s biggest countries?
4) How many people live in the country?
5) Where do most of them live?
6) Name two things the country is famous for.
7) Which language do most people speak in the country?
8) What is the name of the money used in the country?
9) Which religion do most of the people follow in the country?
10) What is the name of the country’s leader?
11) What kind of climate does the country have, and what does this mean?
12) Name three things that are produced in the country.
13) What is the name of the country’s capital city?
14) How long would it take to fly there from Edinburgh?
15) Name a kind of food made in the country.
16) What two things is the food made of?
Order of the above questions at the five-week test: 1, 2, 4, 3, 11, 10, 5, 6, 9, 15, 16, 12, 13, 8, 14, 7.
Appendix C

Supplementary Material for Chapter 6: Enduring Links from Childhood Mathematics and Reading Achievement to Adult Socioeconomic Status
This document contains five additional path diagrams to supplement Figure 6.2 in Chapter 6. In all diagrams, values are path coefficients, with values for males on the left, and for females on the right of each ‘/’. A similar sex-equation process to that described in Chapter 6 was performed on all but the first model; for these last four models, paths that could not be equated between the sexes are marked with an asterisk. In all models, non-significant \( (p > .05) \) paths are dotted.

**Figure C1**

Figure C1 is included to allow the reader to see the model produced before the sex-equation process described in Chapter 6. Prior to this process - which changed some values in the model by a small amount - the direct and indirect associations of mathematics and reading with socioeconomic status (SES) were still found, indicating that they are robust to analytic changes.

**Figures C2 and C3**

In Figures C2 and C3, we model the separate associations of mathematics (Figure C2) and reading (Figure C3) with SES attainment. Chapter 6 describes a discriminant validity analysis carried out following on from these models.

**Figures C4 and C5**

Finally, we explored the sensitivity of our model’s conclusions to the specific variables chosen. If our interpretation of our results, that mathematics and reading promote lifespan social/economic growth, is valid, then the association between mathematics and reading and later-life SES should remain after adding additional plausible causal variables to the model. To this end, we created two new models, each time adding a new variable: oral language ability, followed by aggression.

We created a model including oral language ability for two reasons. First, previous studies (e.g. Conti-Ramsden, Durkin, Simkin, & Knox, 2009), have shown poorer educational outcomes for children with specific language impairment (SLI), indicating that this variable may have similar effects to those already included in our model. Second, SLI is often comorbid with dyslexia (Catts, Adlof, Hogan, & Ellis Weismer, 2005), which raises the possibility that our original model may have misattributed variance to reading ability that was in fact better explained by language
ability. In addition, language ability may underlie some of the variance in mathematics and reading achievement, via its potential effects on classroom learning.

Oral language ability was teacher-rated when cohort members were aged 7, on a scale from 1 (‘expresses well’) to 5 (‘very poor ability’). The variable was reversed for inclusion in the model. The mean score (SD) for the 7707 males assessed was 2.97 (.97), and for the 7306 females, 2.80 (.94). The model is illustrated in Figure C4; the inclusion of oral language ability did not significantly alter the reported paths between mathematics, reading, and attained SES (for the direct effects: male/female change in path weight for mathematics to attained SES $\Delta = -0.02/-0.01$; for reading to attained SES $\Delta = -0.02/-0.02$).

We next examined the potential effects of aggression, as behavioral problems could potentially influence classroom learning and thereby affect educational and other life-course outcomes. Previously, these problems have been shown to be associated with SES attainment (von Stumm, Gale, Batty, & Deary, 2009). It is unclear whether our motivation variable, which did include attitudes towards school, would have fully controlled for classroom behavior.

Teachers rated each cohort member’s age-16 aggression on a scale of 1 (‘timid’) to 5 (‘aggressive’). Again, this variable was reversed before inclusion in the model. The mean score (SD) for the 6322 males assessed was 2.99 (.78), and for the 6009 females, 2.88 (.78). Figure C5 shows a path diagram of the model with the inclusion of this variable; again, the associations between mathematics, reading, and attained SES are not significantly changed from the original model (direct effects: $\Delta$ male/female path weights for mathematics $= -.01/.01$, and reading $= .00/-.01$).

In neither of these new models did the inclusion of additional, theoretically-motivated variables significantly change the effects of childhood reading and mathematics achievement on SES attainment reported in the original model (Figure 6.2) underlining the robustness of these associations.
Figure C1. Associations of age-7 mathematics and reading achievement with attained socioeconomic status, prior to the sex-equation process described in Chapter 6.
Figure C2. Associations of age-7 mathematics achievement with attained socioeconomic status, excluding reading.
Figure C3. Associations of age-7 reading achievement with attained socioeconomic status, excluding mathematics.
Figure C4. Associations of age-7 mathematics and reading achievement with attained SES, with the inclusion of age-7 oral language ability.
Figure C5. Associations of age-7 mathematics and reading achievement with attained SES, with the inclusion of age-16 aggression.
Appendix D

Publication list
The following articles were published and/or submitted during the period in which this thesis was written:

**Under review**


Ritchie, S. J., Bates, T. C., & Deary, I. J. (Under review). Does education boost general intelligence (g), or specific abilities?

Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (Under review). Retrieval practice, with or without mind mapping, facilitates fact learning in primary school children.

**In Press**


**Published**


