The active brain: relationships between physical activity, fitness, cognitive performance and brain structure in young adolescents



Thomas Michiel Wassenaar

Brasenose College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2021

This page is intentionally left blank.

The active brain: relationships between physical activity, fitness, cognitive performance and brain structure in young adolescents

Thomas Michiel Wassenaar

Hilary Term 2021 Brasenose College

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Abstract

Governmental, scientific and public health communities have become increasingly interested in the potential of physical activity to promote neurocognitive health across the lifespan. While research in adults has advanced rapidly in the past 20 years, much less is known about the effects of a physically active lifestyle on cognitive and brain health in young people, especially adolescents. Using magnetic resonance imaging (MRI) and cognitive data from the Fit to Study cluster randomised controlled trial, this thesis therefore examined relationships between neurocognitive health and markers of physical activity and cardiorespiratory fitness in this age group.

A cross-sectional study involving baseline Fit to Study data showed that higher levels of physical activity and fitness are associated with superior performance on domain-specific cognitive assessments (working memory, cognitive flexibility, inhibitory control, visual memory and processing speed) in a large sample of young British adolescents (n = 16,017), but that relationships are small and explain less than 10% of the variance in cognitive outcomes. It also showed that relationships may be stronger in females compared to males and that a physically active lifestyle is linked with more favourable cognitive outcomes. In a sub-sample of adolescents (n = 61), however, no significant relationships between markers of a physically active lifestyle and volumes or multi-modal MRI markers of tissue (micro)structure in the anterior hippocampus were observed.

An intervention study explored whether increasing the amount of vigorous intensity physical activity during school physical education (PE) lessons for one academic year could improve cardiorespiratory fitness and induce neurocognitive changes in the Fit to Study sample (n = 93 schools). The intervention, designed with PE teachers to be scalable and feasible, did not significantly improve fitness or cognitive performance. In addition, no robust intervention-related changes in anterior hippocampal MRI metrics were observed. These findings should, however, be interpreted with caution given low implementation fidelity and high drop-out.

This thesis supports the hypothesis that higher levels of physical activity and cardiorespiratory fitness are associated with superior cognitive performance in young adolescents. However, no physical-activity induced effects on neurocognitive health were observed. Further investigation of *whether* and *how* physical activity may promote positive brain and cognitive changes in adolescents is needed.

This thesis contains approximately 50,000 words.

This page is intentionally left blank.

To my family

This page is intentionally left blank.

Acknowledgements

This thesis could not have been completed without the support of many people, to whom I am truly grateful. At the risk of omitting some, I wish to thank several institutions and individuals in particular:

To the pupils, parents, teachers and schools participating in the Fit to Study cluster randomised controlled trial, without whom none of the work presented in this thesis would have been possible.

To my doctoral advisers, Heidi Johansen-Berg and Claire Sexton, the "dream team", for their endless support, inspiration and advice. To Heidi in particular, for her trust, for always finding time to meet, and for giving me the freedom to do my own thing. To Claire in particular, for her enthusiasm, for asking insightful questions and for keeping my eye on the bigger picture.

To the funders: To Brasenose College, the Clarendon fund and the Nuffield Department of Clinical Neurosciences (NDCN), for providing a scholarship that allowed me to pursue a DPhil in Clinical Neurosciences; To Brasenose College, NDCN and Guarantors of Brain for providing course and travel funding; To the Prins Bernhard Cultuurfonds for financial support during the first year of my DPhil; To the Vice-Chancellors Fund and Goodger and Schorstein Scholarship for providing additional funding to complete my DPhil; To the Medical Sciences Division (MSD) and NDCN for providing a brief COVID-related scholarship extension; To Wellcome and the Education Endowment Foundation for funding the Fit to Study project.

All the analyses presented in this thesis are based on data collected as part of the Fit to Study trial. Fit to Study, like any other project, could not have been conducted without the input and hard work of many individuals, to whom I wish to express my gratitude. To Heidi Johansen-Berg and Helen Dawes, the principal investigators, for securing the funding and providing oversight. To Catherine Wheatley and Kathryn Atherton for running initial pilots of the intervention, and to Kathryn Atherton for setting up the cognitive assessments used in this thesis. To Josh De Leeuw, who subsequently programmed the assessments in jsPsych and Paul Robinson for linking them up with the Fit to Study database. To Catherine Wheatley and Nick Beale, who helped design the main trial and assisted with data collection. To Andy Meaney, Joan Duda, Richard Neil and the late Kate Curtis for their advice on intervention development and testing in schools. To Emma Eldridge, Emily Curtis, Emily Plester, Jack Possee and Oliver Bushnell for their help with data collection. To Piergiorgio Salvan for co-designing and executing the brain imaging sub-study. To Tom Smejka, Eneid Lika, Cyrus Goodger and Patrick Esser, for assisting with data collection of the brain imaging sub-study. To Michael Sanders, Jon Campbell, David Parker and Caroline Young, as well as to weekend "covers" Emily Hinson, Nicola Filippini, Enikő Zsoldos, and Caitlin O'Brien, for collectively running over 150 scan sessions within two summer periods. My own involvement in Fit to Study began in December 2016 when I spent nine months working as a research assistant on the project and was involved in piloting the cognitive assessments and finalising the overall trial design. During my DPhil (2017–2021) I was

involved in the design and execution of the main trial. I had primary responsibility for designing, implementing and analysing cognitive assessments. I also took on primary responsibility for the design and maintenance of the study database, and for ensuring that analysis of missing data across the whole study dataset was dealt with appropriately. I had shared responsibility for designing, collecting and analysing the brain imaging sub-study data and I contributed to the design, implementation and analysis of other study measures. I carried out all of the analyses presented in this thesis.

To all current and former members of the Neuroplastics and Plasticity groups, for making my time in Oxford enjoyable and for being great colleagues and friends. Special thanks to Alberto Lazari and Zeena-Britt Sanders for welcoming me in your posse, sharing the ups and downs of the DPhil, and the weekly snack runs.

To Tom Nichols and the late Dan Lunn for sparking my interest in statistics and providing support when needed. Special thanks to Anderson Winkler, for always being there to advice on stats (and other) queries, and for being an excellent role model for how research should be done.

Various chapters benefited from input from individuals who I wish to acknowledge. Thanks go to: Wilby Williamson, for taking up the tedious job of being the second reviewer on a systematic review (Chapter 1); To Natalie Voets and Mark Jenkinson for discussing hippocampal segmentations (Chapters 5 and 7); To Tom Okell for advising on ASL analyses (Chapters 5 and 7); To Naiara Demnitz who returned the favour of verifying manual adjustments of hippocampal masks (Chapters 5 and 7); To Kamila Szulc-Lerch, Jason Lerch and Mallar Chakravarty for providing me with the resources to extract hippocampal subfields; To Matteo Quartagno, while we have not met in person, I always appreciated and admired your swift responses to (jomo-related) missing data queries; To Karla Diaz-Ordaz for sharing some thoughts on analyses involving missing data (Chapter 6); To Charles Hulme for advising on the latent variable analysis (Chapter 6).

Massive thanks to the Lubberts for being the great friends that you are, the yearly visits, and for keeping up with me being an eternal student (het huiswerk is bijna klaar!).

My biggest thanks go to my parents, Inte and Isabel, and my brother, Pieter, for their unwavering support, and to Paola, for keeping me sane and showing me that there is more to life than work.

Thomas Wassenaar
January 2021

Publications

Throughout my doctorate studies, my supervisors and I, decided to prioritise the preparation of manuscripts for submission to peer-reviewed journals. Several chapters of this thesis are therefore adapted from work that has been published:

- Wassenaar, T. M., Williamson, W., Johansen-Berg, H., Dawes, H., Roberts, N., Foster, C. & Sexton, C. E. A critical evaluation of systematic reviews assessing the effect of chronic physical activity on academic achievement, cognition and the brain in children and adolescents: a systematic review. *International Journal of Behavioral Nutrition & Physical Activity* **17,** 1–18 (2020) [**chapter 1** & **PROSPERO protocol**]

- Wheatley, C. M., Wassenaar, T. & Johansen-Berg, H. *Educational Neuroscience. Development Across the Life Span* (eds Thomas, M. S. C., Mareschal, D. & Dumontheil, I.) 359–385 (Taylor & Francis, New York, 2020) [**chapter 1**][1]

- Wassenaar, T. M., Wheatley, C. M., Beale, N., Salvan, P., Meaney, A., Possee, J. B., Atherton, K. E., Duda, J. L., Dawes, H. & Johansen-Berg, H. Effects of a programme of vigorous physical activity during secondary school physical education on academic performance, fitness, cognition, mental health and the brain of adolescents (Fit to Study): Study protocol for a cluster-randomised trial. *Trials* **20,** 189 (2019) [**chapter 2**]

Other chapters from this thesis are adapted from manuscripts that have been submitted or have been prepared for submission[2]:

- Wassenaar, T. M., Wheatley, C., Beale, N., Salvan, P., Winkler, A., Nichols, T., Dawes, H. & Johansen-Berg, H. Associations between physical activity, cardiorespiratory fitness and cognitive performance in a large sample of young British adolescents. (*submitted*) [**chapter 4**]

- Wassenaar, T. M., Wheatley, C., Beale, N., Salvan, P., Nichols, T., Diaz-Ordaz, K., Dawes, H. & Johansen-Berg, H. The effect of a one-year vigorous physical activity intervention on fitness, cognitive performance and mental health in young adolescents: the Fit to Study cluster randomised controlled trial. (*submitted*) [**chapter 6**]

In addition, I contributed to other work that was outside the scope of this thesis and therefore not included.

- Wheatley, C., Wassenaar, T. M., Beale, N., Hopkins, D., Salvan, P., Dawes, H., Davies, E. & Johansen-berg, H. Associations between physical and environmental variables and physical activity in young British adolescents: longitudinal data from the Fit to Study trial. (*under review*)

[1] I wrote the sections on *how* and *whether* acute and chronic physical activity may affect the brain and cognition

[2] Proposed manuscript details may change following submission of the DPhil thesis

- Demnitz, N., Stathi, A., Withall, J., Stainer, C., Seager, P., DeKoning, J., Esser, P., Wassenaar, T.M., Dawes, H., Brooks, J., Ebmeier, K., Johansen-Berg, H. & Sexton, C. The effects of a 12-month physical activity intervention on hippocampal volume and cognitive function of older adults in the Retirement in Action (REACT) MRI sub-study. (*under review*)

- Beale, N., Eldridge, E., Delextrat, A., Esser, P., Bushnell, O., Curtis, E., Wassenaar, T. M., Wheatley, C., Dawes, H. & Johansen-Berg, H. Exploring activity levels in physical education lessons in the UK: A cross-sectional examination of activity types, fitness levels and attitudes toward PE. (*under review*)

- Salvan. P., Wassenaar, T. M., Wheatley, C., Beale, N., Cottaar, M., Papp, D., Bastiani, M., Fitzgibbon, S., Duff, E., Andersson, J., Winkler, A.M., Douaud, G., Nichols, T.E., Smith, S., Dawes, H. & Johansen-Berg, H. Multi-modal brain markers in early adolescence are linked with a physically active lifestyle. *Journal of Neuroscience*, (2020)

- Wheatley, C., Wassenaar, T., Salvan, P., Beale, N., Nichols, T., Dawes, H. & Johansen-berg, H. Associations between fitness, physical activity and mental health in a community sample of young British adolescents: baseline data from the Fit to Study trial. *BMJ Open Sport & Exercise Medicine* **6,** e000819 (2020)

- Sexton, C. E., Betts, J. F., Dennis, A., Doherty, A., Leeson, P., Holloway, C., Dall'Armellina, E., Winkler, A. M., Demnitz, N., Wassenaar, T., Dawes, H. & Johansen-Berg, H. The effects of an aerobic training intervention on cognition, grey matter volumes and white matter microstructure. *Physiology and Behavior* **223,** 112923 (2020)

- Wheatley, C., Beale, N., Wassenaar, T., Graham, M., Eldridge, E., Dawes, H. & Johansen-Berg, H. Fit to Study: Reflections on designing and implementing a large-scale randomized controlled trial in secondary schools. *Trends in Neuroscience and Education* **20,** 100134 (2020)

- Wassenaar, T. M., Yaffe, K., Van Der Werf, Y. D. & Sexton, C. E. Associations between modifiable risk factors and white matter of the aging brain: insights from diffusion tensor imaging studies. *Neurobiology of Aging* **80,** 56–70 (2019)

- Hobbiss, M. H., Massonnié, J., Tokuhama-Espinosa, T., Gittner, A., De Sousa Lemos, M. A., Tovazzi, A., Hindley, C., Baker, S., Sumeracki, M. A., Wassenaar, T. & Gous, I. "UNIFIED": Bridging the Researcher- Practitioner Divide in Mind, Brain, and Education. *Mind, Brain, and Education* **13,** 298–312 (2019)

- Esser, P., Weedon, B.D., Meaney, A., Delextrat, A., Mahmoud, W., Salvan, P., Wassenaar, T. M., Joshi, S., Liu, Y.C., Kemp, S., Ward, T., Johansen-Berg, H. & Dawes, H. Brain plasticity and motor skill competence development in young people with Developmental Coordination Disorder: Protocol for a controlled experimental paradigm. https://osf.io/w8gj6/ (2019)

Table of Contents

Abstract iii

Acknowledgements viii

Publications x

List of Figures xix

List of Tables xxiii

List of Abbreviations xxix

1 Introduction 1

1.1 The case for a physically active lifestyle during development 2

1.2 Development: a time of great change 5

1.2.1 Developing physical activity and cardiorespiratory fitness . . . 5

1.2.2 Developing brain and cognition 7

1.3 Physical activity, cardiorespiratory fitness, the brain and cognition . . . 12

1.3.1 Physical activity, cardiorespiratory fitness and cognition 14

1.3.2 Physical activity, cardiorespiratory fitness and the brain 21

1.3.3 Putative mechanisms 24

1.4 Aims of this thesis 26

2 Protocol for a cluster randomised trial of a physical activity intervention in young adolescents: Fit to Study 29

2.1 Introduction 30

2.1.1 Hypotheses 32

2.2 Methods 33

2.2.1 Design 33

2.2.2 Independent evaluation 35

2.2.3 Sample selection and recruitment 35

2.2.4 Sample characteristics 37

2.2.5 Intervention 38

2.2.6 Intervention fidelity 41

2.2.7 Blinding . . . 43
2.3 Outcome measures . . . 44
2.3.1 Academic achievement . . . 44
2.3.2 Cognitive and mental heath measures . . . 44
2.3.3 Physical activity and cardiorespiratory fitness . . . 52
2.3.4 Brain imaging sub-study . . . 53
2.4 Data analysis . . . 66
2.4.1 Sample size calculation . . . 67
2.4.2 Statistical analysis . . . 67
2.5 Discussion . . . 68

3 Methodological considerations when measuring and assessing physical activity, the brain and cognition 71
3.1 Measuring physical activity and cardiorespiratory fitness . . . 72
3.1.1 Measuring physical activity . . . 72
3.1.2 Measuring cardiorespiratory fitness . . . 76
3.2 Measuring the brain: Magnetic Resonance Imaging . . . 79
3.2.1 Assessing brain structure . . . 79
3.2.2 Assessing brain function . . . 87
3.2.3 Multimodal MRI and analysis . . . 88
3.2.4 Confounds . . . 89
3.3 Measuring cognition . . . 90
3.3.1 Cognitive tasks in PA research: need for consensus? . . . 90
3.3.2 Cognitive testing: methodological considerations . . . 91
3.4 Data analysis of cluster randomised controlled trials . . . 96
3.4.1 Cluster randomised trials . . . 97
3.4.2 Dealing with clustering in statistical models . . . 98
3.4.3 Missing data . . . 101

4 Associations between cardiorespiratory fitness, physical activity and cognitive function in a large sample of young British adolescents 107
4.1 Introduction . . . 108
4.1.1 Aims . . . 111
4.2 Methods . . . 111
4.2.1 Study design and participants . . . 111
4.2.2 Cardiorespiratory fitness and physical activity assessments . . . 112
4.2.3 Cognitive assessments . . . 112
4.2.4 Data analysis . . . 113
4.3 Results . . . 122
4.3.1 Sample . . . 122

4.3.2 Associations between PA or CRF and cognitive performance . . 125
4.3.3 Multivariate associations between a physically active lifestyle and cognitive performance 127
4.4 Discussion 132
4.4.1 Relationships between PA, CRF and cognitive performance . . 132
4.4.2 Strengths and limitations 135
4.5 Conclusion 136

5 Associations between cardiorespiratory fitness, physical activity, the anterior hippocampus and memory 139
5.1 Introduction 140
5.1.1 Aims 142
5.2 Methods 143
5.2.1 Design and participants 143
5.2.2 Cardiorespiratory fitness and physical activity assessments . . 143
5.2.3 Cognitive assessments 144
5.2.4 MRI data acquisition 144
5.2.5 MRI data analysis 145
5.2.6 Confounds 149
5.2.7 Data analysis 151
5.3 Results 159
5.3.1 Univariate analyses 160
5.3.2 Multivariate relationship between fitness, physical activity and brain MR markers 164
5.4 Discussion 168
5.4.1 Relationships of CRF or PA with anterior hippocampal volume and memory 168
5.4.2 Beyond hippocampal volume 170
5.4.3 Strength and limitations 171
5.5 Conclusion 172

6 The effect of the Fit to Study intervention on cardiorespiratory fitness, cognitive performance and mental health 173
6.1 Introduction 175
6.1.1 Aims 178
6.2 Methods 178
6.2.1 Design 179
6.2.2 Sample 179
6.2.3 Randomisation and blinding 179
6.2.4 Intervention 180

6.2.5 Teacher training . . . 181
6.2.6 Fidelity measures . . . 181
6.2.7 Outcome measures . . . 181
6.2.8 Data analysis . . . 183
6.3 Results . . . 186
6.3.1 Sample characteristics . . . 188
6.3.2 Effects on cardiorespiratory fitness . . . 189
6.3.3 Effects on cognition . . . 191
6.3.4 Effects on mental health . . . 192
6.3.5 Subgroup analysis: moderation by sex, socioeconomic status or baseline CRF . . . 193
6.3.6 Relationship of change in CRF with changes in cognition and mental health . . . 193
6.3.7 Intervention compliance and complier average causal effect . . . 193
6.3.8 Additional analyses . . . 194
6.3.9 Sensitivity analyses . . . 194
6.4 Discussion . . . 195
6.4.1 Findings in context . . . 195
6.4.2 Practical challenges of school-based PA interventions . . . 198
6.4.3 Strengths and limitations . . . 199
6.5 Conclusion . . . 200

7 The effect of the Fit to Study intervention on the anterior hippocampus, cardiorespiratory fitness and relational memory performance 201
7.1 Introduction . . . 202
7.1.1 Aims . . . 204
7.2 Methods . . . 204
7.2.1 Design and participants . . . 205
7.2.2 Blinding . . . 205
7.2.3 Intervention . . . 206
7.2.4 Intervention fidelity . . . 206
7.2.5 Outcomes . . . 207
7.2.6 Data analysis . . . 211
7.3 Results . . . 217
7.3.1 Sample characteristics . . . 219
7.3.2 Intervention fidelity . . . 219
7.3.3 The effect on cardiorespiratory fitness . . . 220
7.3.4 The effect on the anterior hippocampus . . . 222
7.3.5 The effect on relational memory . . . 225

7.3.6 Changes in CRF with changes in hippocampal MRI markers . . 225
7.3.7 Supplementary analyses 226
7.4 Discussion 227
7.4.1 Anterior hippocampus, CRF and cognitive function 227
7.4.2 Findings in context 230
7.4.3 Strengths and limitations 231
7.5 Conclusion 232

8 Conclusions and future directions **233**
8.1 Aims 234
8.2 Summary of findings 235
8.3 Limitations and future directions 236
8.3.1 Physical activity measures 236
8.3.2 Hippocampal measures, hippocampal plasticity and beyond . . 237
8.3.3 Cognitive measures 238
8.3.4 Challenges of large-scale RCTs 239
8.3.5 Designing physical activity interventions 241
8.3.6 Moderators and mediators 242
8.3.7 More than physical activity 243
8.4 General conclusion 244

Appendices

A Supplementary materials to chapter 1: *Introduction* **247**
A.1 Terminology 248
A.2 Methods systematic review of reviews 252
A.2.1 Search strategy and selection criteria 253
A.2.2 Review selection and data extraction 254
A.2.3 Quality assessment 254

B Supplementary materials to chapter 2: *Protocol for a cluster randomised trial of a physical activity intervention in young adolescents: Fit to Study* **257**
B.1 Sample 258
B.1.1 Drop-out post randomisation main trial schools 258
B.1.2 Representativeness brain imaging sub-sample 261
B.1.3 Recruitment brain imaging cohort 2 268
B.2 Measures 269
B.2.1 Main trial 269
B.2.2 Brain imaging sub-study 279
B.3 Sample size 289

C Supplementary materials to chapter 3: *Methodological considerations when measuring and assessing physical activity, the brain and cognition* 291
C.1 Physical activity: alternative objective methods for assessing physical activity . . . 292
C.2 Brain imaging: confounds . . . 292
C.2.1 Scanner-related confounds . . . 292
C.2.2 Subject-related confounds . . . 293
C.3 Cognitive testing: psychometrics . . . 296
C.4 Data analysis . . . 300
C.4.1 Cluster randomised trials . . . 300
C.4.2 Missing data . . . 306

D Supplementary materials to chapter 4: *Associations between cardiorespiratory fitness, physical activity and cognitive function in a large sample of young British adolescents* 319
D.1 Methods . . . 320
D.1.1 Outcome measures . . . 320
D.1.2 Specification curve analysis . . . 322
D.1.3 Missing data . . . 324
D.1.4 Software . . . 328
D.2 Results . . . 328
D.2.1 Data cleaning . . . 328
D.2.2 Descriptives . . . 329
D.2.3 Univariate analysis . . . 334
D.2.4 Specification curve analysis . . . 338
D.2.5 Canonical correlation analysis . . . 360

E Supplementary materials to chapter 5: *Associations between cardiorespiratory fitness, physical activity, the anterior hippocampus and memory* 365
E.1 Methods . . . 366
E.1.1 Cognitive assessments . . . 366
E.1.2 MRI acquisition . . . 367
E.1.3 Hippocampal segmentation . . . 369
E.1.4 Time-of-day effects . . . 375
E.1.5 BLUS residuals . . . 378
E.1.6 Missing data . . . 379
E.1.7 Software . . . 382
E.2 Results . . . 382
E.2.1 Primary analyses . . . 382

E.2.2 Entire hippocampus . . . 386
E.2.3 Multivariate relationship between fitness, physical activity and brain MR markers . . . 391

F Supplementary materials to chapter 6: *The effect of the Fit to Study intervention on cardiorespiratory fitness, cognitive performance and mental health* 401

F.1 Supplementary methods . . . 403
F.1.1 Actigraphy processing . . . 403
F.1.2 Minimum detectable effect size . . . 403
F.1.3 Data cleaning . . . 404
F.1.4 Complier-average causal effects analysis . . . 407
F.1.5 Software . . . 407
F.2 Missing data . . . 408
F.2.1 Missing data patterns . . . 410
F.2.2 Missing data approach . . . 410
F.3 Drop-out of schools at baseline assessments . . . 414
F.3.1 School drop-out pre-baseline . . . 414
F.4 Data cleaning . . . 417
F.5 Baseline PA and PE characteristics . . . 420
F.5.1 Baseline PA and PE characteristics . . . 420
F.5.2 PE lessons across schools . . . 421
F.6 Primary outcomes . . . 422
F.6.1 Fully adjusted models: multiply imputed data . . . 422
F.6.2 Sensitivity analysis: complete-cases . . . 423
F.7 Subgroup analysis . . . 431
F.8 Change - change analysis . . . 437
F.8.1 Primary analysis: multiply imputed data . . . 437
F.8.2 Sensitivity analysis: complete-cases . . . 437
F.9 Intervention fidelity and CACE analysis . . . 442
F.9.1 Intervention fidelity . . . 442
F.9.2 CACE analysis . . . 444
F.10 Additional analysis . . . 445
F.10.1 Multiply imputed data . . . 445
F.10.2 Sensitivity analysis . . . 446

G Supplementary materials to chapter 7: *The effect of the Fit to Study intervention on the anterior hippocampus, cardiorespiratory fitness and relational memory performance* **451**

G.1 Methods . . . 452
G.1.1 Missing data . . . 452
G.1.2 Supplementary analysis . . . 454
G.1.3 Software . . . 455
G.2 Results . . . 455
G.2.1 Demographics . . . 455
G.2.2 Fidelity . . . 456
G.2.3 Primary analysis . . . 457
G.2.4 Sensitivity analyses . . . 460
G.2.5 Supplementary analysis . . . 472

References **475**

List of Figures

1.1 Number of published articles on physical activity and cognition (1970–2020) . . . 13
1.2 Countries where PA interventions have been conducted to assess the effect on cognitive outcomes . . . 20
1.3 Countries where PA interventions have been conducted to assess the effect on brain outcomes . . . 24

2.1 SPIRIT figure Fit to Study trial . . . 34
2.2 CONSORT flow diagram Fit to Study trial . . . 36
2.3 Reaction-time task paradigm . . . 45
2.4 Relational memory paradigm . . . 46
2.5 Colour-shape switch task paradigm . . . 48
2.6 Two-back working memory paradigm . . . 49
2.7 Modified Flanker task paradigm . . . 50
2.8 CONSORT flow diagram Fit to Study brain imaging sub-study . . . 54
2.9 SPIRIT figure Fit to Study brain imaging sub-study . . . 56
2.10 MRI protocol brain imaging sub-study . . . 60
2.11 Object-location task paradigm . . . 63
2.12 Illustration of Godfrey protocol and $\dot{V}O_2$ dynamics . . . 64

3.1 Diffusion tensor imaging: the tensor model . . . 84
3.2 Illustration of clustering in Fit to Study . . . 97
3.3 Growth in popularity of multiple imputation . . . 102
3.4 Multiple imputation principle . . . 103

4.1 Completion of cognitive assessments at baseline . . . 121
4.2 Canonical correlation and cross-loadings of the first mode of covariation 129
4.3 Canonical correlation and cross-loadings of the second mode of covariation . . . 131

5.1 MRI protocol at baseline . . . 144
5.2 Available MRI scans at baseline . . . 149
5.3 Missing data pattern brain imaging sub-study at baseline . . . 158

5.4 Relationship between CRF or MVPA and left and right anterior hippocampal volume . . . 162
5.5 Canonical correlation and cross-loadings of the first mode of covariation (primary analysis) . . . 166
5.6 Canonical correlation and cross-loadings of the first mode of covariation (secondary analysis) . . . 167

6.1 CONSORT flow diagram of Fit to Study cluster randomised controlled trial . . . 187
6.2 Path diagram showing the effects of the VPA intervention on an executive function (EF) latent variable . . . 192

7.1 Experimental design of the brain imaging sub-study . . . 205
7.2 MRI protocol of the Fit to Study brain imaging sub-study . . . 208
7.3 Completion of MRI scans across timepoints . . . 210
7.4 Missing data pattern brain imaging sub-study . . . 214
7.5 CONSORT flow diagram brain imaging sub-study . . . 218
7.6 Distributions of change scores (post-pre) of MRI markers in the anterior hippocampus . . . 223

B.1 Comparison of fitness levels main trial and sub-sample . . . 268

C.1 Illustration of clustering in Fit to Study . . . 300
C.2 The effect of clustering on statistical power . . . 303
C.3 The effect of clustering on type I error rates . . . 305
C.4 Multiple imputation principle . . . 314

D.1 Validation of 20MSR values as an estimate of cardiorespiratory fitness . 322
D.2 Results of the specification curve analysis for relational memory outcomes (data cleaning) . . . 340
D.3 Results of the specification curve analysis for processing speed (simple reaction time task, data cleaning) . . . 341
D.4 Results of the specification curve analysis for inhibitory control (Flanker task, data cleaning) . . . 342
D.5 Results of the specification curve analysis for cognitive flexibility (colour-shape task-switching task, data cleaning) . . . 343
D.6 Results of the specification curve analysis for working memory (two-back task, data cleaning) . . . 344
D.7 Results of the specification curve for relational memory outcomes . . . 350
D.8 Results of the specification curve for processing speed (simple reaction time task) . . . 351

D.9 Results of the specification curve for inhibitory control (Flanker task) . 352
D.10 Results of the specification curve for cognitive flexibility (colour-shape switching task) 355
D.11 Results of the specification curve for working memory (two-back task) 358
D.12 Cross-validated physical and cognitive variates 361
D.13 Canonical correlation and cross-loadings of the first mode of covariation (secondary analysis) 363
D.14 Cross-validated physical and cognitive variates (secondary analysis) . . 364

E.1 Qualitative comparison of automated hippocampal segmentation algorithms 369
E.2 Quantitative comparison of automated hippocampal segmentation algorithms 370
E.3 Illustration of manual edits to hippocampal segmentations 372
E.4 Demarcation of the anterior hippocampus 374
E.5 Variability in the location of the uncal apex 375
E.6 Time-of-day relationships with hippocampal volumes 377
E.7 Cross-validated canonical correlation between the first physical variate and brain variate 392
E.8 Cross-validated canonical correlation between the first physical health variate and brain variate 393
E.9 Canonical correlation and cross-loadings of the first mode of covariation (complete-case) 394
E.10 Cross-validated canonical correlation between the first physical variate and brain variate (complete-case) 396
E.11 Canonical correlation and cross-loadings of the first mode of covariation (complete-case) 397
E.12 Cross-validated canonical correlation between the first physical health variate and brain variate (complete-case). 399

F.1 Missing data patterns Fit to Study 411
F.2 Distributions of outcome variables (complete-case data) 426

G.1 Distribution of uncorrected p-values of partial tests by MRI marker . . 459
G.2 Effect of the intervention on cardiorespiratory fitness (complete-cases) 463
G.3 Effect of the intervention on left and right anterior hippocampal volume (complete-cases) 464
G.4 Effect of the intervention on relational memory performance (complete-cases) 465
G.5 Distributions of difference scores (post-pre) for SVD imputed MRI markers in the anterior hippocampus, by treatment group and hemisphere . 467

G.6 Distributions of difference scores (post-pre) for MRI markers in the anterior hippocampus, by treatment group and hemisphere (complete-cases) 470

List of Tables

1.1 Cognitive outcomes: findings from systematic reviews and meta-analyses 16
1.2 Brain outcomes: findings from systematic reviews and meta-analyses . 22

2.1 School and participant characteristics throughout the trial 39
2.2 Baseline characteristics Fit to Study brain imaging sub-study 57
2.3 Sampling frequency of participants by schools 58

4.1 Physically active and cognitive phenotypes 117
4.2 Characteristics of schools and pupils at baseline 124
4.3 Standardized regression coefficients of multiply imputed data 125

5.1 Baseline characteristics brain imaging sample 160
5.2 Univariate regression analyses between CRF or PA and anterior hippocampal volumes (standardized) 161
5.3 Univariate regression analyses assessing the relationship between MVPA or CRF and memory outcomes (standardized) 163
5.4 Univariate regression analyses assessing relationships between anterior hippocampal volumes and memory outcomes (standardized) 164

6.1 School- and pupil level baseline characteristics 188
6.2 Mean scores for outcome measures by group and (un)standardized mean differences 190

7.1 Baseline characteristics brain imaging sample by treatment group . . . 219
7.2 The effect of the intervention on CRF, anterior hippocampal volumes and memory performance 221
7.3 Subgroup analysis by sex, socioeconomic status and baseline CRF . . . 225
7.4 Relationship between a change in CRF with a change in anterior hippocampal volumes 226

B.1 School and participant characteristics throughout the trial 259
B.2 School level drop-out model 260
B.3 Pupil level drop-out model 261
B.4 School level demographics for the main trial and sub-sample 262

B.5 School level comparison between main trial sample and sub-sample . . 263
B.6 Pupil level demographics for the main trial and sub-sample 264
B.7 Pupil level comparison between main trial sample and sub-sample . . . 265
B.8 Pupil level comparison between main trial sample and sub-sample (single level) 265
B.9 Pupil level characteristics of pupils in brain imaging sub-study schools 265
B.10 Pupil level characteristics predicting inclusion in the brain imaging sub-study 266
B.11 Pupil level characteristics predicting inclusion in the brain imaging sub-study (single level) 266
B.12 Cardiorespiratory fitness and self-reported physical activity of main trial and sub-study participants 267
B.13 Multilevel logistic regression (logit link) of pupil level fitness and physical activity on inclusion in the brain imaging sub-study 267
B.14 Cognitive assessments completed by sub-study samples 282
B.15 Health behaviour in school-aged children questionnaire 284

C.1 Reliability statistics of cognitive assessments 299

D.1 Physical activity questionnaire items 320
D.2 Missing data (proportion) of task related variables 325
D.3 Variables used in the imputation models for each of the primary univariate analysis models 326
D.4 Number (proportion) of participants that completed each assessment pre and post data cleaning 328
D.5 Models testing whether baseline characteristics predict exclusion during cleaning 329
D.6 Total-, cognitive- and complete-case sample demographics 330
D.7 Mean participant scores on all predictor and outcome variables, split by eFSM status and sex 332
D.8 Unstandardized regression coefficients of multiple imputed data 334
D.9 Standardized regression coefficients of complete-case models 335
D.10 Unstandardized regression coefficients of complete-case models 336
D.11 Standardized regression coefficients of fully-adjusted complete-case models 336
D.12 Unstandardized regression coefficients of fully-adjusted complete-case models 337
D.13 The robustness of relationships between PA or CRF and cognitive outcomes to variations in cleaning of cognitive assessments 339

D.14 Specifications (analytical choices) that are tested for each cognitive assessment . . . 347
D.15 Robustness of relationships between PA or CRF and cognitive outcomes to variations in predictors, outcomes, nuisance and cleaning of cognitive assessments . . . 349
D.16 Results of the specification curve analysis of the Flanker task by outcome measure . . . 354
D.17 Results of the specification curve analysis of the task-switching task by outcome measure . . . 357
D.18 Results of the specification curve analysis of the two-back task by outcome measure . . . 359
D.19 Canonical correlations (ρ) and p-values for each mode . . . 360
D.20 Cross-loadings mode 1 (primary analysis) . . . 360
D.21 Cross-loadings mode 2 (primary analysis) . . . 361
D.22 Canonical correlations (ρ) and p-values for each mode (secondary analysis) . . . 362
D.23 Cross-loadings mode 1 (secondary analysis) . . . 362

E.1 MRI scan parameters . . . 368
E.2 Time-of-day relationships with hippocampal volumes . . . 376
E.3 Time-of-day relationships with hippocampal MRI markers of tissue (micro)structure . . . 378
E.4 Overview of imputed variables . . . 380
E.5 Univariate regression analyses between CRF or PA and anterior hippocampal volumes (unstandardized) . . . 382
E.6 Univariate regression analyses assessing the relationship between MVPA or CRF and memory outcomes (unstandardized) . . . 383
E.7 Univariate regression analyses assessing relationships between anterior hippocampal volumes and memory outcomes (unstandardized) . . . 383
E.8 Univariate regression analyses between CRF or PA and anterior hippocampal volumes (standardized, imputed) . . . 384
E.9 Univariate regression analyses between CRF or PA and anterior hippocampal volumes (unstandardized, imputed) . . . 384
E.10 Univariate regression analyses assessing the relationship between MVPA or CRF and memory outcomes (standardized, imputed) . . . 385
E.11 Univariate regression analyses assessing the relationship between MVPA or CRF and memory outcomes (unstandardized, imputed) . . . 385
E.12 Univariate regression analyses assessing relationships between anterior hippocampal volumes and memory outcomes (standardized, imputed) . . . 386

E.13 Univariate regression analyses assessing relationships between anterior hippocampal volumes and memory outcomes (unstandardized, imputed) 386
E.14 Univariate regression analyses involving the entire hippocampus (standardized) . . . 387
E.15 Univariate regression analyses involving the entire hippocampus (unstandardized) . . . 387
E.16 Univariate regression analyses involving the entire hippocampus (standardized) . . . 388
E.17 Univariate regression analyses involving the entire hippocampus (unstandardized) . . . 388
E.18 Univariate regression analyses involving the entire hippocampus (standardized, imputed) . . . 389
E.19 Univariate regression analyses involving the entire hippocampus (unstandardized, imputed) . . . 389
E.20 Univariate regression analyses involving the entire hippocampus (standardized, imputed) . . . 390
E.21 Univariate regression analyses involving the entire hippocampus (unstandardized, imputed) . . . 390
E.22 Cross-loadings mode 1 (primary analysis, imputed data) . . . 391
E.23 Cross-loadings mode 1 (secondary analysis, imputed data) . . . 392
E.24 Cross-loadings mode 1 (primary analysis, complete-case) . . . 395
E.25 Cross-loadings mode 1 (secondary analysis, complete-case) . . . 398

F.1 Missing data (proportion) of school-level variables . . . 408
F.2 Missing data (proportion) of pupil-level variables . . . 409
F.3 Variables included in multiple imputations . . . 412
F.4 Excluded variables from imputation models . . . 414
F.5 School level and pupil level characteristics of drop-outs and those that remained part of the trial . . . 416
F.6 School level predictors of school drop-out pre-baseline assessment: analysis of deviance table (type II Wald χ^2 test) . . . 417
F.7 Pupil level predictors of school drop-out pre-baseline assessment (multilevel logistic regression with logit link function) . . . 417
F.8 Total number of participants who completed each assessment, pre and post cleaning, per group and timepoint. . . . 418
F.9 Models testing whether baseline characteristics predict removal during cleaning . . . 419
F.10 Physical activity characteristics of schools and pupils at baseline (available cases only) . . . 421
F.11 PE lesson characteristics across intervention and control schools . . . 422

F.12 The effect of the intervention on secondary outcomes (fully adjusted models) . . . 423
F.13 Effect of the intervention on secondary outcomes (complete-cases) . . . 424
F.14 The effect of the intervention on secondary outcomes (fully adjusted models) . . . 431
F.15 Subgroup analysis: moderation by sex, eFSM or baseline CRF (multiply imputed data) . . . 434
F.16 Subgroup analysis: moderation by sex, eFSM or baseline CRF (complete-cases) . . . 436
F.17 Models examining whether a change in CRF is related to a change in cognitive outcomes (multiply imputed data) . . . 438
F.18 Models examining whether a change in CRF is related to a change in mental health outcomes (multiply imputed data) . . . 439
F.19 Models examining whether a change in CRF is related to a change in cognitive outcomes (complete-case data) . . . 440
F.20 Models examining whether a change in CRF is related to a change in mental health outcomes (complete-case data) . . . 441
F.21 School level fidelity measures . . . 443
F.22 CACE estimates and confidence intervals (multiply imputed data) . . . 444
F.23 CACE estimates and confidence intervals (complete-case data) . . . 445
F.24 Models examining baseline differences in PE enjoyment and attitudes towards PA (multiply imputed data) . . . 446
F.25 Models examining the effect of the intervention on PE enjoyment and attitudes towards PA (multiply imputed data) . . . 446
F.26 Models examining baseline differences in PE enjoyment and attitudes towards PA (complete-case data) . . . 447
F.27 Models examining the effect of the intervention on PE enjoyment and attitudes towards PA (complete-case data) . . . 448

G.1 Variables used in imputation models . . . 452
G.2 Variables used in the imputation model for multivariate analysis . . . 453
G.3 School-level baseline characteristics brain imaging sample by treatment group . . . 456
G.4 School level and pupil level fidelity data . . . 457
G.5 Means of multiple imputed MRI markers by treatment group . . . 458
G.6 Number and percent of significant partial tests by MRI marker . . . 458
G.7 Associations between changes in $\dot{\mathbf{V}}O_2$peak and changes in R_1 . . . 459
G.8 Associations between changes in R_1 and changes in relational memory performance . . . 460
G.9 Sensitivity analyses of primary univariate outcomes . . . 461

G.10 Means of single imputed MRI markers by treatment group 468
G.11 Means of MRI markers by treatment group (complete-cases) 471
G.12 The effect of the intervention on hippocampal volumes 472
G.13 The effect of the intervention on volumes of hippocampal subfields . . 473

List of Abbreviations

ADHD . . . . . Attention deficit hyperactivity disorder

ANOVA . . . . Analysis of variance

ANCOVA . . . Analysis of covariance

ASL . . . . . Arterial spin labelling

BASIL . . . . . Bayesian Inference for Arterial Spin Labelling

BMI . . . . . Body mass index

BOCF . . . . . Baseline observation carried forward

CA . . . . . Cornu Ammonis

CACE . . . . . Complier-average causal effect

CASQ . . . . . Cleveland Adolescent Sleepiness Questionnaire

CCA . . . . . Canonical correlation analysis

CFA . . . . . Confirmatory factor analysis

CFI . . . . . Comparative fit index

CI . . . . . Confidence interval

CONSORT . . . Consolidated Standard for Reporting Trials

cRCT . . . . . Cluster randomised controlled trial

CRF . . . . . Cardiorespiratory fitness

CV . . . . . Cross-validation

DG . . . . . Dentate gyrus

DWI . . . . . Diffusion-weighted imaging

EEF . . . . . Education Endowment Foundation

EEG . . . . . Electroencephalography

EF . . . . . Executive function

eFSM . . . . . Eligible for free school meals

EOI . . . . . Expression of interest

EPI . . . . . Echo-planar imaging

FA . . . . . Fractional anisotropy

FAST . . . . . FMRIB's Automated Segmentation Tool

FCS . . . . . Fully conditional specification

FIML . . . . . Full information maximum likelihood

FIRST . . . . . FMRIB's Integrated Registration and Segmentation Tool

FLIRT . . . . . FMRIB's Linear Image Registration Tool

fMRI . . . . . Functional magnetic resonance imaging

FMRIB . . . . . Oxford Centre for Functional MRI of the Brain

FNIRT . . . . . FMRIB's Non-linear Image Registration Tool

FOV . . . . . Field of view

FS . . . . . FreeSurfer

FSL . . . . . FMRIB's Software Library

FWE . . . . . Family-wise error

GEE . . . . . Generalised estimating equation

GLM . . . . . General linear model

GLMM . . . . . Generalized linear mixed model

GDC . . . . . Gradient distortion correction

HBSC . . . . . Health Behaviour in School aged Children

HCP . . . . . Human Connectome Project

HIT . . . . . High intensity training

HR . . . . . Heart rate

ICC . . . . . Intraclass correlation coefficient

ICVF . . . . . Intra-cellular volume fraction

IE . . . . . Inverse efficiency

ISOVF . . . . . Isotropic volume fraction

ITT . . . . . Intention-to-treat

IV . . . . . Instrumental variable

IQR . . . . . Interquartile range

JM . . . . . Joint modelling

LMM . . . . . Linear mixed model

LOCF . . . . . Last observation carried forward

MAR . . . . . Missing at random

MANOVA . . . Multivariate analysis of variance

MCAR . . . . . Missing completely at random

MD . . . . . Mean diffusivity

MDES . . . . . Minimum detectable effect size

MICE . . . . . Multiple imputation using chained equations

MIST . . . . . Multimodal Image Segmentation Tool

MK . . . . . Mean kurtosis

MNAR . . . . . Missing not at random

MNI . . . . . Montreal Neurological Institute

MP-RAGE . . . Magnetization-prepared rapid gradient-echo

MRA . . . . . Magnetic resonance angiography

MRI . . . . . Magnetic resonance imaging

MSL . . . . . Mean string length

MVPA . . . . . Moderate-to-vigorous physical activity

NDI . . . . . Neurite density index

NFER . . . . . National Foundation for Educational Research

NODDI . . . . . Neurite Orientation Dispersion and Density Imaging

NPC . . . . . Non-parametric combination

NPD . . . . . National pupil database

ODI . . . . . Orientation dispersion index

OLT . . . . . Object-location task

PA . . . . . Physical activity

PAGAC . . . . . Physical Activity Guidelines Advisory Committee

PAR-Q . . . . . Physical Activity Readiness Questionnaire

PCA . . . . . Principal component analysis

pCASL . . . . . Pseudo-continuous arterial spin labelling

PDSS . . . . . Paediatric Daytime Sleepiness Scale

PDw . . . . . Proton density-weighted

PE . . . . . Physical education

PLD . . . . . Post labelling delay

POMS . . . . . Profile of Mood States

PRISMA . . . . Preferred Reporting Items for Systematic Reviews and Meta-Analysis

P-SDQ . . . . . Physical Self-Description Questionnaire

PTM . . . . . Progress test in maths

RCT . . . . . Randomised control trial

REML . . . . . Restricted maximum likelihood

RER . . . . . Respiratory exchange ratio

RF . . . . . Radio frequency

RMSEA . . . . . Root mean square error of approximation

ROI . . . . . Region of interest

RPE . . . . . Rating of perceived exertion

rpm . . . . . Revolutions per minute

rs-fMRI . . . . resting-state functional magnetic resonance imaging

RT . . . . . Reaction time

SCA . . . . . Specification curve analysis

SCI . . . . . Sleep Condition Indicator

SD . . . . . Standard deviation

SDQ . . . . . Strength and Difficulties Questionnaire

SEM . . . . . Structural equation model

SES . . . . . Socioeconomic status

SMD . . . . . Standardized mean difference

SPIRIT . . . . . Standard Protocol Items: Recommendations for Interventional Trials

SPM . . . . . Statistical Parametric Mapping

SVD . . . . . Singular value decomposition

T_1w . . . . . T_1-weighted

TE . . . . . Echo time

TLI . . . . . Tucker-Lewis Index

TR . . . . . Repetition time

TSLS . . . . . Two-stage least squares

UK . . . . . United Kingdom

UPN . . . . . Unique pupil number

US . . . . . United States

VBM . . . . . Voxel-based morphometry

VBQ . . . . . Voxel-based quantification

$\dot{\mathbf{V}}\mathbf{O_2}\mathbf{peak}$ . . . . Peak oxygen consumption

$\dot{\mathbf{V}}\mathbf{O_2}\mathbf{max}$ . . . . Maximal oxygen consumption

VPA . . . . . Vigorous physical activity

WHO . . . . . World Health Organization

WIN . . . . . Wellcome Centre for Integrative Neuroimaging

This page is intentionally left blank.

1 Introduction

Contents

1.1 The case for a physically active lifestyle during development . 2
1.2 Development: a time of great change 5
1.2.1 Developing physical activity and cardiorespiratory fitness . . 5
1.2.2 Developing brain and cognition 7
1.3 Physical activity, cardiorespiratory fitness, the brain and cognition 12
1.3.1 Physical activity, cardiorespiratory fitness and cognition . . . 14
1.3.2 Physical activity, cardiorespiratory fitness and the brain . . . 21
1.3.3 Putative mechanisms 24
1.4 Aims of this thesis 26

Governmental, scientific and public health communities have become increasingly interested in the potential of physical activity to promote neurocognitive health across the lifespan. While research in adults has advanced rapidly in the past 20 years, much less is known about the effects of a physically active lifestyle on cognitive and brain health in young people, especially adolescents. Adolescence is a period of dynamic social, psychological and neurocognitive development and provides the foundation for future health, yet is also characterised by high prevalence of physical inactivity. Understanding relationships between a physically active lifestyle and neurocognitive health in this age group is therefore important. This thesis aims to characterise such relationships in a large cohort of young British adolescents. In this chapter, a brief introduction into research linking physical activity and cardiorespiratory fitness to cognitive and brain health in young people is provided, and core aims of the thesis are outlined.



The core of this chapter is adapted from:

- Wassenaar, T. M., Williamson, W., Johansen-Berg, H., Dawes, H., Roberts, N., Foster, C. & Sexton, C. E. A critical evaluation of systematic reviews assessing the effect of chronic physical activity on academic achievement, cognition and the brain in children and adolescents: a systematic review. *International Journal of Behavioral Nutrition & Physical Activity* **17,** 1–18 (2020)
- Wheatley, C. M., Wassenaar, T. & Johansen-Berg, H. *Educational Neuroscience. Development Across the Life Span* (eds Thomas, M. S. C., Mareschal, D. & Dumontheil, I.) 359–385 (Taylor & Francis, New York, 2020)

1.1 The case for a physically active lifestyle during development

The importance of a physically active lifestyle for general health and well-being is widely acknowledged and hardly a new idea. Ancient philosophers commented on the health benefits of physical activity (PA) before the start of the common era (Blair & Powell, 2014; Tipton, 2015), and the proverbial "*a healthy mind in a healthy body*" ("*Mens sana in corpore sano*") from the Roman poet Juvenal is still being used today.

Scientific research in the field of PA began in the early 20th century, but it has only been since mid 20th century that research started to provide evidence for the claimed PA-related benefits (Blair & Powell, 2014; Varela *et al.,* 2018).

The effects of physical *activity*, or alternatively, physical *in*activity, on general health are now well-established (Fiuza-Luces *et al.,* 2013; Janssen & LeBlanc, 2010; Ortega *et al.,* 2008a; Pedersen & Saltin, 2015; Warburton *et al.,* 2006). Physical *in*activity is an important risk factor for various chronic diseases (e.g. cardiovascular disease, depression), obesity, and early deaths (Guthold *et al.,* 2018; Kohl *et al.,* 2012; Lee *et al.,* 2012; McMahon *et al.,* 2017), placing a high economic burden on society (Ding *et al.,* 2016). Conversely, higher levels of PA have been associated with lower risk of mortality, beneficial mental, cardiovascular, and musculo-skeletal health outcomes (Bradley *et al.,* 2019; Dale *et al.,* 2019; Ekelund *et al.,* 2016; Lear *et al.,* 2017) and a stronger immune system (Campbell & Turner, 2018; Nieman & Wentz, 2019; Simpson *et al.,* 2020). Moreover, PA may modify an individual's cardiorespiratory fitness (CRF), a physiological trait (Lang & Little, 2018) that has shown to be an important marker of current and future health (Ortega *et al.,* 2008a; Ruiz *et al.,* 2009a), independent of PA (Ekelund *et al.,* 2007). What's more, PA is easily modifiable and PA interventions are highly scalable and cost-effective. For these reasons, targeting PA has been considered a suitable strategy for global health improvement (Bauer *et al.,* 2014; Biddle & Batterham, 2015; Sallis *et al.,* 2016), and PA has received great interest from scientific and public health communities demonstrated by the 2012 and 2016 Lancet series on PA (2012 series, and 2016 series), the United States (US) 2018 Physical Activity Guidelines Advisory Committee Scientific Report (2018 Physical Activity Guidelines Advisory Committee, 2018) and the World Health Organization (WHO) Global Action Plan for Physical Activity (Bull *et al.,* 2020; World Health Organization, 2018).

Of particular interest to scientists and public health organizations is the potential of PA to positively affect the brain and cognitive performance (Erickson *et al.,* 2019). Improvements in brain and cognitive health may provide long term health benefits,

including a lower risk of mortality (Cole *et al.,* 2018; Hayat *et al.,* 2018; Sedaghat *et al.,* 2016), improved mental health (Global Council on Brain Health, 2018; Hatch *et al.,* 2007) and slower age-related cognitive decline (Elliott *et al.,* 2019; Stern, 2012; Whalley *et al.,* 2006), and may be important for success in life, including higher academic (Bull *et al.,* 2008; Bull & Scerif, 2001; Rohde & Thompson, 2007; Shaw *et al.,* 2006) and career attainment (Dreher & Bretz, 1990). Studies conducted over the past 50 years have shown that higher levels of PA and CRF tend to be positively related with measures of the brain and cognition (Hillman *et al.,* 2008; Ludyga *et al.,* 2020; Thomas *et al.,* 2012; Voelcker-Rehage & Niemann, 2013). The hippocampus, a region important for learning and memory, has shown to be particularly sensitive to the effects of PA (Erickson *et al.,* 2011; Firth *et al.,* 2018; Voss *et al.,* 2013). The majority of evidence to date, however, is based on research conducted in animals (e.g. Cotman & Berchtold, 2002; Cotman *et al.,* 2007; Vivar & Van Praag, 2017; Voss *et al.,* 2013) and human older adults (e.g. Colcombe & Kramer, 2003; Erickson *et al.,* 2014; Falck *et al.,* 2019; Firth *et al.,* 2018; Hillman *et al.,* 2008; Kramer *et al.,* 2018; Northey *et al.,* 2018; Sexton *et al.,* 2016; Young *et al.,* 2015). Far less research has considered the relationship between PA or CRF and neurocognitive health during a crucial period of development: adolescence (Blakemore, 2019; Patton *et al.,* 2018; Sawyer *et al.,* 2012; Sawyer *et al.,* 2018).

Adolescence, the period between 10 and 24 years of age (Sawyer *et al.,* 2018), has been described as a (neurobiological) sensitive period of development characterised by enhanced plasticity, during which experience or environmental influences strongly affect the brain and may have long-lasting effects on neurocognitive health (Dow-Edwards *et al.,* 2019; Fuhrmann *et al.,* 2015; Larsen & Luna, 2018; Walhovd *et al.,* 2014b). This period in life therefore presents an opportunity for interventions to positively impact on short- and long-term health outcomes. As 80% of adolescents do not meet PA recommendations of 60 minutes of moderate-to-vigorous PA (MVPA) per day (Guthold *et al.,* 2020) and opportunities to engage in PA during school have reduced significantly (Fedewa & Ahn, 2011), understanding whether PA and CRF are related with measures of cognitive and brain health in this age group has never been more important.

1.2 Development: a time of great change

When studying relationships between PA, CRF, the brain and cognition during development, it is important to realise that these characteristics are not static, but subject to change (Armstrong *et al.,* 2011; Blakemore, 2019; Dow-Edwards *et al.,* 2019; Fuhrmann *et al.,* 2015; Patton *et al.,* 2018; Patton & Viner, 2007). That is, childhood and adolescence are marked by ongoing psychological, biological and behavioural development; complex processes that are under the influence of experience (i.e. the environment) and one's genetic make-up. Here, we briefly provide some insight into this dynamically changing landscape, with a particular focus on those characteristics that are central to this thesis, before reviewing *whether* and *how* PA and CRF may be linked with measures of neurocognitive health during development.

1.2.1 Developing physical activity and cardiorespiratory fitness

Physical activity is commonly defined as "*any bodily movement produced by skeletal muscles that results in energy expenditure*" (Caspersen *et al.,* 1985). This term is not specific to any type of PA and includes all intensities and domains (a detailed breakdown of PA is provided in Appendix A.1). Exercise is a subset of PA that is planned, structured and often repetitive, and is designed to improve or maintain physical fitness, performance or health (2018 Physical Activity Guidelines Advisory Committee, 2018; Caspersen *et al.,* 1985). The global PA guidelines recommend that children and adolescents engage in 60 minutes of MVPA per day (Bull *et al.,* 2020). In 2016, about 81% of adolescents worldwide did not meet these guidelines, with a higher proportion of physically inactive girls (84.7%) than boys (77.6%; Guthold *et al.,* 2020). Compared to 2001, the prevalence of insufficient PA decreased among boys (80.1% in 2001), but not girls (85.1%; Guthold *et al.,* 2020). Beyond temporal trends, there is strong evidence that PA declines with increasing age (Sallis & Saelens, 2000). Both the total volume of PA and average minutes spent in MVPA decline from childhood into adolescence (Dumith *et al.,* 2011; Farooq *et al.,* 2020; Farooq *et al.,* 2018; Nader *et al.,* 2008; Schwarzfischer *et al.,* 2019) as individuals become more sedentary (Corder *et al.,* 2016; Corder *et al.,* 2015; Harding *et al.,* 2015). A recent systematic review (Farooq *et al.,* 2020) reported a

significant annual decline in MVPA of -3.4% in boys and -5.3% in girls, with a particularly strong decline at age 9 (-7.8% in boys and -10.2% in girls).

Vigorous PA (VPA), which may be a more important predictor of health outcomes compared with lower intensity PA (Carson *et al.,* 2014), declines proportionally more than MPA (Corder *et al.,* 2016; Corder *et al.,* 2015), though changes in VPA vary by various factors, including sex, age, and baseline VPA level (Beltran-Valls *et al.,* 2019). A longitudinal study involving British children showed that average daily minutes of VPA decreased from 6.1 (standard deviation [SD] = 6.5) at age 9–10 to 3.0 (SD = 4.2) four years later (aged 13–14 years; Corder *et al.,* 2015). It is likely a combination of biological (e.g. sex, genetics), environmental (e.g. access to facilities), psychosocial (e.g. self-efficacy) and socio-cultural factors (e.g. parental behaviour) that explains the decline in PA during development (Bauman *et al.,* 2012; Biddle *et al.,* 2011; D'Haese *et al.,* 2016; Ferreira *et al.,* 2007; Lightfoot *et al.,* 2018; Lindquist *et al.,* 1999; Sterdt *et al.,* 2014; Uijtdewilligen *et al.,* 2011), though their nature and individual contribution is still unclear.

CRF or aerobic fitness is defined as the "*ability of the respiratory and circulatory systems to supply fuel during sustained physical activity and to eliminate fatigue products after supplying fuel*" (Caspersen *et al.,* 1985, additional details in Appendix A.1). It is often regarded as a proxy for PA, yet is a physiological trait (Lang & Little, 2018; Schutte *et al.,* 2016) that, unlike PA (a behaviour), does not vary on a day-to-day basis (Warren *et al.,* 2010). CRF can be increased by PA (Armstrong *et al.,* 2011) and is typically expressed in terms of the peak oxygen uptake ($\dot{\mathbf{V}}O_2$peak): the highest level of oxygen consumed during exercise. Analyses of temporal trends of $\dot{\mathbf{V}}O_2$peak obtained with a maximal exercise test suggest that CRF levels in children and adolescents have been stable over time (Armstrong *et al.,* 2011; Armstrong & Welsman, 2019d), though those derived from field-based assessments suggest that CRF has declined (Tomkinson *et al.,* 2019). The influence of increased fat mass on field-based measures of CRF may partially explain this discrepancy (Armstrong *et al.,* 2011; Armstrong & Welsman, 2019c; Armstrong &

Welsman, 2019d), but other factors (e.g. experience with tests) could play a role too. In contrast to PA, CRF levels increase with age during development, particularly in boys whose fat-free mass increases considerably with maturation (Armstrong & Welsman, 2019b; Tomkinson *et al.,* 2018). In boys, $\dot{V}O_2$peak was found to increase by 150% between 8–16 years, compared to 80% in girls (Armstrong *et al.,* 2011; Armstrong & Welsman, 1994).

1.2.2 Developing brain and cognition

Initially, studies used neuropsychological test batteries to examine changes in cognitive performance during development. With advances in neuroimaging techniques, such as magnetic resonance imaging (MRI), it has become possible to additionally and concurrently study the developing brain *in vivo* (Brenhouse & Andersen, 2011; Gilmore *et al.,* 2018; Lenroot & Giedd, 2006). Using such techniques, researchers have shown that the brain and cognitive abilities undergo profound changes during the first decades of life.

1.2.2.1 Developing cognition

Cognition, from the Latin *cognoscere*, or "to know", refers to the faculty of knowing, but has, despite various attempts, no clear definition (Allen, 2017; Bayne *et al.,* 2019). Cognition is typically described in broad terms, often referring to a range of mental processes relating to the acquisition, storage, manipulation, and retrieval of information. In this thesis, cognition is treated as an umbrella term that encompasses various more tangible or precisely defined processes, such as processing speed, memory, attention and executive functions (Chang *et al.,* 2017; Donnelly *et al.,* 2016). Rudimentary forms of such cognitive abilities are present in infancy, develop rapidly throughout childhood and continue to be refined during adolescence, evidenced by lower error rates and reaction times, as well as reduced variability in task performance (e.g. Anderson, 2002; Davidson *et al.,* 2006; Diamond, 1990; Hassevoort *et al.,* 2018; Huizinga *et al.,* 2006; Kail, 1991; Luna *et al.,* 2004; Ordaz *et al.,* 2013; Romine & Reynolds, 2005; Tamnes *et al.,* 2012). Moreover, individual differences in cognitive abilities become more heritable and stable

from childhood to adolescence, though remain sensitive to environmental influences (Friedman *et al.,* 2016; Polderman *et al.,* 2007; Tucker-Drob & Briley, 2014). Behavioural interventions, including physical activity, thus have the potential to favourably impact cognitive development in this period.

Executive functions (EF) have been of particular interest to developmental psychologists (Best & Miller, 2010) and PA researchers (e.g. Álvarez-Bueno *et al.,* 2017b; Donnelly *et al.,* 2016; Xue *et al.,* 2019). EF, also known as cognitive control or executive control, is an umbrella term referring to a set of interrelated cognitive processes that are thought to underlie goal-directed behaviour (Best & Miller, 2010; Diamond, 2013; Miyake *et al.,* 2000). To date, there is no consensus regarding the precise component(s) of EF (Baggetta & Alexander, 2016; Packwood *et al.,* 2011), although scientists commonly distinguish three interrelated, but distinct core elements (Diamond, 2013; Miyake *et al.,* 2000): (1) working memory (the ability to maintain and manipulate information online), (2) inhibitory control (the ability to override prepotent or dominant responses), and (3) cognitive flexibility (the ability to shift flexibily between tasks or mental sets), that give rise to higher-order EFs, such as planning and problem solving (but other components have been proposed too, see Appendix A.1 and Anderson, 2002; Baddeley & Hitch, 1974; Friedman *et al.,* 2008; Zelazo & Carlson, 2012). This three-factor structure of EF, however, only appears during adolescence, and seems to be preceded by a single- (early childhood) and dual-component (late childhood) EF structure (Best & Miller, 2010; Brydges *et al.,* 2014; Lee *et al.,* 2013; Theodoraki *et al.,* 2020; Wiebe *et al.,* 2008). The age-related differentiation and specialisation of EFs is, however, still somewhat contested and an area of active research (e.g. Hartung *et al.,* 2020; Huizinga *et al.,* 2006; Theodoraki *et al.,* 2020). Nevertheless, a large body of evidence suggests that individual EFs continue to develop throughout adolescence and follow distinct developmental pathways (Best & Miller, 2010; Best *et al.,* 2009; Constantinidis & Luna, 2019; Huizinga *et al.,* 2006; Lee *et al.,* 2013; Luna *et al.,* 2004; Luna *et al.,* 2015; Magar *et al.,* 2010).

1.2.2.2 Developing brain

In parallel with cognitive maturation, the brain undergoes profound changes during development (e.g. Blakemore, 2012; Giedd & Rapoport, 2010; Gilmore *et al.,* 2018; Silk & Wood, 2011; Walhovd *et al.,* 2014b). Early post-mortem and animal studies have shown that brain development is characterised by a dynamic interplay of progressive (e.g. myelination) and regressive (e.g. synaptic pruning) changes (details in e.g. Lenroot & Giedd, 2006; Mills & Tamnes, 2018; Silk & Wood, 2011; Stiles & Jernigan, 2010; Tau & Peterson, 2010), resulting in adult brains that are stable, efficient and specialised (Dahl *et al.,* 2018; Johnson, 2001). *In vivo* MRI studies have shed light on trajectories of structural and functional brain maturation. Despite various methodological challenges (discussed in e.g. Becht & Mills, 2020; Madhyastha *et al.,* 2018; Sørensen *et al.,* 2021; Tamnes *et al.,* 2018b; Telzer *et al.,* 2018; Vijayakumar *et al.,* 2018a), such studies tend to agree that brain development is protracted, coordinated, tissue-specific, spatially heterogeneous and typically non-linear (e.g. Dow-Edwards *et al.,* 2019; Lenroot & Giedd, 2006; Mills & Tamnes, 2018). To illustrate several of these properties and provide some insight into the dynamic nature of brain development, several aspects of grey and white matter development are mentioned below. Maturation of other brain properties, including those derived from functional MRI (fMRI), such as task-related brain activity (Constantinidis & Luna, 2019; Crone & Steinbeis, 2017; Durston *et al.,* 2006; Luna *et al.,* 2015; Paus *et al.,* 2008) and functional connectivity (Cui *et al.,* 2020; Horien *et al.,* 2019; Li *et al.,* 2019), and those obtained from more advanced MRI sequences (e.g. perfusion Satterthwaite *et al.,* 2012) or analysis methods (e.g. structural covariance networks, DuPre & Spreng, 2017), are equally important, but outside the scope of this thesis.

Structural MRI studies tracking development of brain volume have shown rapid growth in the first two years of life (Gilmore *et al.,* 2018; Knickmeyer *et al.,* 2008), followed by a more gradual increase in volume until age 13 years (Hedman *et al.,* 2012). After mid-adolescence, brain volume decreases and stabilises in the early twenties (Hedman *et al.,* 2012; Mills *et al.,* 2016). Whole brain volume is the sum of grey and white matter volume, which follow distinct developmental trajectories. Grey matter (GM)

is present in the cortex and subcortical structures and is composed of neuronal cell bodies, glial cells, dendrites, (un)myelinated axons and capillaries. Cortical GM volume increases rapidly from birth until early childhood and declines non-linearly during late childhood and adolescence (Giedd *et al.,* 1999; Gogtay *et al.,* 2004; Mills *et al.,* 2016; Paus *et al.,* 2008; Sowell *et al.,* 2003; Tamnes *et al.,* 2017; Tamnes *et al.,* 2013a). Development of cortical GM volume is regionally variable (Gogtay *et al.,* 2004; Tamnes *et al.,* 2013a). For instance, association cortices, including the prefrontal cortex, that subserve higher order cognitive functions (e.g. EF), mature later than somatosensory and visual cortices (Gogtay *et al.,* 2004). Cortical thickness follows a similar trajectory to cortical GM volume, though its decline starts earlier (around age 4, Walhovd *et al.,* 2017), whereas cortical surface area decreases slightly from late childhood (Tamnes *et al.,* 2017; Walhovd *et al.,* 2017). Interestingly, maturation of cortical GM has been associated with the development of cognitive abilities (Shaw *et al.,* 2006; Tamnes *et al.,* 2013b; Vijayakumar *et al.,* 2014; Walhovd *et al.,* 2016; Walhovd *et al.,* 2014b), though longitudinal evidence of such relationships is sparse.

Subcortical regions continue to mature during adolescence (Herting *et al.,* 2018; Wierenga *et al.,* 2014). The hippocampus, for instance, shows subtle volumetric increases to mid-adolescence, followed by a slight decrease into late adolescence and young adulthood (Herting *et al.,* 2018; Narvacan *et al.,* 2017; Tamnes *et al.,* 2018a; Wierenga *et al.,* 2014). Moreover, distinct trajectories have been observed for its subregions (Tamnes *et al.,* 2018a), and its maturation predicts cognitive abilities, such as learning and memory performance (Daugherty *et al.,* 2017; De Master *et al.,* 2014; Tamnes *et al.,* 2018a; Tamnes *et al.,* 2014).

In contrast to volumes of most GM structures, WM volume —predominantly composed of myelinated long-distance axonal projections —linearly increases throughout childhood and and adolescence (Mills *et al.,* 2016), reaching a peak during the fourth decade of life (Lebel *et al.,* 2012). Similarly, markers of white matter *microstructure* show continued, but regionally variable and non-linear, development throughout childhood

and adolescence (Goddings *et al.,* 2020; Lebel & Deoni, 2018; Lebel *et al.,* 2019). The importance of white matter for efficient cognitive functioning is increasingly recognised (Filley & Fields, 2016) and there is some evidence linking maturation of WM to development of EFs (e.g. Darki & Klingberg, 2015; Goddings *et al.,* 2020; Lebel *et al.,* 2019).

Brain development is complex and under the influence of many genetic and (interacting) environmental factors (Foulkes & Blakemore, 2018; Paus, 2013), including sex (Herting *et al.,* 2018; Kaczkurkin *et al.,* 2019), pubertal status (Goddings *et al.,* 2019; Goddings *et al.,* 2014; Herting *et al.,* 2017; Vijayakumar *et al.,* 2018b), and the socioeconomic (e.g. Farah, 2018; Foulkes & Blakemore, 2018; Tooley *et al.,* 2020, but see: Walhovd *et al.,* 2016) and family environment (Walhovd *et al.,* 2016; Whittle *et al.,* 2017). Twin studies have demonstrated high heritability of global and regional GM and WM during development (Blokland *et al.,* 2012; Peper *et al.,* 2007), though recent evidence suggests that environmental influences may play a substantial role in explaining inter-individual differences in developmental structural brain changes (Brouwer *et al.,* 2017). Indeed, studies in adolescents have shown that environmental factors and experiences, such as musical training (Hudziak *et al.,* 2014), alcohol use (Kozareva *et al.,* 2019; Squeglia *et al.,* 2015), and a positive parenting style (Whittle *et al.,* 2017), may influence brain development.

The ability of the brain to change in response to environmental influences or experiences is termed "plasticity". The brain remains plastic throughout life (Lindenberger & Lövdén, 2019), though there are certain periods —typically referred to as sensitive periods —of heightened (neural) sensitivity during which the environment may have a particularly strong impact on the brain and have long-lasting effects on behaviour. Adolescence has been described as a sensitive period (Dow-Edwards *et al.,* 2019; Fuhrmann *et al.,* 2015; Larsen & Luna, 2018; O'Leary *et al.,* 2019b; Thompson & Steinbeis, 2020), which has sparked the attention of public health and research communities (Dahl *et al.,* 2018). This period in life has been termed a "window of opportunity" (Dahl *et al.,* 2018)

during which interventions, such as PA, may positively change the developmental course and improve adolescent health.

1.3 Physical activity, cardiorespiratory fitness, the brain and cognition

Research examining relationships between PA or CRF and neurocognitive functioning in humans has flourished over the past four decades. Most of these studies have focussed on older adults (Hillman *et al.,* 2008; Khan & Hillman, 2014), and have shown that higher PA is associated with a lower risk of cognitive decline and better brain health (Erickson *et al.,* 2015; Erickson *et al.,* 2014; Prakash *et al.,* 2015). This work has been spurred by rodent experiments showing that enriched environments involving running wheels yield profound structural brain changes and improvements in learning and memory (reviewed in e.g. Vivar & Van Praag, 2017; Voss *et al.,* 2013). Based on the findings in rodents and ageing populations, and increased awareness of childhood and adolescence as important developmental periods (see section 1.2 above), the past 20 years have witnessed an upsurge in studies involving young people (Figure 1.1).

The findings of these studies have been summarised extensively in systematic reviews, which have been evaluated by The 2018 Physical Activity Guidelines Advisory Committee (PAGAC) to inform policy (2018 Physical Activity Guidelines Advisory Committee, 2018). The report suggested *moderate* evidence for PA-related beneficial effects on cognitive performance during pre-adolescence, but inconsistent evidence during adolescence (2018 Physical Activity Guidelines Advisory Committee, 2018). These conclusions were recently updated to also suggest beneficial effects on brain structure and function during pre-adolescence and limited but promising evidence for PA on cognition in adolescence (Erickson *et al.,* 2019). In contrast, the United Kingdom (UK) equivalent of the PAGAC concluded there was inconclusive evidence for effects of PA on cognitive performance in children and young people (Jago *et al.,* 2018), whereas a recent systematic review of reviews concluded that the evidence from 25 systematic reviews supports a causal link between PA and cognition in young people (Biddle *et al.,*

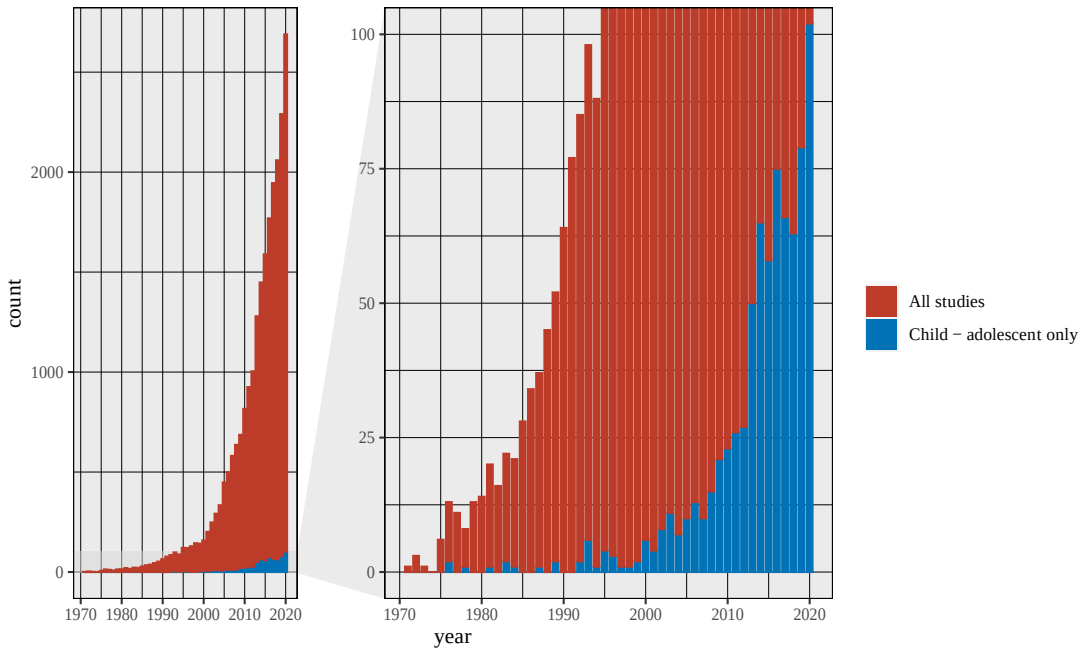


Figure 1.1: Histogram of the number of published articles containing physical activity and cognition in abstract or title (1970–2020). The search was performed on PubMed on 5th January 2021 using the following search query "((physical activity[Title/Abstract] OR exercise[Title/Abstract] OR fitness[Title/Abstract]) AND (cognition[Title/Abstract] OR cognitive[Title/Abstract] OR academic achievement[Title/Abstract] OR brain[Title/Abstract] OR MRI[Title/Abstract])) NOT (animal[Title/Abstract] OR rat[Title/Abstract] OR mice[Title/Abstract])" to retrieve all papers, and "(((physical activity[Title/Abstract] OR exercise[Title/Abstract] OR fitness[Title/Abstract]) AND (cognition[Title/Abstract] OR cognitive[Title/Abstract] OR academic achievement[Title/Abstract] OR brain[Title/Abstract] OR MRI[Title/Abstract]) AND (child[Title/Abstract] OR adolescent[Title/Abstract])) NOT (animal[Title/Abstract] OR rat[Title/Abstract] OR mice[Title/Abstract])" to retrieve child and adolescent papers only. Only papers published between 1970–2020 are displayed. The search was performed using the *RISmed* package (Kovalchik, 2017) in R (version 6).

2019). As a first step in this thesis and to seek clarity among the contrasting reports, we conducted a systematic review of reviews, which aimed to synthesise the evidence for the effect of PA on brain structure and function, academic achievement and cognitive performance in childhood and adolescence (Wassenaar *et al.*, 2020).

A summary of the findings of the systematic review regarding the effect of PA on cognitive and brain outcomes is presented in the next sections. Systematic reviews

on the effects of *chronic* PA interventions (i.e. more than one PA session, over a set period) on cognitive and/or brain outcomes in school-aged children or adolescents were included. Systematic reviews that contained a mixture of observational and interventional evidence, that reported on a single bout of *acute* PA (i.e. a single session) or cardiovascular fitness only, or included case reports, were excluded. The evidence for the effects of acute PA on and neurocognitive function has been summarised elsewhere (e.g. Basso & Suzuki, 2017; Chang *et al.,* 2019; Haverkamp *et al.,* 2020; Hillman *et al.,* 2011; Janssen *et al.,* 2014; Kao *et al.,* 2020; Meijer *et al.,* 2020; Moreau & Chou, 2019; Wheatley *et al.,* 2020a) and was outside the scope of the review and this thesis. Moreover, the evidence predominantly consisted of aerobic PA interventions; resistance training is less common in adolescents and there is a lack of studies exploring its neurocognitive effects (Herold *et al.,* 2019b; Smith *et al.,* 2014b; Soga *et al.,* 2018). The review's methods, including inclusion criteria, are presented in Appendix A.2. While interventional evidence is of greatest interest to policymakers because it allows for causal conclusions about the effect of chronic PA on neurocognitive health, the majority of studies to date have been of cross-sectional nature. Given their contribution to our understanding of relationships involving PA, findings from cross-sectional research are briefly reviewed too.

1.3.1 Physical activity, cardiorespiratory fitness and cognition

Observational studies in the field of PA have traditionally focussed on measures of CRF (Kramer *et al.,* 2018; Van Waelvelde *et al.,* 2020) as a proxy for PA (Sibley & Etnier, 2003) and possible mediator of relationships between PA and cognitive outcomes (Åberg *et al.,* 2009; North *et al.,* 1990). Despite high between-study heterogeneity in assessments and outcome measures, converging evidence suggests that higher CRF is associated with better performance on tasks measuring specific EFs (i.e. inhibitory control, working memory and cognitive flexibility; Van Waelvelde *et al.,* 2020), as well as those measuring visual or relational memory (Chaddock *et al.,* 2010a; Hassevoort *et al.,* 2017; Monti *et al.,* 2012). Most of these studies, however, have been conducted in small samples of

children, typically comparing task performance between high and low fit individuals (Van Waelvelde *et al.,* 2020). Fewer studies have considered adolescents, with current evidence suggesting positive associations (Hogan *et al.,* 2015; Huang *et al.,* 2015; Shigeta *et al.,* 2020; Skog *et al.,* 2020; Westfall *et al.,* 2018), though null results have been reported too (Cancela *et al.,* 2019; Ruiz *et al.,* 2010; Ruotsalainen *et al.,* 2020).

The evidence for relationships involving measures of self-reported or objective PA is less clear and heterogeneity in study designs, samples, and assessments of PA and cognition make firm conclusions difficult to draw. An increasing body of observational studies have shown small, but positive associations between self-reported or objective PA and composite cognitive functions (Esteban-Cornejo *et al.,* 2015; Martínez-Gómez *et al.,* 2011; McPherson *et al.,* 2018; Quan *et al.,* 2018; Ruiz *et al.,* 2010), but inconsistent associations between objective PA (typically MVPA) and domain-specific cognitive abilities, including inhibition (Aadland *et al.,* 2017a; Aadland *et al.,* 2017b; Ishihara *et al.,* 2017; Ishihara *et al.,* 2018; Pindus *et al.,* 2015; Ruotsalainen *et al.,* 2020; Van Der Niet *et al.,* 2015; Wickel, 2017), working memory (Aadland *et al.,* 2017b; Ishihara *et al.,* 2017; Ruotsalainen *et al.,* 2020; Van Der Niet *et al.,* 2015) and cognitive flexibility (Aadland *et al.,* 2017b; Booth *et al.,* 2013; Ruotsalainen *et al.,* 2020; Syväoja *et al.,* 2014b; Van Der Niet *et al.,* 2015). Associations may vary with sex, though whether relationships are stronger in boys (Booth *et al.,* 2013; Quan *et al.,* 2018) or girls (Aggio *et al.,* 2016) is still unclear.

Intervention studies have been conducted to assess causal relationships between PA and cognitive outcomes. Following the systematic search, thirteen systematic reviews were found that evaluated cognitive outcomes (Table 1.1). Reviews included between one (Lees & Hopkins, 2013) and 36 intervention studies (Álvarez-Bueno *et al.,* 2017b; Gunnell *et al.,* 2018), and of 83 published studies, 42 (50%) were included in more than one review (additional details in Wassenaar *et al.,* 2020). Reviews synthesised findings of studies in terms of overall cognitive performance, and / or various sub-domains (e.g. EFs, memory). One review was of high quality, one of low quality and the remaining

11 reviews of critically low quality (details in Wassenaar *et al.*, 2020).

Table 1.1: Cognitive outcomes: findings from systematic reviews and meta-analyses (ordered by quality rating)

Authors	Population	Systematic review results	Meta-analysis results[c]
High quality reviews			
Martin *et al.*, 2018	Overweight or obese children (3–18 years)	High quality evidence for an effect of PA on composite executive functions and non-verbal memory, but not cognitive flexibility, inhibition (low quality), attention or visuo-spatial abilities **Composite executive functions (s = 3):** 2/3 no evidence for an effect, 1/3 showing a positive effect **Inhibition control (s = 1)**: no evidence for an effect **Attention (s = 3):** no evidence for an effect **Working memory (s = 1):** no evidence for an effect **Visuo-spatial abilities (s=3):** no evidence for an effect **Cognitive flexibility (s = 2):** no evidence for an effect **Non-verbal memory (s = 2):** some evidence for a small effect **General intelligence (s = 1):** some evidence for an effect	**Composite executive functions:** *PA:* 0.42 (0.05–0.78), (s = 1); *Exergaming:* 0.58 (-0.02, 1.18), (s = 1) **Inhibition control:** NA **Attention:** 0.46 (-0.16, 1.08), I^2 = 41%, (s = 2) **Working memory:** NA **Visuo-spatial abilities:** NA **Cognitive flexibility:** -0.06 (-0.37, 0.25), I^2 = 0%, (s = 2) **Non-verbal memory:** NA **General intelligence:** NA
Low quality reviews			
Singh *et al.*, 2019	Children and adolescents (3–16 years)	6 high-quality studies: 10/21 (48%) analyses found a significant beneficial intervention effect, leading to inconclusive evidence	NA
Critically low quality reviews			
Álvarez-Bueno *et al.*, 2017b[a]	Healthy children and adolescents (4–18 years)	**Non-executive functions:** 7/7 found improvements; of 4 studies that included multiple intervention groups, two suggested that increases in duration and intensity was associated with greater improvements **Executive functions:** 29/29 found improvements; of 11 studies that included multiple intervention groups, three did not find differences between the groups **Meta-cognition:** 15/15 found improvements; of 6 studies that included multiple intervention groups, none found differences in improvements	**Non-executive functions:** d = 0.23 (0.09, 0.37), I^2 = 21.9% (s = 6, k = 17) **Executive functions:** d = 0.20 (0.10, 0.30), I^2 = 70.0% (s = 22, k = 42) **Working memory:** d = 0.14 (0.00, 0.27), I^2 = 48% (s = 9, k = 13) **Selective attention / inhibition:** d = 0.26 (0.10, 0.41), I^2 = 76.0% (s = 17, k = 24) **Selective attention:** d = 0.13 (-0.07, 0.33), I^2 = 66.8% **Inhibition:** d = 0.38 (0.13, 0.63), I^2 = 68.7% **Cognitive flexibility:** d = 0.11 (-0.10, 0.32), I^2 = 68.7% (s = 4, k = 5) **Meta-cognition:** d = 0.23 (0.13, 0.32), I^2 = 4.7% (s = 10, k = 21) **Higher level executive functions:** d = 0.19 (0.06, 0.31), I^2 = 12.9% (s = 7, k = 13)

Continued on next page

Table 1.1 – continued from previous page

Authors	Population	Systematic review results	Meta-analysis results
De Greeff *et al.*, 2018	Primary school children (6–12 years)	Combined academic achievement and cognition: 9/14 reported positive findings on at least 1 outcome measure, 5 reported no significant findings	**Overall cognitive functions:** g = 0.37 (0.20, 0.55), I^2 = 64.92% (s = 14, k = 18) **Executive functions:** g = 0.24 (0.09, 0.39), I^2 = 34% (s = 12, k = 15) **Working memory:** g = 0.36 (0.10, 0.62), I^2 = 56.79% (s = 6, k = 8) **Cognitive flexibility:** g = 0.18 (0.01, 0.35), I^2 = 4.79% (s = 4, k = 4) **Inhibition:** g = 0.19 (-0.04, 0.42), I^2 = 49.7% (s = 6, k = 7) **Planning:** g = 0.12 (-0.08, 0.32), I^2 <0.01% (s = 4, k = 4)
Gunnell *et al.*, 2018	Healthy children (1–17.99 years)	**Inhibitory control:** *PA vs none (n = 1,248, s = 5):* 3/5 —, 1/5 —↑, 1/5 —↓; *PA vs PA (n = 557, s = 6):* 3/6 ↑, 1/6 —, 2/6 —↑; *Multiple comparisons (n = 181, s = 1):* 1/1 — **Working memory:** *PA vs none (n = 1,804, s = 3):* 1/3 —, 1/3 ↑, 1/3 —↑; *PA vs PA (n = 487, s = 3):* 2/3 —, 1/3 ↑; *Multiple comparisons (n = 181, s = 1):* 1/1 — **Cognitive flexibility:** *PA vs PA (n = 501, s = 2):* 1/2 —↑, 1/2 ↑; *Multiple comparisons (n = 246, s = 2):* 1/2 —, 1/2 —↑ **Unitary constructs:** *PA vs none (n = 549, s = 2):* 1/2 —↑, 1/2 —; *PA vs PA (n = 472, s = 3):* 1/3 ↑, 1/3 —↑, 1/3 — **Attention:** *PA vs none (n = 1,809, s = 8):* 5/8 —, 1/8 ↑, 1/8 —↑; *PA vs PA (n = 156, s = 5):* 1/5 —↑, 2/5 ↑, 2/5 —; *Multiple comparisons (n = 757, s = 1):* 1/1 —↑ **Information processing:** *PA vs none (n = 1,659, s = 5):* 1/5 ↑, 4/5 —; *Multiple comparisons (n = 265, s = 2):* 2/2 —↑; *PA vs PA (n = 448, s = 3):* 3/3 — **Memory:** *PA vs none (n = 44, s = 1):* 1/1 — **Motor speed and learning:** *PA vs PA (n = 508, s = 2):* 1/2 ↑, 1/2 — **Composite cognition:** *PA vs none (n = 1,794, s = 3):* 1/3 —, 2/3 —↑	Not performed given heterogeneity of study designs, PA exposures and outcomes
Verburgh *et al.*, 2014	Children and adolescents (6–17 years), but one study in young adults	Inconsistent results among 5 studies that reported on the effect of chronic PA on executive functions (one in young adults)	**Executive functions** across age groups: d = 0.14 (-0.04, 0.32), Q = 5.1 (s = 5) **Planning:** d = 0.16 (-0.07, 0.39), Q = 0.89 (s = 3)
Jackson *et al.*, 2016	Healthy children (7–12 years)	8/8 studies showed a positive effect of PA on inhibitory control, but none were statistically significant in isolation; other domains of executive function were measured too infrequently to perform a meta-analysis	**Inhibitory control:** d = 0.2 (0.03, 0.37), I^2 = 0%
Li *et al.*, 2017b	Healthy adolescents (13–18 years)	1/2 studies showed a beneficial effect on cognitive function; ff five cognitive function parameters, only one showed significance	NA
Lees & Hopkins, 2013	Children and adolescents (<19 years)	1/1 studies showed positive effects of PA on cognitive performance; another study was included in the data table, but not part of the results section or evidence synthesis	NA

Continued on next page

Table 1.1 – continued from previous page

Authors	Population	Systematic review results	Meta-analysis results
Bustamante *et al.*, 2016	Overweight or obese children and/or adolescents	*Quasi-experimental (s = 4):* each of the four studies showed some benefit on neural, cognitive or academic outcomes; but the PA was confounded with other elements of the intervention and no control groups; little evidence for PA effects *RCT (s = 10, but only 5 independent studies):* high quality RCTs (s = 2) have shown benefits for different executive functions	NA
Suarez-Manzano *et al.*, 2018	Children and adolescents with ADHD (6–18 years)	7/7 studies showed a positive effect of PA on cognition, no study revealed a negative association; the systematic practice of PA between 5–20 weeks, 30–90 min at moderate-vigorous intensity (40–75%) produces a chronic effect that improves cognition in young people with ADHD	NA
Mura *et al.*, 2015	Children (3–18 years)	7/9 studies showed an improvement in global cognitive performance, 1/9 showed no difference and 1/9 worse intelligence; two of these studies found dose-response relationships, with high dose PA performing better than low-dose PA or control; specific cognitive skills improved in almost all studies (6 studies)	NA
Vazou *et al.*, 2019	Typically developing children and adolescents (4–16 years)	**Aerobic only (s = 7):** Significant cognitive outcomes: planning (s = 1), creativity (s = 2), working memory and spatial memory span (s = 4), attentional accuracy and spatial inattention (s = 1), cued recall memory (s = 1) and mathematics fluency (s = 1) **Motor skills (s = 4):** improvements in working memory (s = 1), spatial processing/math/reading/concentration (s = 1 study), lower error on attentional task (s = 1), mixed effects (s = 1) **Cognitively engaging PA (s = 2):** Improved planning (s = 1), and spatial memory, but not verbal memory (s = 1) **Aerobic and motor skill (s = 1):** No difference in inhibition accuracy or reaction time (s = 1) **Motor skills and cognitively engaging (s = 7):** Improved inhibition (s = 1), attention related to arithmetic (s = 1), parent-rated inhibitory behavioural control and reaction time (s = 1), inhibition (s = 1), attention (s = 1), inattention and hyperactivity (s = 1) and attentional accuracy (s = 1) **Aerobic and cognitively engaging (s = 8):** improvements in math fluency (s = 2), math and spelling (s = 1), fluid intelligence (s = 1), cognitive shifting (s = 1), memory recall (s = 1), verbal working memory and inhibition (s = 1), time-on-task in classroom (s = 1) **Aerobic, motor and cognitively engaging (s = 1):** improved receptive attention	**Overall cognitive function:** g = 0.46 (0.28, 0.64), I^2 = 85% (s = 21, k = 28) **PA interventions versus comparison treatments:** *PA vs no treatment:* 0.86 (0.18, 1.55), I^2 = 93% (s = 5) *PA vs academic instruction:* 0.57 (0.32, 0.83), I^2 = 81% (s = 10) *PA vs traditional PE:* 0.09 (-0.07, 0.24), I^2 = 44% (s = 9); *PA combinations vs aerobic PA:* 0.80 (-0.08, 1.67), I^2 = 88% (s = 4) **Qualitatively different PA interventions vs comparison:** see Wassenaar *et al.*, 2020

Abbreviations: d = Cohen's d, g = Hedges' g, k = number of comparisons, n = number of participants, NA = not assessed, PA = physical activity, RCT = randomised controlled trial, s = number of study/studies

[a] This review also analysed *cognitive life skills*, which is different from any of the other typically examined cognitive functions and therefore excluded from this table

[b] PA vs none: PA was compared to a sedentary control condition, Multiple comparisons: studies with multiple treatment and/or control groups, PA vs PA: comparison of multiple types of PA interventions. Colour coding represents combinations of: – = null results, ↓ = unfavourable results, ↑ = favourable results

[c] Results are reported as: standardized mean difference, 95% confidence intervals, heterogeneity statistics if available, the number of studies (s) and number of comparisons (k)

Across all thirteen reviews, eight concluded that PA was beneficial (Álvarez-Bueno *et al.*, 2017b; Bustamante *et al.*, 2016; De Greeff *et al.*, 2018; Jackson *et al.*, 2016; Lees & Hopkins, 2013; Mura *et al.*, 2015; Suarez-Manzano *et al.*, 2018; Vazou *et al.*, 2019) and five reported mixed or inconclusive evidence (Gunnell *et al.*, 2018; Li *et al.*, 2017b;

Martin *et al.,* 2018; Singh *et al.,* 2019; Verburgh *et al.,* 2014). Two meta-analyses showed a positive effect of PA on overall cognitive performance (De Greeff *et al.,* 2018; Vazou *et al.,* 2019), two of three showed PA-related benefits on overall EF (Álvarez-Bueno *et al.,* 2017b; De Greeff *et al.,* 2018; Verburgh *et al.,* 2014) and inhibition (Álvarez-Bueno *et al.,* 2017b; De Greeff *et al.,* 2018; Jackson *et al.,* 2016), and two of three meta-analyses reported non-significant effect sizes for planning or higher-level cognitive functions (De Greeff *et al.,* 2018; Verburgh *et al.,* 2014). Moreover, a meta-analysis additionally reported beneficial effects of PA on non-EF and working memory, but not selective attention or cognitive flexibility (Álvarez-Bueno *et al.,* 2017b). Despite some positive results, the effect sizes are generally small (0.1–0.3) and suffer from substantial heterogeneity. There is some evidence for beneficial effects of PA on cognitive functions in children with excess weight and obesity (Bustamante *et al.,* 2016) and with Attention Deficit Hyperactivity Disorder (ADHD, Suarez-Manzano *et al.,* 2018).

Subgroup analyses of meta-analyses further suggest that cognitive performance may benefit most from PA during curricular physical education (PE; Álvarez-Bueno *et al.,* 2017b), with mixed evidence for enhanced (i.e. quantitative increase in PA) or enriched PA (i.e. qualitative manipulation of PA, e.g. increasing coordinative task requirements; Álvarez-Bueno *et al.,* 2017b; Vazou *et al.,* 2019; Wassenaar *et al.,* 2020). Intervention duration does not seem to be an important moderator of PA-related cognitive changes (De Greeff *et al.,* 2018; Jackson *et al.,* 2016; Verburgh *et al.,* 2014), although one review reported a negative relationship with working memory (Álvarez-Bueno *et al.,* 2017b). No significant relationships were found between study quality and heterogeneity (Álvarez-Bueno *et al.,* 2017b) or effect sizes (Verburgh *et al.,* 2014).

The evidence from these low-quality systematic reviews is inconsistent, with conclusions suggesting both positive and mixed (inconclusive) effects of PA on overall cognitive performance and its sub-domains. High between-study heterogeneity in quality and methods may partially explain the inconsistency in findings. In particular, reviews did not report on intervention fidelity, varied in their presentation of results

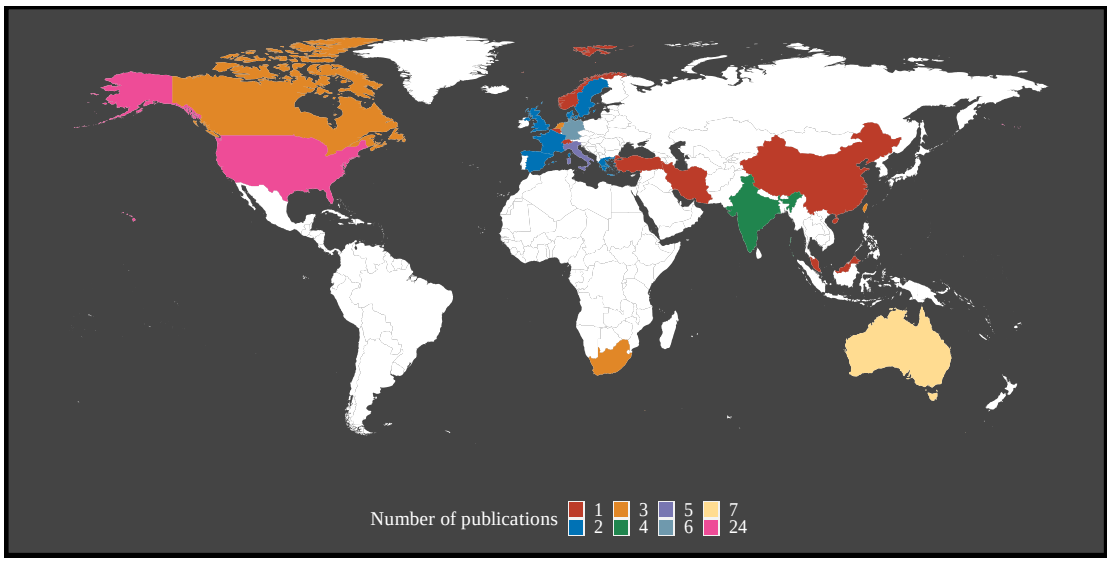


Figure 1.2: Countries where PA interventions have been conducted to assess the effect on cognitive outcomes. Legend shows the number of unique primary studies that have been published in the respective (colour-coded) countries and were included in systematic reviews. Adapted from Wassenaar *et al.* (2020).

(e.g. definition of sub-domains) and hardly distinguished between PA interventions (e.g. dose, type). What's more, most of the evidence is based on low-quality (primary) studies conducted in developed countries (Figure 1.2), that primarily involved children, poorly reported on intervention protocols and did not examine moderators (e.g. sex or socioeconomic status). Hence, well-controlled randomised controlled trials in adolescents are required.

Recently published systematic reviews, the majority of which reported on findings from a similar set of primary studies, support the conclusions drawn above. In particular, they found no significant effect of physically active lessons on cognitive outcomes (Norris *et al.,* 2020), but a significant effect of chronic PA on overall cognition and EFs (Haverkamp *et al.,* 2020; Ludyga *et al.,* 2020; Xue *et al.,* 2019). The effects of PA on sub-domains of EFs were inconsistent (Haverkamp *et al.,* 2020; Heinze *et al.,* 2020; Xue *et al.,* 2019). Similar inconsistencies have been reported for academic outcomes (details in Wassenaar *et al.,* 2020).

Most studies included in the reviews have incorporated interventions that involved large doses of MVPA (e.g. 60 min 4 times per week) that may not be easily scalable. A recent line of research has focussed on vigorous or high intensity PA interventions, that are time-efficient and result in physical adaptations that are comparable to endurance training results (Benda *et al.,* 2015; Gibala *et al.,* 2006). Evidence to date is promising: studies have shown improvements in EFs (Costigan *et al.,* 2016; Davis *et al.,* 2011; Jeon & Ha, 2017; Leahy *et al.,* 2020; Moreau *et al.,* 2017) and psychological health (Costigan *et al.,* 2016; Leahy *et al.,* 2019), but more research is needed.

1.3.2 Physical activity, cardiorespiratory fitness and the brain

Sparse evidence exists for relationships between PA and CRF and measures of brain structure and function in young people, especially adolescents. Observational studies in children have shown that CRF is positively associated with volumes and shapes of subcortical brain structures that support memory and EFs, such as, respectively, the hippocampus (Chaddock *et al.,* 2010a; Esteban-Cornejo *et al.,* 2017; Gorham *et al.,* 2019; Ortega *et al.,* 2019, but see Migueles *et al.,* 2020) and basal ganglia (Chaddock *et al.,* 2010b; Ortega *et al.,* 2019). Evidence in adolescents is sparse and inconsistent, with studies showing both positive (Herting & Nagel, 2012) and no relationships (Ruotsalainen *et al.,* 2019) between CRF or MVPA and hippocampal volumes. Equally mixed are relationships with a range of cortical measures: higher CRF and PA levels have been associated with thicker (Esteban-Cornejo *et al.,* 2019a; López-Vicente *et al.,* 2017; Ross *et al.,* 2015) and thinner cortical thickness (Chaddock-Heyman *et al.,* 2015; López-Vicente *et al.,* 2017), larger surface area (Herting *et al.,* 2016, but see: Esteban-Cornejo *et al.,* 2019a), and greater (Esteban-Cornejo *et al.,* 2017; Herting *et al.,* 2016) or smaller (Ruotsalainen *et al.,* 2020) grey matter volumes in a range of (non-overlapping) parietal, frontal and temporal brain areas.

Studies exploring associations with WM structure have demonstrated that, in healthy and overweight children and adolescents, higher PA (Rodriguez-Ayllon *et al.,* 2020a; Ruotsalainen *et al.,* 2020) and CRF (Chaddock-Heyman *et al.,* 2014; Esteban-Cornejo

et al., 2019b; Herting *et al.,* 2014; Rodriguez-Ayllon *et al.,* 2020b; Ruotsalainen *et al.,* 2020) are associated with more favourable measures of WM volume and microstructure (e.g. higher fractional anisotropy, FA). These associations have been reported on a global level and, less consistently, for various white matter tracts (e.g. corpus callosum). However, high between-study variability in analysis approaches (e.g. choice of regions of interest) and study samples renders conclusions tentative.

Evidence for PA-and CRF-related variability in brain function stems from research using electroencephalography (EEG) and fMRI. In particular, studies using EEG have shown consistent associations between higher CRF and greater allocation of attentional resources —measured by the P3 component of the event-related potential —on EF tasks (Hillman *et al.,* 2009; Kamijo & Masaki, 2016; Kao *et al.,* 2020; Pontifex *et al.,* 2011; Stroth *et al.,* 2009). Moreover, in children, inter-individual differences in CRF have been related to variability in task-evoked brain networks involved in EFs (Chaddock *et al.,* 2012; Voss *et al.,* 2011a), and greater functional connectivity between the hippocampus and frontal areas (Esteban-Cornejo *et al.,* 2021).

Table 1.2: Brain outcomes: findings from systematic reviews and meta-analyses

Author	Population	Systematic review results
CRITICALLY LOW QUALITY REVIEWS		
Gunnell *et al.,* 2018	Healthy children (1–17.99 years)	For brain function, increases, no change or a mixture were interpreted as being supportive of brain function; for brain structure, results were favourable or null **Brain function[a]:** *PA vs none:* *Activation (n = 86, s = 3):* 1/3 no change, 2/3 increased, decreased and no change ([study 1] decreased during anti-saccade task, increased or no change during flanker task; [study 2] no change in frontal or supplementary eye fields, increased in bilateral prefrontal cortex and decreased in bilateral parietal cortex) *Resting-state synchrony (n = 37, s = 1):* 1/1 increases, decreases and no change *Blood flow (n = 30, s = 1):* 1/1 no changes in cerebral blood flow velocity **Brain structure (n = 36, s = 2, same sample):** *PA vs none:* 1/2 improved white matter coherence and myelination, 1/2 no change
Bustamante *et al.,* 2016	Overweight or obese children and/or adolescents	Benefits for neurologic outcomes following PA in high quality studies (RCT, s = 4, 2/4 brain function, 2/4 brain structure), but all from the same group; results from a quasi-experimental study (s = 1) suggest a neural benefit, but the study is of low rigor and suffers from confounding

Continued on next page

Table 1.2 – continued from previous page

Author	Population	Systematic review results
Lubans *et al.*, 2016[b]	Children (7–11 years)	5/6 studies reported significant brain changes (2/6 using EEG, 4/6 using MRI one of which explored brain structure), but there was little overlap between studies

Abbreviations: n = number of participants, PA = physical activity, RCT = randomised controlled trial, s = study/studies
[a] The authors also included findings on changes in brain-derived neurotrophic factor, which are not measured by EEG or MRI and therefore excluded from this table. PA vs none: PA was compared to a sedentary control condition
[b] This review examined brain changes as potential mediators of cognitive changes, rather than exploring brain changes per se

Interventional studies generally support observational evidence of PA-related variation in brain outcomes. Following the systematic search, three reviews were found that systematically explored exercise-related brain changes (Bustamante *et al.*, 2016; Gunnell *et al.*, 2018; Lubans *et al.*, 2016). Reviews included five or six primary studies, and three (of 10 unique publications) were included in all reviews (Davis *et al.*, 2011; Krafft *et al.*, 2014a; Krafft *et al.*, 2014b, details in Wassenaar *et al.*, 2020). All reviews were classified as being of critically low quality.

Across three reviews (Table 1.2), one concluded that PA is beneficial for the brain (Bustamante *et al.*, 2016), one reported mixed favourable and null effects (Gunnell *et al.*, 2018) and one reported a lack of available evidence (Lubans *et al.*, 2016). Among healthy children and adolescents, Gunnell *et al.* (2018) stratified findings from six RCTs by brain function and structure and showed some evidence for changes in activation and resting-state synchrony following a PA intervention, but not blood flow and inconsistent changes in white matter structure. Lubans *et al.* (2016) tabulated brain changes from six studies to explore neurobiological mechanisms of cognitive changes, and showed increases, decreases or no changes following PA interventions in widespread brain areas (as well as cognitive changes). The authors did not synthesise brain findings and reported a lack of overlap between studies (e.g. imaging methods, brain regions). Bustamante *et al.* (2016) focused on children and adolescents with excess weight/obesity and interpreted PA-related brain changes of four high-quality studies as beneficial, but their conclusions lack anatomical specificity.

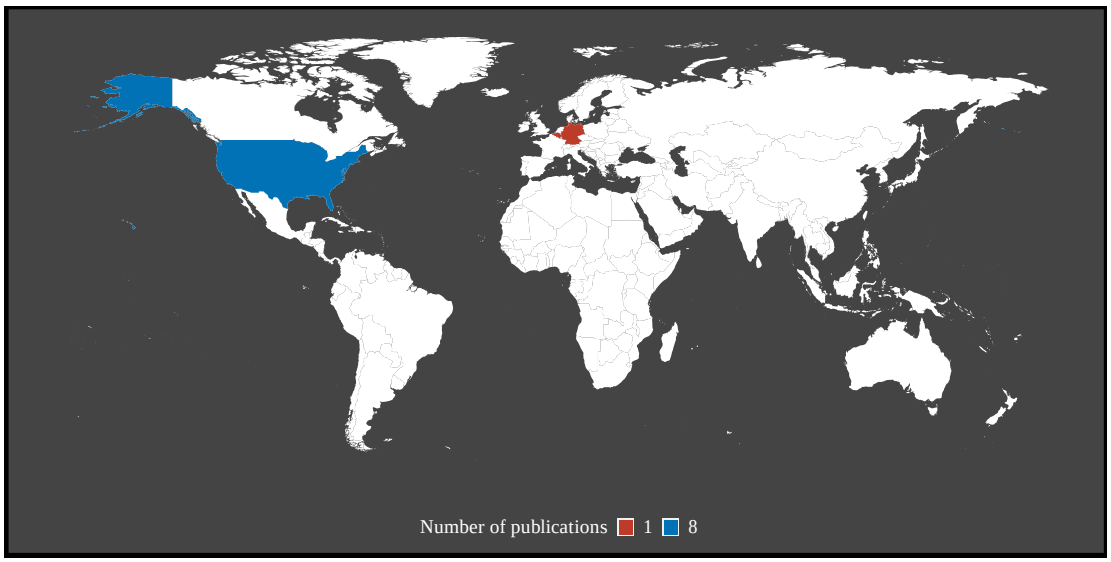


Figure 1.3: Countries where PA interventions have been conducted to assess the effect on brain outcomes. Legend shows the number of unique primary studies that have been published in the respective (colour-coded) countries and were included in systematic reviews . Adapted from Wassenaar *et al.* (2020).

Overall, there is a lack of studies that include brain imaging, in particular T1-weighted (structural) MRI. Consequently, the evidence for PA-related brain changes is based on a small number of studies, 50% of which included the same sample of overweight children and all of which were conducted in developed countries (Figure 1.3). High variability in outcome measures and lack of replication further complicate the interpretation of brain findings. Surprisingly, none of studies that were included in the reviews have examined PA effects on hippocampal metrics (e.g. volume), despite overwhelming evidence from animal studies (Vivar & Van Praag, 2017) and various reports in human adults (Erickson *et al.*, 2011; Firth *et al.*, 2018; Thomas *et al.*, 2016).

Recently published systematic reviews (Meijer *et al.*, 2020; Valkenborghs *et al.*, 2019) generally support conclusions drawn above and show that PA induces favourable changes in EEG-measured neurophysiological function (Meijer *et al.*, 2020).

1.3.3 Putative mechanisms

Various theories and hypotheses have been proposed to explain the PA-induced neurocognitive changes (e.g. cardiovascular fitness hypothesis, North *et al.*, 1990; Spirduso,

1980; neurogenic reserve hypothesis, Kempermann, 2008; catecholamine hypothesis, McMorris, 2016; evolutionary neuroscience hypothesis, Raichlen & Alexander, 2017). While the exact mechanism(s) is (are) still unclear, significant progress has been made in understanding *how* exercise might influence the brain and cognitive function. Research using animals has been crucial to this endeavour, as one can study directly the exercise-induced brain changes on a cellular and molecular level. Neuroimaging studies in humans have started to unravel the neurobiological mechanisms on a systems level, as well as explore psychosocial mediators of the exercise-induced cognitive changes (e.g. sleep and mood; Lubans *et al.,* 2016; Stillman *et al.,* 2016). Comprehensive reviews on these mechanisms have been published elsewhere (e.g. Cabral *et al.,* 2019; Cassilhas *et al.,* 2016; Cotman *et al.,* 2007; Gomez-Pinilla & Hillman, 2013; Hötting & Röder, 2013; Rendeiro & Rhodes, 2018; Vivar & Van Praag, 2017; Voss *et al.,* 2013); a brief overview of several key (neurobiological) processes is provided here.

Animal research has demonstrated exercise-induced neurochemical, structural, mitochondrial and vascular brain changes (Vivar & Van Praag, 2017; Voss *et al.,* 2013), the majority of which have been reported in the hippocampus (Cooper *et al.,* 2018). Exercise modifies the epigenetic regulation of gene expression (Fernandes *et al.,* 2017) and upregulates levels of neurotransmitters (e.g. glutamate, serotonin) and growth factors, such as brain-derived neurotrophic factor (BDNF; e.g. Basso & Suzuki, 2017; Cooper *et al.,* 2018; Vaynman *et al.,* 2004). These molecular events may promote cellular alterations, including increased rates of neurogenesis (O'Leary *et al.,* 2019b; Van Praag, 2005; Van Praag *et al.,* 1999), and angiogenesis (Clark *et al.,* 2009; Van Der Borght *et al.,* 2009), as well as a range of maturational (Cooper *et al.,* 2018), connectivity (Vivar *et al.,* 2016), and morphological changes, such as an increase in spine density and dendritic morphology (Eadie *et al.,* 2005). The effect of exercise on hippocampal neurogenesis has been found to be age-specific: adolescent-initiated running led to an increased expression of plasticity and neurogenesis-related genes, and promoted neurogenesis and neurite complexity to a greater extent than adult-initiated exercise (O'Leary *et al.,* 2019a; O'Leary *et al.,* 2019b). Running also increases astrocyte and oligodendrocyte

precursor cell proliferation (Matsumoto *et al.,* 2011; Perez *et al.,* 2019; Tomlinson *et al.,* 2016). A recent line of evidence suggests that muscle-derived factors (e.g. cathepsin B) and hippocampal activity could be among the first physiological responses to PA, possibly driving PA-related cognitive changes (Rendeiro & Rhodes, 2018).

Research in human adults has found support for the exercise-induced upregulation of BDNF and its effects on cognition (Szuhany *et al.,* 2015; Walsh *et al.,* 2020). A small number of neuroimaging studies have further explored neurobiological mechanisms underlying changes in hippocampal volumes, in an attempt to bridge the gap with the animal literature. PA-related changes in CRF have been related to increased hippocampal perfusion (Chaddock-Heyman *et al.,* 2016; Maass *et al.,* 2015), blood volume (Pereira *et al.,* 2007) and measures of tissue density (Kleemeyer *et al.,* 2016), supporting findings in rodents. A recent study showed, however, that the exercise-related increases in hippocampal volumes may be better explained by increases in myelin instead of vasculature (Thomas *et al.,* 2016). Further research in humans is required to unravel *how* PA exerts its effect on the brain and cognition.

1.4 Aims of this thesis

The overall aim of this thesis is to **provide a better understanding of the relationships between measures of physical activity and cardiorespiratory fitness and markers of cognition and anterior hippocampal structure in young adolescents**.

To address this aim, I use data from a large-scale cluster-randomised controlled trial, described in chapter 2, that examined the effect of a VPA intervention —'Fit to Study —delivered during school physical education (PE) lessons on maths attainment, mental health, cognitive performance, cardiorespiratory fitness and brain health.

Measuring PA, CRF, the brain and cognitive performance is not trivial, nor is the analysis of cluster randomised trials. Chapter 3 provides a brief evaluation and discussion of various methodological considerations when conducting a cluster randomised controlled trial in the field of PA and (cognitive) neuroscience.

The first aim of this thesis is to assess whether PA and CRF are associated with cognitive performance in young adolescents and whether such relationships vary with participants' sex and socioeconomic status (chapter 4), using baseline data of Fit to Study. To extend the current knowledge of such relationships, I additionally use a holistic framework to explore whether a physical phenotype, that includes measures of PA and CRF, is associated with measures of cognitive performance.

To provide insight into the neural correlates of PA and CRF, I next assess (chapter 5) whether these measures are associated with anterior hippocampal volumes. Simply observing inter-individual differences in tissue volume, however, provides little insight into the cellular processes driving these effects. I therefore additionally explore whether a physically active phenotype covaries with MRI measures of (micro)structure in the anterior hippocampus.

While the analyses corresponding to the first two aims provide insight into whether PA and CRF are associated with inter-individual differences in neurocognitive health in young adolescents, their cross-sectional nature limits conclusions about directionality. To assess whether increasing PA is a feasible and effective method for improving neurocognitive health in this age group, I use longitudinal data from the Fit to Study randomised controlled trial. In particular, I first examine whether the one-academic year VPA intervention enhances cognitive performance and CRF (chapter 6). I then assess whether the intervention has a beneficial effect on hippocampal MRI markers (chapter 7).

This page is intentionally left blank.

2

Protocol for a cluster randomised trial of a physical activity intervention in young adolescents: Fit to Study

Contents

2.1 Introduction . . . 30
2.1.1 Hypotheses . . . 32
2.2 Methods . . . 33
2.2.1 Design . . . 33
2.2.2 Independent evaluation . . . 35
2.2.3 Sample selection and recruitment . . . 35
2.2.4 Sample characteristics . . . 37
2.2.5 Intervention . . . 38
2.2.6 Intervention fidelity . . . 41
2.2.7 Blinding . . . 43
2.3 Outcome measures . . . 44
2.3.1 Academic achievement . . . 44
2.3.2 Cognitive and mental heath measures . . . 44
2.3.3 Physical activity and cardiorespiratory fitness . . . 52
2.3.4 Brain imaging sub-study . . . 53
2.4 Data analysis . . . 66
2.4.1 Sample size calculation . . . 67
2.4.2 Statistical analysis . . . 67
2.5 Discussion . . . 68

In this chapter, I describe the study protocol of the Fit to Study project: A cluster randomised controlled trial examining the effect of a vigorous physical activity intervention on academic performance, cardiorespiratory fitness, cognition, mental health and the brain in young adolescents. The methods described in this chapter form the basis for the remaining chapters of this thesis.



The core of this chapter is adapted from:

- Wassenaar, T. M., Wheatley, C. M., Beale, N., Salvan, P., Meaney, A., Possee, J. B., Atherton, K. E., Duda, J. L., Dawes, H. & Johansen-Berg, H. Effects of a programme of vigorous physical activity during secondary school physical education on academic performance, fitness, cognition, mental health and the brain of adolescents (Fit to Study): Study protocol for a cluster-randomised trial. *Trials* **20,** 189 (2019)

2.1 Introduction

Early adolescence is a period of dynamic neurobiological and psychological change (Foulkes & Blakemore, 2018) and provides a foundation for future health (Dahl *et al.,* 2018; Sawyer *et al.,* 2012). Understanding how health and education policies might steer the developmental course in a positive direction is therefore important (Dahl *et al.,* 2018).

Converging lines of research suggest that physical activity (PA) interventions have the potential to stimulate positive brain changes, improve cognitive function and boost academic attainment in this age group. Neuroscience studies using animal models have indicated that, at the molecular level, aerobic exercise increases levels of growth factors responsible for synaptic plasticity, particularly in the hippocampus (Gomez-Pinilla & Hillman, 2013; Voss *et al.,* 2013). At the cellular level, increased growth factor production is thought to promote the development of new blood vessels and neurons, and their integration into existing networks of cells in this region (Cotman *et al.,* 2007; Voss *et al.,* 2013). Neuroimaging studies in adults (Hillman *et al.,* 2008), and increasingly

in children, have found support for PA and fitness-related changes in brain function and structure (Donnelly *et al.,* 2016; Hillman *et al.,* 2017; Valkenborghs *et al.,* 2019). For instance, hippocampal volume is greater in higher fit children (Chaddock *et al.,* 2010a) and adolescents (Herting & Nagel, 2012) and may mediate the relationship between fitness level and memory outcomes (Chaddock *et al.,* 2010a).

Intervention studies in schools have found some evidence that PA programmes can lead to improved cognitive performance, particularly in the domains of attention and executive functions (working memory, cognitive flexibility, and inhibitory control; Álvarez-Bueno *et al.,* 2017b; De Greeff *et al.,* 2018; Moreau *et al.,* 2017; Vazou *et al.,* 2019). Separately, there are associations between objectively-measured PA and academic attainment among adolescents (Booth *et al.,* 2013) and indications that school-based PA interventions can improve educational outcomes in this age group (Álvarez-Bueno *et al.,* 2017a; Donnelly *et al.,* 2016). The balance of evidence among adolescents —based on a small number of studies —favours positive associations between PA, fitness, cognition and attainment (2018 Physical Activity Guidelines Advisory Committee, 2018; Álvarez-Bueno *et al.,* 2017a; Álvarez-Bueno *et al.,* 2017b; Biddle *et al.,* 2019; Lees & Hopkins, 2013), although the variety of study designs, including different types, intensities and durations of PA, make firm conclusions difficult to draw (Wassenaar *et al.,* 2020).

The benefits of PA for physical and mental health are well known (McMahon *et al.,* 2017; World Health Organization, 2018), yet PA levels among young adolescents are low (Sallis *et al.,* 2016), with approximately 20% of adolescents meeting PA guidelines of at least an hour per day of moderate-to-vigorous PA (MVPA), and continue to decline during early adolescence (Inchley *et al.,* 2016). Adolescence is also a peak time for the onset of mental ill-being, including vulnerability to low self-esteem (Harter, 1993; Paus *et al.,* 2008). There is some evidence that physical health and mental well-being and ill-being are also linked with academic attainment (Deighton *et al.,* 2018; Veldman *et al.,* 2014), but whether these factors act as mediators in the relationship between PA and

academic attainment is currently unclear.

School is a recommended setting for promoting PA among adolescents, including those from lower socioeconomic backgrounds, because they spend a large proportion of their time at school and interventions can be delivered without relying on support from families (Kriemler *et al.,* 2011). There is evidence that physical education (PE) interventions to increase activity are an effective method of improving cognitive function and academic performance (Álvarez-Bueno *et al.,* 2017b; Kriemler *et al.,* 2011). However, less than 50% of a typical PE lesson is spent in MVPA; the type of activities that noticeably accelerate the heart rate (Fairclough & Stratton, 2005; Hollis *et al.,* 2017). Short bouts of high-intensity or vigorous physical activity (VPA), during which individuals work at around 80% of their maximum heart rate for between 45 seconds and four minutes, have been shown to deliver equivalent fitness benefits to longer, lower-intensity workouts (Buchheit & Laursen, 2013a; Buchheit & Laursen, 2013b; Costigan *et al.,* 2016). This type of activity is brief enough to be incorporated into PE lessons without disrupting curriculum delivery, and there is evidence that it can improve adolescent fitness (Costigan *et al.,* 2016; Logan *et al.,* 2014).

The primary aim of the Fit to Study trial was to investigate the impact of a one-year VPA intervention delivered during regular school PE on maths attainment. To date there is little school-based evidence to test whether relationships between PA and academic achievement among young adolescents are mediated or moderated by fitness, cognitive function, mental health outcomes and/or the brain. The secondary aim of this trial was therefore to investigate whether the intervention improved fitness, cognitive performance, mental health outcomes, and changes brain structure and function, compared to usual PE lessons delivered over a school year.

2.1.1 Hypotheses

Based on findings from previous studies, several hypotheses were specified prior to the start of the study. The primary hypothesis has been addressed by the independent

evaluator (see section 2.2.2: Independent evaluation).

Primary hypothesis

- A programme of VPA, delivered during PE lessons for one academic year (10-months), will improve maths attainment at the end of the intervention in the intervention group relative to the control group.

The secondary hypotheses have been addressed by the research team. For completeness, all hypotheses are listed below and the ones that are relevant to this thesis are formatted in **bold**.

Secondary hypotheses

- The intervention will increase VPA during PE in the intervention group relative to the control group.
- **The intervention will increase fitness levels of pupils in the intervention group relative to the control group.**
- **The intervention will improve cognitive performance (processing speed, memory and executive functions) in the intervention group relative to the control group.**
- The intervention will improve physical and mental health outcomes in the intervention group relative to the control group.
- **The intervention will change brain structure and function in the intervention group relative to the control group.**

2.2 Methods

2.2.1 Design

The study was a parallel group, superiority cluster randomised efficacy trial of a one academic year (10-months) VPA intervention versus control. Clusters were secondary state schools in the South/Mid-England and participants were Year-8 pupils within those schools. The intervention was incorporated into PE lessons, to minimally disturb the curriculum and provide a scalable intervention that could be easily implemented in regular school PE lessons. Trial data was collected at baseline (end of Year 7, pupils

Figure 2.1: SPIRIT figure of Fit to Study cluster randomised controlled trial.

Activity or assessment	School / individual level	All pupils / subset	Study period			
			Admission to trial / pre-study	Baseline / Allocation	Post-intervention	12-month follow-up
			$-t_1$	t_0	t_1	t_2
Eligibility screen	School	All	✓			
Memorandum of understanding	School	All	✓			
Opt-out consent form	Individual	All	✓			
Randomisation	School	All		✓		
Intervention						
Intervention						
additional 10 min VPA per PE lesson	School	All		⟶	⟶	
Control						
PE as usual	School	All				
Assessments						
Actigraphy						
Physical activity during PE	School PE class	All		✓	✓	
Physical activity during PE	Individual	Sub-set			✓	
Cognitive assessments						
Simple reaction time task	Individual	All		✓	✓	
Modified flanker task	Individual	All		✓	✓	
Visual two-back task	Individual	All		✓	✓	
Colour-shape switch task	Individual	All		✓	✓	
Relational memory task	Individual	All		✓	✓	
Psychosocial measures						
Mental health (SDQ)	Individual	All		✓	✓	
Self-esteem (P-SDQ)	Individual	All		✓	✓	
Sleepiness (PDSS)	Individual	All		✓	✓	
Global health (HBSC)	Individual	All		✓	✓	
PE enjoyment	Individual	All		✓	✓	
Physical activity (HBSC)	Individual	All		✓	✓	
Attitudes / beliefs about physical activity	Individual	All		✓	✓	
Fitness						
20 m Shuttle Run	Individual	Sub-set		✓	✓	
12 min Cooper Run	Individual	Sub-set		✓	✓	
Brain imaging sub-study						
MRI, cognition, mental & physical health, fitness, PA, gait assessments	Individual	Sub-set		✓	✓	✓

Abbreviations: HBSC = Health Behaviour for School Aged Children, PDSS = Paediatric Daytime Sleepiness Scale, PE = physical education, P-SDQ = Physical Self-Description Questionnaire, SDQ = Strength and Difficulties Questionnaire, PA = physical activity, VPA = vigorous physical activity

aged 11–12), and after 12 months (end of Year 8, pupils aged 12–13). Figure 2.1 presents an overview of the schedule for enrolment, the intervention and the assessments, according to the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) guidelines.

A subset of pupils from a subset of schools also took part in a brain imaging sub-study, which is embedded in the trial. Data for the brain imaging sub-study was collected at baseline (end of Year 7, pupils aged 11–12), after 12 months (end of Year 8, pupils aged 12–13) and 24 months (end of Year 9, pupils aged 13–14).

The trial has been approved by the Central University Research Ethics Committee of Oxford University (Registration No: R48879). The main trial protocol was retrospectively registered at ClinicalTrials.gov on 18th of September 2017 (NCT03286725). At this time, study enrolment was complete and data acquisition was in progress. The brain imaging sub-study protocol was retrospectively registered on 19th of July 2018 (NCT03593863). At this time study enrolment and data acquisition were both still in progress. The primary measure, attainment in maths, was recorded separately on the ISRCTN registry (15730512).

2.2.2 Independent evaluation

The trial was independently evaluated by NatCen Social Research. NatCen performed the randomisation and the trial's power calculation, and carried out the primary intention-to-treat analysis, comparing maths performance between intervention and control schools. The protocol and statistical analysis plan for the independent evaluation have been published separately (on the Education Endowment Foundation's [EEF] website: EEF Fit to Study, Husain *et al.,* 2019).

2.2.3 Sample selection and recruitment

A total of 106 schools were recruited and assessed for eligibility (Figure 2.2). To be eligible, schools had to:

- be secondary state- or academy schools, but not grammar schools,
- be mixed or single-gender,
- have a proportion of pupils eligible for free school meals (eFSM) —a proxy for socioeconomic status / deprivation (Hobbs & Vignoles, 2007; Ilie *et al.,* 2017; Taylor, 2018) —preferably more than 15%, which is the average for England (at the time of recruitment, Husain, 2017),

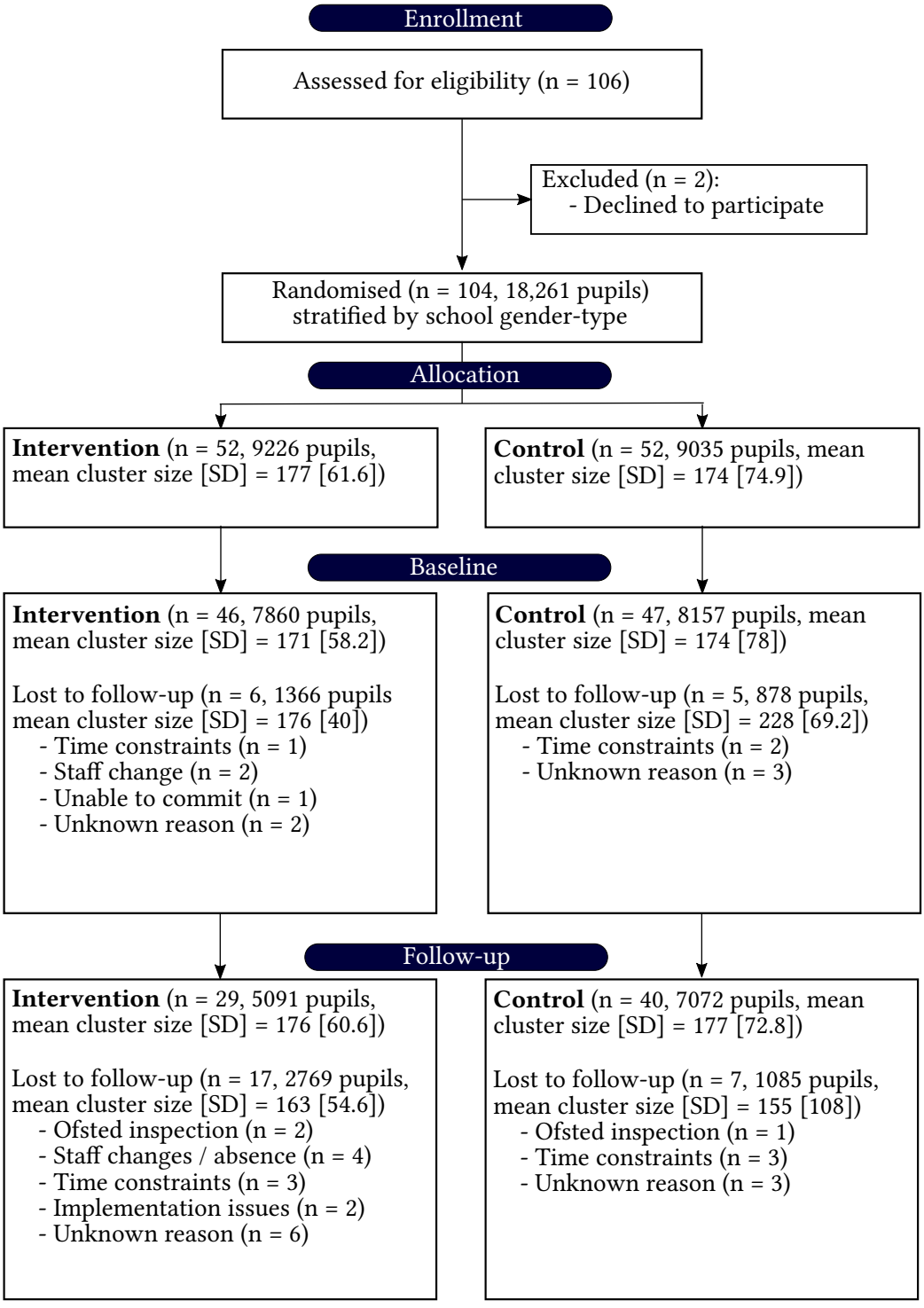


Figure 2.2: CONSORT flow diagram. A total of 106 schools were assessed for eligibility to participate in the trial. Two schools declined to participate and the remaining 104 schools were randomised into an intervention and control group. Prior to baseline assessments, 11 schools dropped out. Another 24 schools were lost to follow-up, for a variety of reasons, including ofsted inspections, staff changes, time constraints, implementations issues and unknown/no reasons.

- be located within a pre-defined set of local authorities, encompassing the following geographical locations: Greater London; Thames Valley; Southampton and Portsmouth; Bristol and Bath; Birmingham and Coventry; Cheltenham/Gloucester; Luton, Bedford Milton Keynes,
- have Year-7 pupils that will move on to Year 8 at the start of the intervention,
- deliver PE as part of their curriculum,
- sign an agreement to send opt-out consent forms to parents/carers of Year-7 pupils and inform the research team of pupils who have opted-out of data storage.

Following the eligibility assessment, two schools declined to participate. Of the remaining 104 schools, all pupils (n = 18,299) in Year 8 at the start of the intervention (academic year 2017–2018) were part of the trial, given that project activities were classed as part of regular school activities. However, pupils who opted-out of data storage (n = 38) were not required to participate in any of the trial's assessments, unless decided otherwise by the PE teacher. The National Foundation for Educational Research (NFER) led recruitment, supported by Oxford Brookes University. NFER initially sent recruitment materials to 1,125 schools, and, to improve response numbers, widened their search area to include another 223 schools. Recruitment presentations were given at County PE Teacher Conferences organised by the Oxfordshire Sports Partnership and the Buckinghamshire and Milton Keynes Sport and Activity Partnership. Those schools that wished to participate and met the inclusion criteria provided the names, dates of birth, sex, eFSM status and unique pupil numbers (UPN) of all Year-7 pupils. Schools were considered formally recruited upon transfer of the data to NFER.

2.2.4 Sample characteristics

Table 2.1 provides an overview of sample characteristics throughout the trial. A total of 18,261 pupils from 104 schools (mean age on September 1st 2017 = 12.5, standard deviation [SD] = 0.3; no. females = 10,014; no. eFSM = 3,068) were randomised to intervention and control groups, stratified by school gender-type. Eleven schools (6

intervention, 5 control) dropped-out post-randomisation but prior to baseline assessments, because of time constraints (n = 3), staff changes (n = 2), inability to commit (n = 1) or unknown reasons (n = 4). These schools were unaware of their treatment allocation at the time of drop-out. A comparison between schools that dropped-out and remained part of the trial showed no significant differences in any of the school-level or pupil level demographics (all $p > 0.05$, details in are provided in Appendix B.1.1). The remaining 93 schools were available for baseline assessments and are considered the *main trial* baseline sample for the remainder of the thesis.

Attrition is common in (cluster) RCTs (Dumville *et al.*, 2006; Hindmarch *et al.*, 2015), and 15 schools (12 intervention, 3 control) were lost to follow-up. Schools withdrew for a variety of reasons, including the time constraints of PE teachers delivering the intervention and the logistical burden of testing. Of the 78 schools that remained part of the trial, nine withdrew from assessments of secondary measures (5 intervention, 4 control) and discontinued the intervention. These schools were, however, willing to complete the primary outcome assessment, and were therefore classified as "*evaluation only*". Of 69 *active* schools and 9 *evaluation-only* schools, five active schools were not available for assessment of the primary outcome. The primary outcome was thus administered to pupils in 73 schools; demographics of this sample are provided in NatCen's evaluation report (Husain *et al.*, 2019). Secondary measures were collected in the 69 *active* schools (shown in Figure 2.2).

2.2.5 Intervention

The intervention consisted of a one-academic year (September 2017–June 2018) PA programme delivered by PE teachers during regular Year-8 PE lessons. In the summer term prior to the trial, Year-8 PE staff from all intervention schools were invited to attend a two-hour face-to-face training session delivered by a team from Oxford Brookes University who designed the intervention. Those unable to attend were asked to join an interactive two-hour online training session delivering the same content. Teachers unable to take part during the summer term attended training at the start of the autumn

Table 2.1: Characteristics of schools and participants throughout the trial.

	Randomisation			Baseline		Post-intervention	
	Overall	**Intervention**	**Control**	**Intervention**	**Control**	**Intervention**	**Control**
School level							
No. schools	104	52	52	46	47	29	40
Gender-status, no. (%)							
Co-ed	84 (80.8)	42 (80.8)	42 (80.8)	37 (80.4)	38 (80.9)	23 (79.3)	32 (80.0)
Female	18 (17.3)	10 (19.2)	8 (15.4)	9 (19.6)	8 (17.0)	6 (20.7)	7 (17.5)
Male	2 (1.9)	0 (0)	2 (3.8)	0 (0)	1 (2.1)	0 (0)	1 (2.5)
School type, no. (%)							
Comprehensive to 16	21 (20.2)	10 (19.2)	11 (21.2)	7 (15.2)	9 (19.1)	6 (20.7)	9 (22.5)
Comprehensive to 18	73 (70.2)	37 (71.2)	36 (69.2)	34 (73.9)	33 (70.2)	18 (62.1)	26 (65.0)
Secondary Modern	10 (9.6)	5 (9.6)	5 (9.6)	5 (10.9)	5 (10.6)	5 (17.2)	5 (12.5)
Establishment type , no (%)							
Academy - Sponsor Led	18 (17.3%)	9 (17.3%)	9 (17.3%)	7 (15.2%)	9 (19.1%)	4 (13.8%)	8 (20.0%)
Academy Converter	3 (2.9%)	2 (3.8%)	1 (1.9%)	2 (4.3%)	1 (2.1%)	0 (0%)	1 (2.5%)
Academy Converter - M	35 (33.7%)	17 (32.7%)	18 (34.6%)	16 (34.8%)	15 (31.9%)	10 (34.5%)	15 (37.5%)
Community School	18 (17.3%)	10 (19.2%)	8 (15.4%)	8 (17.4%)	7 (14.9%)	6 (20.7%)	5 (12.5%)
Foundation School	11 (10.6%)	5 (9.6%)	6 (11.5%)	5 (10.9%)	5 (10.6%)	4 (13.8%)	5 (12.5%)
Free School - M.	5 (4.8%)	3 (5.8%)	2 (3.8%)	3 (6.5%)	2 (4.3%)	2 (6.9%)	1 (2.5%)
Voluntary aided school	12 (11.5%)	5 (9.6%)	7 (13.5%)	4 (8.7%)	7 (14.9%)	3 (10.3%)	4 (10.0%)
Voluntary controlled	2 (1.9%)	1 (1.9%)	1 (1.9%)	1 (2.2%)	1 (2.1%)	0 (0%)	1 (2.5%)
School size, no. (%)							
Large (>200)	32 (30.8)	18 (34.6)	14 (26.9)	14 (30.4)	13 (27.7)	9 (31.0)	11 (27.5)
Medium (100-200)	60 (57.7)	29 (55.8)	31 (59.6)	27 (58.7)	27 (57.4)	18 (62.1)	24 (60.0)
Small (<100)	12 (11.5)	5 (9.6)	7 (13.5)	5 (10.9)	7 (14.9)	2 (6.9)	5 (12.5)
No. of Year 7 form groups							
Median (range)	7 (3–14)	7 (3–12)	7 (3–14)	6 (3–12)	7 (3–14)	6 (4–12)	7 (3–14)
Missing[a], no. (%)	2 (1.9)	1 (1.9)	1 (1.9)	1 (2.2)	1 (2.1)	1 (3.4)	1 (2.5)
Socioeconomic status							
eFSM[b], mean (SD), %	17.4 (10.9)	17.2 (8.62)	17.7 (12.8)	17.0 (9.00)	17.7 (12.9)	16.7 (9.88)	16.9 (12.9)
IMD[c], med (range), decile	5 (1–10)	5 (1–10)	6 (1–10)	5 (1–10)	6 (1–10)	5 (1–10)	7 (1–10)
Location[d], no. (%)							
Rural town and fringe	9 (8.7)	5 (9.6)	4 (7.7)	5 (10.9)	5 (10.9)	3 (10.3)	4 (10.0)
Rural village	1 (1.0)	0 (0)	1 (1.9)	0 (0)	0 (0)	0 (0)	0 (0)
Urban city and town	45 (43.3)	22 (42.3)	23 (44.2)	19 (41.3)	19 (41.3)	13 (44.8)	18 (45.0)
Urban conurbation	49 (47.1)	25 (48.1)	24 (46.2)	22 (47.8)	22 (47.8)	13 (44.8)	18 (45.0)
Pupil level							
No. pupils	18261	9226	9035	7860	8157	5091	7072
Age[e], mean (SD), yr	12.5 (0.3)	12.5 (0.3)	12.5 (0.29)	12.5 (0.3)	12.5 (0.29)	12.5 (0.29)	12.5 (0.29)
Females, no. (%)	10014 (54.8)	5179 (56.1)	4835 (53.5)	4466 (56.8)	4495 (55.1)	3026 (59.4)	3952 (55.9)
eFSM, no. (%), yes	3068 (16.8)	1495 (16.2)	1573 (17.4)	1243 (15.8)	1422 (17.4)	797 (15.7)	1163 (16.4)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, IMD = index of multiple deprivation, M. = mainstream, med = median, no. = number, SD = standard deviation

[a] A row indicating *missing* values was added only when missing values were present.

[b] Percent eFSM indicates the percentage of enrolled pupils that are eligible for free school meals within a school.

[c] Index of multiple deprivation was collected using the school's postcode, through Government records: http://imd-by-postcode.opendatacommunities.org/imd/2019

[d] Location was assessed using Government records: https://www.gov.uk/government/collections/rural-urban-classification

[e] Age at baseline (1 September 2017)

term. During these sessions, the intervention was described to teachers.

The intervention required teachers to deliver specific elements of additional VPA, over and above normal VPA, during all Year-8 PE lessons, for the whole school year. Four minutes of VPA are incorporated into the PE lesson as part of a warm-up and three two-minute infusions per hour of PE are incorporated into the main PE class. In this trial, VPA was defined as activities that raise the heart rate (HR) to 71–85% of the maximum HR (HR_{max}), whereas moderate PA raises the HR to 60–70% of HR_{max}. The intervention elements were designed to be easy to incorporate into PE lessons, to maximise the training effect, and to minimise the risk of injury with a warm-up that progresses from light to moderate-intensity activities and includes bursts of VPA.

- **Warm-up** Teachers were asked to deliver an active ten-minute warm-up at the start of each lesson, beginning with light-intensity movements (e.g. wrist rotations; leg swings) and moving through moderate-intensity activity (e.g. arms rotations; walking with high knee-raise) to incorporate two periods of VPA (e.g. vigorous arm sprints; running on the spot) totalling four minutes. Fifteen young adolescents piloted the warm-up and the infusions while wearing chest strap Polar heart rate monitors, which indicated that heart rates were in the vigorous zone for the specified time during all the intervention elements.

- **Infusions** In addition, three two-minute infusions of VPA would be incorporated into the main PE lesson, assuming a one-hour lesson duration. If lessons were shorter or longer than one hour then the number of infusions was scaled up or down accordingly. Teachers were provided with three example infusions: infusion 1 consisted of fast arm rotations (the aim is to build the intensity by speeding up arm movements concluding with two maximum effort sprints for counts of ten); infusion 2 comprised squats and lunges with an emphasis on technique and a full range of motion (to work the muscles to the maximum, concluding with a maximum effort sprint on the spot for a count of ten); and infusion 3 sprinting on the spot (this must be completed at a vigorous intensity with the two "on the spot" sprint sessions lasting counts of ten), with active recovery periods where necessary during each two-minute session.

Teachers were also invited to create their own warm-up and/or infusions to provide similar levels of VPA to fit their lesson requirements. Researchers explained the rationale for the intervention and discussed implementation strategies.

ACTIVE INGREDIENTS

The intervention aimed to increase the amount of VPA (at 71–85% of HR_{max}), by incorporating into the PE lesson:

- **WARM-UP:** a ten-minute warm-up, incorporating two periods of VPA (4 minutes)
- **INFUSIONS:** three two-minute infusions (VPA) per one-hour PE lesson

Teachers were given access to video demonstrations of the warm-up and infusion material on the trial's website (https://www.fit-to-study.org/interventions/). Heads of PE at intervention schools received emails reminding them to deliver the intervention in September and December 2017, and February 2018. Researchers visited 30 intervention schools between September and December 2017 to answer questions and offer support and advice. PE staff were invited to join an online intervention discussion forum and schools were offered the chance to have intervention lessons filmed. Researchers were also available to answer email and phone queries.

Control schools were not given training and were asked to deliver PE as usual. During the intervention period, heads of PE in these schools received emails in September 2017 and February 2018 with information about testing. School participation was incentivised to minimise drop-out and encourage compliance. Both intervention and control schools received £500 upon completion of the trial, following the academic attainment assessment.

2.2.6 Intervention fidelity

Intervention fidelity, or implementation fidelity, refers to the degree to which an intervention was delivered as intended (also known as treatment integrity, see e.g. Breitenstein *et al.*, 2010; Moncher & Prinz, 1991; Yeaton & Sechrest, 1981). Measurement of

intervention fidelity is imperative in intervention studies, particularly those involving complex behavioural interventions, to ensure the reliability and validity of the intervention and accurate interpretation of treatment effects (Bellg *et al.,* 2004; Gearing *et al.,* 2011; Mars *et al.,* 2013). Intervention fidelity is a multidimensional construct, the core elements and measurement of which have been discussed ever since its introduction in the 1980s (e.g. Bellg *et al.,* 2004; Breitenstein *et al.,* 2010; Dane & Schneider, 1998; Darrow, 2013; Lambert *et al.,* 2017; Mars *et al.,* 2013; O'Donnell, 2008; Schoenwald & Garland, 2013; Schultes *et al.,* 2015). Despite much debate, the importance of measuring intervention fidelity is clear: fidelity mediates study outcomes, with higher fidelity associated with greater treatment effects (Mars *et al.,* 2013; Nelson *et al.,* 2012). Intervention fidelity, or the lack thereof, could help explain research findings. For instance, the absence of an intervention effect may be due to low implementation fidelity rather than actual ineffectiveness of the treatment (Bellg *et al.,* 2004; Breitenstein *et al.,* 2010; Mars *et al.,* 2013). Failure to measure or consider intervention fidelity may result in researchers wrongly concluding that the intervention itself was ineffective.

While reliable measurement and accurate reporting of intervention fidelity is crucial for establishing intervention effectiveness and study replication, it is costly, time-consuming (Perepletchikova *et al.,* 2007) and not common practice in PA research (Wassenaar *et al.,* 2020). In the Fit to Study trial, we assessed intervention fidelity with a broad set of measures:

- **Teacher logs** All Year-8 PE teachers in intervention schools were asked to keep a written log of VPA achieved in each PE lesson. Specifically, they were asked to record whether the class completed four minutes of VPA during the warm-up and to mark how many infusions were delivered. Teachers were informed that for a teacher-created warm-up to be marked as completed it must have lasted 10 minutes, and progressed from light intensity small movement activities to larger moderate intensity activities with short bursts of vigorous intensity activities. For a teacher-created infusion to be marked as completed it must have

lasted 2 minutes, and included two short bursts of VPA (described in section 2.2.5: Intervention).

- **Objective PA during PE** Researchers visited a convenience sample of intervention and control schools (at least 50% of pupils per school) during the intervention period to take objective measurements of PA during PE using the Axivity AX3, a wrist-worn tri-axial accelerometer designed by Open Lab, Newcastle University, UK. Class average minutes of VPA during active lesson time were calculated using established cut-points (Phillips *et al.,* 2013) and standardised to minutes per hour for comparison purposes.

- **Pupil reported compliance** A subjective measure of compliance was collected from participants from all schools during post-intervention testing using a brief three-item survey asking whether (1) PE lessons started with a warm-up, (2) PE lessons included bursts of VPA that raised their heart rate and made them feel out of breath, and (3) they took part in warm-ups or VPA bursts if asked by PE teachers.

- **Post-intervention school survey** NatCen conducted an online survey among PE teachers in intervention schools post-intervention (June –July 2018) to obtain an estimate of the percentage of Year-8 PE lessons in which the intervention was delivered as intended. This retrospective assessment of intervention compliance was conducted in response to low completion rates of teacher logs (Husain *et al.,* 2019).

2.2.7 Blinding

Researchers assessing secondary outcomes were blinded to treatment status of schools, but those who visited schools to collect fidelity measures and provide top-up intervention training were not blinded. Schools were informed of their status following baseline measurements, so that intervention school teachers could receive training and deliver the intervention.

2.3 Outcome measures

Here, we briefly describe the outcome measures that are relevant to this thesis. For a full account of the outcome measures, we refer to Appendix B.2 and the protocol paper (Wassenaar *et al.,* 2019a).

2.3.1 Academic achievement

The primary outcome measure was academic attainment, assessed by the Progress Test in Mathematics (PTM), Level 13 (GL Assessment, 2018). To reduce the test burden, NatCen administered the test to a random selection of half of all form groups per school, post-intervention only (Bartasevicius & Jabin, 2018).

2.3.2 Cognitive and mental heath measures

Cognitive and mental health measures were collected once at baseline and once during the summer term at the end of the intervention. All cognitive assessments and health questionnaires were delivered via the web-browser, which allows for the collection of large and broadly generalisable datasets with minimal resources, without affecting the validity of the measures or compromising data quality (Germine *et al.,* 2012; Gosling *et al.,* 2004; Templer & Lange, 2008). The assessments were programmed in JavaScript using jsPsych (De Leeuw, 2015), a JavaScript library for running behavioural experiments in a web browser that allows for accurate response time measurements (De Leeuw & Motz, 2016). Participants were given a personal ID number to access the assessments. Upon completion of the assessments, summary measures (e.g. mean response time) were created and temporarily stored against participants' personal IDs in a secure database before being transferred to a second, secure research database, against private ID numbers to ensure pseudonymisation.

2.3.2.1 Cognitive performance

Cognitive performance was assessed with tests of processing speed, visual relational memory, and executive functions (working memory, inhibition and cognitive flexibility), that are well-validated in children and sensitive to the effects of PA (Álvarez-Bueno

et al., 2017b; Chaddock *et al.*, 2010a; Chang *et al.*, 2012; Chen *et al.*, 2014b; Hillman *et al.*, 2014). Teachers were instructed to ask pupils to complete the tasks at home, as pilot tests in three schools indicated that the local IT environment (e.g. bandwidth) may not be able to cope with large groups of pupils (>20) completing the tasks at the same time, and it was deemed a logistical burden on schools to set the tasks during school lessons. Nevertheless, some schools decided to complete the tasks during school time as they felt this was easier to manage. The five tasks are briefly described below. More details are provided in Appendix B.2.1.1.

Processing speed A simple reaction time task (Figure 2.3) was used to assess processing speed. Participants were instructed to respond as quickly as possible (pressing the spacebar) to the presentation of a triangle on the screen (40 trials). Responses between 100-3,000 ms were considered valid. The average reaction time across valid trials was the outcome measure.

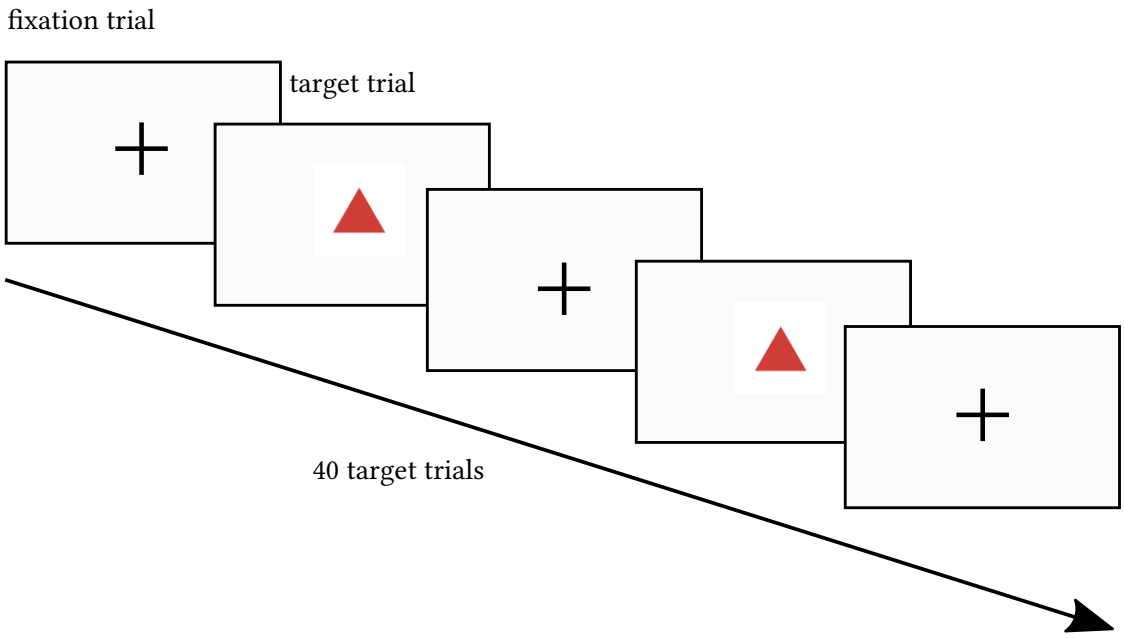


Figure 2.3: Reaction time task paradigm. Participants had to respond as quickly as possible to the appearance of a red cross on the screen.

Relational memory A modified relational memory task (Chaddock *et al.*, 2010a) was used to assess visual relational memory performance (Figure 2.4). In the task, participants were presented with two blocks of 18 study trials (i.e. events) containing scene, face and object stimuli. Within each study trial, each type of stimulus was presented individually and sequentially (1,300 ms). Participants were instructed to memorise the stimuli within each study trial as belonging to a single event. During the recognition phase, participants were shown 36 test trials, each containing a place, object and face stimulus, half of which were intact (i.e. the stimuli had previously been presented in the same study trial). Participants were instructed to indicate whether all stimuli were from the same event, or whether one of the stimuli did not belong to the event. Responses within 200–15,000 ms were considered valid. Accuracy on valid trials was the outcome measure.

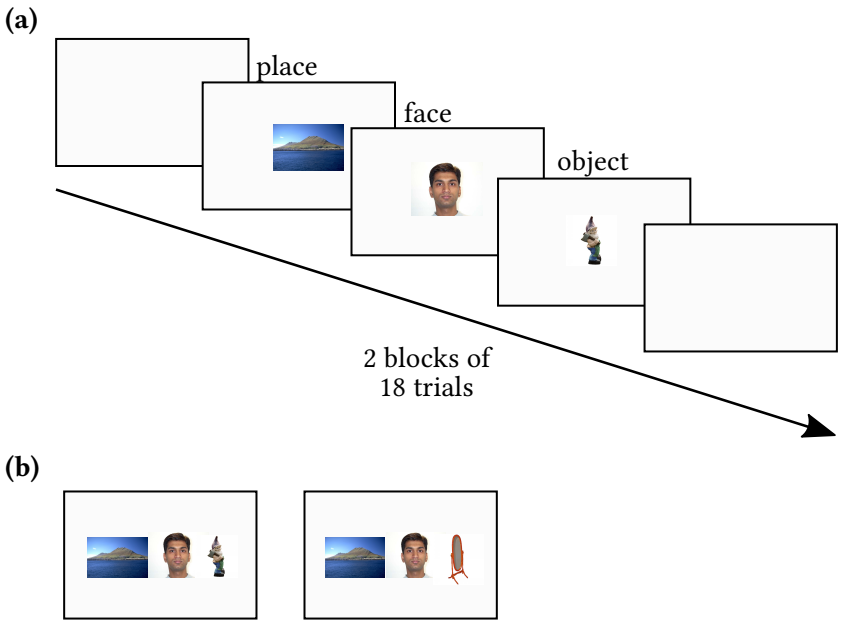


Figure 2.4: Relational memory paradigm. (*a*) In the *encoding phase*, participants were presented with two blocks of 18 study trials containing scene, face and object stimuli. They were instructed to memorise the stimuli as belonging to a single event. (*b*) In the *recognition phase* participants had to indicate whether all items were from the same event or whether one of the items did not belong to the event.

Cognitive flexibility A modified version of the colour-shape switch task (Hillman *et al.,* 2014) was used to assess cognitive flexibility (Figure 2.5). Participants were instructed to learn a set of response actions that were related to the colour or shape of a centrally presented cartoon character: press the "F" key if the character was blue or a circle and the "J" key if the character was green or square. The direction of the character's arms indicated whether to pay attention to its shape or colour. During homogeneous trials (2 blocks of 32 trials), participants had to pay attention to either the colour or the shape. During heterogeneous trials (2 blocks of 64 trials), participants had to flexibly shift their attention towards either the character's shape or colour based on the direction of the character's arms. The direction of the character's arms switched every two trials, giving rise to switch trials and non-switch trials. Responses within 200–6,000 ms were considered valid. Accuracy and average reaction times on switch and non-switch valid trials were recorded.

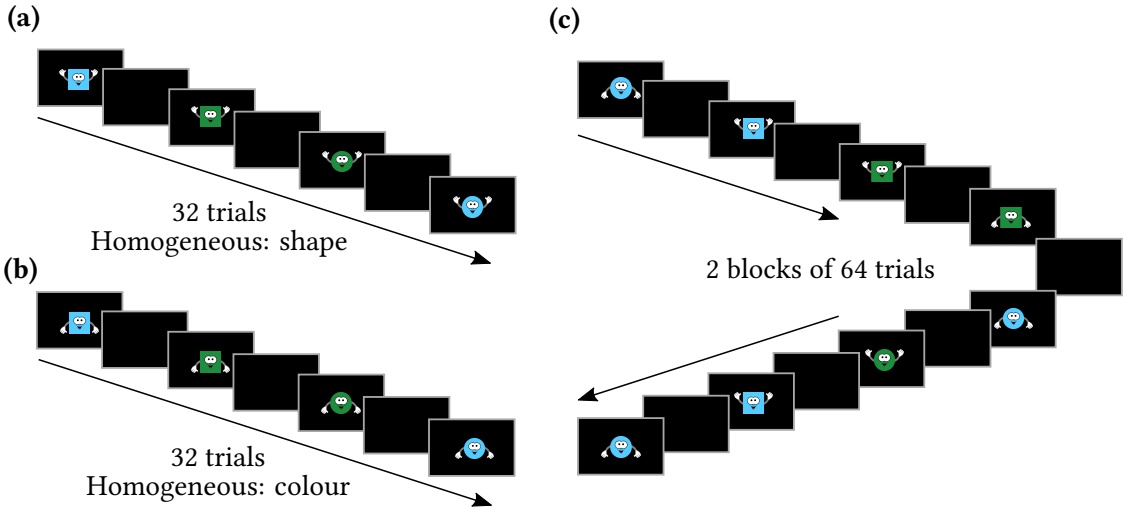


Figure 2.5: Colour-shape switch task paradigm. (*a*) In the homogeneous *shape* condition (here: arms up), participants had to pay attention to the character's shape and press the "F" key if the character was a circle or the "J" key if the character was a square, (*b*) in the homogeneous *colour* condition (arms down), participants had to focus on the character's colour and press the "F" key if the character was blue, or the "J" key if the character was green, (*c*) in the heterogeneous *colour-shape switch* condition, participants had to shift their attention to the colour or shape based on direction of the character's arms: shape if the arms were up, colour if the arms were down. The direction of the characters arms *switched* every two trials. On non-switch trials, the participant had to keep their attention towards the same attribute, whereas on switch trials, the participant had to *switch* between attributes. It is assumed that individuals use top-down voluntary control when switching between competing tasks, requiring them to respond more slowly (reaction time) or less accurately (see e.g. Kiesel *et al.*, 2011; Monsell, 2003). Figure is adapted from Hillman *et al.* (2014).

Working memory A modified version of the visual two-back task (Chen *et al.*, 2016; Chen *et al.*, 2014b; Guiney & Machado, 2013) was used to assess working memory performance (Figure 2.6). Participants were shown pseudo-random sequences of weather symbols and instructed to identify as quickly and accurately as possible whether the stimulus on the current trial was identical to the one presented two trials ago. The task consisted of two blocks of 26 trials each and responses within 150–2,500 ms were considered valid. Accuracy and average reaction time on valid trials were recorded.

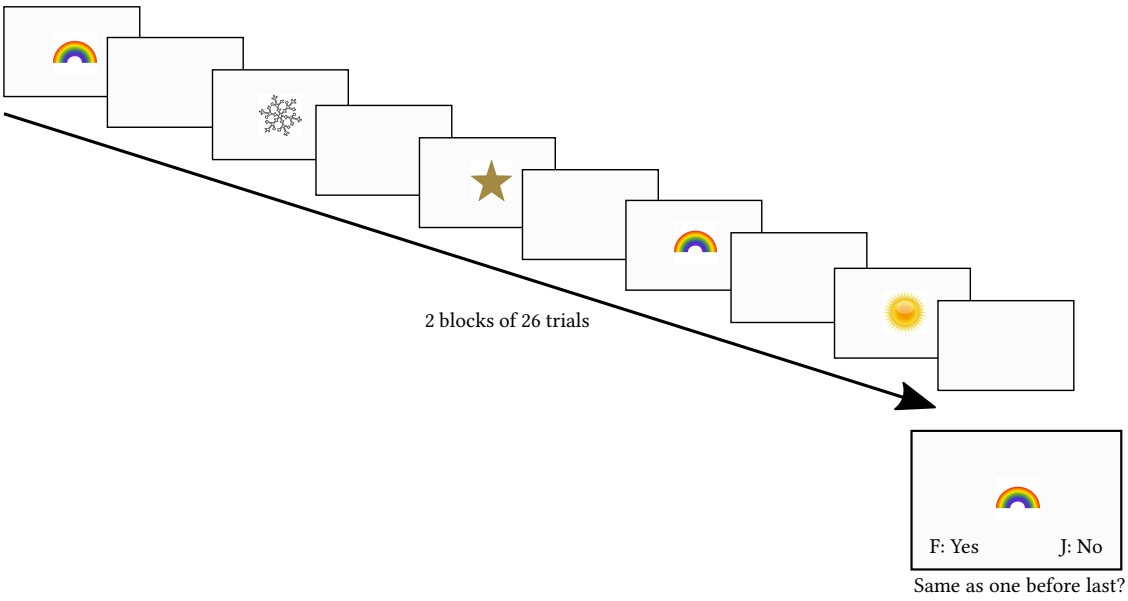


Figure 2.6: Two-back working memory paradigm. Participants were shown pseudo-random sequences of weather symbols and were instructed to identify as *quickly* and *accurately* as possible whether the current stimulus was identical ("F" key) or not ("J" key) to the one that appeared two trials ago.

Inhibitory control A modified Flanker task (Hillman *et al.*, 2014) was used to assess inhibitory control (Figure 2.7). Participants were instructed to respond as quickly and accurately as possible to the direction of a centrally presented target arrow, which was flanked by arrows that were either in the same direction (i.e. congruent) or opposite direction (i.e. incongruent) to the centrally presented target arrow. The task consisted of two blocks of 52 trials each and responses within 200–1,400 ms were considered valid. Accuracy and average reaction times on congruent and incongruent valid trials were recorded.

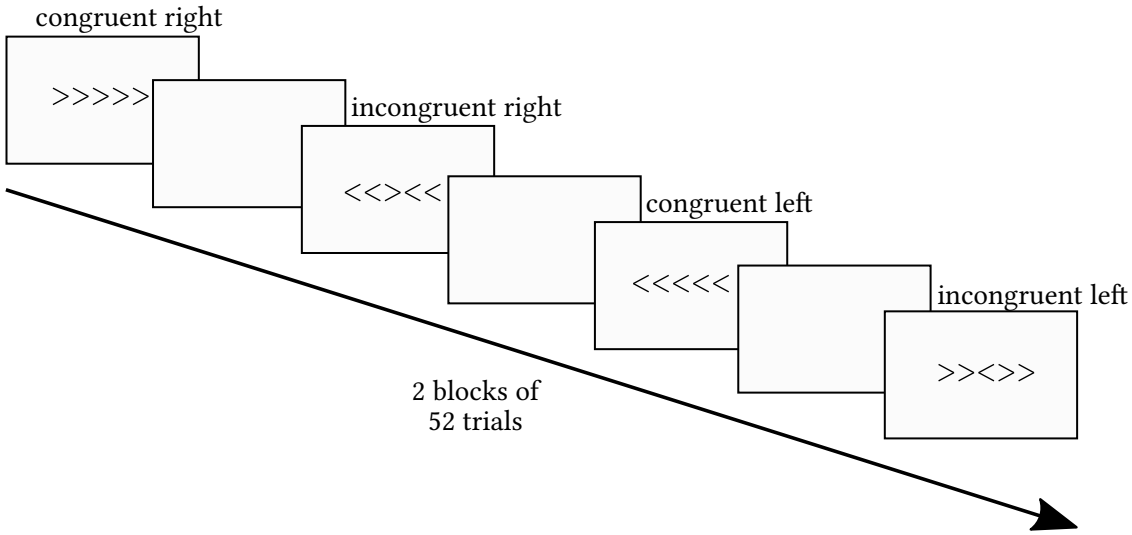


Figure 2.7: Modified Flanker task The task contained four trial types: congruent left, congruent right, and incongruent left and incongruent right. Participants were instructed to press, as quickly and accurately as possible, the "F" key if the centrally located stimulus pointed to the left, and the "J" key if the stimulus pointed to the right.

Randomisation of cognitive assessments The order of cognitive assessments was pseudo-randomised across participants within schools to reduce the probability of task-order effects. All participants started with the reaction time task (i.e. processing speed), followed by the remaining four tasks (for further details: Appendix B.2.1.1).

2.3.2.2 Mental- and physical health

The online mental and physical health questionnaire took approximately 20 minutes to complete and was intentionally kept brief so that it could be easily completed in school. The questionnaire comprised the following core measures:

Physical activity PA was measured with a single item (Scott *et al.*, 2015): *"In the past week, on how many days have you done a total of 60 minutes or more of physical activity, which was enough to raise your breathing rate? This may include sport, exercise, and brisk walking or cycling for fun, or to get to and from places."*.

Habitual physical activity Habitual PA over the past six months was measured with a single item drawn from previous studies (Hagger *et al.*, 2001): *"Thinking about*

the past six months, how often have you been physically active for an hour every day during a typical week in school term?" (1 = never to 7 = always).

Mental health Mental health was assessed with the Strengths and Difficulties Questionnaire (SDQ, Goodman, 1997). The SDQ is a brief behavioural screening questionnaire for 3–16 year olds, and consists of 25 items that comprise five sub-scales (five items each): (1) emotional symptoms, (2) conduct problems, (3) hyperactivity / inattention, (4) peer relationship problems, (5) prosocial behaviour. Each item has three response options: "not true", "somewhat true", "certainly true" (scored as 0, 1, 2). The total difficulties score is computed by adding up the scores of subscales 1 to 4, and ranges from 0–40 (low to high difficulties).

Self-esteem Global and physical self-esteem were measured using, respectively, the global self-esteem scale and physical self-esteem scale of the short version of the Physical Self-Description Questionnaire (P-SDQ, Marsh *et al.,* 2010).

Global health Global health was measured using a single item on a five-point scale drawn from the Health Behaviour in School-Aged Children (HBSC) survey (Inchley *et al.,* 2016): *"In general, how would you say your health is?"* (1 = poor to 5 = excellent).

Sleepiness Daytime Sleepiness was assessed using the Paediatric Daytime Sleepiness Scale (PDSS, Drake *et al.,* 2003)

PE enjoyment PE enjoyment was assessed with a single item (Prochaska *et al.,* 2003): *"I enjoy PE."* (1 = strongly disagree to 7 = strongly agree).

PA-related psychological variables Psychological variables linked with daily MVPA were assessed to investigate whether the intervention changed attitudes and beliefs about daily MVPA and active types of people. An overview of the measures is provided in Appendix B.2.1.2 and (Wassenaar *et al.,* 2019a).

2.3.3 Physical activity and cardiorespiratory fitness

2.3.3.1 Cardiorespiratory fitness

Aerobic fitness was assessed using the 20 m multistage shuttle run (20MSR) test. This well-validated test (Mayorga-Vega *et al.*, 2015) measures participants' maximum aerobic capacity by running back and forth on a 20 meter track. The initial running pace is 8.5 km/h and raises progressively by 0.5 km/min as indicated by a sound ("beep"). The maximum performance was reached when the participant failed to cross the line (within 2 m) at the moment of the beep for two consecutive 20 m distances. The test was performed indoors or outdoors on a non-slip flat surface. The outcome measure was the total number of laps completed. Schools with a policy of not using the multistage fitness test (n = 4, of which 3 control) completed the Cooper Run test (Cooper, 1968) which required pupils to run or walk as far as possible in 12 minutes, around a measured track (between 80 m and 400 m). The total distance run was used as an outcome measure. Fitness measures were collected once at baseline and once during the summer term. The tests were conducted by PE teachers during regular PE lessons to minimise the burden on participants and to avoid disrupting other classes.

2.3.3.2 Physical activity during PE

Objective measures of PA during PE were collected in each school once during regular lessons at baseline (May–July 2017, and some remaining schools in September 2017), and again during lessons in spring term (March–May 2018). Each school was visited at each timepoint to test at least 50% of the pupils in the year group, using the Axivity AX3 accelerometer (Open Lab, Newcastle University, UK). Participants wore the accelerometer on their non-dominant wrist. The devices were attached after participants changed for PE and collected before they changed back into school clothes. Following conversion of raw accelerometer data into 1 s epochs, class average minutes of MPA and VPA during active lesson time were calculated using established cut-points (Phillips *et al.*, 2013) and standardised to minutes per hour.

2.3.4 Brain imaging sub-study

2.3.4.1 Design

A sub-sample of participants was recruited to participate in additional testing to investigate the effect of the PA intervention on brain structure and function. The brain imaging sub-study ran in parallel with the main trial to make optimal use of the intervention. Assessments took place pre-intervention (baseline, t_0), immediately post-intervention (t_1) and 12-months post-intervention (t_2). The brain imaging sub-study consisted of two sub-samples, recruited at different times during the trial:

- **Cohort 1** the first cohort of 61 participants was recruited at the start of the trial and completed brain-study assessments at all three timepoints,
- **Cohort 2** a second cohort of 42 participants was recruited for post-intervention assessments that were similar, but not identical, to the assessments of the first cohort, as well as 1-year follow-up assessments. This cohort was recruited to enable cross-sectional analysis of between-group differences in a second sample and to increase the proportion of pupils from a lower socioeconomic background.

Figure 2.9 presents an overview of the schedule for enrolment, the intervention and the assessments, according to the SPIRIT guidelines (Chan *et al.*, 2013). The sub-study was granted separate ethical approval by the University of Oxford Medical Sciences Inter-Divisional Research Ethics Committee (Registration No: R51313) and was retrospectively registered (19th of July 2018) at ClinicalTrials.gov (ID: NCT03593863).

2.3.4.2 Recruitment

Recruitment for the brain imaging sub-study took place during two waves, resulting in two cohorts. Cohort 1 is central to this thesis, and its recruitment will be described in detail below. Details for cohort 2, including its outcome measures, can be found in Appendix B.1.3 and Wassenaar *et al.* (2019a).

A sample of main trial schools was identified (in May 2017) based on them being (1) within a 75 miles radius from Oxford, and (2) with a lower socioeconomic

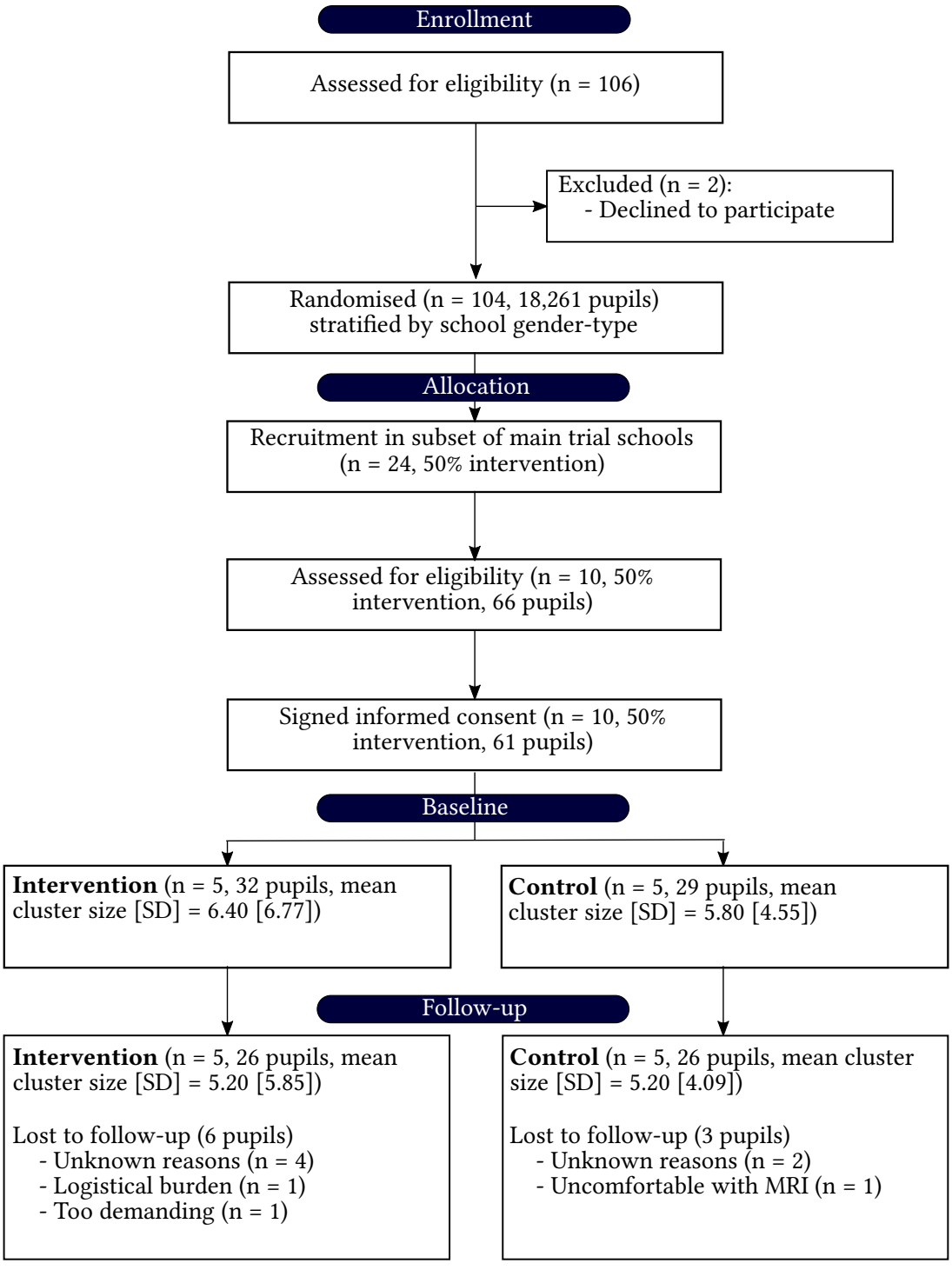


Figure 2.8: CONSORT flow diagram of the brain imaging sub-study (cohort 1). Of 106 main trial schools that were assessed for eligibility, two declined to participate and 104 were randomised into an intervention and control group (n = 52 schools in each arm). Of randomised schools, 24 (50% intervention) were selected for recruitment of brain imaging sub-study participants. A total of 66 participants from 10 schools (50% intervention) were eligible for participation, of which 61 signed an informed consent form. Of 61 participants that were available for baseline assessments, nine dropped out during the course of the study.

background, as indicated by the percentage of free school meal (FSM)-eligible Year-7 pupils and postcode-derived index of multiple deprivation. The trial manager, who was not blinded to treatment status, ensured a balance between intervention and control schools. Selected schools received an advert to send around to all of their pupils, and one of the researchers, blinded to treatment status, visited a sub-sample of four schools to collect expressions of interest from parents and pupils. Individual pupils were considered eligible for the study if the pupil: (1) was part of the Fit to Study main trial, (2) had no learning or motor disabilities, (3) was literate and fluent English-speaking, (4) had no contraindication to MRI, including but not limited to certain metallic implants, metallic parts to the eye or a history of claustrophobia, (5) had no contraindication to a maximal exercise test ($\dot{\mathbf{V}}O_2$max test), as assessed by a modified version of the Physical Activity Readiness Questionnaire (PAR-Q, Canadian Society for Exercise Physiology, 2002), including any known cardiac or respiratory risk (Riebe *et al.,* 2015).

Subjects were enrolled in the study upon completion of an assent form and completion of a consent form by their parent/guardian. Figure 2.8 shows a diagram of the participant flow through the study.

Figure 2.9: SPIRIT figure of Fit to Study brain imaging sub-study

Activity or assessment	School / Individual level	Study period: Main trial pre-study	Admission to MRI study	Baseline	Admission to MRI study	Post-intervention	12-month follow-up
		$-t_2$	$-t_1$	t_0	$t_{0.5}$	t_1	t_2
Randomisation of schools	School	✓					
Eligibility screen	School / Individual		✓		✓		
Opt-in consent	Individual		✓		✓		✓
INTERVENTION							
Intervention							
additional 10 min VPA per PE lesson	School			→			
Control							
PE as usual	School						
ASSESSMENTS							
Physical health							
7-day actigraphy	Individual			✓		✓	✓
Maximal exercise test	Individual			✓		✓	
Gait	Individual			✓		✓	
Resting heart rate	Individual			✓		✓	
Resting blood pressure	Individual			✓		✓	
Height and weight	Individual			✓		✓	✓
Cognitive assessments							
Object-location task	Individual			✓		✓	
Relational memory task	Individual			✓		✓	✓
Colour-shape switch task	Individual			✓		✓	✓
Tower of London	Individual					✓	✓
Implicit association task	Individual			✓			
Mental and physical health questionnaires							
ADHD rating scale IV	Individual			✓		✓	
Pubertal status (PDRS)	Individual			✓		✓	✓
Mood (POMS)	Individual			✓		✓	✓
Sleep (CASQ and SCI)	Individual			✓		✓	✓
Mental health (SDQ)	Individual			✓		✓	✓
General health and behaviour (HBSC)	Individual			✓		✓	✓
Psychological variables linked with daily MVPA	Individual			✓		✓	✓
Brain imaging (MRI)							
T_1-weighted scan	Individual			✓		✓	
Diffusion-weighted scan	Individual			✓		✓	
Quantitative MRI scan	Individual			✓		✓	
Resting-state fMRI scan	Individual			✓		✓	
Arterial spin labelling	Individual			✓		✓	
Angiography	Individual					✓	

Abbreviations: CASQ = Cleveland Adolescent Sleepiness Questionnaire, HBSC = Health Behaviour for School Aged Children, PDRS = Pubertal Development Rating Scale, POMS = Profile of Mood States, SCI = Sleep Condition Indicator, SDQ = Strength and Difficulties Questionnaire, PA = physical activity, VPA= vigorous physical activity

Colour coding represents: ✓ = completed by cohorts 1 and 2, ✓ = completed by cohort 1 only, ✓ = completed by cohort 2 only.

2.3.4.3 Sample description

School- and pupil-level demographics are presented in Table 2.2. Sixty-one pupils (mean age = 12.5, SD = 0.33, % females = 49) from ten schools were recruited at baseline, 32 of which were attending intervention schools. The number of schools per treatment arm was balanced (n = 5 each).

Table 2.2: Baseline characteristics of schools and participants.

	Intervention	Control	Total
School level[a]			
No. schools	5	5	10
Gender type			
Co-ed	5 (100%)	4 (80.0%)	9 (90.0%)
Female	0 (0%)	1 (20.0%)	1 (10.0%)
Male	0 (0%)	0 (0%)	0 (0%)
Type, no. (%)			
Comprehensive to 16	0 (0%)	0 (0%)	0 (0%)
Comprehensive to 18	4 (80.0%)	3 (60.0%)	7 (70.0%)
Secondary Modern	1 (20.0%)	2 (40.0%)	3 (30.0%)
Size, no. (%)			
Large (>200)	1 (20.0%)	0 (0%)	1 (10.0%)
Medium (100-200)	4 (80.0%)	4 (80.0%)	8 (80.0%)
Small (<100)	0 (0%)	1 (20.0%)	1 (10.0%)
No. of form groups, median (range)	6 (6–11)	5 (4–8)	6 (4–11)
Socioeconomic status			
Percent eFSM[b], mean (SD), %	8.62 (4.31)	8.07 (9.16)	8.35 (6.75)
IMD[c], median (range), decile	9 (7–10)	8 (3–10)	8.5 (3–10)
Location[d], no. (%)			
Rural town and fringe	3 (60.0%)	2 (40.0%)	5 (50.0%)
Urban city and town	2 (40.0%)	3 (60.0%)	5 (50.0%)
Urban major conurbation	0 (0%)	0 (0%)	0 (0%)
Individual level			
No. of individuals	32	29	61
Age[e], mean (SD), y	12.4 (0.27)	12.5 (0.39)	12.5 (0.33)
Female, n (%)	16 (50.0%)	14 (48.3%)	30 (49.2%)
Socioeconomic status			
IMD, median (range), decile	8 (2–10)	9 (3–10)	8 (2–10)
eFSM, no. (%), yes	0 (0%)	1 (3.4%)	1 (1.6%)
Pubertal status, mean (SD)	2.13 (0.718)	2.29 (0.509)	2.20 (0.630)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, no. = number, SD = standard deviation

[a] School establishment type is omitted from this table to save space and given low variability in this variable

[b] Percent eFSM indicates the percentage of enrolled pupils that are eligible for free school meals within a school.

[c] Index of multiple deprivation was collected using the school's postcode, through Government records: http://imd-by-postcode.opendatacommunities.org/imd/2019

[d] Location was assessed using https://www.gov.uk/government/collections/rural-urban-classification

[e] Age at the time of baseline (MRI) assessments

Notably, the number of participants per school ranged from 1–17 (Table 2.3). Despite intentions to sample equal cluster sizes, this was infeasible due to several reasons,

including a lack of interest from pupils from certain schools, concerns from parents regarding the study or travel distance, and unavailability during the summer months.

Table 2.3: Sampling frequency of participants by schools

School	Treatment	N	%
S-1	Control	1	1.6
S-2	Control	8	13.1
S-3	Control	1	1.6
S-4	Control	8	13.1
S-5	Control	11	18
S-6	Intervention	1	1.6
S-7	Intervention	1	1.6
S-8	Intervention	4	6.6
S-9	Intervention	9	14.8
S-10	Intervention	17	27.9

Values in the first column *school* are arbitrary.

2.3.4.4 Sample representativeness

The representativeness of the brain imaging sample as sub-sample of the Fit to Study main trial sample was explored by comparing school- and pupil-level baseline demographics and describing any observable differences in their distributions. The comparison included the 93 schools (n = 16,017 pupils) that were part of the trial at the time of baseline assessments. Details of the comparison are provided in Appendix B.1.2.

In brief, schools from which brain imaging sub-study pupils were recruited are representative of the larger Fit to Study sample, but on a pupil level, the brain imaging sub-sample contains fewer participants from a lower socioeconomic background and significantly more high fit pupils.

2.3.4.5 Intervention fidelity

Intervention fidelity was assessed as part of the main trial of Fit to Study and no additional measures were implemented to capture fidelity within the brain imaging sub-

sample. Hence, main trial fidelity information (including teacher- and participant reported fidelity) was used to indicate adherence of schools to the intervention in the subsample.

2.3.4.6 MRI acquisition

Scanning was carried out at the Wellcome Centre for Integrative Neuroimaging, Oxford (WIN, formerly FMRIB), using a 3T Siemens Magnetom Prisma (Erlangen, Germany) scanner with a 32-channel head coil. The MR-protocol comprised both functional and structural sequences, took approximately 55 minutes to complete and included: (1) T_1-weighted (T_1w) structural MRI, (2) resting-state functional MRI (rs-fMRI), (3) diffusion-weighted MRI (DWI-MRI), (4) quantitative FLASH-MRI, (5) pseudo-continuous arterial spin labelling (pCASL), and (6) MR angiography (MRA; post-intervention only).

Brain imaging in paediatric populations comes with various methodological and practical challenges (discussed in chapter 3), including neurodevelopmental changes in neural tissue, as well as behavioural changes (e.g. sleep), head motion, and possible anxiety to the scanner (see e.g. Barkovich *et al.,* 2019; Galván *et al.,* 2012). The primary practical challenge at the stage of acquisition is head motion, which is generally increased in children compared with adults (Dosenbach *et al.,* 2017; Greene *et al.,* 2018; Lerch *et al.,* 2017). In order to minimise head motion, we prepared the participant prior to scanning by explaining scan procedures and upcoming sequences. We further optimised the order of the acquisition (Figure 2.10) in accordance with expected compliance of the pupils. For instance, it was assumed that subject motion would increase with scan time, whereas wakefulness would decrease. Therefore, the T_1-weighted structural scan was acquired first, given its crucial role in all processing pipelines, followed by the resting-state fMRI, which requires participants to stay awake. Moreover, to comfort participants while in the scanner, they watched a documentary during all structural scans (T_1-weighted, diffusion-weighted and quantitative MRI).

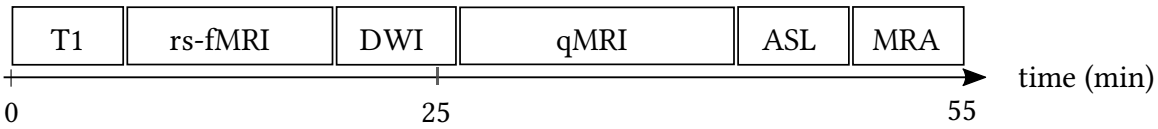


Figure 2.10: MRI protocol. At baseline, all participants completed a shorter version of the protocol, comprising: T_1-weighted MRI, resting-state fMRI, diffusion weighted imaging, quantitative MRI and ASL. At post-intervention, MR angiography was added to the protocol to capture any effects of the VPA intervention on vasculature. The size of the scan blocks approximate scan duration. Abbreviations: T1 = T_1-weighted, rs-fMRI = resting-state fMRI, DWI = diffusion weighted imaging, qMRI = quantitative MRI, ASL = arterial spin labelling, MRA = MR angiography.

Details of each sequence are provided below (in order of acquisition):

- **T_1-weighted structural** A high resolution T_1-weighted scan (3D Magnetization-Prepared RApid Gradient Echo [MP-RAGE], repetition time (TR) = 1,900 ms; echo time (TE) = 3.97 ms; flip angle = 8°; field-of-view (FOV) = 192 mm; voxel size = 1 mm isotropic; scan time = 5.34 min) was acquired for analysis of brain structure and preprocessing of other modalities.

- **Resting-state functional MRI** A multi-band echo-planar imaging (EPI) sequence was acquired (TR = 933 ms; TE = 33.40 ms; FOV = 192 mm; 644 volumes; 72 slices; voxel size = 2 mm isotropic; multi-band acceleration factor = 6; scan time = 10.1 min). Participants were asked to look at a fixation cross, blink normally, try not to fall asleep and not to think about anything in particular. A field map was acquired prior to the resting-state sequence to correct for inhomogeneity distortions.

- **Diffusion-weighted scan** Diffusion-weighted echo-planer images (b-values = 0, 1250, 2500 s/mm^2 , with respectively 11, 60, 60 diffusion-weighted directions; TR = 2,483 ms; TE = 78.20 ms; FOV = 214 mm; voxel size = 1.75 mm isotropic; multi-band acceleration factor = 4; four additional b = 0 s/mm^2 images were acquired with reversed phase encoding, scan time = 5.4 min) were acquired following the resting-state fMRI scan to explore tissue microstructural properties.

- **Quantitative MRI** To retrieve quantitative MRI markers, two 3D multi-echo FLASH datasets were acquired: one predominantly proton density-weighted (PDw, flip angle = 6°), and one predominantly T_1w (flip angle = 21°; FOV = 256 mm;

voxel size: 1 × 1 × 1 mm; TR = 25 ms; first TE = 2.34 ms; eight equally space echoes, echo spacing = 2.3 ms; GRAPPA acceleration factor = 2 in both phase-encoded directions, with 40 reference lines in each direction, scan time = 5.11 min each). Two single-echo, low-resolution (4 mm isotropic) FLASH scans were acquired before each high-resolution scan (FOV identical to the high-resolution scans; TR = 4 ms; TE = 2 ms; one using the 32-channel receive head coil, the other receiving on the body coil). To correct for the effect of radio frequency (RF) inhomogeneities, the local RF field is mapped using a 2D DAM method with a FLASH readout.

- **Arterial spin labelling** A multiple post-labelling delay pseudo-continuous ASL scan (pCASL, with background pre-saturation; post-labelling delays = 0.25, 0.5, 0.75, 1, 1.25 and 1.5 s; labelling duration = 1,400 ms; TR = 4,100 ms; TE = 14 ms; flip angle = 90°; FOV = 220 mm; voxel size = 3.4 × 3.4 × 4.5 mm ; 96 volumes; scan time = 5.34 min) was acquired to quantify absolute resting cerebral blood flow and arterial transit time.

- **MR angiography** MR angiography images (3D time-of-flight sequence, TR = 21.0 ms; TE = 3.58 ms; GRAPPA acceleration factor = 3; flip angle = 18°; FOV = 192 mm; 4 slabs with 32 slices each; voxel size: 0.6 mm isotropic, scan time = 6 min) were collected to explore the brain vasculature. MR angiography data was collected at post-intervention only.

2.3.4.7 Cognitive assessments

Participants completed the object-location memory task (OLT), implicit association task (IAT), relational memory task and colour-shape switch task. The OLT and relational memory task were included to assess hippocampal-dependent memory and are of primary interest to this thesis. The IAT was included to measure biases associated with PA. The relational memory task and colour-shape switch task were taken from the main trial and included to (1) obtain a lab-based, well-controlled measure of the task for cross-validation, and (2) ensure that all brain-study participants have a score on the

task, independent of whether they completed the task as part of the main trial.

Details of the OLT task are provided below. The relational memory task (section 2.3.2.1) and colour-shape switch task (section 2.3.2.1) have been described before. The IAT was outside the scope of this thesis; details can be found in Appendix B.2.2.1 and Wassenaar *et al.* (2019a).

Object-location task

Visual short term memory was assessed using a modified version of the previously described *"what was where"* object-location task (Pertzov *et al.,* 2012; Pertzov *et al.,* 2013). Participants completed a single block of 50 trials. In each trial, participants were shown three fractal objects, randomly located on an interactive touch screen (Figure 2.11). The locations of the fractals were determined by a MATLAB script (MathWorks, Inc). Participants were instructed to remember the identity of objects and their locations. Following a delay of 1 or 8 seconds, two fractals were displayed on the vertical meridian. One of these fractal objects was part of the original memory array (i.e. the target), whereas the other were not (i.e. the foils). Participants were instructed to touch the remembered object and "drag" it to its remembered location. Memory for object identity (accuracy) and object location were computed, as well as the number of *swap* or misbinding errors (i.e. the correct object was selected, but located near the location of one of the other fractals in the memory array). Prior to the start of the task, participants completed a practice block of seven trials with either one or three fractal objects being displayed in the centre of the screen.

2.3.4.8 Cardiorespiratory fitness

Cardiorespiratory fitness was assessed using the gold-standard cardio-pulmonary exercise test (CPET) ($\dot{\mathbf{V}}O_2$max test; Beltz *et al.,* 2016). An incremental step test on a cycle ergometer (Lode Excalibur Sport, Groningen, The Netherlands) was used to measure the volume of oxygen utilised by the body per unit time at peak exertion. Maximal oxygen consumption ($\dot{\mathbf{V}}O_2$max) was reached when oxygen consumption remains at a steady state despite an increase in workload. Participants performed the Godfrey protocol

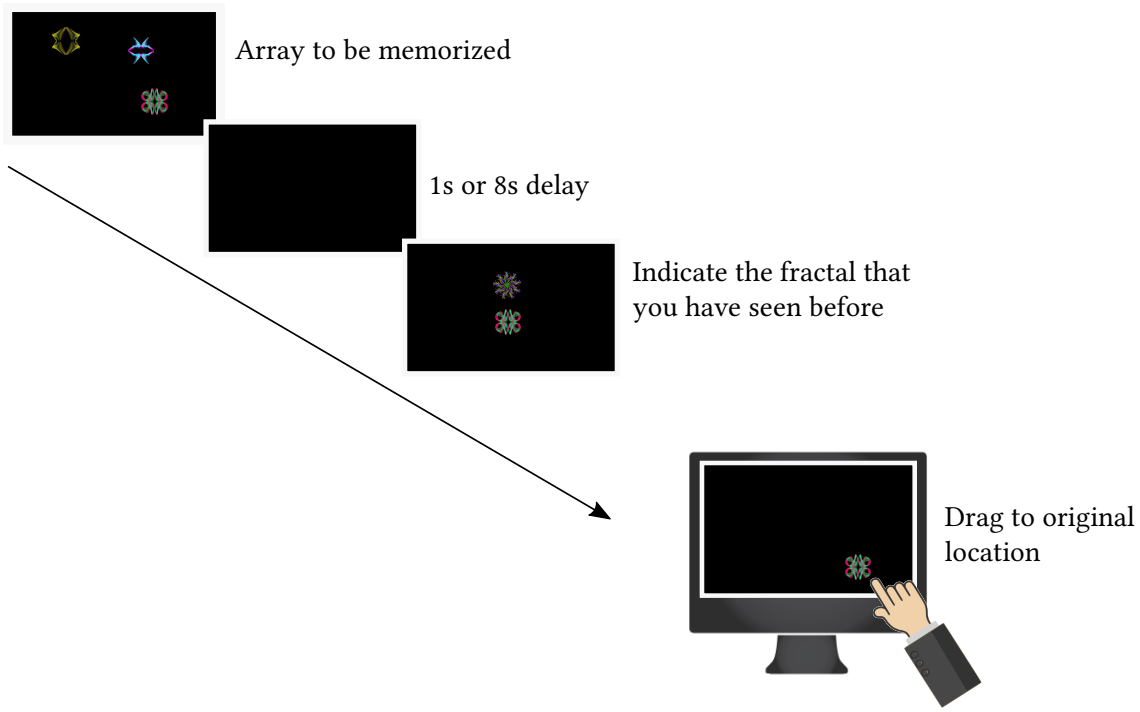


Figure 2.11: Object-location task paradigm. Participants were shown three fractal objects, randomly located on the screen. They were instructed to pay attention to their identity and locations. Following a delay of one or eight seconds, two fractals were shown on the vertical meridian. Participants had to identify the remembered object and drag it to the remembered location.

(Godfrey & Mearns, 1971), which consists of one-minute stages of progressively higher workload (Figure 2.12a) after an initial warm-up phase of two minutes of unloaded cycling. Workload increments are 15 W or 20 W per minute for, respectively, participants under or over 150 cm tall. Participants were instructed to maintain a cadence of 60 revolutions per minute (rpm) throughout. The test was terminated when the subject reached volitional exhaustion or was unable to maintain a cadence of 60 rpm despite verbal encouragement.

The rate of oxygen consumption and heart rate (HR) were measured throughout the test, and rating of perceived exertion was recorded after every increment using the OMNI scale (Utter *et al.*, 2002). HR was measured using a Polar heart rate monitor. Participants were fitted with a face mask to measure oxygen consumption (Figure 2.12b) using an automated indirect calorimetry system (Metalyser, Cortex), which analyses the expired air with every breath. The system was calibrated at the start of the day

prior to the testing sessions using atmospheric pressure, the standard calibration gas (15% O_2 and 5% CO_2) and volume, using the *MSS toolbox* (Metosoft Studio). $\dot{V}O_2$ data were then cleaned (outliers >3 SD removed), smoothed and time-averaged in 9-second intervals for further analysis.

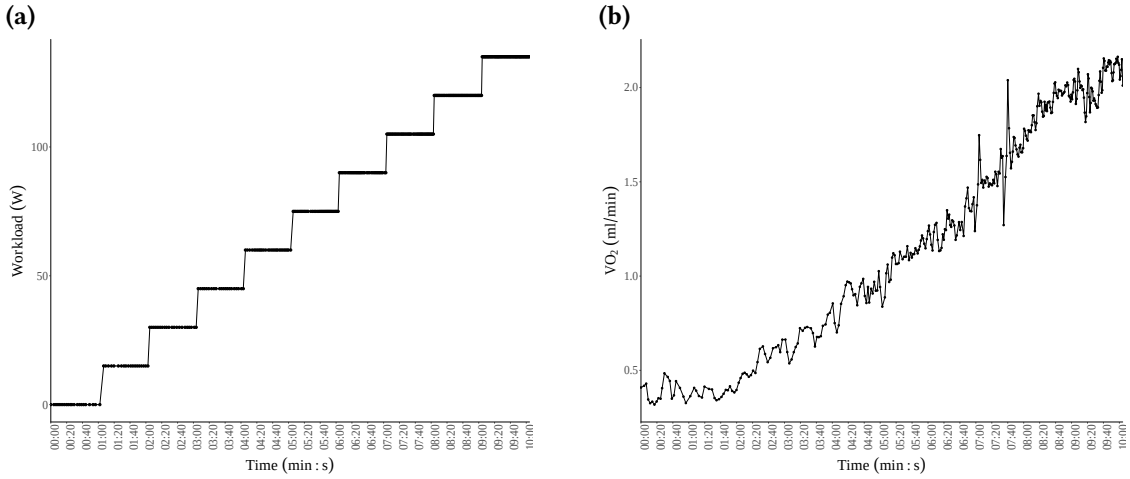


Figure 2.12: Illustration of Godfrey protocol and $\dot{V}O_2$ dynamics. In (*a*) workload is increased after every minute, with in (*b*), a concomitant increase in $\dot{V}O_2$.

Only a minority of children exhibit the classic plateau that is used to define $\dot{V}O_2$max in adults (Winter *et al.,* 2007). Therefore, our key outcome variable is $\dot{V}O_2$peak: the average $\dot{V}O_2$ over the last 30 seconds of the test (Winter *et al.,* 2007). At least two of the following five criteria had to be met for a participant to achieve its peak performance (Cooper & Storer, 2001): (1) HR >180 beats/min (Schulze-Neick *et al.,* 1992), (2) respiratory exchange ratio (RER) >1.06 (Winter *et al.,* 2007), (3) rating of perceived exertion (RPE) at the end of the test >8 (Franklin *et al.,* 2000), (4) subjective signs of exhaustion (Winter *et al.,* 2007), and/or (5) whether the ventilatory threshold (VT2), also known as lactate or anaerobic threshold, was reached. In addition to $\dot{V}O_2$peak as measure of CRF, the maximum workload (i.e. peak power, expressed in Watts) achieved during the test was also extracted.

2.3.4.9 Mental and physical health questionnaires

A comprehensive set of questionnaires was completed online on a computer by all participants, including:

Attention Deficit Hyperactivity Disorder (ADHD) ADHD symptoms were assessed by the ADHD Rating Scale IV (DuPaul *et al.,* 1998), completed by a parent of the participant.

Puberty status Pubertal status was assessed using the Pubertal Development Rating Scale (Peterson *et al.,* 1988), a self-report measure of physical development for youths under the age of 16.

Mood Mood was assessed using the abbreviated Profile of Mood States (POMS, Grove & Prapavessis, 1992). The abbreviated POMS scale is a well-validated questionnaire that contains 40 self-report items on a five-point Likert scale.

Sleep Sleep was assessed using the Cleveland Adolescent Sleepiness Questionnaire (CASQ, Spilsbury *et al.,* 2007) and the Sleep Condition Indicator (SCI, Espie *et al.,* 2014). The CASQ is a self-report scale to measure excessive daytime sleepiness. The SCI is an 8-item self-report measure of insomnia symptoms, based on DSM-V criteria, and captures both sleep problems and daytime functions.

Mental health Mental health was assessed with the Strength and Difficulties Questionnaire (SDQ, Goodman, 1997).

General health General health and behaviour was assessed using various items of the Health Behaviour for School Aged Children (HBSC) survey (2009/2010, Currie *et al.,* 2010), capturing body image, eating behaviour, oral health, PA, leisure time activity, substance use, family structure, peer relationships, self-rated health, and socioeconomic environment. An overview of the HBSC items is included in Appendix B.2.2.3.

PA-related psychological variables Exploratory explicit and implicit measures of psychological variables linked to daily average PA, using the frameworks set out in the Theory of Planned Behaviour (Ajzen, 1985; Ajzen, 1991) and the Prototype Willingness Model (Gerrard *et al.,* 2008). Details are provided in Appendix B.2.2.1.

2.3.4.10 Objective physical activity

Objective PA was assessed using the Axivity AX3 accelerometer (Newcastle, UK). Accelerometers were set to record at 100 Hz and were worn for seven days, starting on Monday (08:00), during a typical school week following the baseline MRI assessment (September or October 2017) and prior to the follow-up assessment (June 2018). Bespoke software was used to convert raw accelerometer data into 1 s epochs and to calculate whether an epoch was spent in sedentary, light, moderate or vigorous PA according to established cut-points (Phillips *et al.*, 2013). Non-wear time was defined as 30 consecutive minutes of zero counts. A valid weekday was defined as 12 consecutive hours of activity, from 08:00 to 20:00, to capture travel to school as well as any after school sport or PA. A weekend valid wear-day was any consecutive 10-hour period between 08:00 and 20:00, scaled up to 12 hours. Participants who had at least three valid weekdays and one valid weekend day were included in the analysis. Missing week and weekend days were imputed by taking the average of sedentary, light, moderate and vigorous PA estimates from, respectively, available week and weekend-day data.

2.3.4.11 Gait

Participants were instructed to walk over a 10 m obstacle-free and flat surface walkway, while wearing a single inertial measurement unit (LPMS-B, life Performance Research, Japan) to capture to obtain spatial and temporal gait parameters (Esser *et al.*, 2011; Esser *et al.*, 2009). Participants performed two consecutive 10 m walks, taking approximately 20 seconds to complete.

2.4 Data analysis

Details on the analysis of data from the Fit to Study cluster RCT are provided in each of the experimental chapters (chapter 4 to chapter 7); a general overview —in line with Wassenaar *et al.* (2019a) —is provided below. More information about cluster randomised trials, including handling of missing data, is provided in chapter 3.

2.4.1 Sample size calculation

Sample size calculations were performed by NatCen for the primary intention-to-treat (ITT) analysis, which compared maths test performance between intervention and control schools (Husain, 2017). The initial sample size calculation demonstrated a range of minimum detectable effect sizes (MDES) varying between 0.18–0.26 SD for various achieved sample sizes (80—100 schools), assuming 80% power, $\alpha = 0.05$ (two-tailed), 90 pupils per school and baseline covariates explaining 50% of variance at pupil and school level (Husain, 2017). Details on the sample size calculation, including updated MDES' and a-posteriori computed MDES' for CRF, cognitive- and mental health outcomes, are provided in Appendix B.3 .

2.4.2 Statistical analysis

In accordance with the Consolidated Standard for Reporting Trials (CONSORT) 2010 statement for cluster randomised controlled trials guidelines (Campbell *et al.,* 2012), the cluster and participant flow will be reported. Descriptive statistics of recruitment and drop-out will be provided on both cluster-and subject level. Further, completeness of, and adherence to, the intervention will be reported on a cluster-level only. Baseline characteristics of the randomised groups will be summarised across schools and/or pupils, and will be presented on a pupil level where appropriate.

The primary ITT analysis comparing maths test performance between intervention and control schools has been conducted by NatCen and is described elsewhere (Bartasevicius & Jabin, 2018; Husain, 2017; Husain *et al.,* 2019). To account for differences between pupils at baseline, maths performance (Key Stage 2 test score) was collected from the National Pupil Database (NPD). Analysis of secondary outcome measures will be conducted on an ITT basis.

Linear mixed models (LMMs), or multilevel models, will be used as these models take the hierarchical structure of the data into account (i.e. pupils within schools, Campbell *et al.,* 2000). The models will be used to assess the effect of the intervention,

as well as to explore associations between secondary outcome measures, using (school-level) random intercepts to capture between school variability. Analyses exploring the effect of the intervention on secondary outcomes will follow recommendations for (cluster) randomised trials (EMA, 2015; Twisk *et al.,* 2018). That is, models will be adjusted for baseline values of the outcome measure (i.e. analysis of covariance, ANCOVA) and include the stratification variable used at randomization (school gender-type, Kahan & Morris, 2012). Analyses used to explore inter-individual differences will use LMMs and adjust for confounding factors where appropriate (e.g. sex, age, socioeconomic status). More details on multilevel models are provided in chapter 3.

Analysis and processing of MRI data collected as part of the brain imaging sub-study was carried out using tools from MrTrix3, MATLAB, FMRIB's Software Library (FSL), Statistical Parametric Mapping (SPM) and Freesurfer, employing parametric and non-parametric statistical analysis methods where appropriate. Processing of the data is described in more detail in the relevant experimental chapters (chapter 5 and chapter 7).

2.5 Discussion

The primary aim of the Fit to Study trial was to investigate the impact of a one academic year (10 months) VPA programme delivered during school PE classes on academic performance (i.e. maths) in secondary Year 8 pupils. The trial secondarily aimed to explore the impact of the intervention on CRF, cognitive performance, mental health and the brain. The findings of this trial will contribute to the growing evidence base of the effects of PA on academic achievement (Biddle *et al.,* 2019; Donnelly *et al.,* 2016; Marques *et al.,* 2018; Sember *et al.,* 2020; Singh *et al.,* 2019; Sneck *et al.,* 2019; Sullivan *et al.,* 2017) and will advance our understanding of the effect of PA on a range of variables —fitness, cognition, mental health and the brain —that are thought to moderate or mediate this relationship. Although many of these variables have individually been linked to PA (Álvarez-Bueno *et al.,* 2017b; Biddle *et al.,* 2019; Owens *et al.,* 2017; Singh *et al.,* 2019) their role in PA-academic achievement relationships is still unclear.

The early years of secondary school are a period of dynamic neurobiological change. Evidence that additional VPA during school PE has a positive impact on the brain, cognitive function and attainment has the potential to help policymakers and educators steer the developmental course in a positive direction. The Fit to Study intervention could offer a feasible method of increasing VPA without disrupting the curriculum. Further potential advantages are its scalability and ease of dissemination, which are important if the intervention is to be used more widely in the future (Reis *et al.,* 2016). The intervention may be incorporated in PE lessons without taking up additional PE time; it should not interfere with the curriculum and does not require extensive training of teachers.

The results could provide further evidence that school is an effective setting for increasing adolescent PA and positively influencing their PA behaviour (Hills *et al.,* 2015) irrespective of their parents' PA behaviour or socioeconomic status. This is important, given that pupils from deprived backgrounds often have lower PA levels (Drenowatz *et al.,* 2010; Eime *et al.,* 2015; Richter *et al.,* 2009) and are more difficult to reach via non-school interventions. The findings could also add to evidence on whether PA interventions during PE, including those that supplement PE lessons with VPA, are effective at increasing PA levels (2018 Physical Activity Guidelines Advisory Committee, 2018; Lonsdale *et al.,* 2013) and improving cognition (Álvarez-Bueno *et al.,* 2017b; Wassenaar *et al.,* 2020).

The findings of the Fit to Study trial will provide valuable information for other research groups and organisations looking to design and implement school-based PA interventions. In addition, the trial has the potential to inform the political and scientific debate regarding the role of schools in promoting PA among adolescents.

This page is intentionally left blank.

3

Methodological considerations when measuring and assessing physical activity, the brain and cognition

Contents

3.1 Measuring physical activity and cardiorespiratory fitness . . . 72
3.1.1 Measuring physical activity . . . 72
3.1.2 Measuring cardiorespiratory fitness . . . 76
3.2 Measuring the brain: Magnetic Resonance Imaging . . . 79
3.2.1 Assessing brain structure . . . 79
3.2.2 Assessing brain function . . . 87
3.2.3 Multimodal MRI and analysis . . . 88
3.2.4 Confounds . . . 89
3.3 Measuring cognition . . . 90
3.3.1 Cognitive tasks in PA research: need for consensus? . . . 90
3.3.2 Cognitive testing: methodological considerations . . . 91
3.4 Data analysis of cluster randomised controlled trials . . . 96
3.4.1 Cluster randomised trials . . . 97
3.4.2 Dealing with clustering in statistical models . . . 98
3.4.3 Missing data . . . 101

Measuring physical activity, cardiorespiratory fitness, the brain and cognitive performance is not trivial, nor is the analysis of cluster randomised trials. This chapter provides a brief evaluation and discussion of various methodological considerations when conducting a cluster randomised controlled trial in the field of physical activity and neuroscience. In particular, I evaluate physical activity and cardiorespiratory fitness measurements and their challenges, briefly review MRI and what it allows us to measure, discuss the measurement of cognitive function, and delve into challenges related to the analysis of cluster randomised trials.

3.1 Measuring physical activity and cardiorespiratory fitness

Physical activity (PA) and cardiorespiratory fitness (CRF) measurements are commonly used to determine trends, examine relationships with health outcomes and to evaluate the effectiveness of interventions (Ding *et al.,* 2020; Lang *et al.,* 2018; Sandercock & Cohen, 2019; Welsman & Armstrong, 2019a). However, the accurate measurement of PA and CRF is an on-going challenge and typically involves a compromise between accuracy level and feasibility. Additional challenges may arise when assessing such measures in young people. In the following sections, I briefly review several common approaches to measure PA and CRF, with a particular focus on their strengths and limitations.

3.1.1 Measuring physical activity

A wide range of methods have been used to quantify PA behaviour in children and adults. These methods include subjective measures based on self-report, as well as objective measures, collected using actigraphy (or "accelerometry"), heart rate (HR) monitors, pedometers, direct observation, indirect calorimetry or doubly labelled water (Armstrong & Welsman, 2006). The extent to which these techniques capture the multidimensional nature of PA (i.e. frequency, duration, intensity and type) varies, which may explain, at least in part, the weak relationships that have often been reported

between the various methods. Moreover, each method has its own set of limitations that should be considered at each stage of the research process (e.g. design, analysis, interpretation).

It is outside the scope of this thesis to discuss each method in depth; excellent evaluations of their respective strengths and weaknesses have been provided in various papers (e.g. 2018 Physical Activity Guidelines Advisory Committee, 2018; Armstrong & Welsman, 2006; Dishman *et al.,* 2001; Sylvia *et al.,* 2014; Vanhees *et al.,* 2005; Warren *et al.,* 2010) and books (e.g. McArdle *et al.,* 2010; Montoye *et al.,* 1996). Self-report measures of PA and actigraphy are discussed in more detail in the next paragraphs, because they have been used extensively throughout this thesis, are suitable for large-scale population studies and are commonly used in the field.

Subjective measures: self-report

Self-report methods, including self- and interviewer-administered questionnaires, diaries and proxy-reports (by teacher or parent), are widely used to assess PA behaviour because of their low cost and ease of use (Armstrong & Welsman, 2006; Shephard, 2003). Such methods could capture both quantitative and qualitative components of PA (e.g. context of PA) and are easy to implement in large population-based studies. Moreover, they can provide information on the time spent in different intensity categories (e.g. light, moderate, vigorous) and are useful to rank respondents in the order of reported activity (Sylvia *et al.,* 2014; Warren *et al.,* 2010). However, self-report methods are limited by the potential for social desirability and recall bias (Sallis, 1991; Sallis & Saelens, 2000), may be costly to process depending on how the questionnaire is administered (paper vs. electronically), and specific challenges arise when self-report surveys are used in children (Warren *et al.,* 2010). Children tend to engage in sporadic physical activities rather than consistent bouts (Bailey *et al.,* 1995) and are less aware of time compared to adults, which makes it more difficult to accurately recall PA characteristics (e.g. duration or frequency).

A great number of self-administered questionnaires —the most common form of PA assessment—have been developed to measure PA behaviour in young people (Chinapaw *et al.,* 2010; Hidding *et al.,* 2018). A recent systematic review identified 89 childhood PA questionnaires, but found no questionnaire with acceptable reliability, validity and methodological quality (Hidding *et al.,* 2018). The single-item PA questionnaire that was used in Fit to Study (Booth *et al.,* 2001; Scott *et al.,* 2015) was reported to be among the most reliable questionnaires for measurement of PA in adolescents (intraclass correlation coefficient [ICC] = 0.75), with moderate convergent validity (r = 0.44, with an accelerometer-based moderate-to-vigorous PA assessment). However, like all other PA measures, it comes with a trade-off: while it is brief and easy to administer, it lacks detail on PA characteristics (e.g. activity type or intensity).

Clearly, self-report measures are a useful tool for PA assessment, but researchers ought to be mindful of their limitations when designing experiments and interpreting results. The lack of questionnaires with favourable measurement properties (or lack of testing thereof) calls for further research in this area. Moreover, the field would benefit from consensus on a smaller set of questionnaires for between-study comparisons to be meaningful. To measure the full extent of PA behaviour, self-report questionnaires may need to be combined with objective measures of PA, such as actigraphy.

Objective measures

Accelerometers are the most popular method for objective PA measurement (Freedson *et al.,* 2005). Accelerometers are wearable monitors that detect accelerations from the body in one to three directions (i.e. uni to tri-axial). Recorded accelerations from traditional accelerometers (e.g. Actigraph and Actical) are typically converted to brand-specific activity counts and captured within a user-specified time period ("epoch", usually between 1–60 s). Using a set of cut-points, or intensity thresholds, the amount and intensity of PA can subsequently be estimated (e.g. time spent in sedentary, light, moderate or vigorous PA). Such information can now also be obtained from the raw acceleration signal, as well as a host of other metrics, including PA energy expenditure

(Van Hees *et al.,* 2009), various sleep related indices (e.g. sleep duration and efficiency; Barreira *et al.,* 2015; Meredith-Jones *et al.,* 2019; Van Hees *et al.,* 2018) and gait characteristics (Brandes *et al.,* 2006).

Accelerometers are increasingly used in large epidemiological studies (Lee & Shiroma, 2014; Troiano *et al.,* 2014), given that they can provide rich, objective and reliable information on various quantitative and qualitative PA characteristics, can capture high-quality data on structured and incidental PA in real time in free-living environments over extended time-periods (days to weeks), and are re-usable, easy to use and increasingly cost-effective (Armstrong & Welsman, 2006; Freedson *et al.,* 2005; Migueles *et al.,* 2017; Sabia *et al.,* 2014; Warren *et al.,* 2010; Yang & Hsu, 2010). Moreover, smartphones with built-in accelerometers have taken PA monitoring to the next level (Althoff *et al.,* 2017; Bort-Roig *et al.,* 2014).

However, accelerometers have several limitations: they do not capture the context of PA, may not be valid to measure specific type of activities (e.g. activities against a resistance force or cycling), and calibration of cut-points has proven to be complex and influence PA intensity estimates (De Almeida Mendes *et al.,* 2018; Freedson *et al.,* 2005; Lee & Shiroma, 2014; Logan *et al.,* 2016; Matthews, 2005; Migueles *et al.,* 2019; Routen *et al.,* 2012; Smith *et al.,* 2017). Moreover, with rapid technological advances in the actigraphy field, less attention has been paid to the standardization of various aspects of data collection and processing, including the type of device, sensor placement, sampling frequency, epoch length, non-wear time and valid wear day definitions, and the type of algorithm used for analysis, which may influence PA estimates (e.g. time in vigorous PA is inversely related to epoch length, Nilsson *et al.,* 2002) and greatly limit the extent to which data across studies can be compared (Banda *et al.,* 2016; Cain *et al.,* 2013; Migueles *et al.,* 2017; Migueles *et al.,* 2019; Reilly *et al.,* 2008).

Some of the decisions regarding data collection and processing are particularly important for studies involving children and adolescents. For instance, compliance

may be higher with wrist-worn compared to waist-worn accelerometers (Fairclough *et al.,* 2016; Scott *et al.,* 2017), short epoch lengths (1–15 s) are recommended to capture the short, sporadic bouts of activity that frequently occur with young people (Cain *et al.,* 2013; Logan *et al.,* 2016), and cut-points need to be specific for this age group (Duncan *et al.,* 2019; Trost *et al.,* 2011). Excellent reviews have been published that evaluate the plethora of data processing decisions (e.g. Cain *et al.,* 2013; Migueles *et al.,* 2017; Migueles *et al.,* 2019; Reilly *et al.,* 2008) and provide recommendations for researchers planning to use accelerometry. Importantly, however, the field is developing rapidly, with attempts to move beyond cut-points and report standardized metrics (Fairclough *et al.,* 2020; Fairclough *et al.,* 2019; Rowlands *et al.,* 2018a), the development of novel (machine-learning) algorithms for PA classification (Clark *et al.,* 2017; Farrahi *et al.,* 2019) and further validation and standardization of existing approaches. Until the field has matured, research using accelerometry should aim to collect high-quality data over multiple days to allow for reanalysis with "better" algorithms in the future.

Other objective PA measurements exist (e.g. HR monitoring, doubly-labelled water), but are less frequently used and come with their own set of limitations. A brief discussion of such measures is provided in Appendix C.1.

3.1.2 Measuring cardiorespiratory fitness

CRF is commonly assessed using objective assessments, although a small number of subjective measures exists too (e.g. Keith *et al.,* 2015; Ortega *et al.,* 2011; Van Heuvelen *et al.,* 1997), one of which has been validated in youth (Ortega *et al.,* 2011; Sánchez-López *et al.,* 2015). Among the objective assessments, laboratory and field-based CRF tests can be distinguished (Tomkinson & Olds, 2008).

Laboratory-based assessments of CRF use incremental cardiopulmonary exercise tests and indirect calorimetry to measure the maximum rate of oxygen volume utilised by the body at peak exertion ($\dot{V}O_2max$), which is considered the "gold" or criterion

standard measure of CRF (Hill & Lupton, 1923; Hill *et al.,* 1924). In children and adolescents, however, $\dot{V}O_2$peak —the highest value of oxygen consumption ($\dot{V}O_2$) attained during the test —rather than $\dot{V}O_2$max is used as the gold-standard measure of CRF, because young populations do not typically reach the plateau in oxygen consumption seen in adults (Armstrong *et al.,* 1996; Rowland, 1993; Welsman & Armstrong, 2019b). To assure that a maximum effort has been delivered, researchers rely on subjective ratings of exhaustion and secondary criteria, such as pre-specified HR values (e.g. 85% of HR_{max}), the respiratory exchange ratio and blood lactate. Such criteria, however, are no panacea and their use may underestimate an individual's true $\dot{V}O_2$peak by 10–22% (Barker *et al.,* 2011). The addition of a supramaximal test may provide a true estimate of an individual's $\dot{V}O_2$max (Barker *et al.,* 2011), but due to time limitations, this is hardly done in practice. Moreover, $\dot{V}O_2$peak estimates tend to be higher (11–14%) using treadmill compared to cycling ergometers (Armstrong & Welsman, 2019b), and vary by an individual's sex, age, maturity status, body composition (e.g. skinfold thickness) and body mass (Armstrong & Welsman, 2019a; Armstrong & Welsman, 2019b). To control for the strong relationships with body mass, $\dot{V}O_{2s}$peak is typically expressed as a ratio with body mass (ml/kg/min), which is a stronger predictor of health compared to unscaled $\dot{V}O_2$peak values (Lang *et al.,* 2018). However, the appropriateness of the ratio-scaled data is being questioned and allometric scaling may be preferred instead (Armstrong & Welsman, 2019d; Tanner, 1949; Welsman & Armstrong, 2019a).

Despite being the gold-standard measure of CRF, the assessment of $\dot{V}O_2$peak is of limited use in population-based studies due to the need for expensive and cumbersome equipment, a controlled environment, qualified examiners and substantial amounts of time required to test individuals. Field-based tests, in contrast, are low cost, flexible in testing locations, require minimal equipment and no specialised personnel, and allow for testing of multiple individuals simultaneously (Tomkinson *et al.,* 2017; Tomkinson & Olds, 2008), providing a viable alternative to estimate $\dot{V}O_2$peak in large-scale studies (Armstrong & Welsman, 2019d; Lang *et al.,* 2018). Numerous field-based tests have been developed for CRF assessment in youth (Artero *et al.,* 2011; Castro-Piñero *et al.,*

2010; Mayorga-Vega *et al.,* 2015; Mayorga-Vega *et al.,* 2016; Tomkinson & Olds, 2008), including time (e.g. 12-min Cooper test), distance (e.g. 1.5 mile run) and endurance-based (e.g. 20 m shuttle run) tests. Such tests have generally shown high reliability and moderate validity (Tomkinson & Olds, 2008), albeit estimates vary widely between tests and with protocols used to assess these metrics (Artero *et al.,* 2011; Castro-Piñero *et al.,* 2010; Mayorga-Vega *et al.,* 2015; Mayorga-Vega *et al.,* 2016; Ortega *et al.,* 2008a; Ruiz *et al.,* 2009b).

Of the available tests, the 20 meter shuttle run (20MSR; Léger *et al.,* 1988) is the most widely used and particularly attractive due to its low cost and ease of administration (Lang *et al.,* 2019), which is why this measure has been used in Fit to Study. Nevertheless, this assessment has various methodological issues too. For instance, various variants of the 20MSR test protocol exist and test performance is often reported in different ways, including the number of completed laps, stages, speed at the last stage, or converted to $\dot{\mathbf{V}}O_2$peak. The conversion to $\dot{\mathbf{V}}O_2$peak is based on estimation equations, at least 17 of which have been reported in the literature (Tomkinson *et al.,* 2017), which vary in their validity, their generalisability and result in substantially different $\dot{\mathbf{V}}O_2$peak estimates. Raw values, such as the number of laps, may therefore be the preferred outcome (Lang *et al.,* 2018), and have been used in Fit to Study. A meta-analysis on the validity of the 20MSR found that 51% of correlation coefficients between 20MSR scores and $\dot{\mathbf{V}}O_2$peak explain less than 50% of the variance in $\dot{\mathbf{V}}O_2$peak, and concluded that the 20MSR provides only an *estimate* of CRF (Mayorga-Vega *et al.,* 2015). Indeed, it is well-known that multiple factors contribute to 20MSR performance, including fat mass, running economy, anaerobic capacity, and motivation (Armstrong & Welsman, 2019c; Armstrong & Welsman, 2019d; Lang *et al.,* 2018; Tomkinson & Olds, 2008). Despite such caveats, however, the 20MSR, and field-based tests in general, are increasingly used for population surveillance (Armstrong & Welsman, 2019d; Lang *et al.,* 2018; Sandercock & Cohen, 2019).

3.2 Measuring the brain: Magnetic Resonance Imaging

Since its invention in the 1970's, magnetic resonance imaging (MRI) has revolutionised the study of the human brain by providing a safe, non-invasive tool to examine this organ's structure and function *in vivo* (Damadian, 1971; Edelman, 2014; Lauterbur, 1973). It enables longitudinal tracking of brain changes in different populations at multiple spatial scales and has contributed significantly to our understanding of how the environment or experience, such as PA, shapes the brain (Lövdén *et al.*, 2013; Tardif *et al.*, 2016).

Existing neuroimaging techniques are, however, not without limitations and there are many challenges surrounding MR image acquisition, analysis and interpretation that should be considered (see e.g. Grover *et al.*, 2015; Lerch *et al.*, 2017; Winston, 2012). Of particular importance is the notion that MRI provides an indirect measure of the tissue properties and cellular changes that researchers typically wish to make inferences on (Zatorre *et al.*, 2012). Hence, to aid the interpretation of MRI findings in Fit to Study, we briefly discuss commonly derived MRI measures for the multi-modal MRI sequences employed in this thesis (T_1-weighted MRI, diffusion-weighted MRI, relaxometry and arterial spin labelling), their limitations and their biological interpretation. In addition, I discuss a set of factors, or confounds, that may affect MRI-derived measures if not considered at each stage of the design process.

3.2.1 Assessing brain structure

Brain structure can be assessed using a wide range of MR sequences (or "contrasts", see e.g. Cercignani & Bouyagoub, 2018; Grover *et al.*, 2015; Tardif *et al.*, 2016; Winston, 2012). The T_1-weighted contrast, or "anatomical scan", has arguably been the most popular method to examine the brain's size and shape. With high contrast-to-noise ratio between grey matter (GM), white matter (WM), and cerebral spinal fluid (CSF), this contrast is ideal for examining macro- or mesoscopic morphological changes (i.e. across multiple voxels, or three-dimensional pixels), yet provides little insight into the

underlying neurobiological mechanisms. With advances in neuroimaging techniques, however, it has become possible extract meaningful information from within a single voxel. Such microstructural imaging techniques aim to quantify specific tissue properties using biophysical models and quantitative MRI, and include diffusion weighted imaging and relaxometry (for other available sequences, see Alexander *et al.,* 2011; Alexander *et al.,* 2017; Cercignani & Bouyagoub, 2018; Edwards *et al.,* 2018; Novikov *et al.,* 2018a; Novikov *et al.,* 2018b; Weiskopf *et al.,* 2015; Winston, 2012).

3.2.1.1 T_1-weighted MRI: morphology

Several techniques have been used to perform morphometric analysis of T_1-weighted data (Mietchen & Gaser, 2009; Van Der Kouwe *et al.,* 2008), including voxel-based morphometry (VBM, Ashburner & Friston, 2000) and volumetry (region of interest, ROI, analysis). VBM is an automated technique that involves a voxel-wise comparison of local tissue concentration, typically GM, between groups or timepoints (Ashburner & Friston, 2000; Good *et al.,* 2001). It is a popular, unbiased method to assess GM differences across the whole-brain (Lövdén *et al.,* 2013), but relies heavily on accurate inter-subject registration (Bookstein, 2001) and tissue segmentation (Tardif *et al.,* 2016), and is affected by small changes in the processing pipeline (e.g. smoothing; Henley *et al.,* 2010; Senjem *et al.,* 2005). Moreover, VBM-derived maps are hard to interpret because (1) any observed differences may reflect variations in both brain shape and size (Hutton *et al.,* 2009; Zatorre *et al.,* 2012), and (2) it is unclear what VBM measures: its main metric reflects the probability of a voxel to be GM and is commonly expressed as density, volume or concentration, yet has no clear neurobiological interpretation (Lövdén *et al.,* 2013).

Volumetry involves manually or automatically delineating a pre-defined ROI. It provides a more meaningful and interpretable measure compared to VBM, and has therefore been used in this thesis. Manual segmentation is commonly regarded as the gold standard (Konrad *et al.,* 2009; Lerch *et al.,* 2017; Morey *et al.,* 2009), but requires trained personnel and is highly time-consuming, limiting its scalability. Moreover,

manual segmentations may show considerable inter- and intra-rater variability if no protocols are in place. Automated or semi-automated techniques resolve many of these issues, but can be computationally expensive and vary in their segmentation accuracy (Dill *et al.,* 2015). Both types of segmentation methods (manual and automated) have been used to obtain the volume of the hippocampus, a popular ROI in the exercise-neuroscience field (Firth *et al.,* 2018; Voss *et al.,* 2013).

Hippocampal volumetry

The hippocampus is a complex anatomical structure (Destrieux *et al.,* 2013; Duvernoy, 2005; Thomas *et al.,* 2008) and is notoriously hard to segment on T_1-weighted scans (Boccardi *et al.,* 2011; Konrad *et al.,* 2009). Accuracy of hippocampal segmentations varies with MR acquisition parameters (e.g. field strength), postacquisition processing and the definition of anatomical boundaries (Geuze *et al.,* 2005). Regarding the latter, there is considerable variation in how anatomical boundaries are defined across manual segmentation protocols (Konrad *et al.,* 2009), resulting in varying estimates of hippocampal volumes between studies and poor replicability (Geuze *et al.,* 2005). Harmonised protocols have been developed (Boccardi *et al.,* 2015), but accurate reporting of the manual delineation procedure remains key.

Manual segmentation of the hippocampus can take up to two hours (Carmichael *et al.,* 2005) and is unfeasible in large imaging studies, such as UK BioBank (Miller *et al.,* 2016). Automated segmentation algorithms have been developed that substantially reduce the segmentation time and aim to improve accuracy and reproducibility (Dill *et al.,* 2015; González-Villà *et al.,* 2016), of which FSL's FIRST (Patenaude *et al.,* 2011) and FreeSurfer (FS, Fischl *et al.,* 2002) are most frequently used. The performance of these algorithms has been compared with gold-standard manual segmentations in studies involving older adults, showing: (1) FS and FIRST typically overestimate hippocampal volumes (Morey *et al.,* 2009; Mulder *et al.,* 2014; Pipitone *et al.,* 2014), and (2) have shown variable agreement (i.e. Pearson correlations) with manually derived volumes (FS: 0.61–0.84, Cherbuin *et al.,* 2009; Sánchez-Benavides *et al.,* 2010; FIRST: 0.47–0.69, Nugent

et al., 2013; Pardoe *et al.,* 2009). What's more, the performance of such algorithms varies with age (Wenger *et al.,* 2014). In a study involving children (N = 147, mean age = 8.47), both algorithms substantially overestimated hippocampal volumes, but FS had marginally stronger agreement (r = 0.61–0.69) with manual segmentations than FIRST (r = 0.51–0.59). When using such algorithms in paediatric cohorts, it is recommended to visually inspect the output and, if needed, manually correct segmentations (Schoemaker *et al.,* 2016).

What do these methods measure?

Methods such as VBM and volumetry are highly sensitive to macroscopic brain changes, yet it is unclear what such changes reflect on a microscopic level. These techniques rely on the presence of image intensity boundaries of the T_1-weighted signal, which may be affected by a host of cellular processes, including: neurogenesis, angiogenesis, gliogenesis, myelination and changes in neuronal morphology (e.g. spine density) (Anderson, 2011; Roth *et al.,* 2010; Zatorre *et al.,* 2012). To establish the cellular basis of macroscopic changes, one could (1) replicate studies in animal models (e.g. mice), which allow for histology (for examples of such work in plasticity research: Lerch *et al.,* 2011; Sagi *et al.,* 2012; Sampaio-Baptista *et al.,* 2013; Scholz *et al.,* 2015), but are invasive and require experience in working with animals, or (2) complement conventional T_1-weighted sequences with MR contrasts that are more biologically specific (e.g. Lorio *et al.,* 2016), commonly referred to as microstructural imaging methods.

3.2.1.2 Microstructural imaging

Microstructural imaging is flourishing (Cercignani & Bouyagoub, 2018; Edwards *et al.,* 2018). This is due, at least in part, to its ability to quantify MRI parameters of brain tissue that relate more directly to properties of cellular compartments. To date, most of the work has focussed on WM, but microstructural MR techniques have great potential for GM too (e.g in the cortex, Edwards *et al.,* 2018, and hippocampus, Sagi *et al.,* 2012; Sierra *et al.,* 2015). Here, we discuss two the microstructural imaging methods that have been acquired as part of Fit to Study and used in this thesis: diffusion weighted imaging and relaxometry.

Diffusion weighted MRI

Diffusion weighted imaging is sensitive to barriers in tissue that hinder or restrict the diffusion of water (Le Bihan & Johansen-Berg, 2012; Le Bihan *et al.,* 2001). If no barriers are present (e.g. CSF), water diffuses uniformly in all directions, i.e. isotropic; if diffusion is restricted due to, e.g. membranes or fibres, water follows the long axis of the barriers, i.e. anisotropic. By fitting a model to diffusion measurements at each voxel, useful parameters can be obtained that provide information about the microstructure. The tensor model (Figure 3.1) is most commonly fitted to the data (Basser *et al.,* 1994a; Basser *et al.,* 1994b; Mori & Zhang, 2006; Pierpaoli *et al.,* 1996) and yields parameters that are sensitive to changes in tissue microstructure, such as fractional anisotropy (FA) and mean diffusivity (MD), but that are not specific to underlying tissue characteristics and therefore difficult to interpret. It assumes unidirectional diffusion, which is a poor reflection of the complex tissue organization, particularly in regions where fibers are not aligned (e.g. crossing, Jones *et al.,* 2013)

Alternative biophysical models have been developed that model the signal as a function of several compartments (see Alexander *et al.,* 2017; Assaf *et al.,* 2019 for an overview of such multicompartment models), which are assumed to reflect particular tissue properties. The NODDI (neurite orientation dispersion and density imaging) model (Tariq *et al.,* 2016; Zhang *et al.,* 2012) is a popular model that aims to quantify the density and dispersion of neurites (axons and dendrites). It distinguishes three compartments within a voxel (intra-cellular, extra-cellular and CSF), to yield three parameters: the intra-cellular volume fraction (ICVF, or neurite density index), isotropic volume fraction (ISOVF, or free water volume fraction) and orientation dispersion index (ODI, spatial configuration of neurites). Notably, while compartment models like NODDI have increased biological interpretability, they make assumptions that may not always hold true in practice (Jelescu & Budde, 2017).

The tensor model and NODDI, while fitted to the same data, provide different information. Metrics from both models have therefore been used in this thesis.

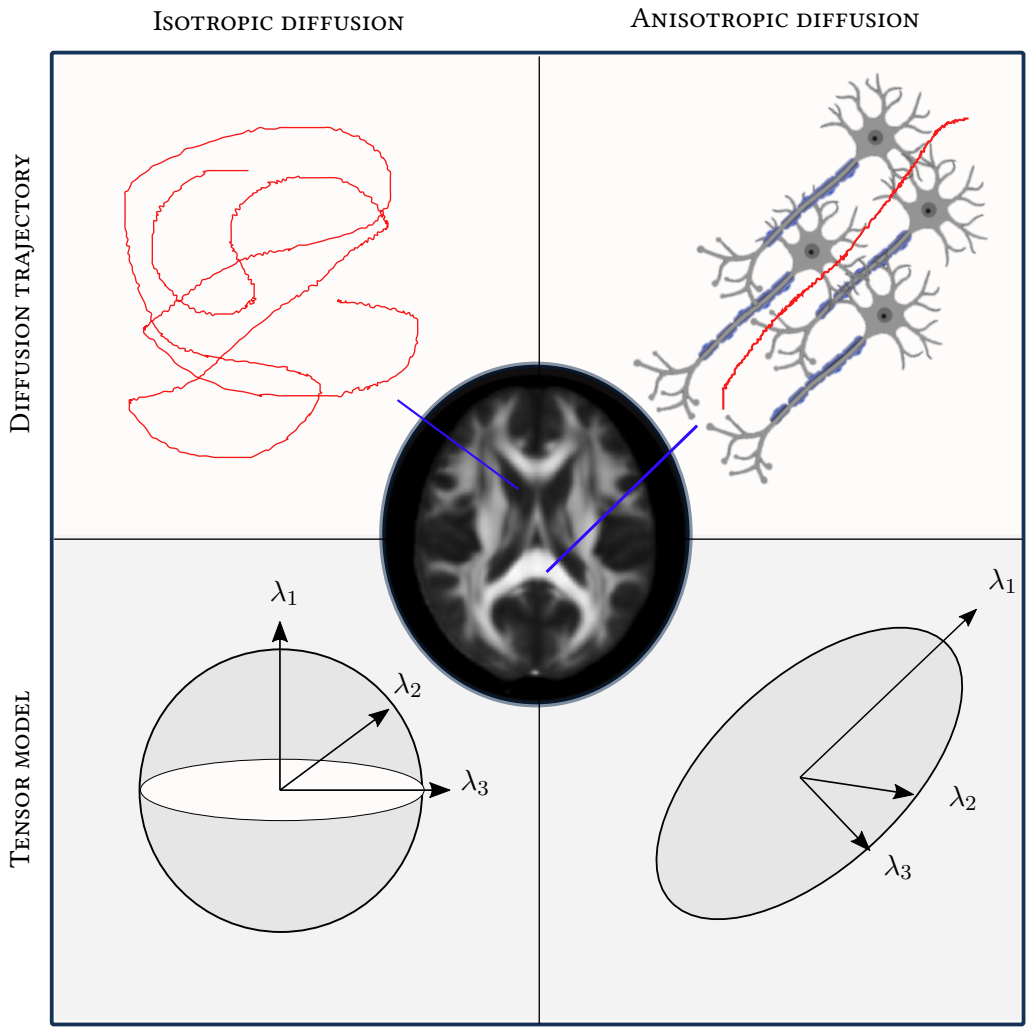


Figure 3.1: Isotropic and anisotropic diffusion captured by the tensor model. (*Top row*) The diffusion trajectory of water differs in the presence or absence of barriers. (*Bottom row*) The tensor model describes water diffusion at each voxel. In the tensor model, water diffusion is characterized by its three principal eigenvectors and their associated eigenvalues (λ_1, λ_2, λ_3). From the eigenvalues, the DTI parameters FA, MD, as well as radial diffusivity (RD) and axial diffusivity (AxD) can be computed. FA reflects the shape of the tensor and has higher values with more ellipsoid shape tensors (i.e., with anisotropic diffusion, *bottom right*). MD is the magnitude of diffusion and can be computed by averaging the eigenvalues of the tensor (MD $= \frac{\lambda_1+\lambda_2+\lambda_3}{3}$). RD reflects the diffusion perpendicular to the long axis of the tensor and can be computed by averaging eigenvalues λ_2 and λ_3 (RD $= \frac{\lambda_2+\lambda_3}{2}$). AxD reflects the diffusion along the long axis of the tensor and is equal to λ_1. Figure and legend adapted from Wassenaar *et al.* (2019b).

Relaxometry: T_1 and T_2^*

Relaxometry is an umbrella term referring to a group of quantitative MR methods that measure T_1, T_2 and T_2^* relaxation times (Alexander *et al.,* 2011; Deoni, 2010; Does, 2018). T_1, or the longitudinal relaxation time, describes the time it takes for a system of hydrogen atoms to return to its equilibrium state (i.e. along the main magnetic field, $\mathbf{B}_0$) following a radio frequency (RF) pulse. T_1 is determined by macromolecular content (e.g. lipds) and the microstructural tissue environment (e.g. water), and is shorter for myelin and longer in water. T_2, or transverse relaxation, describes the time constant at which atomic spins de-phase, following the excitation pulse. T_2 decay depends on spin-spin interactions, and is independent from T_1, yet occurs at the same time and is typically much shorter. T_2 decay is faster than would be predicted in model systems, which is caused by magnetic field inhomogeneities. This effective T_2 is termed T_2^*. Compared to conventional T_1-weighted imaging, relaxometry aims to map T_1 and T_2^* separately, is more specific to particular microstructural characteristics and is reliable across scan sites (Deoni *et al.,* 2008; Weiskopf *et al.,* 2013), but has a long acquisition time. Multiparameter mapping (Helms *et al.,* 2008a; Helms *et al.,* 2008b; Weiskopf *et al.,* 2013) may be used to more efficiently capture T_1 and T_2^*, and has therefore been used to capture microstructural metrics in Fit to Study. Relaxation times are commonly expresses as relaxation rates: $R_1 = \frac{1}{T_1}$ and $R_2^* = \frac{1}{T_2^*}$.

What do these methods measure?

Microstructural diffusion or relaxometry metrics are typically compared across individuals at the voxel level, following registration to a common template, or within a ROI. Moreover, in WM, these metrics can be sampled along WM tracts, constructed using tractography (Jbabdi & Johansen-Berg, 2011), or projected to a WM skeleton, followed by a voxel-wise comparison (Smith *et al.,* 2006). But what do changes in these measures actually reflect?

Tensor-derived metrics such as FA and MD are not biologically specific (Assaf *et al.,* 2019; Hutchinson *et al.,* 2018; Jones *et al.,* 2013). FA for instance, has shown to be sen-

sitive to a wide range of cellular components, including intra-voxel orientation dispersion, myelination, membrane permeability, axon diameter and axon density (Beaulieu, 2009; Zatorre *et al.,* 2012), whereas MD has shown to be affected by processes such as synaptogenesis, neurogenesis, angiogenesis, dendritic sprouting and astrocyte swelling (Johansen-Berg *et al.,* 2012; Tardif *et al.,* 2016).

The NODDI based metrics are more specific to tissue characteristics compared to FA or MD. However, the neurite density index (NDI) —commonly interpreted as an indication of axon and dendritic density (Mah *et al.,* 2017) —has also been associated with myelination (Grussu *et al.,* 2017), and may vary with changes in the extracellular space (Mah *et al.,* 2017). Despite a paucity of validation studies, most of which concern WM, increasing evidence suggests that ODI is a valid measure of (histology-based) fiber orientation dispersion (Grussu *et al.,* 2017; Mollink *et al.,* 2017; Sato *et al.,* 2017; Schilling *et al.,* 2018) and NDI may reflect neuron density (Schmitz *et al.,* 2019; Wang *et al.,* 2019), though a negative association between NDI and markers of neurite integrity was found in the hippocampus (Crombe *et al.,* 2018).

Quantitative maps of R_1 and R_2^* lack biological specificity and have shown to be sensitive to both iron and myelin (Edwards *et al.,* 2018). R_1 predominantly reflects variation in myelin, though iron contributes too (Edwards *et al.,* 2018; Stüber *et al.,* 2014). R_1 may also be sensitive to neuronal fibers (Edwards *et al.,* 2018), such as axons (Harkins *et al.,* 2016), and a general linear model describing R_1as a function of macromolecules, iron and water content has been proposed (Callaghan *et al.,* 2015). R_2^* is related to local susceptibility changes and has shown to be a sensitive marker of iron (Bagnato *et al.,* 2018; Haacke *et al.,* 2005; Langkammer *et al.,* 2010; Sun *et al.,* 2015). Compartments that contribute to the R_2^* contrast include iron-rich oligodendrocytes, astrocytes, microglia, and myelinated fibers, as well as blood vessels.

In this thesis, those markers with the highest potential to inform about GM (hippocampal) tissue microstructural characteristics and exercise-related neurobiological

mechanisms (e.g. vasculature, neurogenesis / tissue density and myelination) were selected, including MD, NDI, R_1 and R_2^* (for reasons, see also chapter 5). Metrics that are predominately sensitive to tissue geometry, such as FA and ODI, were excluded.

3.2.2 Assessing brain function

Studies assessing brain function typically use the blood-oxygen level dependent (BOLD) contrast (Ogawa *et al.,* 1992), which relies on regional differences in cerebral blood flow from which neural activity is inferred. The BOLD signal is, however, ambiguous and reflects a combination of cerebral blood flow, oxygen metabolism and cerebral blood volume (Buxton, 2010; Ekstrom, 2010). These processes can be studied independently. For instance, arterial spin labelling (ASL) is a popular method to obtain a quantitative and reliable assessment of cerebral blood flow (CBF, Almeida *et al.,* 2018), despite having a lower signal-to-noise ratio (SNR) and temporal resolution than BOLD. ASL-derived perfusion estimates were therefore used in this thesis.

3.2.2.1 Arterial spin labelling

ASL measures cerebral perfusion without the need for an intravenous contrast agent (Detre *et al.,* 1992; Williams *et al.,* 1992). It magnetically labels water in blood flowing into the brain using a (inversion) RF pulse (Jezzard *et al.,* 2017). After a post-labelling delay (PLD), during which the labelled blood travels to the the tissue bed (the arterial transit time), an image is acquired. A measure of perfusion can be obtained by subtracting the labelled image from a control image, where the blood was not labelled. The traditional use of a single PLD has been complemented with more efficient and reliable labelling schemes, such as pseudo-continuous ASL (pCASL), that have improved SNR (Gevers *et al.,* 2011; Jezzard *et al.,* 2017; Okell *et al.,* 2013). Following preprocessing, using tools such as BASIL (Chappell *et al.,* 2009), a quantitative perfusion map is obtained, which expresses perfusion in the quantity of blood delivered per unit of tissue per unit of time (i.e. ml blood / 100 g tissue / min). The perfusion map can be compared across individuals at the voxel-level or within a ROI.

What do these methods measure?

In contrast to BOLD fMRI or macrostructural metrics such as VBM, perfusion is a meaningful and interpretable measure in itself. Moreover, perfusion has been used as a marker of angiogenesis (Tardif *et al.,* 2016; Zatorre *et al.,* 2012). While the formation of new capillaries cannot be observed with non-invasive imaging techniques, it is assumed that vessel growth results in detectable increases in blood volume and flow. Importantly, however, an observed increase in perfusion is not sufficient to conclude that new vessels have formed; other factors, such as changes in neurovascular coupling (Tardif *et al.,* 2016) could explain increased perfusion too.

3.2.3 Multimodal MRI and analysis

Conventional T_1-weighted MR sequences have proven sensitive to macrostructural brain changes, but lack biological specificity. The greater specificity of quantitative MRI may be harnessed to characterise cellular mechanisms that underlie macro-scale changes. However, replacing or complementing T_1-weighted scans with a single quantitative MR contrast is not sufficient: despite gains in specificity, there is no one-to-one mapping with cellular properties. Used in combination, however, such MR contrasts have the potential to shed light on microstructural properties that drive macroscopic changes (Cercignani & Bouyagoub, 2018; Walhovd *et al.,* 2014a).

Rather than analysing each MR modality individually, multimodal data can be pooled and analysed simultaneously using a wide range data-driven (e.g. joint inference, machine learning) and model-based methods (see e.g. Cercignani & Bouyagoub, 2018). Of particular interest to this thesis are a set of multivariate methods that allow for joint inference across multiple modalities (Winkler *et al.,* 2016). In particular, non-parametric combination (NPC; Pesarin, 1990) is an elegant way of combining test-statistics of multiple independent statistical tests to test a joint null hypothesis. In contrast to the classical multivariate analysis of variance (MANOVA), it is more powerful and based on minimal assumptions (Winkler *et al.,* 2016). NPC has therefore been used in this

thesis; an example of its use can be found in chapter 7.

Canonical correlation analysis (CCA, Hotelling, 1936) is another attractive multivariate method for inference across modalities. CCA seeks to identify modes of covariation linking two sets of variables (X, Y). That is, it finds a *linear* combination of variables in one set (X), composed of e.g. multi-modal MRI markers, that maximally correlates with a linear combination of variables in the other set (Y), composed of e.g. PA markers. These linear combinations of variables in X and Y are called *canonical variates*, which are latent (like factors in factor analysis). CCA has been used extensively to interrogate complex brain-behaviour relationships in large multimodal datasets (Alnæs *et al.,* 2020; Miller *et al.,* 2016; Smith *et al.,* 2015). An example of CCA in the context of brain imaging data is provided in chapter 5.

Notably, NPC and CCA are among a wide range of multivariate techniques that could be used to analyse multimodal data (Sui *et al.,* 2012). Reasons for selecting those techniques will be provided in the relevant chapters, but critically depend on the research question and dataset (e.g. type of data, sample size).

3.2.4 Confounds

MRI measures may be affected by a great number of factors. Such factors, or *confounds*, can be broadly classified into those that are scanner related and those that originate from subjects. A thorough discussion of both types of confounds is provided in Appendix C.2. If confounds vary systematically across subjects or timepoints, they reduce the reliability of the findings and may induce spurious brain-behaviour relationships or intervention effects, leading to incorrect conclusions. Scanner related confounds are usually addressed on the scanner or using *post hoc* processing strategies, whereas those related to subjects are typically "controlled for" in the statistical analysis. A subset of confounds can additionally be controlled for through clever experimental design. For instance, variations in scanner hardware and software can be prevented by testing all

participants on the same scanner.

In this thesis, we controlled for scanner-related confounds, such as magnetic field inhomogeneities and eddy currents through post-hoc processing strategies. Details of these processing steps are provided in the relevant chapters (chapter 5 and chapter 7). We additionally adjusted the statistical analysis for subject-related confounds. These were selected based on: (1) availability and (2) having known relationships with MRI-derived estimates, and included participant's age (Barnes *et al.,* 2010), sex (Herting *et al.,* 2018; Paus *et al.,* 2017; Persson *et al.,* 2014), pubertal status (Goddings *et al.,* 2014; Herting *et al.,* 2017; Vijayakumar *et al.,* 2018b), socioeconomic status (Farah, 2018; Hackman *et al.,* 2010; Yu *et al.,* 2018) and head size. Head size has been related to regional variations in brain size and shape and is commonly treated as a confounding variable in volumetric analyses (Barnes *et al.,* 2010; De Jong *et al.,* 2017; Miller *et al.,* 2016; Nordenskjöld *et al.,* 2015; O'Brien *et al.,* 2011; Voevodskaya *et al.,* 2014).

3.3 Measuring cognition

Measuring cognitive performance is not straightforward. It requires inferring unobservable, psychological constructs and abilities from behavioural observations, usually involving (neuro)psychological test batteries and careful experimentation. Entire fields of study, including (cognitive) psychology and neuroscience, have been devoted to understanding cognitive processes and their objective measurement. The aim of this section is not to cover all of what has been the topic of separate lines of research; rather, a small selection of methodological considerations and deliberations that are of importance to researchers in the exercise-neuroscience field are discussed.

3.3.1 Cognitive tasks in PA research: need for consensus?

A large number of cognitive assessments exist. For instance, a recent scoping review identified 338 individual measures of executive functions (EF) across 705 studies involving adolescents (Nyongesa *et al.,* 2019). Such measures assess performance in one or multiple EF domains. For instance, the Stroop task, Simon task, Flanker task,

antisaccade task, delay-of- gratification task, go/no-go task, and stop-signal task all purportedly assess inhibitory control, whereas fluency tasks (e.g. category, verbal), the Wisconsin Card Sorting Task, Trail Making Test and task-switching paradigms are typically used assess cognitive flexibility, and the digit-span, Corsi-block and complex-span tests have been used to assess working memory (for an overview of EF tasks, see e.g. Diamond, 2013).

Clearly, a PA researcher has a wide range of cognitive tasks at their disposal. With such abundance of assessments, it may be no surprise that a great variety of cognitive assessments have been used. Indeed, a recent systematic review identified a total of 60 unique cognitive tasks among 109 studies to assess cognitive performance in children and adolescents (Wade *et al.,* 2020), the majority of which tapped into EFs. While the great variety of tasks used in PA research may provide insight into the specificity of PA effects, it may hinder the synthesis of research findings and obscure overall effects (Wade *et al.,* 2020; Young *et al.,* 2015), particularly if tasks are combined (e.g. in a meta-analysis) that measure slightly different aspects of a cognitive domain. Various reviews have called for consensus on a smaller number of valid and reliable cognitive assessments that span scientifically and clinically important cognitive domains (e.g. Singh *et al.,* 2019; Wade *et al.,* 2020; Young *et al.,* 2015), in order to improve comparability between studies. Fit to Study made an attempt at doing this by using cognitive assessments that have been used in previous PA research and that span cognitive domains that are hypothesised to be affected by PA (i.e. EFs, memory and processing speed, Chaddock *et al.,* 2010a; Chang *et al.,* 2012; Donnelly *et al.,* 2016; Gunnell *et al.,* 2018).

3.3.2 Cognitive testing: methodological considerations

When assessing the relationship between PA or CRF and cognitive functioning, it is imperative that cognitive assessments are selected that tap into the cognitive domains of interest and are able to detect the hypothesised and potentially small effects of interest. Working with cognitive tasks is challenging and requires making informed decisions

on a wide range of topics. Here, I briefly address a small selection of these topics that deserve attention among researchers working in the PA-neuroscience field, including the problem of task impurity, assessing task performance, and the administration of tasks (other considerations are reviewed in e.g. Pontifex *et al.*, 2019; Schmitt *et al.*, 2005). A brief discussion of psychometric properties is provided in Appendix C.3.

3.3.2.1 Task impurity

Task impurity is a particular challenge for tasks measuring EFs (Burgess, 1997; Friedman & Miyake, 2017; Shallice, 1996), which have been used extensively in the PA field (Wade *et al.*, 2020). It refers to the issue that EFs involve lower cognitive processes in addition to the core EF processes, with scores on EF tasks thus reflecting both EF and non-EF (e.g. visual processing) performance. In other words, a task score is composed of variance attributable to the EF process, as well as variance attributable to non-EF processes (and some error variance). Moreover, even tasks that are considered to measure a specific EF process, such as the N-back task for working memory, may assess both *specific* working memory and *common* EF processes.

One way to deal with the task impurity problem, at least partly, is to create a difference score that isolates the putative EF process of interest (Paap & Oliver, 2016). A downside of this approach is, however, that such scores typically have low reliability (Draheim *et al.*, 2019, and see Appendix C.3). Another way of dealing with task impurity is by collecting multiple measures of the EF component of interest. Importantly, however, these measures should have little or no *non-EF* variance in common. That is, they should share the specific EF component, but require involvement of different lower-level cognitive processes. Latent variable analysis can then be used to extract the common variance across those tasks, which reflects a purer measure of the specific EF component (Friedman & Miyake, 2017; Friedman *et al.*, 2008; Miyake *et al.*, 2000).

While it is desirable to include multiple tasks that tap into the same cognitive construct (e.g. inhibitory control), this may not always be feasible and is rarely done

in practice. With computerized cognitive assessments typically taking between 5–15 min to complete, the number of tasks that can be assessed is usually small and dictated by the time available in the test protocol. For instance, it would have been desirable to include multiple measures of each cognitive domain (e.g. inhibitory control, working memory) in Fit to Study, but this was not feasible given time constraints, and sacrificing some cognitive domains to improve measurement of another was not in line with the study aims.

3.3.2.2 Assessing performance: speed versus accuracy

Performance on many commonly administered cognitive tasks, particularly if they are computerized, can be expressed in terms of speed (reaction time, RT) and accuracy (i.e. proportion correct) of responses. It has long been recognised that these measures are intertwined and interact in complex ways (Heitz, 2014). Typically, participants tend to reduce response speed if accuracy is emphasised, or vice versa, a phenomenon known as the *speed-accuracy trade-off* (Heitz, 2014). However, with instructions commonly emphasising both speed and accuracy ("respond as quickly *and* accuracy as possible") — which, given the speed-accuracy trade-off, is contradictory (Heitz, 2014) —participants likely respond differently and may adjust speed and/or accuracy to a different degree. Manipulations of task parameters (e.g. response deadlines) and instructions could be used to control speed-accuracy trade-offs (Heitz, 2014), although it is well-known that instructions are only temporally effective (Schouten & Bekker, 1967) and such manipulations are hardly used in practice, especially in the field of PA-neuroscience.

The analysis of tasks involving speed and accuracy has taken many forms and is an area of continuous investigation (Draheim *et al.,* 2016; Draheim *et al.,* 2019; Heitz, 2014; Hughes *et al.,* 2014; Liesefeld & Janczyk, 2019; Vandierendonck, 2018). Typically, researchers wish to express cognitive task performance with a single metric, in order to model relationships with one or multiple (independent) variables using classical statistical techniques, including linear regression, pair-wise correlations, or analysis

of variance (ANOVA). To this end, either RT or accuracy is chosen as outcome measure. This facilitates the analysis: results are easier to be interpreted and no multiple testing correction is required (Liesefeld & Janczyk, 2019). However, it assumes that a difference in performance (e.g. between participants) manifests itself solely through RT or accuracy and ignores possible speed-accuracy trade-offs. Between those measures, accuracy may be preferred, given that it may more easily account for such trade-offs and tends to be reliable *if* performance is not at ceiling (Draheim *et al.,* 2019). Moreover, it may be a more meaningful outcome in populations that tend to respond impulsively, such as children (Davidson *et al.,* 2006), and it seems to be more sensitive to the effects of PA than RT based outcomes (Álvarez-Bueno *et al.,* 2017b).

A number of integrative measures of speed and accuracy have been proposed too. These run from two-vector methods that combine and convert RT and accuracy scores into a single measure based on arbitrary accuracy thresholds (e.g. >or <80%), such as in the National Institutes of Health toolbox (Akshoomoff *et al.,* 2014; Weintraub *et al.,* 2013), to measures that directly combine accuracy and RT, such as the the inverse-efficiency (IE) score (Townsend & Ashby, 1978), rate correct score (Woltz & Was, 2006), a bin score (Hughes *et al.,* 2014), linear integrated speed-accuracy score (Vandierendonck, 2018), and the balanced integration score (Liesefeld & Janczyk, 2019). While such integrated scores take both RT and accuracy into account, they too have their drawbacks. For instance, some are only appropriate in certain circumstances (e.g. IE is biased with a high proportion of errors), they may be more or less sensitive to speed-accuracy trade-offs and they vary in their weighting of speed and accuracy. What's more, they are more difficult to interpret compared to RT or accuracy outcomes, and one may nevertheless need to inspect relationships with RT and accuracy individually (Bruyer & Brysbaert, 2011). In addition, specific approaches exist to analyse speed-accuracy trade-offs (Heitz, 2014; Kyllonen & Zu, 2016), but typically rely on speed-accuracy manipulations and usually provide multiple indicators of performance. Data reduction approaches, such as principal components analysis or factor analysis, could be used across RT and accuracy data to obtain a single measure, but may result in loss

of important information and present with interpretational difficulties.

Rather than examining a single outcome, one could examine speed and accuracy (or derived measures) simultaneously using multivariate analysis techniques. Such techniques, including multivariate ANOVA (MANOVA), structural equation models (Tomarken & Waller, 2005) and CCA, allow one to model relationships between multiple outcomes (or dependent variables) and multiple predictors (although strictly speaking, CCA does not distinguish between predictors and outcomes). CCA is a particularly attractive method when one is interested in testing relationships between combinations of variables, not individual variables themselves, as may well be the case with accuracy and speed data. An example of how CCA can be used to interrogate multivariate relationships between measures of PA and cognitive performance is provided in chapter 4.

3.3.2.3 Task administration: web-based cognitive testing?

Cognitive performance has traditionally been assessed with pen-and-paper based tests, under well-controlled conditions by a trained psychologist. With technological advances, computerized cognitive assessments have become available, which address several limitations of conventional paper-based tests, including a higher consistency in test administration, precise control over stimuli, the ability to record multiple responses (e.g. speed and accuracy), and lower cost of administration (i.e. no need for trained personnel and lower cost of data handling). However, if specific software or knowledge is required to run the cognitive assessments, test administration may still be limited by the need for dedicated hardware and trained personnel. Such barriers reduce the number and diversity of individuals (e.g. socioeconomic status, geographical location) that can be tested and challenge the feasibility of cognitive testing in population-based studies like Fit to Study and UK Biobank (Fawns-Ritchie & Deary, 2020).

Web-based assessments provide a promising alternative, by allowing researchers to study large and diverse samples of individuals simultaneously, irrespective of their

geographical location. Online test administration, however, has its own challenges, particularly when such testing is unsupervised, including possible distractions, technical variability in hardware and software, the lack of appropriate hardware (e.g. computer) and internet access (e.g. 1 in 10 people in the UK did not have internet access in 2018, Office for National Statistics 2018, accessed 14th September 2020), and subject-specific factors, such as motivation, fatigue and technical expertise, that the researcher may not be aware of (Templer & Lange, 2008).

Despite these challenges, it has been shown that web-based testing in self-selected samples does not compromise data quality (Germine *et al.,* 2012; Templer & Lange, 2008), although a recent study indicates that precision and accuracy of response times may not be consistent across devices, browsers, operating systems and common experiment platforms (e.g. jsPsych or Gorilla; Anwyl-Irvine *et al.,* 2020). Moreover, additional challenges may arise when testing large samples of young people in schools. For instance, variability in IT equipment in schools participating in Fit to Study, including outdated hardware or software, and institutional firewalls, may complicate online assessments (Wheatley *et al.,* 2020b). As an alternative, dedicated laptops, tablets or smartphones could be used, but this has significant cost and human resource implications. While web-based testing is certainly no panacea (yet), it offers an acceptable trade-off between accuracy and feasibility. Technological advances and increased availability of computers may take away several of the IT-related limitations in the future.

3.4 Data analysis of cluster randomised controlled trials

Cluster randomised controlled trials (cRCT), like Fit to Study, are increasingly common in the exercise-neuroscience field. Here, we briefly review some considerations, specifically those related to the analysis, of cRCTs.

3.4.1 Cluster randomised trials

In cRCTs, also called group or community RCTs, intact groups or clusters of individuals are randomised (Cornfield, 1978). The unit of randomisation can be diverse, including for instance families, classrooms, schools, hospitals, study sites or communities. In Fit to Study, schools are randomised to intervention and control, and pupils are nested within schools (Figure 3.2). Additional clustering arises from repeated measurements over time.

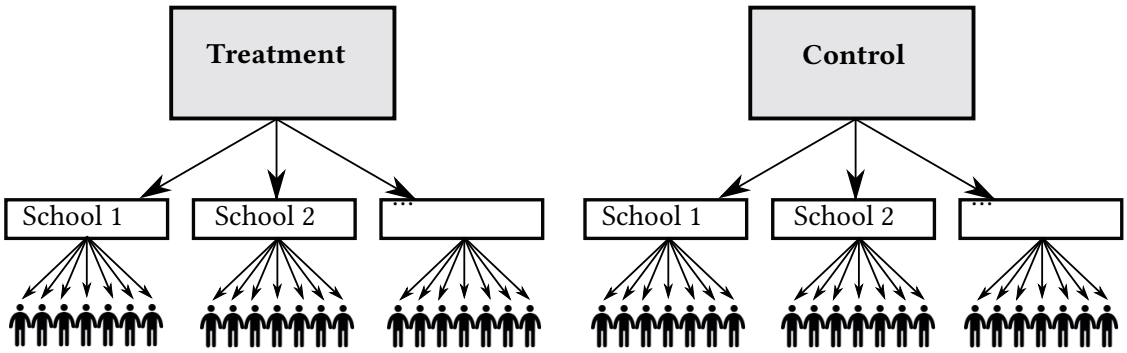


Figure 3.2: Illustration of clustering in Fit to Study. Schools were nested within study arms and pupils were nested within schools The school level is commonly denoted level 2, whereas level 1 often refers to pupils. However, if repeated measurements are collected, time may be nested within pupils who are nested within schools (i.e. three levels, not shown).

Cluster RCTs have been used extensively in education and healthcare settings (Ukoumunne *et al.,* 1999), and they are the method of choice if (1) the intervention is designed to operate on a group level, as is the case for Fit to Study (increasing ecological validity), (2) the intervention cannot be delivered to individuals without the risk of experimental contamination (e.g. changes to the physical or social environment), or (3) practical considerations prevent the intervention to be delivered to individuals (Donner & Klar, 2004; Murray *et al.,* 2004; Turner *et al.,* 2017a; Turner *et al.,* 2017b). The delivery of an intervention to groups of individuals increases administrative efficiency, may reduce implementation costs and is thought to enhance compliance (Donner & Klar, 2004).

3.4.2 Dealing with clustering in statistical models

The non-independence of observations due to clustering needs to be accounted for in statistical models to appropriately control the type I error rate (an in depth discussion, with simulations, is provided in Appendix C.4.1). The mixed-effects model is one of the most popular and versatile methods to take nested structure of cRCTs into account, but is not the only available technique. Alternatives include generalised estimation equations and cluster robust standard errors. A general, non-technical overview of these methods is provided below, with a particular emphasis on the mixed-effects model.

3.4.2.1 Mixed effects model

Mixed-effects models, also known as hierarchical models, multilevel models, random-effects models, or variance-component models, are a group of statistical models that explicitly account for dependence structure in the data (details are provided in Baayen *et al.,* 2008; Gelman & Hill, 2006; Harrison *et al.,* 2018; Hayes, 2006; Laird & Ware, 1982; Peugh, 2010; Raudenbush & Bryk, 1986; Snijders & Bosker, 2012). Such models are widely used in the analysis of cRCTs, where participants are nested within clusters (e.g. schools), as well as in repeated measures (longitudinal) studies, where an individual's responses over time are nested within individuals. Compared to traditional statistical methods, such as ordinary-least squares regression or analysis of variance (ANOVA) that assume that data (or residuals) are independent and identically distributed (*iid*), mixed-effects models capture the dependencies via random-effects parameters, resulting in higher power, nominal type-I error rates and more accurate estimates of the effects (Harrison *et al.,* 2018).

One of the most important concepts in mixed effects models is that of *fixed* and *random* effects. Definitions vary widely and whether something is classified as fixed or random may change depending on the analysis of interest (Gelman & Hill, 2006). A fixed effect is usually called fixed if all possible conditions of interest are present (e.g. treatment groups), and its estimate represents the overall (or population-average) effect of the variable on the outcome, regardless of the higher level unit(s) in which it

is nested. A random effect variable usually represents a random subset of all possible conditions, captures random variability in the data, and shows how much the estimates for a particular cluster differ from the fixed effect estimate. It tends to be a grouping variable one wishes to generalise over (Singmann & Kellen, 2019), such as schools in Fit to Study.

Random-effects can be introduced in two ways: through the intercept and slope (Harrison *et al.*, 2018). One of the simplest multilevel or mixed-effects models is the random-intercept model, which assumes that intercepts vary across clusters, but constrains slopes to be equal. Hence, the relationships between the predictors and outcome is the same for each cluster. Imagine, for instance, that one models the linear relationship between CRF and cognition in the Fit to Study dataset, whereby CRF is centred on the cohort average. By fitting a random intercept model, the variability in intercepts indicates that cognitive performance for a pupil of average CRF varies between clusters, but by constraining the slopes to be equal one assumes that a change in cognition with CRF is the same across clusters. In contrast, a random intercept *and* slope model allows slopes (i.e. a change in cognition with CRF) to vary across clusters too.

Mixed models are extremely flexible and can accommodate multiple (independent) random effects (e.g. nested or crossed random effects). Moreover, one can use such models to inspect variance components. For a simple random intercept model, for instance, the variance of the outcome is partitioned into two parts (but more if multiple random effects): the *within* and *between* cluster variance. Such variance measures can be used to derive the intra-class correlation coefficient (ICC, equation is provided in Appendix C.4.1), which has implications for the design of studies (e.g. power) and may be of interest in itself. For instance, in school-based cRCTs, the ICC may provide insight into the contribution of schools to pupil-level outcomes (Sellström & Bremberg, 2006).

Mixed effects models are not without assumptions (Gelman & Hill, 2006; Snijders & Bosker, 2012). The frequently used *linear* mixed effects model (LMM), where fixed

effects and random effects (and observation-level residual variance) contribute linearly to the outcome, assumes, amongst others, that the within cluster (i.e. observation-level) residuals and random effects are normally distributed (McNeish *et al.,* 2016; Snijders & Bosker, 2012). However, not all data (or residuals) are normally distributed, and while LMMs are robust to violations of normality (Schielzeth *et al.,* 2020), researchers could either perform transformations or use generalised linear mixed models (GLMMs) that allows for outcome variables with non-normal (error) distributions (e.g. binary variables are modelled with a binomial distribution, Bolker *et al.,* 2009). What's more with mixed effects models, the sample size should be sufficiently large to allow accurate estimation of coefficients. There are no strict sample size requirements, and recommendations vary by model complexity, intra-class correlations and coefficients of interest, but a minimum of five clusters is required for accurate estimation of fixed effects at the lowest level in cross-sectional settings with a continuous outcome (McNeish & Stapleton, 2016). A large sample size is, however, preferable, and should be aimed for when designing experiments with clustered data.

3.4.2.2 Alternatives

The mixed effects model is not the only type of model that can deal with clustered data. In particular, when clustering is not of interest but rather a nuisance, a different class of methods may be used that does not explicitly model multiple levels. Such methods, also referred to as population-averaged or marginal models, include cluster-robust standard errors (e.g. Pustejovsky & Tipton, 2018) and generalised estimating equations (GEE, Liang & Zeger, 1986; Zeger *et al.,* 1988, for an overview see, McNeish *et al.,* 2016; Turner *et al.,* 2017b). In brief, such methods use the regression coefficients of a single level (linear) model and adjust the standard errors of the regression estimates to account for clustering. In GEE, the regression coefficients are additionally corrected for clustering. Compared to the mixed-effects model, population-averaged models require a smaller number of assumptions to be met, are less affected by small cluster sizes and yield coefficients that are easier to interpret, but require a comparatively larger number of clusters to produce robust estimates (i.e. typically >50, but small sample

size corrections exist, see Pustejovsky & Tipton, 2018), may be more affected by highly unbalanced clusters and do not provide information on random effects (McNeish *et al.,* 2016).

In most chapters of this thesis, the mixed effects model was used given its versatility and superior performance with a small number of clusters. Only in chapter 6 cluster robust standard errors were used, in line with a latent variable analysis that is central to that chapter.

3.4.3 Missing data

Missing data are a pervasive issue in many fields of research, but are often not adequately dealt with. Inadequate handling of missing data can lead to loss of power, less precision, and biased inferences (Rubin & Little, 2002; White *et al.,* 2011). Typically, three missing data mechanisms, or "reasons" (Enders *et al.,* 2018b), are distinguished (Rubin, 1976) that have important implications for handling of the missing data: (1) Missing Completely At Random (MCAR), if none of the variables are related to or predict missingness, (2) Missing At Random (MAR), if the probability of missingness is unrelated to unobserved data conditional on the observed data, and (3) Missing Not At Random (MNAR), if the probability of missingness depends on unobserved and observed data. It is important to note that these are assumptions about missingness, not properties of the data themselves, which affect the accuracy of the missing data technique. For instance, multiple imputation typically requires a MAR mechanism for estimates to be accurate. An introduction to missing data, including ad-hoc missing data methods, is provided in Appendix C.4.2.

The Fit to Study dataset, described in chapter 2, contains missing data in most of its measures. The hierarchical structure of the data, multitude of variable types and wide range of research questions, requires that appropriate missing data techniques are used to reduce the probability of biased results. Here, we provide some background

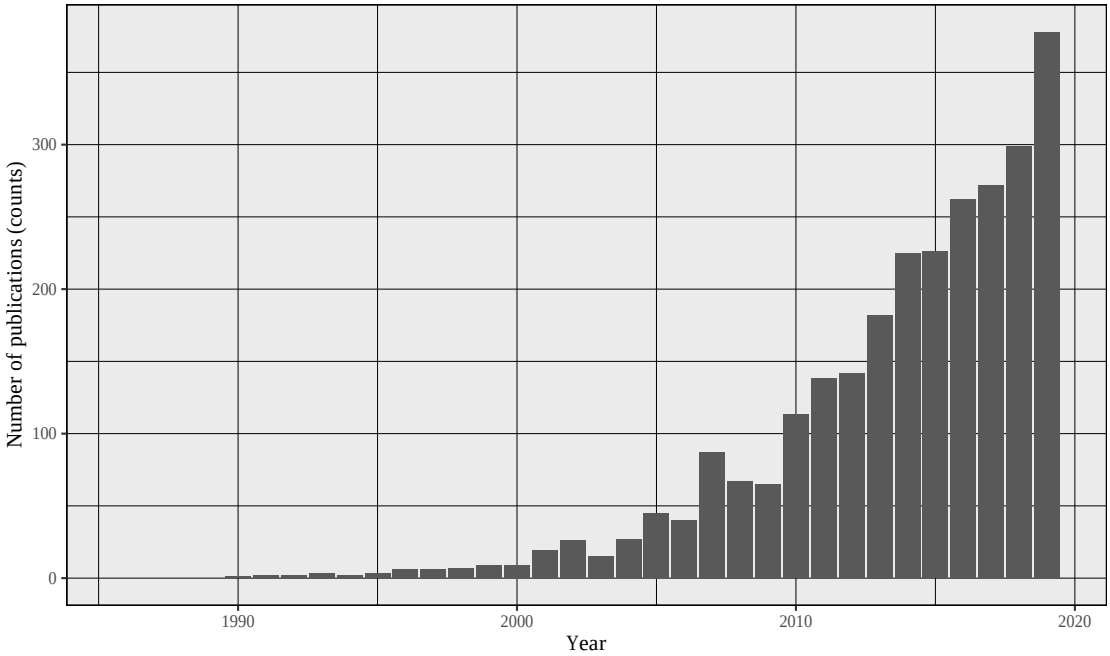


Figure 3.3: Number of publications per year containing "multiple imputation" in the title or abstract. The number of publications per year, up to and including 2019, that contained "multiple imputation" in the abstract or title was extracted from PubMed (19 September 2020) using the RISmed package in R (Kovalchik, 2017). No publications were retrieved in the period between 1950–1985.

to (multilevel) multiple imputation, which is the missing data procedure that has predominantly been used in this thesis.

3.4.3.1 Multiple imputation

Multiple imputation was developed by Donald Rubin in the 1970's (Scheuren, 2005). It is a common strategy to deal with missing data (Rubin, 1987; Schafer, 1997) and has gained tremendous popularity in recent years, demonstrated by the exponential increase in number of publications (Figure 3.3).

In brief, multiple imputation consists of three phases: (1) an imputation phase, (2) an analysis phase, and (3) a pooling phase (Figure 3.4). In the imputation phase, several copies (M) of the data are created in which missing entries are replaced by imputed values. The imputation procedure typically uses a regression model (i.e. the imputation model) to specify a distribution from which plausible synthetic values are drawn. That

is, M random draws are made from the posterior (predictive) distribution of the missing data, given the observed data, to generate M imputed datasets (Enders *et al.*, 2018b). Following the imputation phase, each of the multiple imputed datasets are analysed using standard statistical techniques (e.g. a mixed effects model), resulting in m sets of parameter estimates. These estimates will likely differ between imputed datasets, and reflect the uncertainty in the imputations. In the pooling phase, the estimates are combined, or pooled, across imputations, according to Rubin's rules (Rubin, 1987). More details on multiple imputation, including Rubin's rules, common challenges in missing data, and an overview of available imputation models and software, is provided in Appendix C.4.2.3.

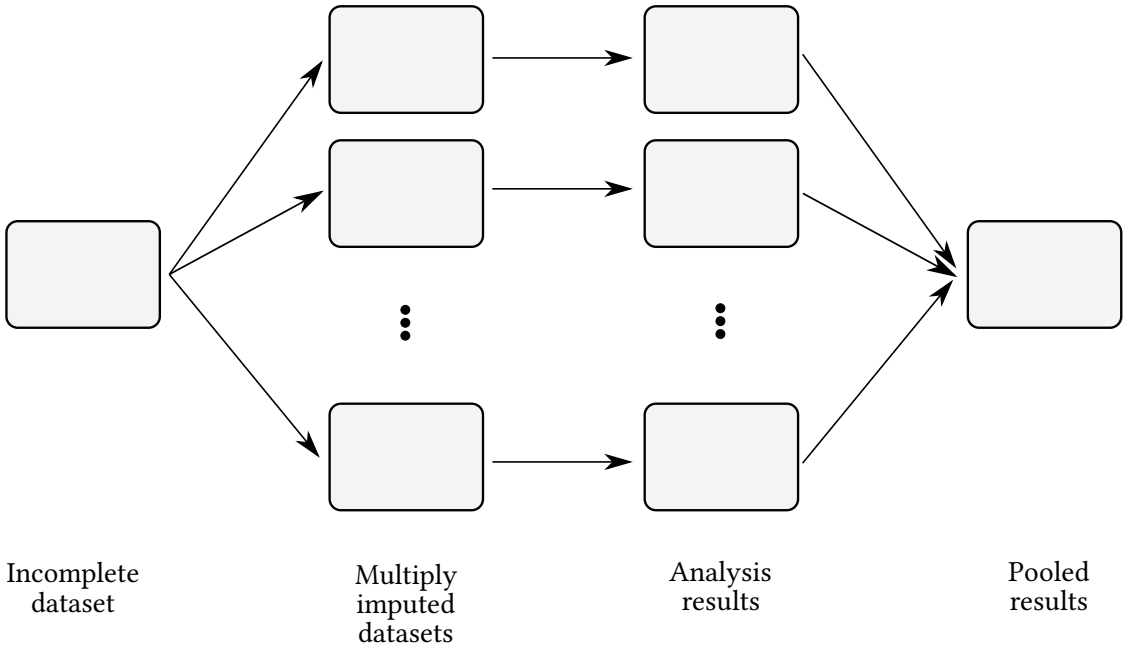


Figure 3.4: Principle of multiple imputation. From left to right, from a single incomplete dataset several multiply imputed datasets are created. Each is analysed individually and the results are pooled to obtain unbiased estimates with appropriately adjusted variance. Figure adapted from Van Buuren (2018).

An important notion in multiple imputation literature is that the imputation model should be more general than the (substantive) analysis model, often referred to as *congeniality* (Meng, 1994, or *compatibility*, albeit its definition is grounded in mathematics, for details see e.g. Enders *et al.*, 2020; Van Buuren, 2018). If the analysis model contains particular relations between variables, such as power terms or interactions,

they should be included in the imputation model to obtain unbiased results (Von Hippel, 2009). Failing to include such associations biases the parameters estimates towards zero. Moreover, a multilevel imputation model should be used if the data is clustered and analysed with a multilevel model (Audigier *et al.,* 2018; Goldstein *et al.,* 2009; Grund *et al.,* 2016; Grund *et al.,* 2018a; Grund *et al.,* 2018b; Huque *et al.,* 2018; Mistler & Enders, 2017; Van Buuren, 2018; Yucel, 2008). Variables that are not part of the analysis model, but are either related to missingness or explain a large amount of variance in variables of interest, should be included as auxiliary variables. Such variables aid imputations and make the MAR assumption more plausible (Graham, 2009; Hughes *et al.,* 2019; Lang & Little, 2018).

Of particular interest to this thesis is multiple imputation when the analysis model contains interactions. For instance, we assess whether the relationship between CRF and cognition is moderated by sex or socioeconomic status (chapter 4) and examine whether the effect of the intervention on cognitive outcomes is moderated by baseline CRF (chapter 6). From the perspective of congeniality, such interactions should be included in the imputation model. A number of methods have been used to impute interaction terms, including those that impute the main effects and passively compute the interaction (i.e. passive imputation, or *impute then transform*), those that add the interaction term as just another variable (i.e. JAV, Von Hippel, 2009), and predictive mean matching, but such methods tend to introduce bias (Seaman *et al.,* 2012). Model-based imputation, whereby the imputation model is tailored to a particular analysis model using a fully Bayesian approach (available in e.g. Blimp or jomo, or JointAI Erler *et al.,* 2016), appropriately deals with interactions terms (Enders *et al.,* 2020) and is therefore preferable. As an alternative, one could impute separately within subgroups if one of the variables of interaction term is categorical and the dataset is sufficiently large. Covariance structures are then allowed to vary within subgroups and any interaction terms are preserved (Enders & Gottschall, 2011).

In this thesis, a combination of missing data approaches were used, which assume that data is MAR. In particular, for the studies involving the Fit to Study main trial data (chapter 4 and chapter 6), we use multiple imputation. The cross-sectional analyses in chapter 4 contain complex three-way interactions, for which model-based imputation is used, whereas the multiple imputations for chapter 6 are split by treatment group to allow the covariance structures to vary within each group. The sample size of the brain imaging sub-study, combined with the rich set of measures, renders large-scale multiple imputations complex and in some cases unfeasible (e.g. due to non-convergence). We therefore used a pragmatic approach to deal with missing cases in those chapters. In particular, the primary cross-sectional analyses presented in chapter 5 contain many variables, which resulted in non-converging imputation models. A single imputation method was therefore used in this chapter. The primary intervention analyses presented in chapter 7, however, used multiple imputation (split by treatment group), specific to each research question.

This page is intentionally left blank.

4

Associations between cardiorespiratory fitness, physical activity and cognitive function in a large sample of young British adolescents

Contents

4.1 Introduction . . . 108
4.1.1 Aims . . . 111
4.2 Methods . . . 111
4.2.1 Study design and participants . . . 111
4.2.2 Cardiorespiratory fitness and physical activity assessments . . . 112
4.2.3 Cognitive assessments . . . 112
4.2.4 Data analysis . . . 113
4.3 Results . . . 122
4.3.1 Sample . . . 122
4.3.2 Associations between PA or CRF and cognitive performance . . . 125
4.3.3 Multivariate associations between a physically active lifestyle and cognitive performance . . . 127
4.4 Discussion . . . 132
4.4.1 Relationships between PA, CRF and cognitive performance . . . 132
4.4.2 Strengths and limitations . . . 135
4.5 Conclusion . . . 136

Increasing evidence suggests that higher levels of physical activity and cardiorespiratory fitness benefit cognitive performance. Findings in developing populations, however, have been less consistent. In this chapter, I use use a large community sample of young British adolescents (n = 16,017) and a combination of univariate and multivariate analyses, to demonstrate that physical activity and fitness are positively associated with a range of domain-specific cognitive outcomes, but that relationships are small and explain less than 10% of the variance in cognitive performance. Sex moderates the association between fitness and inhibitory control and processing speed, with stronger relationships in girls than boys. I additionally show that the choice of predictors, outcomes and nuisance variables may impact on observed associations, and that a physically active lifestyle is associated with more favourable cognitive outcomes.

4.1 Introduction

Regular physical activity (PA) and higher levels of cardiorespiratory fitness (CRF) have well-established health benefits (Warburton *et al.,* 2006), including a lower risk of mortality (Högström *et al.,* 2014; Lang *et al.,* 2018; Lear *et al.,* 2017) and beneficial cardiovascular health outcomes (Fiuza-Luces *et al.,* 2013; Lang & Little, 2018; Ortega *et al.,* 2008b). Moreover, there is increasing evidence showing favourable effects of a physically active lifestyle on the brain and cognitive performance (2018 Physical Activity Guidelines Advisory Committee, 2018; Erickson *et al.,* 2019), which is of scientific and public health interest given the importance of cognitive abilities for general health and success in life (e.g. academic performance, Bull & Scerif, 2001). The evidence for relationships between PA and CRF and cognitive functioning, however, has long been based on research conducted in animals (Voss *et al.,* 2013) and human (older) adults (Hillman *et al.,* 2017). Over the past decade, an increasing number of studies have investigated such relationships in developing populations, including adolescents, which have been spurred, in part, by the high prevalence of physical inactivity (80% among adolescents, Guthold *et al.,* 2020) in this population and the recognition that experiences early in life are the foundation for future health (Dahl *et al.,* 2018; Sawyer *et al.,* 2012).

There is some evidence that PA is related with cognitive performance in young people, though the relationship is complex and heterogeneity in study designs, samples, and assessments of PA and cognition make firm conclusions difficult to draw. Observational studies have generally found small, but positive associations between self-reported or objective PA and composite measures of cognitive functioning (Esteban-Cornejo *et al.,* 2015; Martínez-Gómez *et al.,* 2011; McPherson *et al.,* 2018; Quan *et al.,* 2018; Ruiz *et al.,* 2010), but inconsistent associations (mostly null) between objective PA and domain-specific cognitive abilities, including composite executive functions (EF, Aadland *et al.,* 2017b; Mora-Gonzalez *et al.,* 2019), inhibition (Aadland *et al.,* 2017a; Aadland *et al.,* 2017b; Ishihara *et al.,* 2017; Ishihara *et al.,* 2018; Pindus *et al.,* 2015; Ruotsalainen *et al.,* 2020; Van Der Niet *et al.,* 2015; Wickel, 2017), working memory (Aadland *et al.,* 2017b; Ishihara *et al.,* 2017; Ruotsalainen *et al.,* 2020; Van Der Niet *et al.,* 2015), cognitive flexibility (Aadland *et al.,* 2017b; Booth *et al.,* 2013; Ruotsalainen *et al.,* 2020; Syväoja *et al.,* 2014b; Van Der Niet *et al.,* 2015), visual memory (Syväoja *et al.,* 2014b), and processing speed (Pindus *et al.,* 2015; Syväoja *et al.,* 2014b). There is also mixed evidence that relationships vary with sex, with studies showing positive (Martínez-Gómez *et al.,* 2011) and negative associations (Aggio *et al.,* 2016) in girls only, stronger relationships for boys (Booth *et al.,* 2013; Quan *et al.,* 2018), and others showing no moderation by sex (Aadland *et al.,* 2017b; McPherson *et al.,* 2018; Ruiz *et al.,* 2010; Wickel, 2017). Findings from PA intervention studies have been equally inconsistent (Wassenaar *et al.,* 2020): both no changes (Singh *et al.,* 2019; Verburgh *et al.,* 2014) and PA-related improvements (Donnelly *et al.,* 2016; Haverkamp *et al.,* 2020; Xue *et al.,* 2019) in cognitive abilities have been reported.

CRF is often regarded as a proxy for PA (Sibley & Etnier, 2003), yet is a distinct construct (Lang & Little, 2018) and an important marker of current and future health (Ortega *et al.,* 2008a; Ruiz *et al.,* 2009a), independent of PA (Ekelund *et al.,* 2007). CRF can be increased through PA (Armstrong *et al.,* 2011), which is thought be a requirement for cognitive improvements to occur (Åberg *et al.,* 2009; North *et al.,* 1990), and has been the focus of cross-sectional studies in the PA field (Kramer *et al.,* 2018; Van

Waelvelde *et al.,* 2020). Despite high between-study heterogeneity in assessment protocols, converging evidence suggests that higher CRF is linked with better performance on tasks measuring specific EFs (Van Waelvelde *et al.,* 2020), as well as those measuring visual or relational memory (Chaddock *et al.,* 2010a; Hassevoort *et al.,* 2017; Monti *et al.,* 2012). Most of these studies, however, have been conducted in small samples of children, typically comparing task performance between high and low fit individuals (Van Waelvelde *et al.,* 2020). Fewer studies have considered adolescents, with current evidence suggesting positive associations (Hogan *et al.,* 2015; Huang *et al.,* 2015; Shigeta *et al.,* 2020; Skog *et al.,* 2020; Westfall *et al.,* 2018), though null results have been reported too (Cancela *et al.,* 2019; Ruiz *et al.,* 2010; Ruotsalainen *et al.,* 2020). Moreover, few studies have explored moderating factors that could inform the design of targeted interventions, like socioeconomic status or sex, despite some indications that CRF may be related to EFs in boys only (Aadland *et al.,* 2017b; Drollette *et al.,* 2016, but see: Mora-Gonzalez *et al.,* 2019) and to academic achievement in boys (Kwak *et al.,* 2009) or girls (Grissom, 2005) and those from a higher socioeconomic background (Grissom, 2005).

To date, studies have typically focussed on either CRF or PA when assessing relationships with cognitive function. Significant progress has been made in each line of research, though their relationship with specific cognitive abilities in adolescents remains unclear. In this study, we aimed to assess relationships between CRF or PA and performance on a comprehensive set of cognitive tasks using a combination of univariate and multivariate methods in a large community sample of young adolescents. Our main objective was to assess whether PA and CRF were associated with domain-specific measures of EFs (working memory, inhibitory control and cognitive flexibility), processing speed and relational memory, and whether these associations were moderated by sex and/or socioeconomic status. We secondarily examined whether a physically active lifestyle, including higher levels of CRF and PA, was associated with cognitive performance.

4.1.1 Aims

The following primary aim was specified prior to the start of the analysis:

Primary aims

1. examine whether CRF and PA are related to performance on tasks measuring relational memory, processing speed, and executive functions, and whether these relationships vary with sex and/or socioeconomic status

We additionally prespecified the following secondary objective to further explore relationships between measures of PA and CRF and metrics of cognitive performance:

Secondary aims

1. examine whether a physically active lifestyle, including higher levels of CRF and PA, is associated with cognitive performance

4.2 Methods

Full details of the methods of the Fit to Study trial are provided in chapter 2 and (Wassenaar *et al.,* 2019a). A brief description is provided below.

4.2.1 Study design and participants

The present cross-sectional analysis included baseline cognitive, CRF and PA data collected between May and October 2017 as part of the Fit to Study cluster randomised controlled trial.

A total of 104 schools (18,261 Year-8 pupils) were recruited. Prior to baseline assessments, but post-randomisation, 11 schools dropped-out from the trial. The schools that dropped-out were not significantly different from those that remained in the trial (see chapter 2.2.4). We therefore considered the full sample consisting of 93 schools (n = 16,017, mean age [SD] = 12.5 [0.3], 8,961 (55.9%) girls; 2,665 (16.6%) eligible for free school meals [eFSM]).

4.2.2 Cardiorespiratory fitness and physical activity assessments

4.2.2.1 Cardiorespiratory fitness

CRF was measured using the 20 meter Shuttle Run (20MSR) test. This well-validated and reliable test (Castro-Piñero *et al.,* 2010; Liu *et al.,* 1992; Mayorga-Vega *et al.,* 2015) measures the maximum aerobic capacity by participants running back and forth along a 20-meter track, progressively faster, until exhaustion. The total number of laps completed is considered to be an estimate of aerobic fitness (Tomkinson *et al.,* 2017). Schools with a policy of not using the 20MSR completed the 12min run / walk Cooper Run test instead (Cooper, 1968).

4.2.2.2 Physical activity

Self-reported PA was assessed using a computer-based, online questionnaire programmed in jsPsych (De Leeuw, 2015). Single item measures were used to assess PA in the past week (Scott *et al.,* 2015) and past 6 months (i.e. habitual PA, Hagger *et al.,* 2001; details in the Appendix D.1.1.1). Attitude to PA was the average score of three items assessed on a (7-point) visual analog scale (boring —fun; stressful —relaxing; dissatisfying — satisfying), with higher scores indicating better outcomes (Michie *et al.,* 2008).

4.2.3 Cognitive assessments

Cognitive performance was assessed at home or in school with online tests of processing speed, visual relational memory, and EFs (working memory [two-back], inhibition [modified Flanker task] and cognitive flexibility [colour-shape switching], Diamond, 2013) programmed in jsPsych (De Leeuw, 2015; De Leeuw & Motz, 2016). The cognitive battery took approximately 50 minutes to complete. The task order was pseudo-randomised across participants within schools: all participants started with the task assessing processing speed, followed by the remaining four tasks in random order.

Details of the cognitive battery have been provided in the protocol (Wassenaar *et al.,* 2019b) and in chapter 2. In brief, the battery consisted of the following tasks:

- A simple reaction time task was used to assess processing speed. The average reaction time (RT) across valid trials was the primary outcome measure.
- A modified version of a previously described relational memory task (Chaddock *et al.*, 2010b) was used to assess visual relational memory performance. The proportion of correct responses on valid trials (i.e. accuracy) was the primary outcome measure.
- A modified Flanker task (Hillman *et al.*, 2014) was used to assess inhibitory control. The primary outcome was the proportion of correct responses on incongruent trials.
- A visual two-back task (Chen *et al.*, 2014a; Guiney & Machado, 2013) was used to assess working memory performance. Response accuracy was the primary outcome measure.
- A modified colour-shape switching task (Hillman *et al.*, 2014) was used to assess cognitive flexibility. The primary outcome was the proportion of correct responses on switch trials.

4.2.4 Data analysis

4.2.4.1 Data cleaning

The online questionnaire and cognitive assessments were completed at school or at home without supervision from teachers or researchers, resulting in a large proportion of careless responses. We used a conservative approach to data cleaning, that aimed to remove careless responders and data collection errors.

Cognitive assessments were cleaned per assessment. For each assessment, we identified participants with fewer than 30% valid trials (i.e. out of the total number of trials), under the assumption that a proportion (>30%) of trials is required to achieve a reliable estimate of each cognitive construct. Furthermore, for the relational memory and visual two-back task, we identified participants who consistently pressed a single response

button throughout the task, as indicated by zero counts for false alarms and hits, or by zero counts for misses and correct rejections. All identified participants were marked as careless responders and removed from the dataset. We additionally inspected the primary outcome measures of each task and removed participants with values that were outside the expected range (i.e. 0–100 for accuracy outcomes, and mean RT outside the pre-defined RT bounds, see chapter 2, section 2.3.2.1) or were missing.

Questionnaire measures included in this study were part of a larger battery that consisted of three blocks Likert-scale based items. Long strings of identical answers (i.e. same position on the scale) within a block, accounting for reversed items, were identified. The mean string length (MSL) per block was computed, and data were removed if participants gave the same answer throughout a block and recorded a string of MSL+/- 3SD in at least one more block. In addition, we removed participant scores that were outside the range of individual items (e.g. score >5 on 5–point Likert scale) due to data collection errors.

The CRF assessment was completed during PE under supervision of the PE teacher and no additional cleaning was performed. Schools that completed the Cooper Run test (n = 4) were excluded from the analysis and treated as missing, because the Cooper and 20MSR scores cannot easily be combined. A discussion of these tests is provided in Appendix D.1.1.2, as well as a validation of the 20MSR as measure of CRF.

4.2.4.2 Statistical analysis

Primary univariate analysis

All continuous outcomes were assessed using linear mixed models (LMM) that included a random intercept to capture variability between schools.

The primary analysis assessed the relationships between either PA or CRF and cognitive outcomes, and whether these relationships were moderated by sex and/or eFSM status, using model-based multilevel multiple imputation assuming data was

missing at random (MAR, details on multiple imputation are provided below and in Appendix D.1.3). Predictor variables were subjective PA during the past week (i.e. PA) or the cumulative number of laps on the 20MSR (i.e. CRF). Outcome variables were the primary outcomes of the cognitive assessments (see Cognitive assessments), generally comprising an accuracy-based measure of the most challenging condition of each task (Ludyga *et al.,* 2020), given the notion that response accuracy may be more sensitive to the effects of PA (Álvarez-Bueno *et al.,* 2017b) and a more meaningful measure in children than average RT (Davidson *et al.,* 2006; Kao *et al.,* 2017; Raine *et al.,* 2018). For the colour-shape switch task and modified Flanker task, it is common to use the difference in performance between the most complex and simple condition as outcome measure. Such difference scores (e.g. interference scores or switch costs), however, have notoriously low test-retest reliability (Hedge *et al.,* 2018a), in contrast to their component scores, and it has been recommended that these are not used for individual differences research (Draheim *et al.,* 2019; Hedge *et al.,* 2018a; Hedge *et al.,* 2018b). For this reason, we specified response accuracy on the most complex condition as primary outcome measure.

Models included a three-way interaction between sex, eFSM and the PA or CRF predictor variable, as well as their corresponding two-way interactions (sex× PA/CRF, eFSM × PA/CRF, sex × eFSM). The models additionally included participant's age as a covariate of no interest. We secondarily adjusted for an additional set of nuisance variables, including where an assessment was completed (at home or in school) and when an assessment was completed (summer, holidays or autumn; for the 20MSR, this variable had two levels: summer or autumn). We fitted each model and inspected the p-value of the interaction terms. Non-significant interactions (starting with the highest order) were removed from the model and the model was refit. For significant interactions, we computed the estimated marginal means of the linear trends.

All continuous variables were standardized (mean = 0, standard deviation = 1) to facilitate comparisons of estimates between and within models. Diagnostic checks,

including normality of residuals (level-1 and 2), were performed (non-imputed data). The residuals of the models involving the reaction time task were skewed, we therefore transformed the outcome (RT) using a rank inverse normal transformation.

Two-sided inferences with p <0.05 were considered statistically significant. The marginal (fixed effects only) and conditional explained variance (R^2) were extracted (Nakagawa *et al.,* 2017). Bonferroni correction was applied to correct for multiple comparisons (across the 10 tests, adjusted α = 0.005).

As a sensitivity analysis, we repeated the primary univariate analyses using complete-cases only. We additionally inspected the robustness of CRF and PA –cognition relationships to variations in (1) the cleaning procedure, and (2) the selected predictor, outcome and nuisance variables, using specification curve analysis (SCA), an approach that explores theoretically justifiable, statistically valid and non-redundant models (details in Appendix D.1.2).

Multivariate PA - cognition relationships: canonical correlation analysis

We also used a multivariate approach, canonical correlation analysis (CCA), to identify modes of covariation that associate physically active lifestyles and cognitive performance. CCA has previously been applied to investigate the association between a set of PA metrics to a set of body composition (Hart, 2017) and mental health measures (Wheatley *et al.,* 2020c), as well as to explore complex brain-behaviour relationships (Alnæs *et al.,* 2020; Miller *et al.,* 2016; Salvan *et al.,* 2020; Smith *et al.,* 2015). CCA is a multivariate approach that characterizes associations between two sets of variables (X and Y) by seeking for linear combinations of variables in one set (X, e.g. PA variables) that maximally correlate with a linear combination of variables in the second variable set (Y, e.g. cognitive measures).These linear combinations are called canonical variables, and their ordered correlations are called canonical correlations (ρ). Canonical variates are latent variables and the number of variates is equal to the number of variables in the smallest variable set. CCA aims to maximise the correlation of variate pairs (also

known as modes), with the constraint that each subsequent pair is uncorrelated (i.e. orthogonal) with all the previous pairs. CCA is closely related to principal components analysis (PCA), a multivariate technique that is often used for data reduction purposes (see e.g. Miller *et al.,* 2016). In contrast to PCA, which aims to reduce a *single* dataset to a number of latent factors of variation (Meng *et al.,* 2016; Wang *et al.,* 2018), CCA is designed to find latent dimensions of variability across *two* datasets (Meng *et al.,* 2016).

Here, the first variable set (X) consisted of indicators of a *physically active lifestyle* (PA lifestyle) phenotype (past week PA, habitual PA, attitudes towards PA and CRF; Table 4.1). This set additionally included the interactions between a PA lifestyle variable – created by averaging across standardized (mean = 0, SD = 1) PA and CRF variables – and sex and/or eFSM (i.e. two two-way interactions and one three-way interaction), as well as the main effect of sex and eFSM. The second variable set (Y) consisted of the cognitive outcome measures of all five tasks. Both accuracy and RT were included to capture any potential speed-accuracy trade-offs.

Table 4.1: Physically active and cognitive phenotypes used in the canonical correlation analysis

PA lifestyle phenotype	Cognitive phenotype
PA (past week)	Simple reaction time - RT
PA (habitual)	Relational memory - accuracy
Attitudes to PA	Flanker - congruent accuracy
CRF	Flanker - incongruent accuracy
PA lifestyle × sex	Flanker - congruent RT
PA lifestyle × eFSM	Flanker - incongruent RT
PA lifestyle × sex × eFSM	Task-switch - switch accuracy
sex	Task-switch - non-switch accuracy
eFSM	Task-switch - switch RT
	Task-switch - non-switch RT
	Two-back - accuracy
	Two-back - RT

Abbreviations: CRF = cardiorespiratory fitness, eFSM = eligible for free school meals, PA = physical activity,

Prior to running the CCA, we deconfounded the PA lifestyle and cognitive variable sets using set-specific nuisance variables. Nuisance variables for the PA lifestyle variable set included: participant's age, where (school/home) and when (summer, holidays, or autumn) questionnaire and CRF assessments were conducted. Nuisance variables for the cognitive variable set included: age, where (school/home) and when (summer, holidays, or autumn) each cognitive assessment was conducted. Typically, CCA is run on the residuals of the models that removed the effects (i.e. regressed out) of the nuisance variables. It has, however, recently been shown that such residualisation induces dependencies in the data that violate the assumptions of the subsequent permutation test. Moreover, simple residualisation does not take the (existing) hierarchical structure of the data into account. We therefore used Theil's method (Theil, 1965; Theil, 1968) to project the data to a lower dimensional space to obtain best linear unbiased scalar (BLUS) residuals, which are exchangeable (Magnus *et al.,* 2005; Winkler *et al.,* 2020). Data was mean-centered prior to residualisation.

A recently proposed permutation test was used to assess statistical significance of variate pairs (i.e. modes, Winkler *et al.,* 2020). In brief, a simple permutation test ignores the variability that was explained by previous canonical variables when testing the significance of the current canonical variable pair, leading to increased false positive rates except for the first set of canonical variables (or canonical correlation). The modified permutation test uses a stepwise estimation procedure, which removes variability that has already been explained (for details see: Winkler *et al.,* 2020), thereby controlling the false-positive rate. To additionally account for the hierarchical (i.e. school) structure in the data, we used a block-aware permutation test: data was freely exchangeable within, but not between, blocks (i.e. schools; Winkler *et al.,* 2015). The canonical correlation coefficients and family-wise error rate (FWER) adjusted p-values were obtained. Two-sided inferences with p <0.05 were considered statistically significant.

We subsequently computed the canonical cross-loadings for those variate pairs that resulted in significant canonical correlations. Cross-loadings (or *structure coefficients*)

were computed as the correlation (r) between the scores of the original variable set with the canonical variate(s) of the opposite set. To improve the interpretation of the cross-loadings, we deconfounded the respective variable sets. In particular, for the first variable set that reflected a PA lifestyle phenotype, we used a stepwise, selective residualisation procedure, comprising: (1) residualising the PA and CRF variables (main effects) by regressing out sex, eFSM and the default set of confounds (age, where and when assessments were completed), (2) residualising the lifestyle $\times$ sex and lifestyle $\times$ eFSM interactions by regressing out sex, eFSM, lifestyle, the main effects of PA and CRF and the default set of confounds, (3) residualising the lifestyle $\times$ sex $\times$ eFSM interaction by regressing out sex, eFSM, lifestyle, the main PA and CRF variables, the two two-way interactions and the default set of confounds. For the cognitive variable set, we regressed out the default set of confounds (age, when and where the cognitive assessments were completed). We considered of interest the cross-loadings of variables that had absolute values equal to or greater than 0.1 (i.e. small effect). In addition to the cross-loadings, we computed the redundancy index (Stewart & Love, 1968), which quantifies the amount of variance in one variable set explained by the other variable set.

Cross-validation (CV) was used to assess the generalisability of the findings. In CV, the dataset is typically partitioned into k subsets of data. One subset (i.e. the test set) is held out of the analysis and is used for validation once the model is trained on the remaining subsets. Here we used leave-one-school-out CV to examine the magnitude of the canonical correlations out of the training sample (i.e. the cross-validated canonical correlation), while taking into account the hierarchical structure in the data (Poldrack *et al.*, 2020).

As a secondary analysis, we assessed whether PA and fitness variables alone covaried with the cognitive phenotype, by treating sex and eFSM as nuisance variables (in addition to the nuisance variables stated above).

Missing data Of 16,017 pupils available for baseline assessments, 8,467 (52.9%, from 83 schools) completed one or more cognitive assessments (8,433 following data cleaning, displayed in Figure 4.1). Of participants who attempted one or more tasks, 3,128 completed all five assessments, whereas the remainder completed between one and four tasks in various (19) configurations. Moreover, 10,697 (66.8%) completed the 20MSR and 9,805 (61.2%) completed the questionnaire (9,775 following cleaning). A total of 10 schools did not attempt the tasks (resulting in exclusion of all 1,812 pupils from those schools) and a further 5,772 pupils from 83 were excluded based on not attempting the task or them being careless responders. No reasons for missing data on an individual pupil level were collected. An overview of missing data per outcome variable and predictor is provided in Appendix D.1.3.

Pupils with missing cognitive and questionnaire data were more likely to be boys and from lower socioeconomic status. For the primary confirmatory analyses, we therefore assumed that data was missing at random (MAR). We additionally used data of those pupils available from the post-intervention timepoint, which was highly correlated with their performance at baseline.

We used model-based multilevel imputation available in Blimp to multiply impute the missing data in the presence of interaction terms (Enders *et al.*, 2020). For each analysis model, we created an imputation model that contained all variables that were part of the analysis model, as well as variables that were correlated (r >0.4) with variables in the analysis model or related with missingness in variables in the analysis model. We additionally included the treatment indicator to take possible treatment effects (of post-intervention variables) into account. The full set of variables included in each imputation model is provided in Appendix D.1.3. Convergence of the MCMC chains (4 chains, at least 10,000 burn-in iterations) was inspected visually and using the potential scale reduction factor (PSR; accepted if PSR <1.05). A total of 80 imputed datasets was created for each analysis model (2,000 iterations between imputations). Model estimates were obtained in each dataset and pooled using Rubin's rules. Non-significant

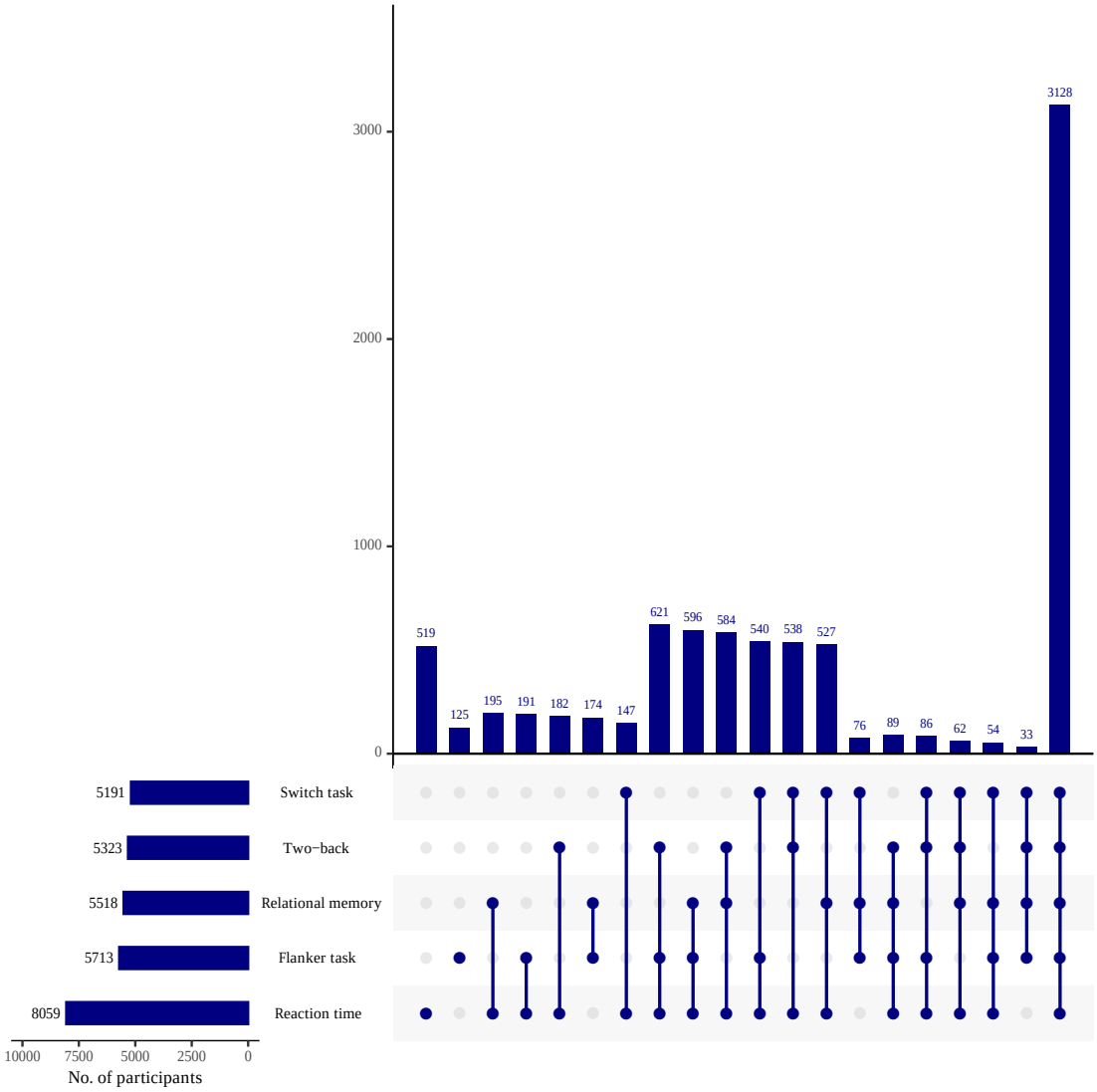


Figure 4.1: Completion of cognitive assessments. *Left bar plot* The total number of participants that completed each cognitive assessment. For instance, 8059 participants completed the reaction time task. *Right bar plot* The number of participants who completed one or more cognitive assessments. Linked dots in combination with the bar plot show the sets of cognitive assessments (dots) that were completed by a subset of participants (bars). For instance, 3128 participants completed all five cognitive assessments, indicated by the right most bar and string of connected dots. This figure was created using the *UpSetR* package in R (Conway *et al.*, 2017).

interactions were removed from the model and we assessed whether removal reduced the predictive power using a multivariate Wald test (D2). We obtained the marginal and conditional explained variance (R^2), as the averaged explained variance across imputed datasets. If the model contained significant interactions, we computed the averaged estimated marginal means of linear trends across imputed datasets.

While we had intended to run the multiple imputations for the secondary analyses too (i.e. including nuisance variables related to where and when an assessment was completed), the models did not converge. Hence, it was deemed inappropriate to analyse those imputed datasets. We, however, explored the effect of these nuisance variables in the sensitivity analyses using the complete-case dataset, as well as in the SCA.

The CCA and SCA were run using complete cases (i.e. available cases per analysis) only, given that it was computationally unfeasible to involve large scale interactions in a multiple imputation setting involving many outcomes or multiply impute the data for each model (i.e. specification) in each SCA (one per cognitive assessment). An overview of analysis software is provided in Appendix D.1.4.

4.3 Results

4.3.1 Sample

Characteristics of the sample are presented in Table 4.2. Of 16,017 participants available for baseline assessments, 8,467 (52.9%) completed one or more cognitive assessments (8,433 following data cleaning), 10,697 (66.8%) completed the 20MSR and 9,805 (61.2%) completed the questionnaire (9,775 following cleaning; details in Appendix D.2.1).

Of pupils who completed the questionnaire, approximately 20% reported that they were physically active for at least 60 min per day for the past seven days. Moreover, pupils completed an average (SD) of 37.5 (20) laps on the 20MSR. Additional school

and pupil level descriptives, as well as mean participant scores on all predictor and outcome variables are provided in Appendix D.2.2.

Table 4.2: Characteristics of schools and pupils

	Total sample
School level	
No. schools	93
Gender status, no. (%)	
Co-ed	75 (80.6%)
Female	17 (18.3%)
Male	1 (1.1%)
School type, no. (%)	
Comprehensive to 16	16 (17.2%)
Comprehensive to 18	67 (72.0%)
Secondary Modern	10 (10.8%)
School establishment, no. (%)	
Academy - Sponsor Led	16 (17.2%)
Academy Converter	3 (3.2%)
Academy Convertor - Mainstream	31 (33.3%)
Community School	15 (16.1%)
Foundation School	10 (10.8%)
Free School - Mainstream	5 (5.4%)
Voluntary aided school	11 (11.8%)
Voluntary controlled school	2 (2.2%)
School size, no. (%)	
Large (>200)	27 (29.0%)
Medium (100-200)	54 (58.1%)
Small (<100)	12 (12.9%)
No. form groups	
Median (range), group	6 (3–14)
Missing[a]	2 (2.2%)
Ofsted rating, no. (%)	
Outstanding	22 (23.7%)
Good	53 (57.0%)
Requires Improvement	10 (10.8%)
Serious Weaknesses	1 (1.1%)
Special Measures	6 (6.5%)
Missing	1 (1.1%)
Socioeconomic status	
Percent eFSM[b], mean (SD)	17.3 (11.1)
IMD, median (range), decile	5 (1–10)
School location[c], no. (%)	
Rural town and fringe	9 (9.7%)
Rural village	0 (0%)
Urban city and town	39 (41.9%)
Urban major conurbation	45 (48.4%)
Pupil level	
No. pupils	16.017
Age[d], mean (SD), yr	12.5 (0.3)
Girls, no. (%)	8961 (55.9%)
eFSM, no.(%), yes	2665 (16.6%)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, IMD = index of multiple deprivation, no. = number, SD = standard deviation

[a] A row indicating *missing* values was added only when missing values were present.

[b] Percent eFSM indicates the percentage of enrolled pupils that are eligible for free school meals within a school.

[c] Location was assessed using https://www.gov.uk/government/collections/rural-urban-classification

[d] Age at baseline (1 September 2017)

4.3.2 Associations between PA or CRF and cognitive performance

We assessed whether the relationships between CRF or PA and cognitive outcomes were moderated by participants' sex and socioeconomic status (eFSM). Table 4.3 provides standardized regression coefficients for each predictor and outcome variable. Details of the models, including unstandardized regression coefficients, are provided in Appendix D.2.3.

Table 4.3: Standardized regression coefficients of multiply imputed data

Outcome	Predictor	β[a]	95% CI	p	$R^2_{(m,c)}$[d]
Relational memory (% correct)	CRF	0.08	0.05, 0.11	<0.001	0.01, 0.04
Relational memory (% correct)	PA	0.03	0.003, 0.06	0.08	0.01, 0.04
Processing speed (RT, trans)[b]	CRF×sex[c]	0.09	0.04, 0.16	0.001	0.03, 0.09
Processing speed (RT, trans)[b]	PA	-0.07	-0.09, -0.04	<0.001	0.01, 0.08
IC (Flanker, % correct)	CRF×sex[c]	0.08	-0.13, -0.02	0.005	0.07, 0.12
IC (Flanker, % correct)	PA	0.05	0.02, 0.08	0.001	0.05, 0.1
CF (Task switch, % correct)	CRF	0.08	0.05, 0.12	<0.001	0.02, 0.07
CF (Task switch, % correct)	PA	0.06	0.03, 0.09	<0.001	0.02, 0.07
WM (two-back, % correct)	CRF	0.09	0.06, 0.13	<0.001	0.02, 0.08
WM (two-back, % correct)	PA	0.05	0.02, 0.08	0.002	0.01, 0.08

Abbreviations: CRF = cardiorespiratory fitness, CF = cognitive flexibility, IC = inhibitory control, RT = reaction time, PA = physical activity, trans = transformed, WM = working memory
[a] Standardized (mean = 0, SD = 1) regression coefficients. Rounded to two decimal places, but three decimal places are presented when rounding would have resulted in zero's
[b] Reaction times were transformed using an inverse normal transformation prior to imputations
[c] Reference level is female
[d] Marginal (m) and conditional (c) explained variance

4.3.2.1 PA – cognition relationships

Relationships between PA and cognitive outcomes were not significantly moderated by sex and/or eFSM (p >0.005). Significant positive associations were observed between PA and working memory performance (β_{PA} = 0.05, 95% CI: 0.02, 0.08, p = 0.002), cognitive flexibility (β_{PA} = 0.06, 95% CI: 0.03, 0.09, p <0.001) and inhibitory control (β_{PA} = 0.05, 95% CI: 0.02, 0.08, p = 0.001), whereas PA was negatively associated with processing speed (β_{PA} = -0.07, 95% CI: -0.09, -0.04, p <0.001). No significant association was observed between PA and relational memory performance (β_{PA} = 0.03, 95% CI: -0.003, 0.06, p = 0.08).

4.3.2.2 CRF – cognition relationships

Relationships between CRF and cognitive outcomes were not significantly moderated by sex and eFSM (i.e. no significant three-way interactions), nor by eFSM alone (p >0.005). The positive association between CRF and inhibitory control (i.e. Flanker task performance), and between CRF and processing speed (i.e. RT performance) was significantly moderated by sex (inhibition: $\beta_{sex \times CRF}$ = -0.076, 95% CI: -0.13, -0.02, p = 0.005; processing speed: $\beta_{sex \times CRF}$ = 0.092, 95% CI = 0.04, 0.15, p = 0.001), with stronger relationships for girls (inhibition: β_{CRF} = 0.17, p <0.001; processing speed: β_{CRF} = -0.2, p <0.001) compared to boys (inhibition: β_{CRF} = 0.09, p <0.001; processing speed: β_{CRF}= -0.11, p <0.001). Positive associations were observed between CRF and cognitive flexibility (β_{CRF} = 0.08, 95% CI: 0.05, 0.12, p <0.001), working memory (β_{CRF} = 0.09, 95% CI: 0.06, 0.13, p <0.001), and relational memory (β_{CRF} = 0.08, 95% CI: 0.05, 0.11, p <0.001).

4.3.2.3 Sensitivity analysis

In the complete-case analysis, the relationship between CRF and inhibitory control was no longer significantly moderated by sex ($\beta_{sex \times CRF}$ = -0.08, uncorrected-p = 0.01), but sex appeared as a significant moderator of the association between CRF and cognitive flexibility ($\beta_{sex \times CRF}$ = -0.14, p <0.001). The relationships between CRF or PA and the remaining cognitive assessments remained in the same direction as those of the primary analysis, albeit with small variations in the (un)standardized coefficients. Additional adjustment for where and when assessments were conducted did not change the results. Details of the models are provided in Appendix D.2.3.2.

Specification curve analysis (SCA) was used to assess the robustness of the associations between PA or CRF and cognitive outcomes to variations in cleaning of the cognitive assessments. Justifiable adaptations in the cleaning approach resulted in small variations in the observed standardized regression coefficients (β) for CRF and PA (details in Appendix D.2.4.1), but generally did not change conclusions regarding directionality or significance of the relationships.

We then inspected the robustness of the relationships to other justifiable choices in predictors (binary versus continuous), outcomes (e.g. accuracy and RT), and nuisance variables (e.g. task-specific or demographic), in addition to the cleaning procedure used. Details of the SCA, including specification curves and median regression coefficients are provided in the Appendix D.2.4.2. Key observations include:

- Across cognitive assessments, higher CRF and PA are generally associated with more favourable cognitive outcomes.
- Cognitive assessments with multiple outcomes (two-back task, Flanker task, and colour-shape task-switching task) demonstrate consistently positive relationships of CRF and PA with accuracy outcomes (of component scores), but no clear relationships with RT outcomes.
- RT difference scores (interference scores) of the Flanker task (inhibitory control) and colour-shape switching task (cognitive flexibility) were positively related with CRF and PA, suggesting that higher CRF and PA are related to larger difference scores.

4.3.3 Multivariate associations between a physically active lifestyle and cognitive performance

CCA was used to assess the multivariate relationship between a physically active lifestyle and cognitive performance. CCA identified two significant modes of covariation between the physically active lifestyle phenotype and the measures of cognitive performance. The redundancy index showed that 2.2% of variance in PA lifestyle variables was explained by variance in cognitive variates, whereas 2.5% of variance in cognitive variables was explained by variance in PA lifestyle variates.

In the first mode of covariation, the active lifestyle variate, including sex, eFSM and their interactions with PA lifestyle variables, was positively correlated with the cognitive variate (Figure 4.2a, ρ = 0.31, p = 0.001, ρ_{CV} = 0.26; additional details are provided in Appendix D.2.5.1). This mode was characterised by high cross-loadings

for CRF (r = 0.19), habitual PA (r = 0.14), past week PA (r = 0.13) and eFSM status (r = -0.2) in the PA lifestyle variate, and high cross-loadings for accuracy outcomes of the EF and relational memory assessments, including working memory (r = 0.23), inhibitory control ($r_{congruent}$ = 0.16), cognitive flexibility (r_{switch} = 0.14, $r_{non-switch}$ = 0.15) and relational memory (r = 0.1), as well as processing speed (r = -0.14), in the cognitive variate (Figure 4.2a). Overall, this mode therefore suggests that fitter and more active adolescents tended to show higher accuracy scores and faster reaction times on cognitive tests. The high loading for eFSM status reflects lower scores on the PA and cognitive variate for those from a lower socioeconomic background (Figure 4.2a)

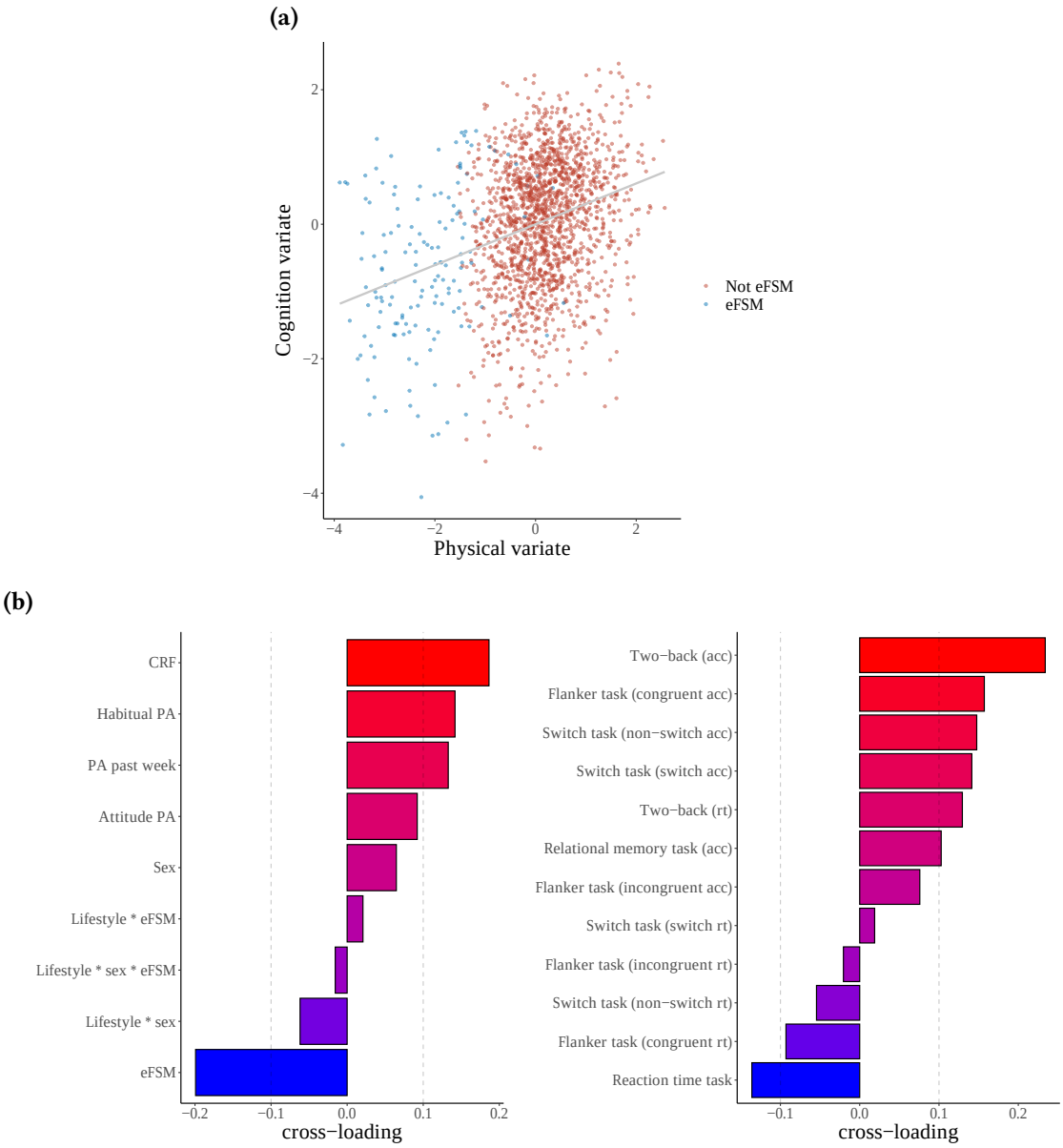


Figure 4.2: Canonical correlation and cross-loadings of the first mode of covariation. (*a*) shows the subject scores on the first physical and cognition variate, colour coded by eFSM status. Participants from a lower socioeconomic background (eFSM, in blue) had lower scores on the physical and cognition variate compared to those from a higher socioeconomic background. (*b*) displays the cross-loadings between the physically active lifestyle variables and the cognition variate (*left*) and between the cognitive variables and the PA lifestyle variate (*right*). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest.

The second mode of covariation demonstrated a positive correlation between the PA lifestyle variate and the cognitive variate (Figure 4.3a, ρ = 0.24, p = 0.001, ρ_{CV} = 0.21). Sex had the highest negative cross-loading in the PA lifestyle variate (r = -0.22), whereas components of inhibitory control had high positive cross-loadings in the cognitive variate (Figure 4.3b, accuracy: $r_{incongruent}$ = 0.21; reaction time: $r_{incongruent}$ = 0.17, $r_{congruent}$ = 0.16). Overall, this mode therefore reflects sex differences in cognitive performance, with generally higher accuracy and reaction times on the Flanker task in girls compared to boys, suggestive of a speed-accuracy trade-off.

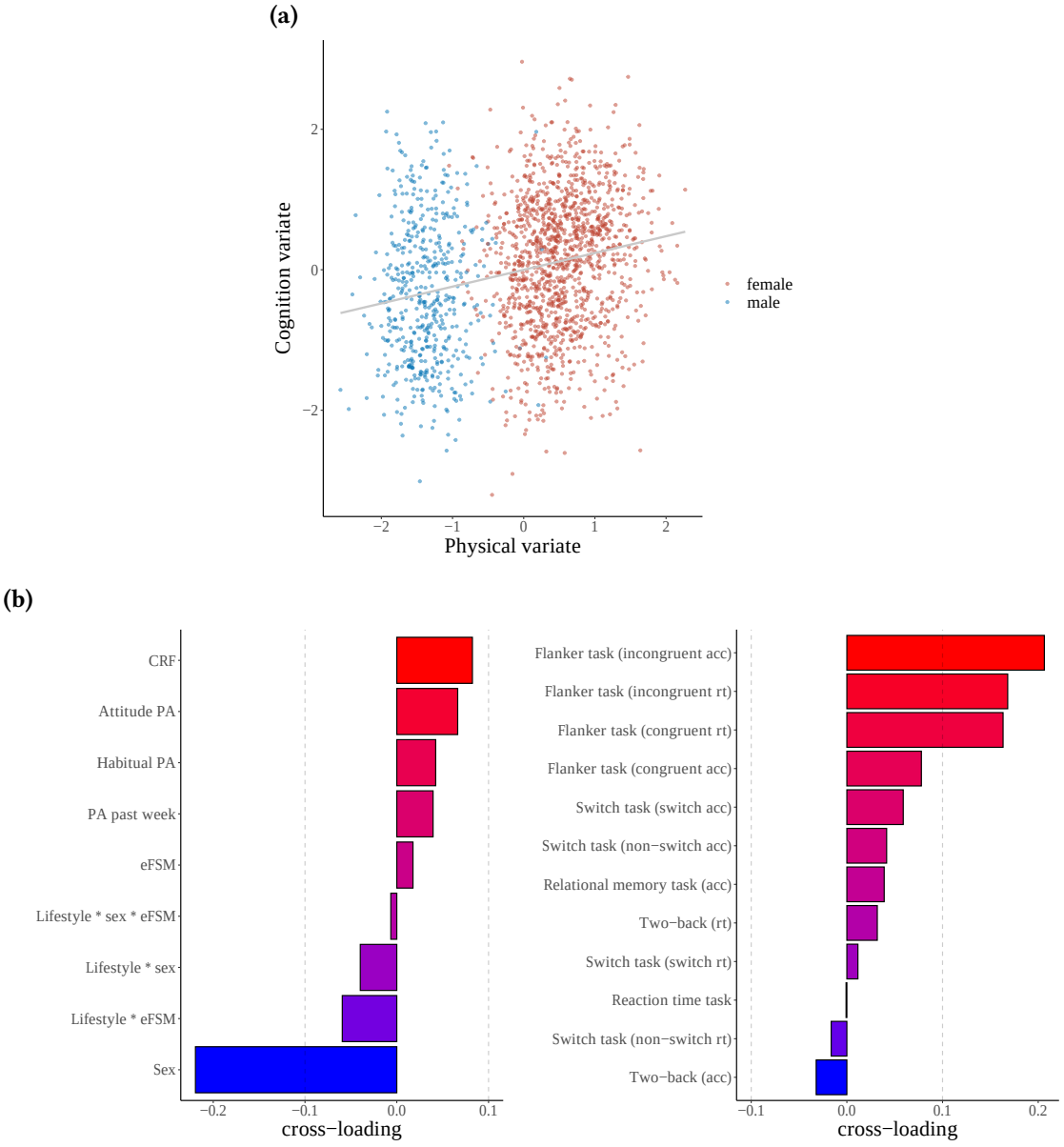


Figure 4.3: Canonical correlation and cross-loadings of the second mode of covariation. (*a*) shows the subject scores on the first physical and cognition variate, colour coded by sex (blue = boys, red = girls). Girls had higher scores on the physical and cognition variate compared to boys. (*b*) displays the cross-loadings between the physically active lifestyle variables and the cognition variate (left) and between the cognitive variables and the PA lifestyle variate (right). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest.

We secondarily assessed whether a physically active phenotype, adjusted for sex, eFSM and the default set of nuisance variables, covaried with cognitive measures. CCA identified a single significant positive mode of covariation (ρ = 0.24, p = 0.001, ρ_{CV} = 0.21), with 2.8% of variance in PA variables being explained by variance in the cognitive variate, and 1% of variance in cognitive variables being explained by the PA variate. In line with the outcomes of the primary CCA, all PA lifestyle variables had high (>0.1) cross-loadings on the cognitive variate, whereas accuracy on the executive function tasks and reaction time of the processing speed task had high cross-loadings (0.1–0.16 on the PA variate, details in Appendix D.2.5.2).

4.4 Discussion

In a large cohort of British adolescents, we assessed whether PA and CRF were associated with processing speed, relational memory and core EFs, and examined whether these associations were moderated by sex and socioeconomic status using a combination of univariate and multivariate approaches. We observed associations between PA or CRF and pre-specified cognitive measures, albeit relationships were small, with models explaining less than 10% of the variance in cognitive performance. Sex moderated the relationship between CRF and inhibitory control, and between CRF and processing speed, with stronger relationships in girls compared to boys. The observed relationships varied marginally with data cleaning approach, but greatly with choice of outcome measures, showing robust positive relationships when accuracy outcomes were considered. The results of the canonical correlation analysis were in line with and extended the univariate analyses by showing that a physically active lifestyle, including higher PA and CRF, is associated with more favourable cognitive outcomes.

4.4.1 Relationships between PA, CRF and cognitive performance

There is a widespread assumption that PA and cognitive performance are positively related, but evidence from studies involving young people is mixed. Studies in this age group have shown positive relationships between objective and self-reported PA and

composite measures of cognition, but there is no converging evidence of significant relationships between PA and domain-specific measures of inhibitory control (Aadland *et al.,* 2017a; Aadland *et al.,* 2017b; Ishihara *et al.,* 2017; Ishihara *et al.,* 2018; Pindus *et al.,* 2015; Ruotsalainen *et al.,* 2020; Van Der Niet *et al.,* 2015; Wickel, 2017), working memory (Aadland *et al.,* 2017b; Ishihara *et al.,* 2017; Mora-Gonzalez *et al.,* 2020; Pindus *et al.,* 2015; Pindus *et al.,* 2016; Ruotsalainen *et al.,* 2020; Van Der Niet *et al.,* 2015; Wickel, 2017), cognitive flexibility (Aadland *et al.,* 2017b; Booth *et al.,* 2013; Ruotsalainen *et al.,* 2020; Syväoja *et al.,* 2014b; Van Der Niet *et al.,* 2015), visual memory (Syväoja *et al.,* 2014b) and processing speed (Pindus *et al.,* 2015; Syväoja *et al.,* 2014b). In contrast, our findings suggest that self-reported PA is positively associated with such domain-specific measures of EF and processing speed, but these relationships were very small, and unlikely to be detected with typically used sample sizes (tens to hundreds of pupils).

In addition, the majority of the previously reported positive relationships between PA and cognitive performance have been found with subjective- rather than objectively measured PA. It is possible that, because self-reported and objectively-measured PA are typically poorly related (LeBlanc & Janssen, 2010), our self-report measure of PA is capturing something that is differentially associated with cognitive performance. This discrepancy is notable in other, similar fields: consistent positive relationships with self-reported, but not objectively-measured PA have been reported for studies examining relationships between PA and academic achievement (Marques *et al.,* 2017; Syväoja *et al.,* 2014a).

Our findings that higher CRF is associated with better performance on tasks assessing EFs, relational memory and processing speed align with previous studies conducted in children and adolescents (Chaddock *et al.,* 2010a; Herting & Nagel, 2012; Huang *et al.,* 2015; Van Waelvelde *et al.,* 2020; Westfall *et al.,* 2018). They strengthen the evidence for a positive association between CRF and cognitive abilities in (young) adolescents, highlight the importance of maintaining or acquiring healthy CRF levels (Lang *et al.,* 2019; Lang & Little, 2018; Tomkinson *et al.,* 2018), and support calls for CRF surveillance

in this population (Lang & Little, 2018; Sandercock & Cohen, 2019).

Of particular interest is that sex, but not socioeconomic status, significantly moderated relationships between CRF, but not PA, and cognitive outcomes. Primary analyses demonstrated that the relationship of CRF with inhibitory control and processing speed were stronger in girls than boys. Sensitivity analyses did not replicate the results for inhibitory control, despite being in the right direction, but suggested that the relationship between CRF and cognitive flexibility was stronger in girls instead. While the moderation results should therefore be interpreted with caution, it is striking that relationships tended to be stronger in girls, not in boys as reported previously (Aadland *et al.,* 2017b; Drollette *et al.,* 2016). Such sex differences may be due to a host of factors, including differences in physical or cognitive maturation, differential sensitivity of assessments (Aadland *et al.,* 2017b), lower CRF of girls leaving more room for improvement (Yildirim *et al.,* 2011), or differences in motivation. The latter may have played a major role in this study, as boys (and pupils from a lower socioeconomic background) were less likely to complete cognitive assessments and were more often identified as careless responders. Future studies are encouraged to further explore sex and socioeconomic status as effect moderators.

4.4.1.1 Beyond simple linear regressions

While univariate linear regressions have proven extremely useful for establishing relationships, they are typically limited to a single outcome and predictor variable of interest. Canonical correlation analysis (CCA) may provide complementary information, given that it characterises associations between two sets of (inter-related) variables. In this study, we used CCA to demonstrate that a physically active lifestyle, including higher levels of PA and CRF, as well as higher SES is associated with higher accuracy and lower reaction time —in other words, "better" performance —on tasks measuring EFs, relational memory and processing speed, in line with findings from independent cross-sectional studies. Moreover, CCA also demonstrated that sex was associated with

inhibitory control, with girls showing better performance, in contrast with previous reports (Chojnacki *et al.,* 2018; Pontifex *et al.,* 2011; Westfall *et al.,* 2018). While a sex-difference in cognitive maturation cannot be excluded, a difference in motivation may be the more likely explanation.

Both our univariate and multivariate analyses highlight the potential benefits of PA and higher CRF for cognitive performance, although neither allows for conclusions regarding causality. Moreover, models containing CRF and PA explained less than 10% of the variance in cognitive outcomes and standardized regression estimates, while significant, were small (<0.2). Such small effect sizes are in line with previous cross-sectional (Álvarez-Bueno *et al.,* 2020; Shigeta *et al.,* 2020; Westfall *et al.,* 2018) and intervention studies (Álvarez-Bueno *et al.,* 2017b; Xue *et al.,* 2019), and raise questions about the practical value of CRF and PA – cognition associations. However, considering that cognitive development is influenced by a host of genetic and (socio-)environmental factors, most of which cannot be easily modified, even small effects of CRF and PA, factors which can potentially be shifted through interventions, may be meaningful.

4.4.2 Strengths and limitations

Obtaining small, yet significant effects is a feature of large-scale studies, which are becoming increasingly common in the field of PA and cognition (Resaland *et al.,* 2016; Shigeta *et al.,* 2020; Solis-Urra *et al.,* 2019). Large-scale studies may more accurately represent the population if sampled broadly and randomly, and allow for the identification of tiny effects that are not statistically significant in smaller samples. Our study is, to the best of our knowledge, the largest study to date to capture valid measures of PA and CRF, as well as a wide-range of specific cognitive abilities in a community sample of adolescents. We demonstrated the feasibility of web-based cognitive assessments at scale in this population, and used a comprehensive set of univariate and multivariate

analyses to show robust relationships between measures of physical health and cognitive performance.

Testing at scale is not without challenges. In particular, the cross-sectional design of this study —which is typical of most large datasets —makes it difficult to distinguish causes and effects. Rare interventional studies, such as the Fit to Study cluster randomised trial (Wassenaar *et al.,* 2019a), could establish directionality. Moreover, assessments are required to be simple and brief. As such, self-report, rather than objective, measures of PA were collected, which may be subject to social-desirability and recall bias, particularly in young people. Objective PA measurements avoid such biases, but have their own challenges (Migueles *et al.,* 2017), including the logistics of distributing measurement devices to thousands of pupils. Smartphone-based accelerometers may prove useful in future studies. The cognitive assessments were selected to capture a broad-range of cognitive domains rather than a precise measurement of single domain, which would have required multiple measures of e.g. inhibitory control (Friedman & Miyake, 2017). Large-scale testing in schools and homes is inevitably less well-controlled than smaller, lab-based studies. Consequently, high amounts of missing data and careless responses were present, which may bias inferences if not appropriately dealt with. We used model-based multilevel multiple imputation for the primary analyses to deal with missing data, and assessed the robustness of the results to data cleaning with SCA. The secondary SCA and CCA were, however, conducted on complete-case data, given that it was computationally unfeasible to use multiply-imputed data.

4.5 Conclusion

In summary, higher levels of PA and CRF are associated with better performance on tasks measuring EFs and processing speed in adolescents. CRF, but not PA, was significantly positively related with relational memory performance, and more strongly related with inhibition and processing speed in girls than boys, suggesting sex-specific

relationships that may have implication for the design and analysis of intervention studies. Adolescents who lead a physically active life show better cognitive performance. While these findings do not support causal conclusions, they highlight the need for global PA action in young people.

This page is intentionally left blank.

5

Associations between cardiorespiratory fitness, physical activity, the anterior hippocampus and memory

Contents

5.1 Introduction . . . 140
5.1.1 Aims . . . 142
5.2 Methods . . . 143
5.2.1 Design and participants . . . 143
5.2.2 Cardiorespiratory fitness and physical activity assessments . . . 143
5.2.3 Cognitive assessments . . . 144
5.2.4 MRI data acquisition . . . 144
5.2.5 MRI data analysis . . . 145
5.2.6 Confounds . . . 149
5.2.7 Data analysis . . . 151
5.3 Results . . . 159
5.3.1 Univariate analyses . . . 160
5.3.2 Multivariate relationship between fitness, physical activity and brain MR markers . . . 164
5.4 Discussion . . . 168
5.4.1 Relationships of CRF or PA with anterior hippocampal volume and memory . . . 168
5.4.2 Beyond hippocampal volume . . . 170
5.4.3 Strength and limitations . . . 171
5.5 Conclusion . . . 172

The previous chapter has shown that higher levels of physical activity and cardiorespiratory fitness are associated with superior cognitive outcomes, but did not provide insight into the neural correlates of a physically active lifestyle. Converging evidence from animal and human research suggests that the hippocampus, a region vital for learning and memory, is particularly sensitive to the effects of physical activity: higher levels of physical activity and fitness have typically been related to greater hippocampal volume, particularly its anterior part. Less is known about such relationships in young people. Using data from the Fit to Study brain imaging sub-study (n = 61), this chapter demonstrates that neither physical activity nor cardiorespiratory fitness are significantly related to anterior hippocampal volume or hippocampal-dependent memory performance. I additionally demonstrate that a physically active lifestyle is not significantly associated with MRI markers of tissue (micro)structure in the anterior hippocampus in this cohort of young adolescents.

5.1 Introduction

Higher levels of physical activity (PA) and cardiorespiratory fitness (CRF) have been associated with superior cognitive performance (demonstrated in chapter 4) and may positively impact brain health (Hillman *et al.,* 2008; Valkenborghs *et al.,* 2019). The hippocampus, a region vital for learning and memory (Cohen & Eichenbaum, 1993; Eichenbaum *et al.,* 2016), has been shown to be particularly sensitive to the effects of PA (Kandola *et al.,* 2016; Rendeiro & Rhodes, 2018; Stillman *et al.,* 2020). Research using rodents has consistently shown exercise-induced molecular and cellular changes in this region, including enhanced synaptic plasticity, neurogenesis, angiogenesis and an upregulation of neurotrophic factors, such as brain-derived neurotrophic factor (BDNF; Vivar & Van Praag, 2017; Voss *et al.,* 2013). A large number of neuroimaging studies in humans have complemented these findings, showing that PA and CRF are related with larger volumes of the hippocampus (Erickson *et al.,* 2009; Firth *et al.,* 2018; Hamer *et al.,* 2018; Maass *et al.,* 2015; Raichlen *et al.,* 2019), especially its anterior portion (Erickson *et al.,* 2011; Killgore *et al.,* 2013; Stillman *et al.,* 2018; Thomas *et al.,* 2016), and that hippocampal volumes may mediate the relationship between CRF and memory performance (Chaddock *et al.,* 2010a; Erickson *et al.,* 2011; Erickson *et al.,* 2009). Most

of the evidence in humans, however, is based on studies conducted in adults, with a particular focus on ageing populations. Far less research has considered relationships between PA or CRF and hippocampal volumes in children and adolescents. While studies involving healthy (Chaddock *et al.,* 2010a; Gorham *et al.,* 2019) and overweight (Esteban-Cornejo *et al.,* 2017; Migueles *et al.,* 2020) children have found some support for the positive associations reported in adults (Chaddock *et al.,* 2010a; Esteban-Cornejo *et al.,* 2017; Gorham *et al.,* 2019, but see Migueles *et al.,* 2020), current evidence in adolescents is sparse and inconsistent, with studies showing both positive (Herting & Nagel, 2012) and no relationships (Ruotsalainen *et al.,* 2019) between CRF or moderate-to-vigorous PA (MVPA) and hippocampal volumes.

The typically reported differences or changes in hippocampal volume, albeit theoretically (and clinically) meaningful (Barnes *et al.,* 2009; Fotuhi *et al.,* 2012), provide little insight into the cellular processes driving these effects (Anderson, 2011; Roth *et al.,* 2010). While the candidate mechanisms reported by animal studies (e.g. neurogenesis, Van Praag, 2005; Van Praag *et al.,* 1999) cannot be directly assessed in humans, advances in brain imaging have enabled researchers to look beyond macrostructural indices of tissue volume derived from conventional anatomical scans. For instance, higher PA or CRF have been related to greater hippocampal perfusion (Chaddock-Heyman *et al.,* 2016; Maass *et al.,* 2015) and blood volume (Pereira *et al.,* 2007), as well as increased tissue density (i.e. lower mean diffusivity, MD, Kleemeyer *et al.,* 2016) and neuronal integrity (Den Ouden *et al.,* 2018; Pajonk *et al.,* 2010), which were thought to underlie increases in hippocampal volumes (Kleemeyer *et al.,* 2016; Maass *et al.,* 2015). Focussing on a single imaging modality, however, may limit conclusions regarding possible cellular mechanisms. For instance, lower MD may be due to angiogenesis and neurogenesis (Kleemeyer *et al.,* 2016), and an association with perfusion does not exclude processes like neurogenesis to take place. More insight may be gained from studies combining multiple modalities. A multimodal imaging study in middle-aged adults indicated, for instance, that exercise-induced increases in hippocampal volume are better explained

by increases in myelin instead of vasculature (Thomas *et al.*, 2016).

To further the evidence in young adolescents, the present cross-sectional study assessed the relationship between CRF or MVPA, the anterior hippocampus and memory performance, using data from 61 participants collected as part of Fit to Study's brain imaging sub-study. The primary aim was to examine whether CRF and PA were related with anterior hippocampal volume and memory performance in our cohort of early adolescents. We secondarily explored whether a physical phenotype, encompassing measures of PA and CRF, was associated with (micro)structural MRI markers in the anterior hippocampus.

5.1.1 Aims

The following primary aim was specified prior to the start of the analysis:

Primary aims

1. examine whether CRF and PA are related with anterior hippocampal volume and memory performance

We additionally prespecified the following secondary objective to further explore relationships between measures of PA and CRF and MRI metrics in the anterior hippocampus:

Secondary aims

1. perform exploratory multivariate analysis of the relationship between measures a physical phenotype, encompassing measures of PA and CRF, and MRI (micro)structural markers in the anterior hippocampus

We additionally examined whether hippocampal volumes were related to memory performance and repeated the primary univariate analyses with the entire hippocampus to ensure comparability with previous studies. We also explored multivariate relationships between a physical health phenotype, as defined in Salvan *et al.* (2020), and MRI markers in the anterior hippocampus.

5.2 Methods

5.2.1 Design and participants

Sixty-one healthy individuals from 10 schools participating in the Fit to Study trial enrolled in the brain imaging sub-study. Full methodological details are provided in chapter 2 and in Wassenaar *et al.* (2019a). The present study used *baseline* data of the Fit to Study brain imaging sub-study.

5.2.2 Cardiorespiratory fitness and physical activity assessments

5.2.2.1 Cardiorespiratory fitness

CRF ($\dot{V}O_2$peak) was assessed by a maximal graded exercise test on a cycle ergometer (Lode Excalibur, Groningen, The Netherlands), using the *Godfrey* protocol (Godfrey & Mearns, 1971). The test was terminated when the participant reached volitional exhaustion or was unable to maintain a cadence of 60 rounds per minute despite verbal encouragement.

5.2.2.2 Physical activity

Objective PA was assessed using the Axivity AX3 accelerometer (Newcastle, UK). Accelerometers were set to record at 100 Hz and were worn for seven days, starting on Monday (08:00), during a typical school week following the baseline MRI assessment. Bespoke software was used to convert raw accelerometer data into 1 s epochs and to calculate the daily average minutes of sedentary, light, moderate or vigorous PA, according to established cut-points (Phillips *et al.*, 2013).

Self-reported PA was assessed using a single item (Scott *et al.*, 2015) asking: "*In the past week, on how many days have you done a total of 60 minutes or more of physical activity, which was enough to raise your breathing rate?*".

5.2.3 Cognitive assessments

Hippocampal-dependent relational memory performance was assessed using an object-location task (OLT) and the relational memory paradigm used in chapter 4. Details are provided in chapter 2 and Appendix E.1.1.

1. **Object-location task** A modified version of a previously described OLT (Pertzov *et al.,* 2012; Pertzov *et al.,* 2013) was used to assess relational memory performance. The proportion of misbinding / swap errors was used as the outcome measure.

2. **Relational memory paradigm** A modified version of a previously described relational memory task (Chaddock *et al.,* 2010b) was used to assess visual relational memory performance. The proportion of correct responses on valid trials (i.e. accuracy) was the primary outcome measure.

5.2.4 MRI data acquisition

Details of the MRI protocol and scan parameters are described in chapter 2 and Appendix E.1.2. In brief, all MR data (T_1-weighted images, resting-state fMRI, diffusion-weighted images, quantitative MRI and arterial spin labelling) were acquired on a 3 Tesla (Prisma, Erlangen, Germany) MRI scanner at the Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB) using a 32-channel head coil. The (baseline) protocol took approximately 50 minutes (Figure 5.1).

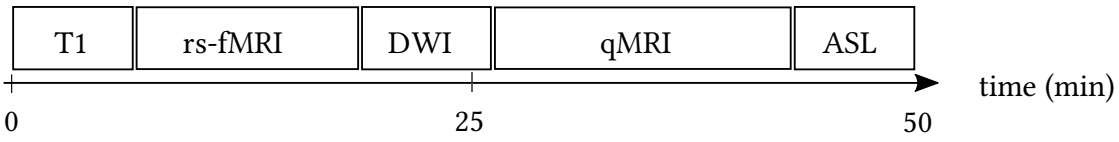


Figure 5.1: MRI protocol at baseline. Participants underwent a 50 minute MRI scan, comprising a T_1-weighted scan (T1), a resting-state fMRI (rs-fMRI), a diffusion-weighted scan (DWI), a quantitative MRI sequence (qMRI) and an arterial spin labelling (ASL) scan. The size of the scan blocks approximate scan duration.

Participants watched a documentary during all structural scans, and were instructed to blink normally, not to fall asleep, and to look at a cross on the screen during the functional MRI scans. This chapter includes all sequences but the resting-state fMRI scan.

5.2.5 MRI data analysis

Image analysis was performed using tools from the FMRIB Software Libraby (FSL version 6; Jenkinson *et al.,* 2012; Smith *et al.,* 2004), Statistical Parametric Mapping (SPM, http://www.fil.ion.ucl.ac.uk/spm/), and FreeSurfer (version 6.0; Dale *et al.,* 1999).

5.2.5.1 T_1-weighted processing

T_1-weighted images were corrected for gradient distortions using tools developed by FreeSurfer and Human Connectome Project (HCP) project teams (https://github.com/Washington-University/Pipelines). Images were then processed using the FSL_ANAT anatomical processing pipeline available within FSL. Briefly, this pipeline performs bias field correction using FAST (Zhang *et al.,* 2001), registration to MNI152 standard space T_1 template using FLIRT (Jenkinson *et al.,* 2002), followed by FNIRT (Jenkinson *et al.,* 2012), and registration-based brain extraction. Tissue-type segmentation was then applied using FAST and subcortical structures were modelled using FIRST (Patenaude *et al.,* 2011). The volumetric scaling factor from SIENAX (Smith, 2002) was used as a proxy for head size (Miller *et al.,* 2016). Grey matter density estimates were derived using a modified voxel-based morphometry (VBM) pipeline. To this end, GM partial volume estimates derived from FAST were non-linearly registered to MNI152 standard space, multiplied by the Jacobian determinant and smoothed with a 2 mm Gaussian kernel.

5.2.5.2 Hippocampal segmentation

FSL's Multimodal Image Segmentation Tool (MIST, Visser *et al.,* 2016) was used for hippocampal segmentations, following a qualitative and quantitative comparison with FreeSurfer and FIRST (see Appendix E.1.3.1 for a discussion of segmentation methods). MIST has been shown to outperform existing methods in the segmentation of subcortical structures (Visser *et al.,* 2016) and consists of two stages: a training stage, in which it learns the intensity and shape of the structure using unlabelled training data, and a segmentation stage, in which the trained model is used to segment the images. While MIST can use multiple image modalities simultaneously (e.g. T_1, FA, FLAIR), it cannot deal with missing data. Nevertheless, MIST has been shown to outperform FreeSurfer

and FIRST segmentations, even if only T_1 images were considered (Visser *et al.,* 2016). Due to missing modalities for some subjects, we trained MIST using gradient- and bias-field corrected T_1-weighted images only. The training set contained 49 baseline scans (chosen at random) and 49 post-intervention scans to give equal weight to both timepoints.

Visual inspection of the MIST-derived hippocampal segmentations indicated a considerable degree of variability in the accuracy of segmentations of the hippocampus across subjects. All hippocampi were therefore manually edited. A protocol for manual adjustments and inter- and intra-rater reliability statistics (all ≥ 0.9) is provided in Appendix E.1.3.2. A single experienced rater (TW) adjusted the hippocampi of all participants across both timepoints (n = 108, see chapter 7). Manual corrections were done in random order and blinded to group allocation to avoid experimenter bias.

Anterior hippocampal segmentation The hippocampus can be subdivided into an anterior and posterior portion, which differ in function, structure and (functional) connectivity with the rest of the brain (Poppenk *et al.,* 2013; Strange *et al.,* 2014; Zeidman & Maguire, 2016). Various methods exist to divide the hippocampus along its long axis into its anterior (or head) and posterior (tail and/or body) region, including those that are based on preferential functional connections with other brain regions (e.g. Libby *et al.,* 2012; Voets *et al.,* 2014), and those that use anatomical boundaries. Among the latter, researchers have used landmark-based segmentation (e.g. the uncal apex, Malykhin *et al.,* 2007; Poppenk & Moscovitch, 2011; Weiss *et al.,* 2005), MNI-coordinate based segmentation (e.g. Y = -21, Fjell *et al.,* 2019; Langnes *et al.,* 2019; Li *et al.,* 2018), percentile-based segmentation (e.g. anterior 25%, Hackert *et al.,* 2002; Thomas *et al.,* 2016), and centre-of-gravity based segmentation (e.g. Chaddock-Heyman *et al.,* 2016; Erickson *et al.,* 2011; for a discussion, see Poppenk *et al.,* 2013).

In line with recommendations by Poppenk *et al.* (2013), we used the uncal apex as internal anatomical landmark for delineation. The uncus is particularly developed

in humans and primates (Duvernoy, 2005), and is clearly visible on T_1-weighted scans (Poppenk *et al.*, 2013; Zeidman & Maguire, 2016). We aligned the T_1-weighted scan with the MNI-152 template using FSL's FLIRT and then defined the anterior hippocampus as the region anterior to the most posterior coronal slice in which the uncal apex was visible. The demarcation was performed by a single rater, blind to group allocation and in random order. Details of this method, including inter-and intra-rater reliability statistics, are provided in Appendix E.1.3.3.

5.2.5.3 Diffusion-weighted images

Diffusion-weighted images were corrected for eddy currents, EPI distortions, and inter- and intra volume subject head motion using FSL's eddy tool (Andersson & Sotiropoulos, 2015; Andersson & Sotiropoulos, 2016). Diffusion data was then corrected for gradient distortions (Glasser *et al.*, 2013) and subsequently fed into two complementary pipelines. First, the diffusion tensor model, including the kurtosis model, was fit to the data using FSL's DTIFIT to obtain fractional anisotropy (FA), mean diffusivity (MD) and mean kurtosis (MK) maps. In addition to DTI fitting, the Bingham-Neurite Orientation Dispersion and Density Imaging (NODDI) model (Tariq *et al.*, 2016) was fit to the data using cuDIMOT toolbox in FSL (Hernandez-Fernandez, 2018). The Bingham-NODDI model generates maps of microstructural parameters, including the intra-cellular volume fraction (ICVF or fintra), the isotropic (free water) volume fraction (ISOVF) and the orientation dispersion index (ODI).

5.2.5.4 Quantitative MRI

Quantitative MRI data (i.e. multi-echo PDw and T_1w images) were corrected for gradient distortions and subsequently processed using the Voxel-Based Quantification (VBQ) toolbox available in SPM to obtain R_1 ($1/T_1$) and R_2^* ($1/T_2^*$) maps (Callaghan *et al.*, 2014). The VBQ processing steps include correction for B_1^+ (B_1 transmit) and B_1^- (RF receive sensitivity) field inhomogeneities.

5.2.5.5 Pseudo-continuous ASL

Pseudo-continuous ASL scans were corrected for bias-field and gradient distortions and subsequently preprocessed using Bayesian Inference for Arterial Spin Labelling (BASIL) tool available in FSL. BASIL uses a variational Bayes approach to fit the general kinetic model on a voxel-by-voxel basis (Chappell *et al.,* 2009) to generate quantitative GM perfusion maps (in ml blood / 100 g tissue / min).

5.2.5.6 Between modality registrations

Diffusion-weighted images, as well as R_1, R_2^* and perfusion maps, were registered to T_1-weighted native space using FLIRT with boundary-based cost function (Greve & Fischl, 2009; Jenkinson *et al.,* 2002).

5.2.5.7 Quality control

All raw and preprocessed images have been manually checked for gross problems affecting image quality, including motion artifacts and registration errors. Two diffusion-weighted scans, three quantitative MRI scans, and five ASL scans were considered of too low quality –mainly due to motion artifacts –for analysis and were therefore removed from the dataset. A total of 59 T_1-weighted scans, 58 diffusion-weighted and resting-state fMRI scans, 51 quantitative MRI scans and 49 ASL scans were available for analysis (Figure 5.2). Note, a decreasing number of scans is available for sequences that were acquired towards the end of the protocol, suggesting a negative effect of scan duration on data acquisition and quality.

5.2.5.8 MRI-derived markers

Following preprocessing and quality control of MRI data, the volume of the left and right (anterior) hippocampus was extracted in subject's native space. In addition, the hippocampus was used as a mask to derive MRI measures of tissue perfusion and microstructure (R_1, R_2^*, FA, MD, ISOVF, ICVF/fintra and ODI). To prevent interpolation of meaningful MRI-map values (such as R_1, perfusion), we spatially aligned hippocampal masks to each modality using the transformation matrix derived from

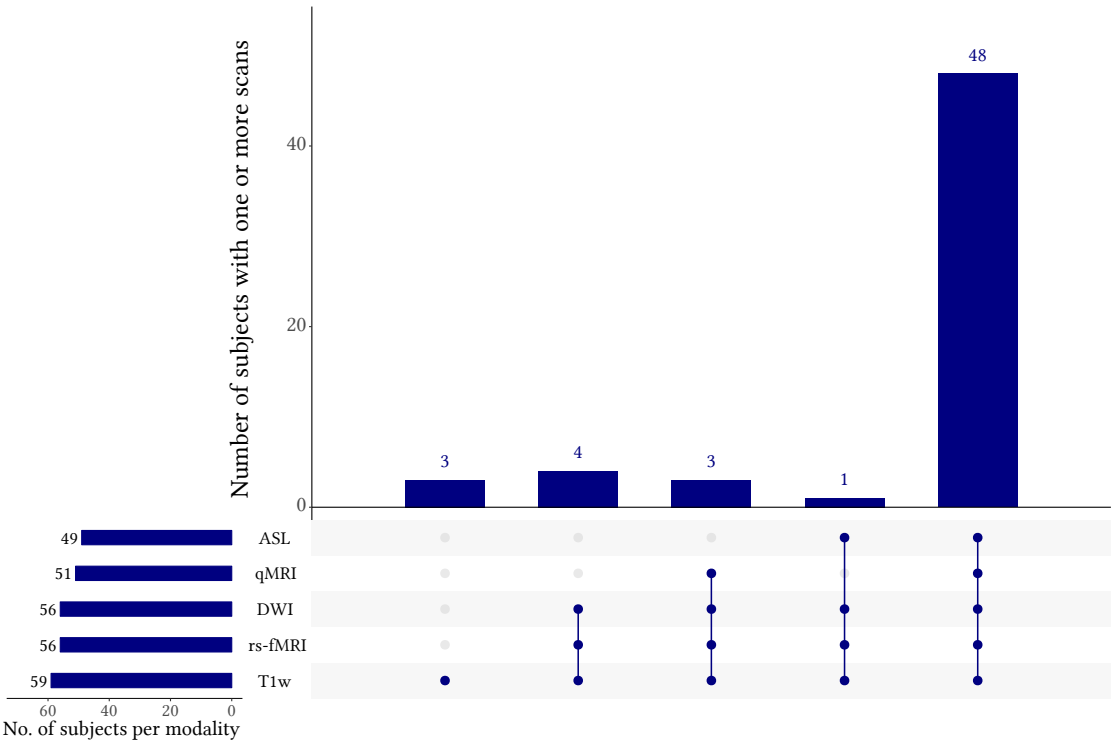


Figure 5.2: Available MRI scans at baseline. *Left bar plot* The number of available scans per sequence (following quality control). For instance, a T_1-weighted scan was available for 59 out of 61 participants. *Right bar plot* The number of participants with one or more complete modalities. Linked dots in combination with the bar plot show the sets of modalities (dots) that were present for a subset of participants (bars). For instance, complete MRI data was available for 48 participants.

FLIRT. We subsequently extracted metrics of central tendency (mean, median) and dispersion (standard deviation, interquartile range) within each hippocampal mask.

5.2.6 Confounds

5.2.6.1 Behavioural confounds

In line with chapter 4, we included age (at the day of MRI), sex and socioeconomic status (postcode-derived multiple index of deprivation) as confounds. We also controlled for pubertal status, as measured by the Pubertal Development Rating Scale (Peterson *et al.,* 1988). Puberty has well-documented effects on brain structure and function in adolescence (Goddings *et al.,* 2014; Herting *et al.,* 2017; Satterthwaite *et al.,* 2014; Vijayakumar *et al.,* 2018b), including hippocampal volume (Bramen *et al.,* 2011; Herting

et al., 2017).

Only few previous studies have considered markers of physical health, including body-mass index (BMI), blood pressure (diastolic, systolic) and heart rate (HR), that are related (i.e. share variance) with brain and behaviour, as confounds (e.g. Opel *et al.,* 2019; Spartano *et al.,* 2019). These physiological variables could theoretically be considered as mediators of relationships between PA/CRF and brain structure or cognition if the study was interventional, and treating them as confounds could underestimate relationships (Suttorp *et al.,* 2015). The cross-sectional nature of the data prevents claims regarding directionality, however, we acknowledge that these physical variables are different in nature from the regular set of confounds. To explore the unique relationship of CRF or PA with anterior hippocampal volume and cognition, we additionally adjusted for this set of physical health variables in univariate analyses.

5.2.6.2 MRI related confounds

MRI-related confounds, also termed *nuisance*, can be broadly classified into those that are scanner related and those that originate from subjects. Scanner related confounds, including magnetic field ($\mathbf{B}_0$ and $\mathbf{B}_1$) inhomogeneities, and gradient nonlinearities, are addressed in post-processing steps. Commonly identified human-related confounds include subject motion, head size, hydration levels, seasonality and diurnal fluctuations (discussed in chapter 3).

In this study, we controlled for head size only, because no information was collected on hydration levels, all participants were scanned within a two-month period in summer 2017, motion-related artefacts were removed from the analysis during quality control (see section 5.2.5.7: Quality control), and time-of-day did not show significant relationships with any of the brain MRI metrics (all FWE-corrected $p > 0.05$, Appendix E.1.4). We adjusted for head size using the volumetric scaling factor from SIENAX (Alfaro-Almagro *et al.,* 2018; Miller *et al.,* 2016; Smith, 2002) by including it as a covariate in the analysis (O'Brien *et al.,* 2011).

5.2.7 Data analysis

5.2.7.1 Univariate analysis

Linear mixed effects models (LMM) were used to examine relationships between CRF or PA, left and right anterior hippocampal volume and relational memory performance. All models included a random intercept to capture between school variability and were fit with restricted maximum likelihood (REML) with Kenward-Roger correction —which is more appropriate for small sample sizes (McNeish, 2017) —using the *lme4* and *lmerTest* packages in R (Bates *et al.,* 2015; Kuznetsova *et al.,* 2017). When the models did not converge due to e.g. a singular fit (usually indicating that the random effects structure does not fit the data), the random intercept was removed and the model was refit as a single level linear model using the *stats* package. Diagnostic checks, including homogeneity of variance, linearity and normality of residuals (level-1 and 2) were performed.

The primary analyses assessed the relationship between CRF or daily average minutes of MVPA and left or right anterior hippocampal volume, and memory performance. We secondarily assessed whether left and right anterior hippocampal volumes were related to relational memory performance. All models were fit using the default set of variables of no interest: age, sex, socioeconomic status and pubertal status (model 1). For models that included anterior hippocampal volums as predictor or outcome, a proxy for head size was also included. Of 59 participants who completed the relational memory task, 30 completed version 1 and 29 version 2. Performance between versions was not significantly different (β_{unst} = -1.84, p = 0.55). A total of 17 participants performed the relational memory task as part of the main trial of Fit to Study, but their performance was not significantly different from those who completed the task for the first time (β_{unst} = -2.7, p = 0.42). However, to take possible learning effects into account, we added a dummy variable indicating whether a participant had completed the task before their participation in the sub-study. We subsequently additionally controlled for (i.e. model 2) markers of physical health (BMI, systolic and diastolic blood pressure and resting HR) to explore the unique variance between CRF or MVPA and hippocampal and

cognitive outcomes (i.e. by removing shared variance with physical health markers).

While it is common to additionally run the analysis with a region-of-no-interest (e.g. the nucleus accumbens, Chaddock *et al.,* 2010a; Erickson *et al.,* 2011; total GM volume, Herting & Nagel, 2012; thalamus, Erickson *et al.,* 2011) to demonstrate specificity of the associations, we decided that this was not needed for two reasons: (1) the aim was to replicate previous results of relationships with the hippocampus, rather than showing specificity to the hippocampus, (2) there is evidence to suggest that associations between CRF or PA and the brain are not specific to the hippocampus and may instead be diffuse (Batouli & Saba, 2017; Salvan *et al.,* 2020).

In line with chapter 4, all continuous variables were standardized (mean = 0, SD = 1) prior to fitting the models in order to (visually) compare the magnitude of regression coefficients (β estimates) across models (NB. this serves only as an indication; no conclusions can be drawn from such an informal comparison). Unstandardized regression coefficients of all models are presented in Appendix E.2. For single-level models the total amount of variance in the outcome explained by the predictors (R^2) is provided, whereas for LMMs, the proportion of variance in the outcome explained by the fixed factors only (i.e. marginal R^2, or R^2_m) and by the fixed and random factors combined (conditional R^2, or R^2_c) is provided (Nakagawa *et al.,* 2017). For all statistical tests, a two-sided p <0.05 was used and Bonferroni-correction was applied to correct for multiple comparisons. An overview of analysis software is provided in Appendix E.1.7.

5.2.7.2 Multivariate analysis: cross-validated canonical correlation analysis

We performed exploratory multivariate analysis linking brain and behaviour using canonical correlation analysis (CCA, Hart, 2017; Hotelling, 1936; Miller *et al.,* 2016; Sherry & Henson, 2005; Smith *et al.,* 2015; Wang *et al.,* 2018). In particular, we examined whether a set of physical activity and fitness variables (i.e. a behaviour phenotype) is

associated with a set of hippocampal MRI markers of tissue (micro)-structure (i.e. a brain phenotype).

Variable selection

Compared with the main trial (chapter 4), where the number of available variables for CCA was small, the brain imaging sub-study contains a rich set of measures capturing brain and behaviour. In contrast to common variable selection methods, which are usually data-driven and identify variables based on the presence of high correlation values or through other forms of clustering or classification, we decided to take a more supervised approach to variable selection for two reasons. First, we were interested in only a subset of variables that are known to be markers of physical or brain health. Second, the interpretation of multivariate techniques like CCA is often complex, particularly when (seemingly) theoretically unrelated variables are part of the model. Therefore, we selected a subset of PA and fitness variables that show moderate to high cross-correlations and are theoretically informative, comprising:

- $\dot{V}O_2$peak (ml/kg/min), as a measure of cardiorespiratory fitness,
- maximum workload (Watt) achieved on the maximal graded exercise test, as a measure of strength and fitness,
- self-reported PA (no. of days), as a measure of (sustained) MVPA,
- weekly minutes of sedentary, moderate and vigorous PA, as an objective measure of time spent physically active in the various intensity zones.

In addition, we also identified BMI, systolic- and diastolic blood pressure, and resting HR as measures of physical health. In line with univariate analyses, these factors were not included in the initial CCA, but added in a secondary step.

MRI markers were selected based on whether they contained signal in the hippocampus and were sensitive to underlying tissue types. Specifically, each MRI marker's distribution of mean, median, standard deviation and interquartile range values across

subjects was inspected and modalities for which there was more variability (indicated by the SD and IQR) than signal were excluded (ISOVF, kurtosis). Then, histological, animal and human studies linking MRI markers to (GM) tissue were reviewed (details in chapter 3), and those markers with the highest potential to inform about tissue microstructural characteristics and exercise-related neurobiological mechanisms (e.g. angiogenesis, neurogenesis, myelination) were selected. We excluded metrics that are predominately sensitive to tissue geometry, such as FA and ODI. The metrics that were included comprised:

- voxel-based morphometry (VBM): a proxy for GM volume,
- mean diffusivity (MD): a measure of water diffusion and tissue density,
- neurite density (ICVF or fintra): a measure of neurite density,
- perfusion: a measure of blood flow (expressed in ml / 100 g / min),
- R_1: the longitudinal relaxation rate, a measure that has been used as a myelin marker,
- R_2^*: the transverse relaxation rate, a measure that is sensitive to both iron and myelin.

Importantly, none of the MRI metrics mentioned above are specific to any particular tissue type and there is a lack of validation studies (for a discussion, see chapter 3).

Deconfounding and dimensionality reduction

Behavioural and brain MRI markers were deconfounded prior to running the CCA, to ensure that CCA results are not driven by nuisance variables. Regressing out many nuisance variables in small samples with linear regression may affect power, potentially inflating the false positive rate, and causes the residuals to be non-independent, which violates the assumptions of the subsequent permutation test (Winkler *et al.*, 2020). While this dependency diminishes with increasing sample size, it requires adjustment

of the residuals in small samples. In addition, the data contains a hierarchical, school-based, dependence structure. To account for these dependencies in deconfounding, we used a method that reduces the data from N subjects with P nuisance variables to N-P observations (Huh & Jhun, 2001), and ensured that residuals were uncorrelated by computing BLUS residuals (Best, Linear, Uncorrelated and Scalar, Magnus *et al.,* 2005; Theil, 1965). An R script to compute BLUS residuals is provided in Appendix E.1.5). To minimise the loss of data in this small dataset, we reduced the number of nuisance variables using principal components analysis (PCA). The first two principal components (explaining 62.2 % variance) were used to obtain deconfounded brain and behaviour datasets.

Cross-validated canonical correlation analysis We used CCA to find a *linear* combination of variables in the physical behaviour phenotype that maximally correlated with a linear combination of variables in the brain phenotype (CCA details are provided in chapter 4). Block-aware permutation testing was used to assess statistical significance of the canonical correlations while taking the hierarchical structure of the data into account. We assessed the generalisability of the CCA results using leave-one-school-out cross-validation (CV).

While it is tempting to inspect and interpret all significant canonical variates, this practice is, in most cases, problematic (Sherry & Henson, 2005). That is, variate pairs max explain very little variance between the variable sets, which result in possible over-interpretation of findings (Sherry & Henson, 2005). In addition, CCA is often used with large sample sizes (see e.g. Miller *et al.,* 2016; Smith *et al.,* 2015), and large samples can result in statistically significant findings that may not be practically significant. Conversely, a researcher should also be careful if small samples are used (Pugh & Hu, 1991). Barcikowski & Stevens (1975) indicated that about 20 participants per variable are required in order to accurately interpret *only* the largest canonical correlation. Other researchers suggested that as few as 10 participants per variable should be sufficient to interpret CCA results (Marascuilo & Levin, 1983), but more recent papers suggest

that the number of subjects need to be greater than the number of variables in the largest variable set (Wang *et al.,* 2018, but see: Helmer *et al.,* 2020). Based on these considerations, we decided to only test and interpret the first canonical correlation.

In line with the univariate analyses, the primary CCA analysis considered the set of physical and MRI hippocampal markers defined above, adjusted for the default set of confounds (PCA across: age, sex, pubertal development, socioeconomic status and head size). To account for the deconfounding-related random selection of (P) subjects that are removed, we created 1000 sets of N-P subjects and ran 1000 permutations for each of them, thereby creating a distribution of family-wise error corrected p-values. The final statistical significance level was computed as the average p-value of the distribution. To partially replicate analyses performed in Salvan *et al.* (2020) for hippocampal markers, we repeated the CCA analysis but with a broader behavioural phenotype, that additionally comprised markers of physical health (BMI, resting HR, and systolic- and diastolic blood pressure).

5.2.7.3 Missing data

The missing data pattern (Figure 5.3) and relationships between variables with missing data were inspected. Participants had complete data for age, sex and the OLT. For the majority of the variables, missingness was below 10%, but qMRI, ASL and actigraphy-based PA estimates had between 10-23% missing data. Reasons for missing actigraphy data include removal of the monitor because it interfered with their daily tasks or because of discomfort (e.g. for sleeping). Missing MRI data is commonly due to poor scan quality or participant non-compliance. A single participant dropped-out of the trial following cognitive assessments.

Univariate data For univariate analyses, complete-case data was used, because imputation models did not converge, rendering imputed values unreliable. Given that actigraphy data was missing for 16% (10 cases) of the sample and only 2% (1 case) of the sample had missing CRF data, we conducted the analyses for CRF and MVPA on available cases. We additionally repeated the analyses on a single imputed dataset (N

= 61, details below). The results of these additional analyses are presented in Appendix E.2.1.2.

Multivariate analysis Only 36 (57%) participants had complete data on markers for physical and brain health (i.e. MRI), which considerably reduces power. Moreover, small samples are a particular concern to approaches such as CCA. The primary multivariate analysis was therefore conducted on imputed data (assuming data was missing at random, MAR). Given the complexity and size of the dataset (i.e. large number of variables compared to the number of participants), the missing data was imputed once using a multilevel singular value decomposition (SVD) method available in *missMDA* (Husson *et al.*, 2019). It accurately predicts missing values by iteratively performing a SVD on a dataset that is decomposed into its within and between groups variability. The number of dimensions (or components) of the SVD is determined using cross-validation (here, between = 7, within = 4). Baseline and post-study variables, including hippocampal MRI metrics and an intervention group indicator, were included in the imputations to preserve relationships in the data and leverage the rich set of measures. A list of included variables is provided in Appendix E.1.6. As a sensitivity analysis, we repeated the primary analysis with complete-case data only.

5.2.7.4 Supplementary analysis

Given that previous studies in young people have focussed on the entire hippocampus rather than its anterior part (Chaddock *et al.*, 2010a; Herting & Nagel, 2012), we repeated the univariate analyses with left and right hippocampal volume.

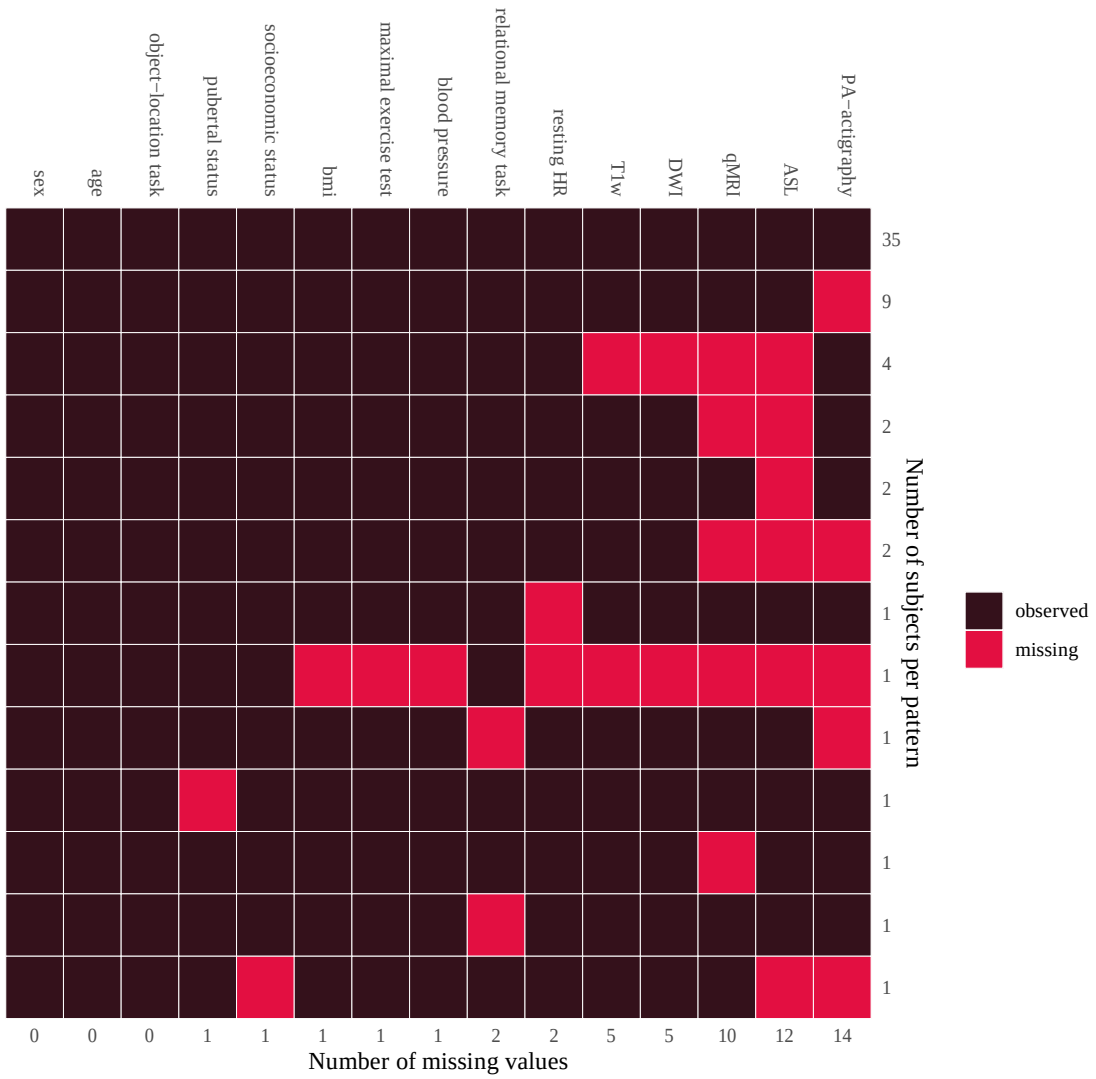


Figure 5.3: Missing data pattern brain imaging sub-study (baseline data). A non-monotone pattern of missing values is present across demographics, behaviour and brain MRI variables. *X-axis* displays the number of missing values for column-wise variables. *Y-axis* displays the number of subjects that are present in each missing data pattern presented row-wise. The number of rows reflect the number of missing data patterns present in this subset of the data. For instance, the *top row* indicates the complete-case scenario (n = 35), in which all variables are observed, whereas the bottom row indicates a pattern of missing data (missing variables: socioeconomic status, ASL and PA-actigraphy) that was present in one participant. Abbreviations: ASL = arterial spin labelling, BMI = body mass index, DWI = diffusion-weighted imaging, qMRI = quantitative MRI, T1w = T_1-weighted imaging. This figure was created using the *JointAI* package (Erler *et al.*, 2019).

5.3 Results

Participant characteristics are displayed in Table 5.1. A total of 61 participants were included (mean age = 12.5, SD = 0.33), 30 (49%) of whom were female. The representativeness of this sample with respect to the total Fit to Study sample was assessed in chapter 2.3.4.4. In brief, schools of the sub-sample were representative of the main trial sample, yet participants were generally more fit compared to the main trial sample.

Table 5.1: Baseline characteristics brain imaging sample

	Characteristics	Missing n (%)
Demographics		
Schools, *n*	10	0 (0)
Age[a], mean (SD), y	12.48 (0.33)	0 (0)
Female, *n* (%)	30 (49.2)	0 (0)
Socioeconomic status		
IMD, median (range), decile	8 (2–10)	1 (2)
eFSM, no. (%), yes	1 (1.6%)	1 (2)
Pubertal status, mean (SD)	2.2 (0.63)	1 (2)
Fitness, physical activity and physical health		
Fitness		
$\dot{\mathbf{V}}O_2$peak, mean (SD), ml/kg/min	37.2 (8.41)	1 (2)
Workload maximum, mean (SD), W	136 (25.8)	1 (2)
Physical activity		
Daily average MVPA, mean (SD), min / day	47.5 (22.1)	14 (23)
Sedentary behaviour, mean (SD), min / wk	4300 (663)	14 (23)
Moderate PA, mean (SD), min / wk	278 (121)	14 (23)
Vigorous PA, mean (SD), min / wk	46.2 (37.1)	14 (23)
Self-report PA, median (range), days / wk	5 (2–7)	0 (0)
Physical health		
BMI, mean (SD), kg/m^2	19.38 (3.8)	1 (2)
Resting systolic blood pressure, mean (SD), mm Hg	106 (10.3)	1 (2)
Resting diastolic blood pressure, mean (SD), mm Hg	72.3 (7.40)	1 (2)
Resting heart rate, mean (SD), bpm	75.5 (12.7)	2 (3)
Cognitive assessments		
Object location task		
Identification accuracy, mean (SD), %	84.2 (10.5)	0 (0)
Localisation errors, mean (SD), %	11.2 (3.96)	0 (0)
Neirest neighbour errors, mean (SD), %	5.46 (1.37)	0 (0)
Swap errors, mean (SD), %	14.0 (8.26)	0 (0)
Associates task		
Accuracy, mean (SD), %	69.8 (11.8)	2 (3)
MRI hippocampal parameters		
Left anterior hippocampus volume, mean (SD), mm^3	1430 (290)	2 (3)
Right anterior hippocampus volume, mean (SD), mm^3	1540 (273)	2 (3)
Left hippocampus volume, mean (SD), mm^3	2710 (362)	2 (3)
Right hippocampus volume, mean (SD), mm^3	2760 (356)	2 (3)

Abbreviations: BMI = body-mass index, bpm = beats-per-minute, eFSM = eligible for free school meals, MRI = magnetic resonance imaging, *n* = number, IMD = index of multiple deprivation, y = year, PA = physical activity, W = Watt, wk = week

[a] Age at time of (baseline) scan

5.3.1 Univariate analyses

5.3.1.1 CRF, MVPA and anterior hippocampal volume relationships

CRF and daily average minutes of MVPA were not significantly associated with left (β_{CRF} = 0.21, 95%CI: -0.1, 0.52, p = 0.18; β_{MVPA} = 0.04, 95%CI: -0.29, 0.36, p = 0.83) or right (β_{CRF} = 0.13, 95%CI: -0.18, 0.45, p = 0.40; β_{MVPA} = -0.05, 95%CI: -0.39, 0.29,

p = 0.77) anterior hippocampal volume, accounting for age, sex, socioeconomic status, pubertal status and head size (model 1, Table 5.2 and Figure 5.4). Additionally controlling for physical health markers resulted in a similar pattern of results (left anterior hippocampus: β_{CRF} = 0.18, 95%CI: -0.21, 0.56, p = 0.43 and β_{MVPA} = 0.02, 95%CI: -0.32, 0.36, p = 0.90; right anterior hippocampus: β_{CRF} = 0.14, 95%CI: -0.23, 0.52, p = 0.45 and β_{MVPA} = -0.03, 95%CI: -0.39, 0.34, p = 0.88, model 2). Sensitivity analysis using imputed data did not change the findings (details in Appendix E.2.1.2).

Table 5.2: Analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and left or right anterior hippocampal volume (standardized).

	Left anterior hippocampus					Right anterior hippocampus				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
$\dot{V}O_2$peak	56	0.21	-0.1, 0.52	0.18	0.26	56	0.13	-0.18, 0.45	0.40	0.26
MVPA[a]	44	0.04	-0.29, 0.36	0.83	0.14, 0.2	44	-0.05	-0.39, 0.28	0.77	0.2
Model 2										
$\dot{V}O_2$peak	56	0.18	-0.21, 0.56	0.43	0.25, 0.29	56	0.14	-0.23, 0.52	0.45	0.31
MVPA[a]	44	0.02	-0.32, 0.36	0.9	0.17, 0.29	44	-0.03	-0.39, 0.34	0.88	0.25

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

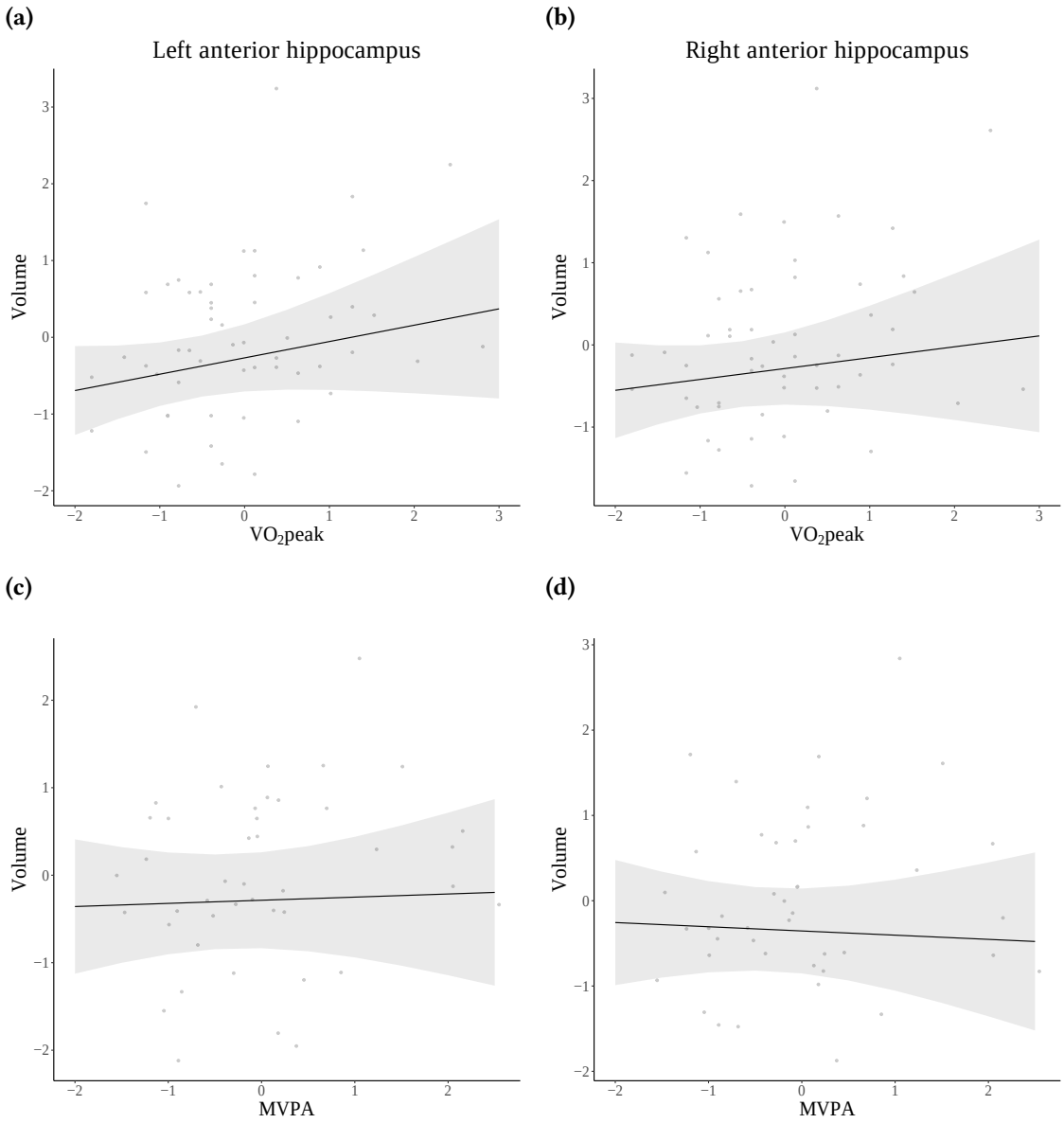


Figure 5.4: Relationship between CRF or MVPA and left and right anterior hippocampal volume (estimated marginal means). Scatter plot of the association between CRF ($\dot{\mathbf{V}}O_2$peak) and left (*a*) and right (*b*) anterior hippocampal volume, as well as between MVPA and left (*c*) and right (*d*) anterior hippocampal volume, adjusted for age, sex, pubertal status, socioeconomic status and head size. The predicted (marginal) effects are displayed with 95% confidence intervals, based on the standard error, and created with the *ggeffects* package in R (Lüdecke, 2018). Dots represent raw values for CRF, MVPA and anterior hippocampal volumes and have been added for visualisation purposes.

5.3.1.2 CRF, MVPA and relational memory relationships

CRF or daily average minutes of MVPA were not significantly associated with the proportion of swap errors on the object location task (β_{CRF} = -0.26, 95%CI: -0.55, 0.03, p = 0.12; β_{MVPA} = 0.15, 95%CI: -0.16, 0.46, p = 0.35) or response accuracy (β_{CRF} = 0.14, 95%CI: -0.17, 0.46, p = 0.36; β_{MVPA} = -0.01, 95%CI: -0.37, 0.35, p = 0.95) on the relational memory paradigm, controlling for age, sex, socioeconomic status and pubertal status (Table 5.3, model 1). Additionally controlling for physical health markers did not change these results (all p >0.05, model 2). Sensitivity analysis using imputed data did not change the findings (details in Appendix E.2.1.2).

Table 5.3: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and memory outcomes (standardized).

	Object-location task					Relational memory paradigm				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
$\dot{V}O_2$peak	57	-0.26	-0.55, 0.03	0.12	0.16, 0.21	55	0.14	-0.17, 0.46	0.36	0.1
MVPA[a]	45	0.15	-0.16, 0.46	0.35	0.14, 0.29	44	-0.01	-0.37, 0.35	0.95	0.11
Model 2										
$\dot{V}O_2$peak	57	-0.12	-0.51, 0.28	0.59	0.18, 0.29	55	0.01	-0.39, 0.41	0.95	0.2
MVPA[a]	45	0.16	-0.18, 0.49	0.37	0.14, 0.28	44	-0.01	-0.38, 0.35	0.95	0.25

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

5.3.1.3 Anterior hippocampal volume and relational memory

Left or right anterior hippocampal volumes were not significantly related with performance on the object location task (β_{left} = 0.12, 95%CI: -0.17, 0.40 , p = 0.45; β_{right} = 0.04, 95%CI: -0.24, 0.33, p = 0.77) or relational memory paradigm (β_{left} = 0.15, 95%CI: -0.17, 0.47, p = 0.35; β_{right} = 0.05, 95%CI: -0.28, 0.38, p = 0.75), controlling for age, sex, socioeconomic status and pubertal status (Table 5.4, model 1). Additionally controlling for physical health markers did not change these results (all p >0.05, model 2). Sensitivity analysis using imputed data did not change the findings (details inAppendix E.2.1.2).

Table 5.4: Univariate analyses assessing relationships between left and right anterior hippocampal volumes and memory outcomes (standardized).

	Object-location task (swap errors)					Relational memory paradigm (accuracy)				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
Left ant. hip.	56	0.12	-0.17, 0.40	0.45	0.15, 0.24	54	0.15	-0.17, 0.47	0.35	0.1
Right ant. hip.	56	0.04	-0.24, 0.33	0.77	0.15, 0.23	54	0.05	-0.28, 0.38	0.75	0.08
Model 2										
Left ant. hip.	56	0.10	-0.21, 0.4	0.55	0.18, 0.31	54	0.16	-0.17, 0.49	0.34	0.2
Right ant. hip.	56	0.02	-0.28, 0.32	0.90	0.18, 0.31	54	0.09	-0.25, 0.42	0.61	0.19

Abbreviations: CI = confidence interval, ant. hip. = anterior hippocampus
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

5.3.1.4 Additional analyses: The entire hippocampus

The majority of previous studies have considered the whole hippocampus in their analyses (e.g. Chaddock *et al.,* 2010a; Herting & Nagel, 2012). To allow for comparisons between studies and replicate their results, we repeated the univariate analyses with the volume of the entire hippocampus. No significant relationships were found between CRF or daily average minutes of MVPA and left hippocampal volume (β_{CRF} = 0.12, p = 0.48, β_{MVPA} = -0.09, p = 0.59), or right hippocampal volume (β_{CRF} = 0.11, p = 0.46, β_{MVPA} = -0.16, p = 0.31). Left or right hippocampal volume was not significantly associated with relational memory performance on either of the tasks (p >0.05). Details of the models, as well as sensitivity analyses, are presented in Appendix E.2.2.

5.3.2 Multivariate relationship between fitness, physical activity and brain MR markers

5.3.2.1 Physical activity, fitness and brain MR markers

We did not find a significant mode of covariation linking individual differences in PA and fitness with differences in hippocampal MRI markers (ρ = 0.69, p = 0.29, ρ_{CV} = 0.11), independent of age, sex, socioeconomic status, pubertal status and head size. Figure 5.5 shows the canonical correlation and cross-loadings of the first mode of covariation (CV canonical correlation is shown in Appendix E.2.3.1). About 4.5% of variance in

PA lifestyle variables was explained by variance in the brain variate, whereas 3% of variance in cognitive variables was explained by the PA lifestyle variate.

The analysis was repeated with complete-cases (n = 36) only. While the estimates changed, no significant mode of covariation was detected (p = 0.97). Details are provided in Appendix E.2.3.2.

5.3.2.2 Secondary analysis of a physically active lifestyle and brain MR markers

We secondarily aimed to broaden our physical phenotype by also including markers of physical health: BMI, blood pressure (systolic and diastolic) and resting HR. About 5% of variance in PA lifestyle variables was explained by variance in the brain variate, whereas 5.6% of variance in brain variables was explained by the PA lifestyle variate, yet no significant mode of covariation was present in the data (ρ = 0.78, p = 0.14, ρ_{CV} = 0.21; Figure 5.6; CV canonical correlation is shown in Appendix E.2.3.1). In line with these results, secondary analysis with complete-cases only (n = 36) showed no significant mode of covariation (p = 0.77). Details of these analyses are presented in Appendix E.2.3.2.

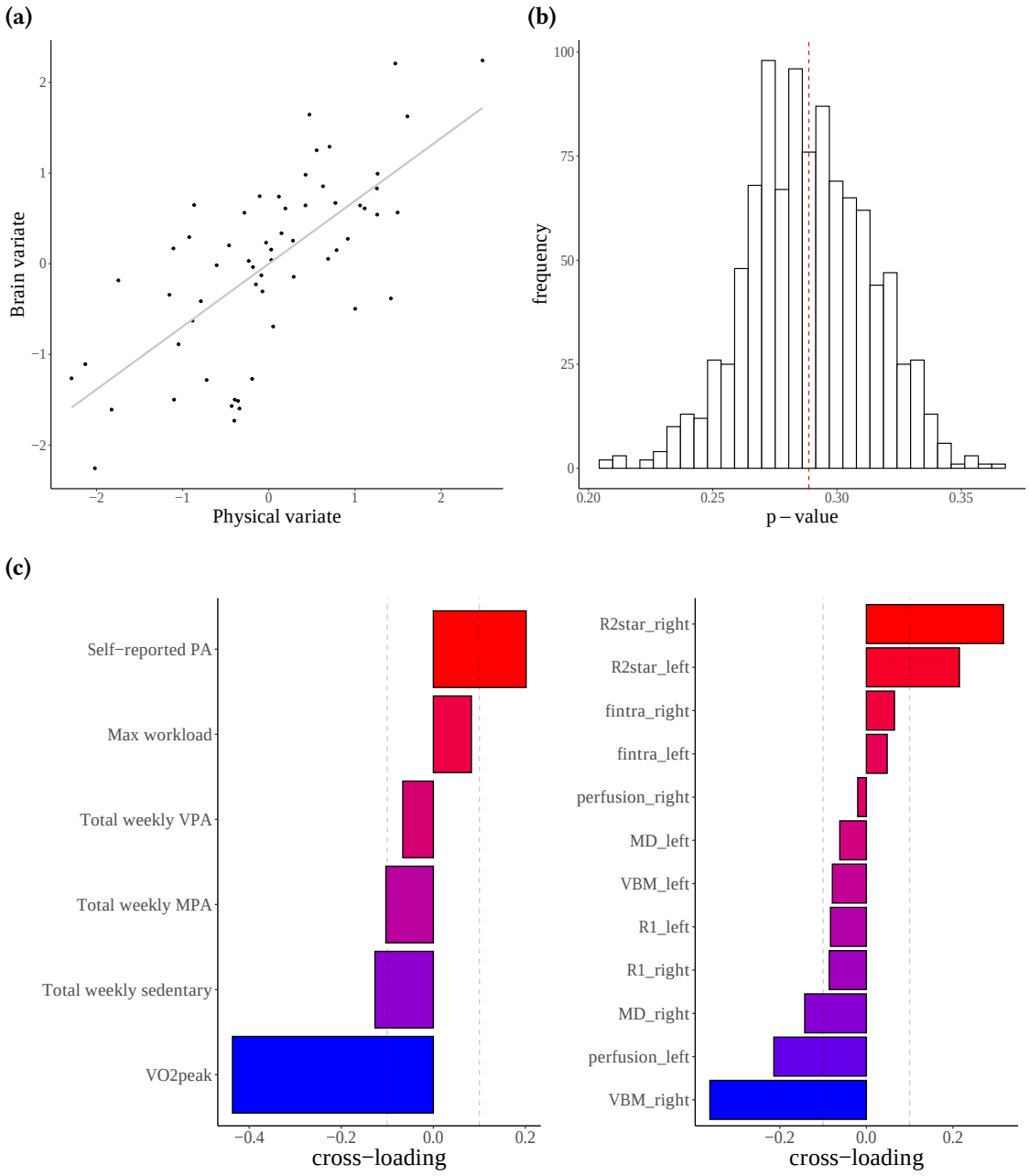


Figure 5.5: Canonical correlation and cross-loadings of the first mode of covariation (primary analysis). (*a*) Subject scores on the first physical and brain variate. (*b*) Distribution of p-values. Red dashed line corresponds to median p-value. (*c*) Cross-loadings between the physically active lifestyle variables and the brain variate (left) and between the brain variables and the PA lifestyle variate (right). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest. However, this mode of covariation did not reach statistical significance (p = 0.29).

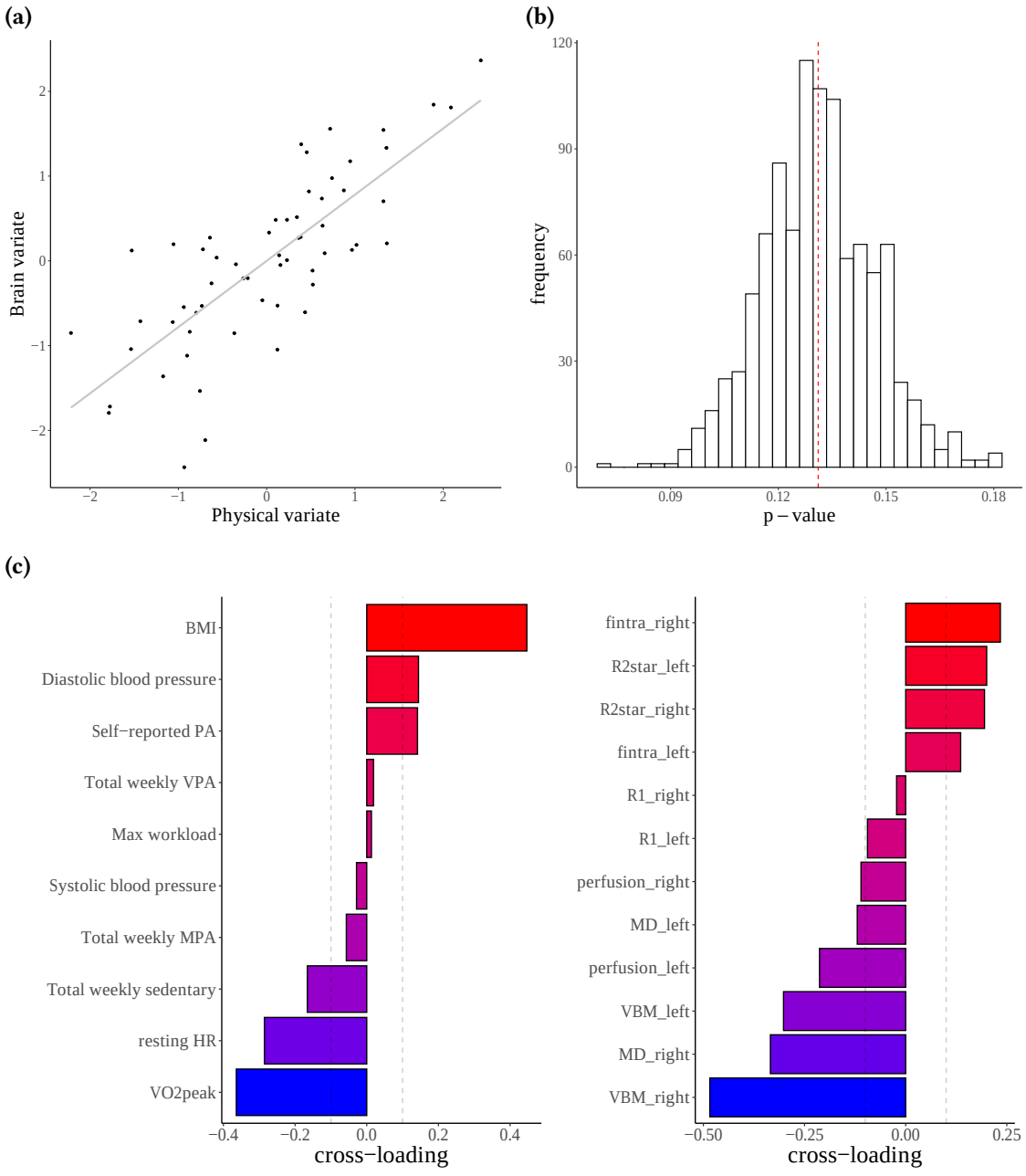


Figure 5.6: Canonical correlation and cross-loadings of the first mode of covariation (secondary analysis). (*a*) Subject scores on the first physical and brain variate. (*b*) Distribution of p-values. Red dashed line corresponds to median p-value. (*c*) Cross-loadings between the physically active lifestyle variables and the brain variate (left) and between the brain variables and the PA lifestyle variate (right). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest. However, this mode of covariation was not significant ($p = 0.14$).

5.4 Discussion

This study aimed to examine whether PA and CRF were associated with anterior hippocampal MRI measures and hippocampal-dependent memory performance. Univariate analyses demonstrated no significant associations between CRF or MVPA and left or right anterior hippocampal volumes, or performance on memory tasks. In addition, a physically active lifestyle was not significantly linked with MRI markers of tissue (micro)structure in the anterior hippocampus.

5.4.1 Relationships of CRF or PA with anterior hippocampal volume and memory

To the best of our knowledge, our study was the first to examine associations between CRF or MVPA and *anterior* hippocampal volumes in young people. This portion of the hippocampus tends to be particularly sensitive to PA and CRF in adults (Demirakca *et al.,* 2014; Erickson *et al.,* 2011; Killgore *et al.,* 2013; Stillman *et al.,* 2018; Thomas *et al.,* 2016). Unlike studies in adults, however, PA and CRF were not significantly related with left or right anterior hippocampal volumes in our cohort of adolescents. Moreover, no significant associations were found when the volume of the entire hippocampus was considered. These null results add to the sparse, but expanding, and inconsistent literature on relationships between PA or CRF and hippocampal volumes in children and adolescents: both positive (Chaddock *et al.,* 2010a; Esteban-Cornejo *et al.,* 2017; Gorham *et al.,* 2019; Herting & Nagel, 2012) and no significant associations (Migueles *et al.,* 2020; Ruotsalainen *et al.,* 2019) have been reported.

In addition, no significant associations between objective measures of MVPA or CRF and performance on hippocampal-dependent relational memory tasks were observed. The results involving CRF are surprising, given that previous studies, including chapter 4, have shown small, but robust, positive associations between CRF and relational memory performance in young people (Chaddock *et al.,* 2010a; Hassevoort *et al.,* 2017; Monti *et al.,* 2012). The absence of significant associations with objectively-assessed MVPA, on the other hand, aligns with findings presented in chapter 4 for self-reported

PA, yet contrasts with results from a previous study in young adults suggesting that higher MVPA is associated with more favourable memory outcomes (Stillman *et al.*, 2016).

While absence of evidence is not evidence of absence, the discrepancy in findings in the literature is striking and could be due to several factors. First, assessments of hippocampal volume, PA, CRF and memory differ between studies, which complicates a direct comparison of findings and may impact on results. Hippocampal volumes, for instance, have typically been assessed using FSL's FIRST or FreeSurfer (though Herting & Nagel, 2012 used manual segmentations), each of which yields different estimates (for a discussion, see chapter 3) that additionally vary with age (Wenger *et al.*, 2014). Moreover, CRF was assessed using treadmill-based maximal exercise tests in previous studies (Chaddock *et al.*, 2010a; Herting & Nagel, 2012), compared with a cycle-based assessment in this study. CRF estimates tend to be higher (11–14%) using treadmill compared to cycling ergometers (Armstrong & Welsman, 2019b), and may be differentially impacted by an individual's body mass. The memory assessments used in this study have previously shown associations with CRF in children (Chaddock *et al.*, 2010b) and hippocampal volumes in older adults (Liang *et al.*, 2016), though low sensitivity or test-retest reliability of these assessments cannot be excluded, which may reduce observed correlation values (Hedge *et al.*, 2018a).

Second, our sample was, on average, more fit compared to the main trial sample (for details see chapter 2), and less fit compared to a previous study showing positive associations between CRF and hippocampal volumes in adolescents (Herting & Nagel, 2012). Moreover, we treated CRF as a continuous variable in analyses, in contrast to Chaddock *et al.* (2010b), in which a high and low fit group were created based on their CRF performance. It may be that associations only appear when the tail ends of the CRF distribution are considered. Finally, while the size of our sample was comparable to previous studies (N = 40 to 60), it may have been too small to detect the typically small brain-behaviour associations. For instance, 193 participants are required to detect an

association (r) of 0.2 —typical for CRF or PA-cognition relationships —with 80% power and an alpha level of 0.05 (assessed using the *pwr* package in R). Correlation values for CRF-hippocampal volume relationships, however, ranged between 0.4–0.6 (Chaddock *et al.,* 2010b; Herting & Nagel, 2012), which can be detected with commonly reported sample sizes of 20–45 individuals. Replication of these associations in large-scale brain imaging studies, such as the ABCD study (Casey *et al.,* 2018), is warranted.

Given the involvement of the (anterior) hippocampus in relational memory processes (Giovanello *et al.,* 2004; Konkel & Cohen, 2009) and the positive associations between hippocampal volumes and memory assessments reported in previous studies (Chaddock *et al.,* 2010a; Herting & Nagel, 2012), we expected the high active and fit participants to perform better on tasks of memory. Findings from a meta-analysis, however, suggested that larger hippocampal volumes were associated with worse memory performance in children and adolescents (Van Petten, 2004). In our study, no significant associations were observed, in line with a recent study suggesting that only in extreme cases may variation in hippocampal volumes explain variation in memory performance (Clark *et al.,* 2020).

5.4.2 Beyond hippocampal volume

Guided by findings in rodents (e.g. Cooper *et al.,* 2018; Vivar & Van Praag, 2017; Voss *et al.,* 2013), most human PA studies have focussed on the hippocampus (Firth *et al.,* 2018; Stillman *et al.,* 2020), using its volume as a proxy for the cellular properties or processes one would like to make inferences on (e.g. neurogenesis or angiogenesis). While volumetric measures have proven to be sensitive to macroscopic brain changes, it is unclear what changes (or differences) in hippocampal volume reflect on a microscopic level. For instance, PA is known to increase a host of processes in the hippocampus, including cell proliferation and survival, neurogenesis, gliogenesis, synaptogenesis and angiogenesis, which may all influence volumetric measurements. Advanced imaging methods, such as diffusion-weighted imaging, quantitative MRI and arterial spin labelling, particularly when combined (Cercignani & Bouyagoub, 2018), could provide

information that is biologically more meaningful than simple volumetric assessments. Studies using such modalities have shown increased perfusion (Maass *et al.,* 2015), blood volume (Pereira *et al.,* 2007), tissue density and integrity (Den Ouden *et al.,* 2018; Kleemeyer *et al.,* 2016; Pajonk *et al.,* 2010) and indicators of myelination (Thomas *et al.,* 2016) in the hippocampus with higher PA, which broadly supports findings from animal studies. Our study was the first to explore multivariate relationships of PA and CRF with multimodal MRI in the anterior hippocampus in adolescents. In contrast to previous studies, however, we found no significant association between a physically active lifestyle and markers of perfusion, myelination and tissue density.

While it is possible that our study was underpowered to detect multivariate brain-behaviour relationships (Helmer *et al.,* 2020 suggests 50 samples per feature, i.e. 900 participants for 18 features), it is worth pointing out that we observed a significant association between *whole-brain* MRI patterns of brain structure, microstructure and function, and a physically active phenotype in the same individuals (Salvan *et al.,* 2020). Interestingly, part of this pattern of results indicated that higher hippocampal perfusion and lower R_1 and R_2^* were related to being physically fit and active (no diffusion-based metrics were used). These findings corroborate earlier findings in adults (Maass *et al.,* 2015; Thomas *et al.,* 2016), and warrant further investigation in intervention studies. Equally important, our earlier study suggests that, rather than being specific to the hippocampus, PA and CRF may affect the brain more globally. Hence, future studies would benefit from a whole-brain, multi-modal approach to interrogate relationships with PA.

5.4.3 Strength and limitations

Key strengths of this study include its use of traditional univariate analyses combined with advanced multivariate analyses, its focus on young adolescents, which have received little attention in the literature (Wassenaar *et al.,* 2020), as well as its rich and detailed set of measures, with the multi-modal MRI dataset being particularly novel in this age group.

This study also has various limitations. First, our sample of adolescents was more fit and contained fewer pupils from a lower socioeconomic background compared to the Fit to Study main trial sample (for details, see chapter 2); it may thus not be representative of a community-sampled cohort. Second, 17 participants had previously completed the relational memory task as part of the main trial. While their performance was not significantly different from those completing the task for the first time, learning or changes in motivation may have affected task performance. To take such effects into account, we additionally controlled the analysis for prior completion of this task. Third, the cross-sectional design renders causal conclusions untenable. Well-sampled, multi-modal randomised-controlled trials (RCTs) are required to assess directionality of relationships. Surprisingly, no such studies exist in young people that examined the effect of PA on (anterior) hippocampal volume (but see chapter 7). Fourth, single imputation was used to deal with missing data in the analyses, which does not take the uncertainty of imputations into account and may result in overly optimistic p-values. Multiple imputation would have been preferable, yet was not feasible given the large number of variables compared to the number of participants. Future studies wishing to run multi-modal analyses would benefit from larger samples and careful data collection procedures to avoid missing data.

5.5 Conclusion

In conclusion, CRF and PA are not significantly associated with anterior hippocampal volumes or relational memory performance in our sample of young adolescents. A physically active lifestyle is not significantly linked with MRI markers of tissue (micro)structure in the anterior hippocampus. Well-sampled cross-sectional studies are required to further elucidate whether PA or CRF are related with hippocampal outcomes in this age group. Furthermore, randomised controlled trials, incorporating multi-modal imaging protocols, are warranted to establish causality and to provide insights into neurobiological mechanisms driving PA-related plasticity. The Fit to Study randomised controlled trial is one such trial, the results of which are discussed in the next chapters.

6

The effect of the Fit to Study intervention on cardiorespiratory fitness, cognitive performance and mental health

Contents

6.1 Introduction . . . 175
6.1.1 Aims . . . 178
6.2 Methods . . . 178
6.2.1 Design . . . 179
6.2.2 Sample . . . 179
6.2.3 Randomisation and blinding . . . 179
6.2.4 Intervention . . . 180
6.2.5 Teacher training . . . 181
6.2.6 Fidelity measures . . . 181
6.2.7 Outcome measures . . . 181
6.2.8 Data analysis . . . 183
6.3 Results . . . 186
6.3.1 Sample characteristics . . . 188
6.3.2 Effects on cardiorespiratory fitness . . . 189
6.3.3 Effects on cognition . . . 191
6.3.4 Effects on mental health . . . 192
6.3.5 Subgroup analysis: moderation by sex, socioeconomic status or baseline CRF . . . 193
6.3.6 Relationship of change in CRF with changes in cognition and mental health . . . 193
6.3.7 Intervention compliance and complier average causal effect . 193
6.3.8 Additional analyses . . . 194
6.3.9 Sensitivity analyses . . . 194

6.4 Discussion 195
6.4.1 Findings in context 195
6.4.2 Practical challenges of school-based PA interventions 198
6.4.3 Strengths and limitations 199
6.5 Conclusion 200

In the previous chapters, I examined cross-sectional relationships between physical activity, cardiorespiratory fitness and cognitive outcomes. Such studies are useful to establish the existence of relationships between sets of variables and explore inter-individual differences, but do not allow for conclusions to be drawn regarding causality. That is, high active pupils may present with better cognitive performance, or, pupils with better cognitive performance may be more inclined to be active. Randomised controlled trials, like Fit to Study, are the gold standard experimental design to disentangle cause and effect. Current evidence from such trials, however, is inconsistent. In this chapter, I examined the effect of the one-academic year Fit to Study intervention on cardiorespiratory fitness, cognitive performance and mental health in the cohort of young adolescents (n = 16,017). The intervention did not significantly improve fitness, cognitive performance or mental health, but these findings should be interpreted with caution given low implementation fidelity and high drop-out. Well-controlled, large-scale, school-based trials that examine the effectiveness of HIT-style physical activity interventions to enhance cognitive and mental health outcomes are warranted.

6.1 Introduction

Adolescence is marked by rapid, dynamic social, psychological and neurocognitive development, during which the brain matures and higher-order cognitive processes, such as executive functions (EF), are refined (Dahl *et al.,* 2018; Larsen & Luna, 2018; Sawyer *et al.,* 2018). It is also the peak onset time for mental health issues (Paus *et al.,* 2008), including an increased social and biological vulnerability to mood disorders and low self-esteem. Notably, adolescence has been described as a (neurobiological) critical or sensitive period of development (Fuhrmann *et al.,* 2015; Larsen & Luna, 2018), characterised by enhanced plasticity, during which experience or environmental influences strongly affect the brain and may have long-lasting effects on behaviour. This period in life therefore presents an opportunity for interventions to positively impact on short- and long-term health outcomes, and has been of increasing interest

to researchers and policymakers (Dahl *et al.,* 2018; Patton *et al.,* 2016).

Interventions involving regular physical activity (PA) have received much attention given that PA is easily accessible, modifiable, cost-effective, scalable and has well-established health benefits, including improvements in cardiometabolic health (2018 Physical Activity Guidelines Advisory Committee, 2018). Converging lines of evidence suggest that regular PA changes the brain by stimulating processes such as neurogenesis and angiogenesis (Voss *et al.,* 2013), and could improve cognitive and mental health in young people (Biddle *et al.,* 2019). In particular, intervention studies involving primarily pre-adolescent children have found some evidence that PA improves cognition, particularly in the domains of attention and EFs (working memory, cognitive flexibility and inhibitory control), as well as attainment in maths, but considerable between-study heterogeneity makes firm conclusions difficult to draw (Álvarez-Bueno *et al.,* 2017a; Álvarez-Bueno *et al.,* 2017b; Gunnell *et al.,* 2018; Singh *et al.,* 2019; Wassenaar *et al.,* 2020). A smaller body of work suggests PA can treat depression, reduce anxiety, and improve self-esteem (Biddle *et al.,* 2019; Ekeland *et al.,* 2005). There is also emerging evidence that childhood PA is predictive of cardiovascular, mental, cognitive, and brain health later in life (McKercher *et al.,* 2014; Middleton *et al.,* 2010; Perez *et al.,* 2019; Twisk *et al.,* 2000). However, approximately 80% of adolescents do not meet PA recommendations of 60 minutes of moderate-to-vigorous PA (MVPA) per day (Guthold *et al.,* 2020; Hallal *et al.,* 2012) and PA levels decline dramatically during teenage years (Dumith *et al.,* 2011). Hence, there is scope for interventions to increase PA levels.

Schools are an ideal setting to promote PA among adolescents, including those from lower socioeconomic backgrounds, because pupils spend a large proportion of their waking hours at school and schools have access to the required facilities (Booth & Okely, 2005). Findings from school-based PA interventions have been inconsistent, with recent reviews reporting no meaningful change in PA across the (school) day (Jones *et al.,* 2019; Love *et al.,* 2019), but enhanced (moderate-to-vigorous) PA during

the actual intervention (Jones *et al.,* 2019; Lonsdale *et al.,* 2013). More promising results have been reported for school-based interventions aimed to increase cardiorespiratory fitness (CRF, Dobbins *et al.,* 2009; Dobbins *et al.,* 2013; Kriemler *et al.,* 2011; Pozuelo-Carrascosa *et al.,* 2018, self-esteem (Liu *et al.,* 2015), and cognitive performance, with meta-analyses showing that interventions delivered during curricular physical education (PE) are particularly effective at improving cognition (Álvarez-Bueno *et al.,* 2017a; Álvarez-Bueno *et al.,* 2017b). The majority of these interventions, however, have involved large doses of PA (e.g. 60 min 4 times per week) that may not be easily implemented in, and scaled up to, all schools, where crowded school curricula require PA interventions to be time-efficient (Hills *et al.,* 2015; Naylor *et al.,* 2015).

High intensity (interval) training (HIT) is a time efficient and effective strategy to improve cardiovascular health and CRF in young people (Costigan *et al.,* 2015; Eddolls *et al.,* 2017; Logan *et al.,* 2014). HIT definitions vary (Biddle & Batterham, 2015), but typically involve short or longer bouts (from <45 s to 2–4 min) of high-intensity exercise interspersed by periods of rest or light activity (Buchheit & Laursen, 2013a; Buchheit & Laursen, 2013b). HIT can be easily implemented in the school day (Costigan *et al.,* 2015) with physical adaptations that are comparable to endurance training results (Benda *et al.,* 2015; Gibala *et al.,* 2006). While interventions involving working at or near the maximum intensity (e.g. >90% $\dot{V}O_2$peak, or max heart rate [HR]) may be limited in their scalability and effectiveness (Biddle & Batterham, 2015), there is some promising evidence from less intense HIT interventions showing improvements in EFs (Costigan *et al.,* 2016; Jeon & Ha, 2017; Leahy *et al.,* 2020; Venckunas *et al.,* 2016) and mental well-being (Costigan *et al.,* 2016; Leahy *et al.,* 2019). However, these interventions were brief (6–14 weeks), relatively well-controlled with small samples (1–6 schools), and did not investigate academic outcomes. To assess the effectiveness and scalability of HIT, large-scale trials are desirable.

The Fit to Study cluster randomised trial investigated the effect of a one- academic year HIT-style intervention during regular PE on maths attainment, as well as mental

health, cognitive performance and CRF in young adolescents (12–13 years old) from 104 schools (52 intervention) in South/Mid England (Wassenaar *et al.,* 2019a). An independent evaluator, NatCen Social Research, analysed the primary outcome, attainment in maths, and found no significant benefit of the HIT style intervention (Husain *et al.,* 2019). Here, we report the prespecified secondary outcomes of the Fit to Study cluster randomised controlled trial. We hypothesised that the intervention would improve CRF, cognitive performance and mental health outcomes in this cohort. We also explored whether the effect of the intervention on the prespecified outcomes was moderated by sex, socioeconomic status, or baseline CRF level. We additionally examined whether changes in CRF were related to changes in cognitive and mental health outcomes.

6.1.1 Aims

The following primary aim was specified prior to the start of the analysis:

Primary aims

1. examine whether the HIT-style VPA intervention had an effect on CRF, cognitive performance and mental health in young adolescents

We additionally prespecified the following set of secondary aims, that aimed to further explore any potential intervention effects:

Secondary aims

1. examine whether the effect of the HIT-style VPA intervention on CRF, cognitive performance and mental health was moderated by sex, socioeconomic status or baseline CRF level
2. explore whether changes in CRF were related to changes in cognitive and mental health outcomes

6.2 Methods

Details on the methods of the Fit to Study trial can be found in chapter 2 and in Wassenaar *et al.* (2019a).

6.2.1 Design

The Fit to Study project was a parallel group, superiority cluster-randomised efficacy trial of a one academic year (10–months) HIT-style VPA intervention versus control involving Year 8 pupils (aged 12–13) from 104 secondary state schools (52 intervention). The intervention was incorporated into regular PE lessons to minimally disrupt the curriculum and ensure scalability. Assessments took place at baseline (end of Year 7, pupils aged 11–12) and after 12 months (end of Year 8, pupils aged 12–13).

6.2.2 Sample

A total of 106 eligible schools responded to an invitation to participate in the study Figure 6.3. Following the eligibility assessment, two schools declined to participate, and, from those remaining, 81 Year-7 pupils opted out of data storage and were not required to complete the trial's assessments. Schools that met inclusion criteria and wished to participate provided the sex, age and eFSM status of all Year-7 pupils. Schools were considered formally recruited upon transfer of the data and memorandum of understanding to NFER. A total of 104 schools (n = 18,261 pupils) were randomised into an intervention and control group.

Measures of self-reported PA in the past week (0–7 days, Scott *et al.,* 2015) and habitual PA over the past week (range 1–7, with 7 the most active, Hagger *et al.,* 2001) were used to characterise the sample's baseline PA levels.

6.2.3 Randomisation and blinding

Schools (n = 104) were allocated to an intervention group or PE as usual control group (1:1 ratio) using stratified block randomisation. Stratification was arranged by school gender type (single-sex or co-educational). NatCen Social Research performed the randomisation using a random number generator in Stata (version 12, Husain, 2017).

Schools were informed of their assigned group following baseline assessments to minimise bias and so that intervention PE teachers could receive training and deliver

the intervention. Given the nature of the intervention, it was not feasible for PE teachers or pupils to be blind to group allocation, but pupils or parents were not specifically told of their assigned group by the research team. Researchers visiting schools for top-up training and collection of fidelity measures were not blinded.

6.2.4 Intervention

The intervention consisted of a one-academic year (10 months: September 2017–June 2018) HIT-style VPA programme delivered by PE teachers during regular Year 8 PE lessons. PE teachers were instructed to incorporate two elements of additional VPA into their regular Year-8 PE lessons, consisting of: (1) four minutes of VPA as part of an active 10-min warm-up, and (2) three 2-min (VPA) infusions per hour of PE, where VPA was defined as activities that raised the HR to 71–85% of the maximum HR. We defined the intervention as HIT-*style* because it meets several HIT criteria (Buchheit & Laursen, 2013a; Buchheit & Laursen, 2013b), yet other HIT definitions exist that emphasise a higher intensity (Martland *et al.*, 2020).

The warm-up was delivered at the start of each lesson and began with light-intensity movements (e.g. wrist rotations), followed by moderate-intensity activities (e.g. arm rotations) to incorporate two 2-min bursts of VPA (e.g. vigorous arm sprints; running on the spot). The three 2-min infusions (per hour of PE) were incorporated in the main PE lesson. The infusions included, for instance, fast arm rotations, squats and lunges, and sprinting on the spot, interspersed with brief (active) rest periods if needed during each infusion. Teachers were invited to create their own warm-up and/or infusions that met the required intensity and duration of PA.

Schools in the control arm were asked to deliver PE as usual. To minimise drop-out and ensure compliance, school participation was incentivised with each school receiving £500 upon completion of primary outcome assessment post-intervention.

6.2.5 Teacher training

PE teachers from intervention schools were trained prior to the start of the intervention by attending an online or face-to-face 2-hour training session delivered by a team from Oxford Brookes University. Following the training session, teachers from intervention schools were given access to the trial's website containing videos demonstrating the intervention elements. Researchers visited a sample of 30 intervention schools between December 2017 and February 2018 to offer support and answer questions. They offered additional support and top-up training to PE teachers throughout the intervention period.

6.2.6 Fidelity measures

Intervention fidelity was assessed using a prespecified set of measures, including (1) teacher log books collected throughout the year in intervention schools, (2) an objective measure of average VPA per hour of PE using Axivity AX3 accelerometers in a convenience sample of participants per school (at least 50% of the Year group, collected once during the intervention period; data processing details are provided in chapter 2 and Appendix F.1.1), and (3) pupil-reported compliance with the intervention using a brief three-question survey administered post-intervention in all schools asking whether (i) PE lessons started with a warm-up, (ii) PE lessons included bursts of VPA that raised their HR and made them feel out of breath, and (iii) they took part in warm-ups or VPA bursts if asked by PE teachers. Due to low compliance with the log books, NatCen Social Research additionally collected the teacher-reported percentage of PE lessons in which the intervention was delivered as intended (post-intervention, in intervention schools only).

6.2.7 Outcome measures

Pupils were assessed before the intervention started (pretest, t_0) and immediately following the intervention (posttest, t_1). Secondary measures included assessments of CRF, cognitive performance, and mental health.

6.2.7.1 Cardiorespiratory fitness

CRF was assessed using the 20-meter shuttle run (20MSR) test (Léger *et al.,* 1988). The outcome measure was the total number of laps completed. Schools with a policy of not using the 20MSR completed the 12 min Cooper Run test instead (Cooper, 1968).

6.2.7.2 Cognitive performance

Cognitive functioning was assessed with online, computer-based tests of processing speed, visual relational memory and core EFs (working memory, inhibition, and cognitive flexibility, Diamond, 2013). The cognitive assessments were programmed in JavaScript using jsPsych (De Leeuw, 2015) and took approximately 50 minutes to complete. The task order was pseudorandomised across participants within schools. Each participant started with the reaction time task, followed by the remaining tasks in random order. Teachers and pupils were instructed to complete the assessments at home, but a proportion of schools decided to complete the tasks during school time.

Details of the cognitive battery have been provided in the protocol (Wassenaar *et al.,* 2019b) and in chapter 2. In brief, the battery consisted of the following tasks:

- A simple reaction time task was used to assess processing speed. The average reaction time across valid trials was the primary outcome measure.
- A modified version of a previously described relational memory task (Chaddock *et al.,* 2010b) was used to assess visual relational memory performance. The proportion of correct responses on valid trials (i.e. accuracy) was the primary outcome measure.
- A modified Flanker task (Hillman *et al.,* 2014) was used to assess inhibitory control. The congruent and incongruent response accuracies were the primary outcome measures.
- A visual two-back task (Chen *et al.,* 2014a; Guiney & Machado, 2013) was used to assess working memory performance. Response accuracy was the primary outcome measure of this task.

- A modified colour-shape switching task (Hillman *et al.*, 2014) was used to assess cognitive flexibility. The switch and non-switch response accuracies were the primary outcome measures.

6.2.7.3 Mental health

Psychosocial problems were measured with the Strength and Difficulties Questionnaire (SDQ, Goodman, 1997), which consists of 25 items measuring five sub-scales: (1) emotional symptoms, (2) conduct problems, (3) hyperactivity / inattention, (4) peer-relationship problems, (5) pro-social behaviour. Items are scored from 0 ("not true") to 2 ("certainly true"). We used combined conduct and hyperactivity scores (externalising, range 0-20) and peer and emotional scores (internalising, range 0-20) as primary outcome measures because there is evidence that in low-risk samples, the more focused subscales might not tap into distinct aspects of mental health (Goodman *et al.*, 2010).

Self-esteem was assessed with the global and physical self-esteem subscales of the short version of the Physical Self-Description Questionnaire (Marsh *et al.*, 2010). Higher scores indicate better outcomes.

6.2.8 Data analysis

6.2.8.1 Sample size calculation

Sample size calculations were performed by NatCen for the primary intention-to-treat (ITT) analysis, which compared maths test performance between intervention and control schools (Husain, 2017). Details on the sample size calculation, including an a-posteriori computed minimum detectable effect size for CRF, cognitive- and mental health outcomes, are provided in Appendix F.1.2.

6.2.8.2 Data cleaning

The computer-based questionnaire and cognitive assessments were completed in school or at home, without supervision from researchers or teachers. A conservative approach to data cleaning was used, that focussed on removing careless responders or data collection errors from the dataset. We additionally removed schools (n = 4, 561 pupils)

that completed the Cooper Run Test. Details of the cleaning procedures are provided in chapter 4 and in Appendix F.1.3.

6.2.8.3 Statistical analysis

The primary analyses were performed on an ITT basis on multiply imputed data under the assumption data was missing at random (MAR). To assess the effect of the intervention on CRF, cognitive outcomes, mental health and self-esteem, we followed recommendations for (cluster) randomised trials: models included baseline values of the outcome measure (ANCOVA) and included the stratification variable used at randomization (school gender-type, Kahan & Morris, 2012). Standard errors were corrected for the clustering of pupils within school using robust (sandwich) estimators (type CR2), that have shown good performance even with a small number of clusters (Pustejovsky & Tipton, 2018).

The primary analyses assessed the effect of the intervention on CRF, mental health (self-esteem and internalising- and externalising symptoms) and cognitive performance. We used confirmatory factor analysis (CFA) to create an EF latent defined by accuracy measures of the colour-shape switching task, Flanker task and two-back task. We used a latent variable model to capture the common variance shared by these different measures. Structural equation models (SEM) were constructed with Full Information Maximum Likelihood (FIML) estimators to allow for missing data. Standard errors were adjusted for clustering (Huber-White). We subsequently additionally adjusted the primary analysis models for the period (i.e. summer term, holidays or autumn term) and location (i.e. school or home) the tasks were completed.

In line with the analysis of the primary outcome (Husain *et al.*, 2019), exploratory subgroup analyses were conducted to assess whether any effect of the intervention on secondary outcomes (EF latent variable, relational memory, processing speed, in-/externalising symptoms and global/physical self-esteem) was moderated by sex, socioeconomic status or high or low baseline CRF. Baseline CRF levels were dichotomised

into high- and low-fit category using sex-specific normative values (high fit males >39 laps, high fit females >28 laps; Tomkinson *et al.,* 2017).

We also explored whether a change in CRF was related to a change in cognitive measures or mental health, and whether this relationship was moderated by treatment status, controlling for age, sex, socioeconomic status (eFSM), school gender-type (single sex or co-educational), and location (home or school) and period of completion (summer term, holidays or autumn term) of assessments.

The primary ITT analysis will provide a conservative (or underestimated) treatment effect when schools do not adhere to the intervention (Hernán & Hernández-Díaz, 2012). We therefore estimated a complier-average causal effect (CACE) by fitting an instrumental variable model using the two-stage least squares method, to estimate the average treatment effect in the population of compliers (details are provided in Appendix F.1.4). Intervention schools in which the intervention was delivered in >50% of PE lessons were classified as compliant.

We used multilevel multiple imputation by chained equations (MICE) to impute missing data under the assumption that data was MAR (details in Appendix F.2). We separately imputed the data for the CACE analyses, given the large amount of missingness in the fidelity measure. As a sensitivity analysis, we repeated all the analyses with complete-case data (i.e. available cases per analysis model).

Two-sided inferences with p <0.05 were considered statistically significant. The adjusted standardized and unstandardized mean differences are provided. The standardized mean difference score was computed using the pooled unconditional standard deviation of the outcome (Husain *et al.,* 2019; White, 2005). Bonferroni-correction was applied to correct for multiple comparisons where appropriate. An overview of analysis software is provided in Appendix F.1.5.

6.3 Results

104 schools, totalling 18,261 pupils participated in the study (Figure 6.3), and were randomised into an intervention and control group. Prior to baseline assessments but post-randomisation, 11 schools withdrew from the trial. The schools were unaware of their allocated group at the time of drop-out and school or pupil level demographics did not significantly predict drop-out (Appendix F.3). Analyses were therefore limited to the schools that were part of the trial at baseline assessments (n = 93, 16,017 pupils). A total of 24 schools (17 intervention) were further lost to follow-up. Reasons for withdrawal included time-constraints (n = 6), staff changes / absence (n = 4), Ofsted inspection (n = 3), implementation issues (n = 2) or unknown reasons (n = 9).

Of 16,017 pupils at baseline, 2,182 (∼14%) completed all secondary outcomes, 8,114 (∼51%) completed the CRF assessment, 6,414 (∼40%) completed the questionnaire and 6,174 (∼39%) completed one or more cognitive assessments, following data cleaning, at post-test. Excluded participants during data cleaning were more likely to be male and from a lower socioeconomic background (see Appendix F.4).

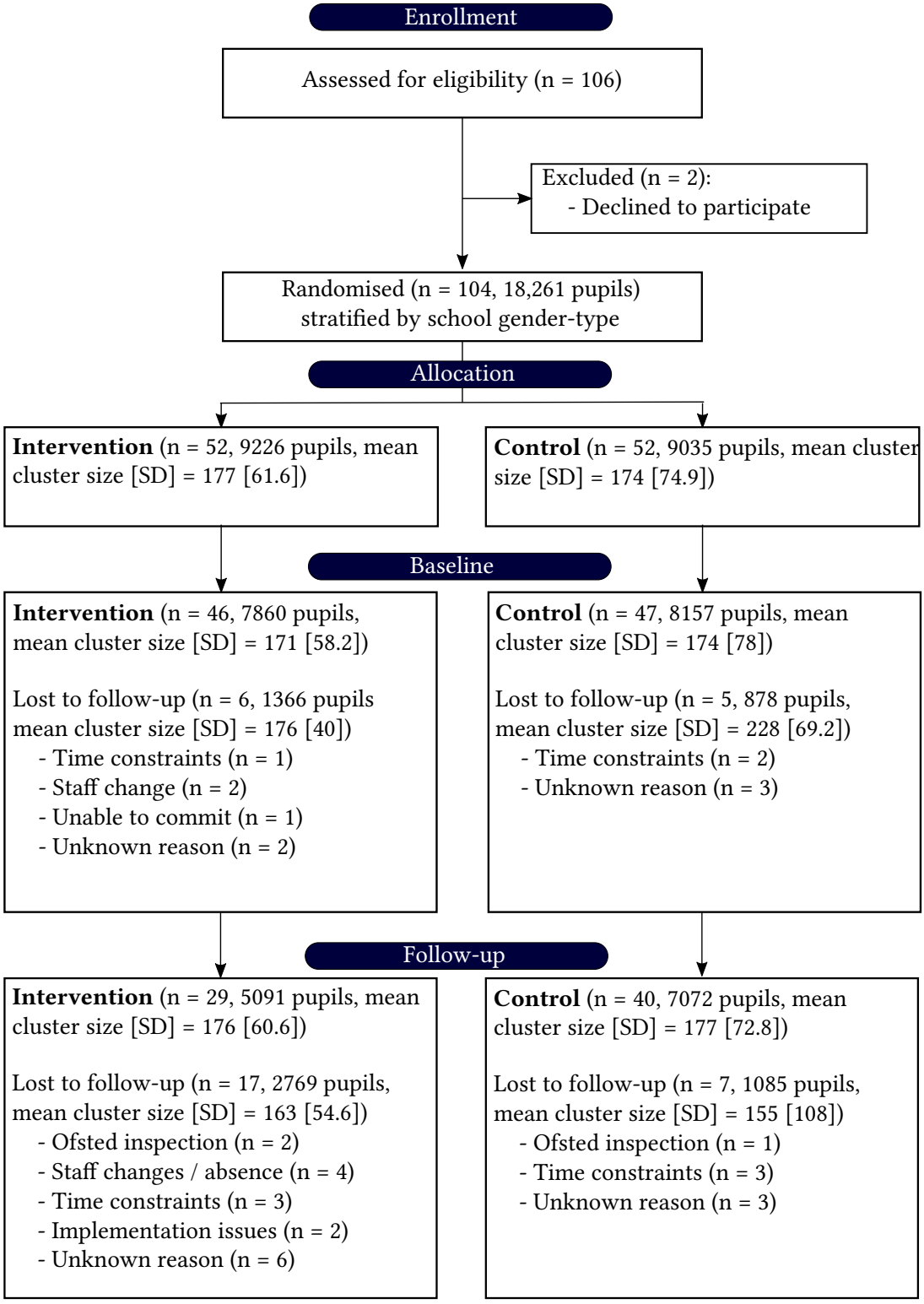


Figure 6.1: CONSORT flow diagram of Fit to Study cluster randomised controlled trial.

6.3.1 Sample characteristics

Descriptive statistics for school- and pupil level characteristics at baseline (n = 93) are presented in Table 6.1. The groups are approximately similar on all measures. Approximately 20% of the sample reported that they were physically active for at least 60 min per day for the entire week (i.e. meeting the PA guidelines), which is comparable between the groups (Appendix F.5). Moreover, across schools, on average, 14 minutes per hour of PE is spent in MVPA, and approximately 4 minutes in VPA.

School curricula in intervention and control schools included, on average, a similar number of PE lessons (1.61 vs 1.59, range: 1–2), a comparable duration of PE lessons (72.3 vs 72.2 min), and total duration of PE per week (110 vs 107 min; details in Appendix F.5).

Table 6.1: School- and pupil level baseline characteristics.

	Intervention	Control
School level		
No. schools	46	47
Gender status, no. (%)		
Co-ed	37 (80.4%)	38 (80.9%)
Female	9 (19.6%)	8 (17.0%)
Male	0 (0%)	1 (2.1%)
School type, no. (%)		
Comprehensive to 16	7 (15.2%)	9 (19.1%)
Comprehensive to 18	34 (73.9%)	33 (70.2%)
Secondary Modern	5 (10.9%)	5 (10.6%)
Establishment type, no. (%)		
Academy - Sponsor Led	7 (15.2%)	9 (19.1%)
Academy Converter	2 (4.3%)	1 (2.1%)
Academy Converter - Mainstream	16 (34.8%)	15 (31.9%)
Community School	8 (17.4%)	7 (14.9%)
Foundation School	5 (10.9%)	5 (10.6%)
Free School - Mainstream	3 (6.5%)	2 (4.3%)
Voluntary aided school	4 (8.7%)	7 (14.9%)
Voluntary controlled school	1 (2.2%)	1 (2.1%)
School size, no. (%)		
Large (>200)	14 (30.4%)	13 (27.7%)
Medium (100-200)	27 (58.7%)	27 (57.4%)
Small (<100)	5 (10.9%)	7 (14.9%)
Number of form groups, no. (%)		
Median (range)	6 (3–12)	7 (3–14)
Missing[a], no. (%)	1 (2.2%)	1 (2.1%)
Ofsted rating, no. (%)		
Outstanding	8 (17.4%)	14 (29.8%)

Continued on next page

Table 6.1 – continued from previous page

	Intervention	Control
Good	29 (63.0%)	24 (51.1%)
Requires Improvement	4 (8.7%)	6 (12.8%)
Serious Weaknesses	1 (2.2%)	0 (0%)
Special Measures	3 (6.5%)	3 (6.4%)
Missing, no. (%)	1 (2.2%)	0 (0%)
Socioeconomic status		
Percent eFSM pupils[b], mean SD	17.0 (9.0)	17.7 (12.9)
IMD[c], median (range), decile	5 (1–10)	6 (1–10)
School location[d], no (%)		
Rural town and fringe	5 (10.9%)	4 (8.5%)
Rural village	0 (0%)	0 (0%)
Urban city and town	19 (41.3%)	20 (42.6%)
Urban major conurbation	22 (47.8%)	23 (48.9%)
Pupil level		
No. pupils	7860	8157
Age[e], mean (SD), years	12.5 (0.296)	12.5 (0.293)
Females, no (%)	4466 (56.8%)	4495 (55.1%)
eFSM, no. (%), yes	1243 (15.8%)	1422 (17.4%)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, IMD = index of multiple deprivation, med = median, no. = number, SD = standard deviation

[a] A row indicating *missing* values was added only when missing values were present.

[b] Percent eFSM indicates the percentage of enrolled pupils that are eligible for free school meals within a school.

[c] Index of multiple deprivation was collected using the school's postcode, through Government records: http://imd-by-postcode.opendatacommunities.org/imd/2019

[d] Location was assessed using Government records: https://www.gov.uk/government/collections/rural-urban-classification

[e] Age at baseline (1 September 2017)

6.3.2 Effects on cardiorespiratory fitness

CRF levels increased from baseline to post-test across groups (Table 6.2, and distribution plots of complete-cases in Appendix F.6.2). However, no significant difference in CRF levels were observed between the intervention and control group at posttest (standardized mean difference [SMD] = 0.02, 95%CI: -0.11, 0.16).

Table 6.2: Mean scores for outcome measures by group, with (un)standardized mean differences.

	Intervention group		Control group		Mean difference (95% CI)[a]	
	Baseline M (SD)	Post M (SD)	Baseline M (SD)	Post M (SD)	Unstandardized	Standardized[b]
20MSR						
CRF, laps	36.11 (20.75)	41.36 (21.95)	37.85 (20.65)	42.37 (22.28)	0.48 (-2.52, 3.48)	0.02 (-0.11, 0.16)
Reaction time task						
RT[c], ms	378.7 (86.65)	378.58 (99.46)	380.21 (89.7)	376.07 (95.94)	3.63 (-7.04, 14.3)	0.04 (-0.07, 0.15)
Relational memory task						
Accuracy, %	59.75 (11.93)	59.53 (13.05)	59.76 (12.24)	60.78 (13.08)	-1.37 (-3, 0.26)	-0.1 (-0.23, 0.02)
two-back task						
Accuracy, %	54.43 (18.46)	55.24 (20.9)	53.4 (18.74)	55.48 (20.99)	-0.96 (-3.38, 1.47)	-0.05 (-0.16, 0.07)
RT[c], ms	857.67 (164.84)	802.26 (175.03)	844.17 (171.04)	799.45 (174.9)	-4.25 (-23.09, 14.58)	-0.02 (-0.13, 0.08)
Flanker task						
Accuracy congruent, %	82.83 (17.14)	84.21 (16.57)	81.91 (16.81)	84.55 (16.44)	-0.92 (-2.38, 0.53)	-0.06 (-0.14, 0.03)
Accuracy incongruent, %	60.22 (20.67)	62.2 (20.33)	60.02 (19.55)	63.06 (19.65)	-1.05 (-2.86, 0.76)	-0.05 (-0.14, 0.04)
RT[c] congruent, ms	490.26 (92.13)	476.82 (84.22)	490.35 (94.51)	479.19 (81.26)	-2.49 (-10.37, 5.39)	-0.03 (-0.13, 0.07)
RT[c] incongruent, ms	552.9 (127.12)	537.89 (108.67)	548.15 (126.16)	541.25 (109.06)	-5.74 (-15.41, 3.93)	-0.05 (-0.14, 0.04)
Colour-shape switching task						
Accuracy non-switch, %	73.27 (17.21)	75.18 (17.61)	72.88 (17.25)	75.61 (17.48)	-0.76 (-2.64, 1.12)	-0.04 (-0.15, 0.06)
Accuracy switch, %	69.51 (16.89)	71.47 (17.65)	69.36 (17.15)	71.56 (17.59)	-0.32 (-2.16, 1.52)	-0.02 (-0.12, 0.09)
RT[c] non-switch, ms	1075.67 (352.18)	974.95 (320.46)	1051.07 (337.52)	970.87 (318.08)	-5.86 (-34.33, 22.61)	-0.02 (-0.11, 0.07)
RT[c] switch, ms	1452.99 (602.23)	1331.63 (562.05)	1404.21 (569.5)	1321.97 (544.16)	-13.18 (-66.27, 39.91)	-0.02 (-0.12, 0.07)
Psychosocial problems						
Internalising score3	5.1 (3.56)	5.47 (3.72)	5.07 (3.49)	5.25 (3.6)	0.19 (-0.18, 0.57)	0.05 (-0.05, 0.16)
Externalising score3	6.34 (3.85)	6.73 (3.78)	6.29 (3.74)	6.45 (3.8)	0.24 (-0.13, 0.61)	0.06 (-0.03, 0.16)
Self-esteem						
Global	4.39 (0.94)	4.23 (1.03)	4.43 (0.9)	4.3 (0.97)	-0.05 (-0.22, 0.12)	-0.05 (-0.22, 0.12)
Physical	4.36 (1.31)	4.07 (1.41)	4.44 (1.27)	4.14 (1.36)	-0.02 (-0.22, 0.19)	-0.01 (-0.16, 0.14)

Abbreviations: 20MSR = 20m shuttle run, CRF = cardiorespiratory fitness, RT = reaction time

[a] Adjusted mean difference, adjusted for clustering, baseline values of the outcome variable and school gender type.

[b] The outcome was standardized (mean = 0, SD = 1), prior to fitting the baseline and stratification-variable adjusted model.

[c] Lower scores represent better performance.

6.3.3 Effects on cognition

The primary cognitive outcome measure was an EF latent variable defined by accuracy measures of the colour-shape switching task, Flanker task and two-back task. We used a latent variable model to capture the common variance shared by these different measures. The model used is shown in Figure 6.2 and provides an adequate fit to the data (χ^2 (54) = 739.43; p <0.001; RMSEA [Root Mean Square Error of Approximation] = 0.04 [90% CI: 0.04, 0.04]; CFI [Comparative Fit Index] = 0.97; TLI [Tucker-Lewis Index] = 0.96). The EF latent variable shows considerable longitudinal stability (standardized coefficient = 0.82) demonstrating that this construct has good reliability, despite a lack of factorial invariance (i.e. the factor loadings on the EF factor vary across time), possibly because of a differential change in individual assessments due to the intervention. The intervention group showed no significant difference in EF skills compared to the control group at posttest (SMD = 0.017, 95% CI: -0.14, 0.17).

In line with the findings for the EF latent variable, no significant differences were observed in tasks that are supposed to measure the sub-domains of EF, including working memory (two-back task), inhibitory control (Flanker task), or cognitive flexibility (colour-shape switching task) between the intervention and control group at posttest (SMD in Table 6.2).

An improvement in relational memory performance and processing speed (reaction time task) was observed across groups from pre to post-intervention. However, no significant difference was observed in relational memory performance (SMD = -0.1, 95% CI: -0.25, 0.02) or processing speed (SMD = 0.04, 95% CI: -0.07, 0.15) between the intervention and control group at posttest.

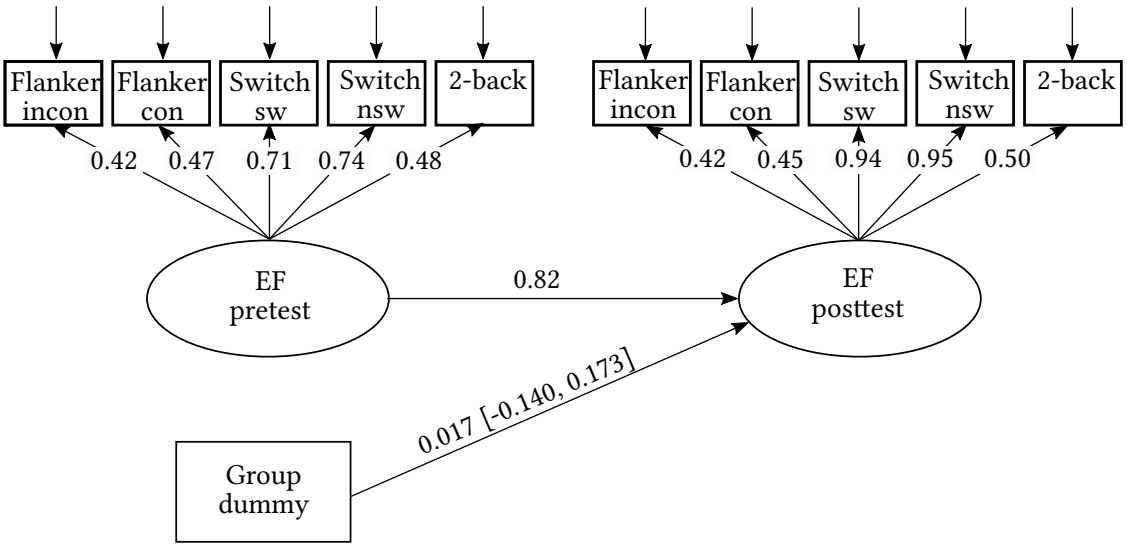


Figure 6.2: Path diagram showing the effects of the VPA intervention on an executive function (EF) latent variable. All coefficients are standardized except for the path coefficient for Group which is y-standardized (equivalent to Cohen's *d*). The arrows from the EF latent variables to the manifest variables show the factor loadings, which represent the regression slope of the factor score on the observed score. The 95% robust (Huber-White) CI is shown for the Group coefficient.The stratification variable (School Gender, dummy coded) was included as a set of additional covariates but they are not shown in the model. The following correlations between residuals were included in the model: $\text{Flanker}_{\text{incon, t1}}$ with $\text{Flanker}_{\text{incon, t2}}$ (0.34), $\text{Flanker}_{\text{con, t1}}$ with $\text{Flanker}_{\text{con, t2}}$ (0.29), $\text{Switch}_{\text{nsw, t1}}$ with $\text{Switch}_{\text{nsw, t2}}$ (0.09), $\text{Flanker}_{\text{con, t1}}$ with $\text{Flanker}_{\text{incon, t1}}$ (0.39) $\text{Switch}_{\text{nsw, t1}}$ with $\text{Switch}_{\text{sw, t1}}$ (0.74), $\text{Flanker}_{\text{incon, t2}}$ with $\text{Flanker}_{\text{con, t2}}$ (0.41), $\text{2-back}_{\text{t2}}$ with $\text{2-back}_{\text{t1}}$ (0.38), but are not shown in the diagram. Abbreviations: EF = executive function, incon = incongruent, con = congruent, nsw = non-switch, sw = switch.

6.3.4 Effects on mental health

The primary ITT analysis demonstrated no significant difference in internalising (SMD = 0.05, 95%CI: -0.05, 0.16) and externalising scores (SMD = 0.06, 95%CI: -0.03, 0.16) between intervention and control groups at posttest. Moreover, across treatment groups, a decrease in global- and physical self-esteem was observed from baseline to posttest. However, pupils in the intervention group did not differ significantly in their global (SMD = -0.05, 95%CI: 0.22, 0.12) or physical self-esteem (SMD = -0.01, 95%CI: -0.16, 0.14) compared to control at posttest.

6.3.5 Subgroup analysis: moderation by sex, socioeconomic status or baseline CRF

The effect of the intervention on cognitive and mental health outcomes was not significantly moderated by participants' sex, socioeconomic status or baseline CRF levels, following correction for multiple comparisons (α = 0.002, with 25 comparisons; Appendix F.7). Uncorrected analyses showed that two (of 25) comparisons had a p-value just below the conventional significance threshold (at p = 0.04, and p = 0.03).

6.3.6 Relationship of change in CRF with changes in cognition and mental health

One-year changes in CRF were not significantly related to changes in cognition or indicators of mental health (all $p_{\text{corrected}}$ >0.05, details are provided in Appendix F.8), and these relationships were not moderated by treatment group ($p_{\text{corrected}}$ >0.05 for interactions).

6.3.7 Intervention compliance and complier average causal effect

School-level and pupil level fidelity measures were collected towards the end of the intervention or post-intervention (details in Appendix F.9). Of 46 intervention schools at baseline, 15 (33%) received face-to-face training, 26 (57%) attended an online training programme and 5 (11%) did not receive training (and were lost to follow-up). Objective measures of PA collected with accelerometers during a single PE lesson during the intervention indicated that pupils in intervention schools spent, on average, more lesson time in VPA compared to control (full lesson: 3.65 vs 3.09 minutes/hour, p = 0.21; active lesson: 4.91 vs 4.38 minutes/hour, p = 0.07), but this difference was not statistically significant. Pupil-reported measures of compliance suggested that PE lessons in both intervention and control schools often started with a warm-up and incorporated infusions (median = 4, with 5 corresponding to "always"). Moreover, pupils in both treatment and control schools indicated that they would take part in

a warm-up and infusions if asked by a PE teacher (median: 4.75 vs 5, respectively).

Of 29 intervention schools that were part of the trial at follow-up, 22 provided information on the percentage of PE lessons in which the intervention was delivered as intended. Of these schools, 17 (77%) reported having delivered the intervention as intended in at least 50% of PE lessons across the year and only two schools (9%) delivered the intervention in over 90% of PE lessons. CACE estimates were similar in directionality to and in line with those obtained from the ITT analysis, and are provided in Appendix F.9.

6.3.8 Additional analyses

We inspected whether there were baseline differences in self-reported PE enjoyment and attitudes towards PA between the groups, which could make the groups more or less susceptible to intervention effects. The groups did not significantly differ in PE enjoyment or PA attitudes at baseline (p = 0.57, details in Appendix F.10). Moreover, no significant differences were observed in PE enjoyment or attitudes towards PA between the intervention and control group at posttest (p = 0.58, details in Appendix F.10).

6.3.9 Sensitivity analyses

The primary ITT models, testing for the effect of the intervention on CRF, cognitive and mental health outcomes, were additionally adjusted for where (home or school) and when (summer, holidays or autumn) participants completed the assessments. Adjusting for these confounding variables did not change the conclusions (i.e. all p >0.05 or 95%CI included zero), but minor changes in the size and directionality of treatment effects were observed (details in AppendixF.6.1).

We additionally repeated the primary ITT (Appendix F.6.2), subgroup (Appendix F.7.0.2), change-change (Appendix F.8.2), CACE (Appendix F.9.2.2) and additional analyses (Appendix F.10.2) on complete-case data. These analyses did not change the conclusions of the primary analyses (all p >0.05).

6.4 Discussion

In a large cohort of British adolescents, we assessed the effect of the Fit to Study scalable HIT-style VPA intervention on secondary outcome measures including CRF, cognitive performance, and mental health. In line with findings for the primary outcome (attainment in maths, Husain *et al.,* 2019) but in contrast to our hypotheses, the HIT-style VPA programme, incorporated into regular PE lessons for one academic year, did not significantly improve CRF, cognitive performance (EF, relational memory and processing speed), or mental health at posttest. Moreover, the effect of the intervention was not significantly moderated by sex, socioeconomic status or baseline CRF levels. Finally, one-year changes in CRF were not significantly related to changes in cognitive or mental health outcomes. Crucially, however, only a small proportion of schools complied with the intervention.

6.4.1 Findings in context

Previous studies that examined the effect of HIT and VPA interventions have shown robust effects on CRF in adolescents (Costigan *et al.,* 2015; Logan *et al.,* 2014; Owens *et al.,* 2017). Less research has examined the effects of such interventions on cognitive- and mental health, but studies to date have shown improvements in EFs (Costigan *et al.,* 2016; Davis *et al.,* 2011; Jeon & Ha, 2017; Leahy *et al.,* 2020; Moreau *et al.,* 2017) and psychological health (Costigan *et al.,* 2016; Leahy *et al.,* 2019). Despite the increasing body of evidence indicating the effectiveness of high-intensity exercise to improve health outcomes, we did not observe significant improvements in any of the outcome measures following a one-academic year HIT-style intervention. Moreover sex, socioeconomic status or baseline CRF did not significantly moderate the effects of the intervention on outcomes, in contrast to meta-analyses showing evidence that males (Ludyga *et al.,* 2020) or females (Barha *et al.,* 2017a) and those with lower baseline CRF may benefit most from high intensity PA interventions (Weston *et al.,* 2014).

It has been hypothesised that changes in CRF —reflective of physiological changes— may mediate PA-related effects on cognitive- and mental health outcomes (Åberg *et al.,*

2009; North *et al.,* 1990). In our study, both groups improved their CRF from pre to post-intervention. However, no group mean difference in CRF was found and, in line with findings from a recent meta-analysis (Tuvey *et al.,* 2019), no significant relationships between changes in CRF and changes in EF, memory, reaction time or mental health were observed.

Given that existing evidence supports the positive impact of HIT and VPA on CRF and suggests associated improvements in cognitive function and psychological health, it is important to interpret Fit to Study's null results in the context of the trial: the absence of evidence is not evidence of absence (Alderson, 2004). Crucially, the Fit to Study intervention was limited by poor intervention fidelity, a common issue in the field of PA research (Tarp *et al.,* 2016), that provides an alternative explanation for the lack of the trial's effectiveness. Fewer than half the schools in the intervention group managed to deliver the intervention in at least 50% of PE lessons during the year. Objective measures of VPA collected during single PE lessons demonstrated that the levels of VPA achieved were lower than the target level: approximately 4-5 min/hour of the active lesson was spent in VPA. Although we detected a trend towards higher VPA in the intervention schools, this failed to reach significance. Clearly, without a significant change or difference in exposure level (i.e. PA), no changes or differences in outcomes are to be expected. To explore the effect of compliance, we estimated the treatment effect in the population of compliers, which consisted of schools delivering the intervention in >50% of lessons. These analyses did not change the results of the primary intention-to-treat analysis, although it is possible that a dose greater than 50% is required to observe a complier effect.

One of the main criticisms of HIT-type interventions is the risk of poor adherence and attrition (Biddle & Batterham, 2015; Perri *et al.,* 2002), most likely due to the high intensity (and perceived exertion) inherent to HIT. However, post-intervention teacher surveys, reported previously (Husain *et al.,* 2019), suggested that not the intensity, but

time constraints, lesson disruption (i.e. flow and objectives), lack of space (in combination with class size), seasonal variation and declining engagement due to lack of perceived improvements and repetitiveness of intervention elements, were important determinants of poor adherence in Fit to Study. Indeed, a recent meta-review reported mean adherence rates to be over 80% for HIT interventions (Martland *et al.,* 2020). It is important to note, however, that fidelity is often poorly reported or not reported at all (Lambert *et al.,* 2017; Martland *et al.,* 2020), with possible overestimation of adherence due to publication bias.

Moreover, although Fit to Study was evidence-based, its design prioritised the intervention being brief, inexpensive, simple, feasible and scalable. The intervention was therefore implemented in regular PE lessons, delivered by PE teachers, and adapted to suit the school PE curriculum, consisting of a potential total weekly dose of 20 min of HIT-style VPA. The target duration and/or intensity was therefore low compared to the majority of interventions that have shown HIT-related improvements in CRF, cognitive, and mental health. In particular, cognitive and mental improvements have been observed following 3–5 HIT sessions per week (Costigan *et al.,* 2016; Jeon & Ha, 2017; Leahy *et al.,* 2019; Moreau *et al.,* 2017), while increased CRF levels have been reported following interventions delivered at higher intensities (>85% HR_{max}, Costigan *et al.,* 2015). Moreover, unlike previous studies, the Fit to Study HIT-style infusions may have been interspersed with substantial (i.e. >30 s—3 min) recovery periods. Indeed, recent reviews suggest that interventions with intensities >85%HR_{max}, two-to-three times per week, with longer high intensity intervals (∼4 min) and active recovery periods (of approx. 3 min), lasting more than 7 weeks may provide the optimal stimulus for health improvement (Calverley *et al.,* 2020; Eddolls *et al.,* 2017; Martland *et al.,* 2020; Weston *et al.,* 2014). While the Fit to Study intervention was delivered for 10 months, it may be that the actual dose of HIT was too low or fragmented to observe clear effects, though low intervention compliance renders strong statements about dose tentative.

6.4.2 Practical challenges of school-based PA interventions

Large-scale cluster RCTs embedded in school settings, like Fit to Study, have high ecological validity, but face many practical challenges that reduce the methodological quality of the trial and may introduce bias (see e.g. Naylor *et al.,* 2015; Van Sluijs & Kriemler, 2016 for reviews). A thorough implementation evaluation of Fit to Study has been published elsewhere (Husain *et al.,* 2019; Wheatley *et al.,* 2020b) and various reviews have been published on school-related barriers (Hills *et al.,* 2015; Naylor *et al.,* 2015); here we highlight challenges related to the scale and intervention approach of Fit to Study.

The scale of the Fit to Study trial was required to provide the statistical power to detect small treatment effects (in the primary outcome) and allowed for a representative sample of young adolescents, yet resulted in less control over the intervention (e.g. training, implementation) and poorer measurement of implementation fidelity. It was not possible for all PE teachers to attend the teacher training sessions, primarily due to time constraints. While teacher training videos were provided to help cascade instructions to colleagues, this process may have caused variability in teacher engagement and the quality of intervention delivery within and across schools. To measure such variability in implementation fidelity, objective and self-report measures were put in place. However, self-report measures are subject to response bias (i.e. social desirability, Breitenstein *et al.,* 2010; Lambert *et al.,* 2017), whereas objective measures of adherence, such as those derived from actigraphy or HR monitors, have major logistical and financial limitations when used at scale (Lee & Shiroma, 2014). Smartphone-based applications may prove useful for fidelity measurement in future trials (Savage, 2015; Smith *et al.,* 2014a).

The Fit to Study trial was designed to examine the efficacy of PA to improve maths attainment (secondarily: cognitive and mental health), and hence, whether HIT-style VPA is a viable public health strategy for cognitive enhancement. As a consequence, we used a population-based approach (Biddle & Batterham, 2015; Van Sluijs & Kriemler,

2016), rather than a targeted (e.g. sex-specific) approach, and designed the intervention and its outcomes to be scalable and feasible such that all pupils could participate. Such an inclusive approach prevents isolation and stigmatization (Pössel *et al.,* 2018), yet it is unlikely that an intervention works for all, despite feasibility testing in a small number of schools (unpublished). Indeed, there is evidence that girls may respond better to PA interventions than boys (Yildirim *et al.,* 2011), whereas high intensity PA may show greater cognitive benefits in boys (Ludyga *et al.,* 2020). Similarly, the intervention was easily incorporated into lesson plans of some schools, but proved incompatible with lesson objectives and timetables in others (Husain *et al.,* 2019). Schools that struggled with delivering all intervention elements were allowed to omit the third infusion occasionally (Husain *et al.,* 2019), in order to keep schools, teachers and pupils engaged. Such modifications are not desirable, but suggest that flexibility in intervention delivery may be required. Lack of time was the major barrier (Husain *et al.,* 2019), not only for implementation of the intervention, but also for conducting secondary outcome assessments. Variability in IT-facilities further complicated completion of assessments, resulting in great amounts of missing data.

6.4.3 Strengths and limitations

Strengths of the study include its novelty in implementing a HIT-style intervention at scale in a real-world setting, it being a highly scalable intervention, which was delivered at low cost (Husain *et al.,* 2019), its broad range of secondary outcome measures, its statistical techniques to deal with missing data, and its large cohort of young adolescents, who reported baseline PA levels that were representative of the UK (Sport England, 2018) and the world (Guthold *et al.,* 2020), ensuring generalisability.

The study also has various limitations. First, poor intervention fidelity and large amounts of missing data render any conclusions tentative and results should be interpreted as hypothesis generating, rather than hypothesis testing. Moreover, while we collected self-report data on compliance with the intervention, we did not measure whether individuals reached the required intensity because distribution of HR monitors

to all participants was unfeasible. It is paramount that future studies put in place measures that capture the multidimensional nature of fidelity (Mars *et al.,* 2013; Taylor *et al.,* 2015). To assess the effect of the intervention on VPA during PE, we collected objective measures of VPA using actigraphy during a single PE lesson at baseline and post intervention. However, we decided to use this measure as an indicator of fidelity only, given that (1) this measure was collected on a class level only in contrast to all other outcome measures considered in this study, and (2) a researcher was present during the measurement, possibly biasing the amount of VPA upwards in intervention schools. No active control group was used in this efficacy trial, and hence, no conclusions regarding the optimal dose can be drawn. Finally, there is some evidence that HIT interventions may have greater benefits in pubertal children compared to pre-pubertal children (Eddolls *et al.,* 2017), yet due to the scale of the trial and sensitive nature of pubertal questionnaires, no information on pubertal status was collected.

6.5 Conclusion

The one-academic-year HIT-style VPA intervention had no significant effect on CRF, cognitive performance, or mental health in young adolescents. Large amounts of missing data and poor fidelity with the intervention limit the extent to which conclusions can be drawn regarding the causal relationship of PA and health outcomes in adolescents. Future well-controlled trials testing for the effects of HIT-style interventions are warranted.

7

The effect of the Fit to Study intervention on the anterior hippocampus, cardiorespiratory fitness and relational memory performance

Contents

7.1 Introduction . . . 202
7.1.1 Aims . . . 204
7.2 Methods . . . 204
7.2.1 Design and participants . . . 205
7.2.2 Blinding . . . 205
7.2.3 Intervention . . . 206
7.2.4 Intervention fidelity . . . 206
7.2.5 Outcomes . . . 207
7.2.6 Data analysis . . . 211
7.3 Results . . . 217
7.3.1 Sample characteristics . . . 219
7.3.2 Intervention fidelity . . . 219
7.3.3 The effect on cardiorespiratory fitness . . . 220
7.3.4 The effect on the anterior hippocampus . . . 222
7.3.5 The effect on relational memory . . . 225
7.3.6 Changes in CRF with changes in hippocampal MRI markers . 225
7.3.7 Supplementary analyses . . . 226
7.4 Discussion . . . 227
7.4.1 Anterior hippocampus, CRF and cognitive function . . . 227
7.4.2 Findings in context . . . 230
7.4.3 Strengths and limitations . . . 231

7.5 Conclusion 232

A rich literature in animals and human older adults suggests that physical activity may benefit cognitive- and brain health. The previous chapter examined effect of the Fit to Study vigorous physical activity intervention on cognitive outcomes in a large cohort of British adolescents. In this chapter, I examined the effect of the intervention on anterior hippocampus in a small sub-sample (n = 61) of the Fit to Study cohort. The intervention had no significant effect on anterior hippocampal volume, cardiorespiratory fitness or memory outcomes. I did observe a significant effect on MRI markers within the anterior hippocampus (p = 0.02), which was driven by a physical activity related increase in R_1 in the intervention group. Poor measurement of intervention fidelity and lack of replication in sensitivity analyses, however, suggest that this finding needs to be interpreted with caution and highlight the need for replication in future studies.

7.1 Introduction

Governmental and public health communities have become increasingly interested in physical activity (PA) as a low-cost lifestyle intervention to promote neurocognitive health across the lifespan. There is evidence that PA benefits cognitive abilities and brain health in older adults (Erickson *et al.,* 2019), with some of the most compelling findings for the hippocampus (Stillman *et al.,* 2020), a region important for learning and memory. Experimental studies have demonstrated PA-related increases in hippocampal volumes, especially in its anterior portion (Erickson *et al.,* 2011; Thomas *et al.,* 2016). Less is known about the effects of PA on neurocognitive health in young people, with a particular lack of studies in adolescents. Randomised controlled trials (RCT) assessing the effects of PA on cognitive performance in this age group have been inconclusive (Wassenaar *et al.,* 2020); those involving neuroimaging to assess its effects on the brain have shown changes in white matter integrity and brain function (Meijer *et al.,* 2020; Valkenborghs *et al.,* 2019), but none have examined hippocampal volumes. To date, there has been only one quasi-experimental study that did not find an effect of doubling the number of PE classes on hippocampal volumes in young adolescents (Bunketorp

Käll *et al.,* 2015), though no baseline measures were taken and participants were not randomised. Hence, whether a PA intervention can impact on hippocampal volumes and cognitive abilities in adolescents is still unclear.

While PA-related volumetric changes of the hippocampus have been found in young (Wagner *et al.,* 2015), middle-aged (Thomas *et al.,* 2016) and older adults (Erickson *et al.,* 2011), the cellular processes underlying these changes remain largely unknown. Some insights may be gained from animal studies, which have shown a host of exercise-induced cellular changes in the hippocampus (Cooper *et al.,* 2018; Gomez-Pinilla & Hillman, 2013; Vivar & Van Praag, 2017; Voss *et al.,* 2019; Voss *et al.,* 2013), including an increased rate of neurogenesis (Creer *et al.,* 2010; Van Praag, 2005; Van Praag *et al.,* 1999), enhanced angiogenesis (Black *et al.,* 1990; Ding *et al.,* 2006; Morland *et al.,* 2017; Van Der Borght *et al.,* 2009), various morphological adaptations (e.g. dendritic complexity, Eadie *et al.,* 2005) and elevated oligodendrocyte precursor proliferation (Matsumoto *et al.,* 2011). In humans, exercise-induced changes in cardiorespiratory fitness (CRF) have been related to increased hippocampal perfusion (Chaddock-Heyman *et al.,* 2016; Maass *et al.,* 2015), blood volume (Pereira *et al.,* 2007), and measures of tissue density (Kleemeyer *et al.,* 2016), which may underlie changes in hippocampal volumes. Studies focussing on single modalities such as perfusion, however, only shed light on a single putative mechanism. Multi-modal imaging studies may be better suited to disentangle the contributions of different biological processes to volumetric changes. While the cross-sectional multi-modal imaging study reported in chapter 5 did not show significant associations between metrics of PA and a range of MRI markers in the hippocampus, a previous experimental study showed that exercise-related increases in anterior hippocampal volumes may be better explained by increases in myelin instead of vasculature (Thomas *et al.,* 2016).

To address the gaps in the literature and extend analyses presented in chapter 5, we examined the effect of the Fit to Study HIT-style VPA intervention on the anterior hippocampus in the brain imaging sub-sample. In particular, we examined the effect of

the intervention on anterior hippocampal volume, CRF, and hippocampal-dependent memory. We additionally examined the effect of the HIT-style VPA intervention on MRI markers of tissue microstructure within the anterior hippocampus. We also explored whether the effect of the intervention on anterior hippocampal outcomes was moderated by sex, socioeconomic status or baseline CRF levels, and whether changes in CRF were related to changes in anterior hippocampal metrics.

7.1.1 Aims

The following primary aims were specified prior to the start of the analysis:

Primary aims

1. examine the effect of the HIT-style VPA intervention on anterior hippocampal volume, CRF, and relational memory performance
2. examine the effect of the HIT-style VPA intervention on MRI markers of tissue microstructure within the anterior hippocampus

We additionally prespecified the following set of secondary aims, that aimed to further explore any potential intervention effects:

Secondary aims

1. explore the whether the effect of the HIT-style VPA intervention on anterior hippocampal volume and MRI markers within the anterior hippocampus was moderated by sex, socioeconomic status or baseline CRF level
2. explore whether changes in CRF were related to changes in left or right anterior hippocampal volume, and to changes in anterior hippocampal MRI markers

We additionally examined the effect of the VPA intervention on the volume of the entire hippocampus and hippocampal subfields.

7.2 Methods

Details on the design, intervention, cognitive assessments and MRI acquisition are provided in chapter 2 and in Wassenaar *et al.* (2019a); details of MRI processing and analysis are described in chapter 5.

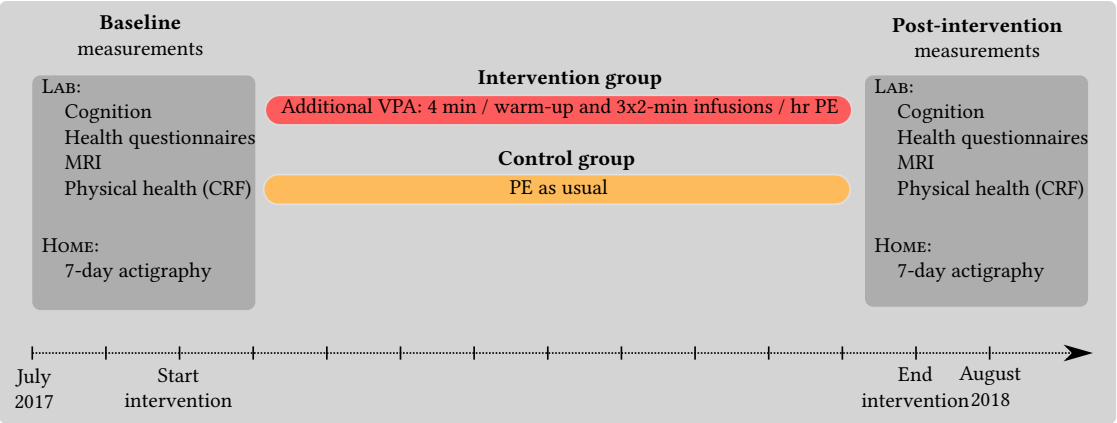


Figure 7.1: Experimental design of the brain imaging sub-study. At baseline (July 2017), participants completed a set of cognitive assessments, health questionnaires, a MRI scan and assessment of physical health in the lab. Their PA behaviour was collected using 7-day actigraphy. The same set of measures were collected directly post-intervention, between June and September 2018.

7.2.1 Design and participants

Sixty-one healthy individuals (31 males; mean age = 12.5 years, SD = 0.33) from 10 schools participating in the Fit to Study trial enrolled in the brain imaging sub-study in June-July 2017.

The brain imaging sub-study ran in parallel with the main trial to make optimal use of the intervention. Assessments took place pre-intervention (baseline, t_0), immediately post-intervention (t_1) and 12-months post-intervention (t_2). In this study, we focus on the data collected pre-intervention and immediately post-intervention to examine the immediate effect of the intervention on CRF, brain health and cognitive performance (Figure 7.2.1).

7.2.2 Blinding

Schools were informed of their treatment status following baseline measurements, so that intervention PE teachers could receive training and deliver the intervention. Given the nature of the intervention, it was not feasible for PE teachers or pupils to be blind to group allocation. Researchers visiting schools for top-up training and collection of fidelity measures were not blinded. The trial manager was involved in recruitment of

schools for the brain imaging and therefore not blinded. Researchers involved in (eligibility) screening of sub-study participants, as well as those collecting questionnaire-, cognitive- and MRI measures were blinded to treatment status. Researchers assisting with the CRF test were not blinded to treatment status of schools, but were unaware of school allocation of participants and therefore considered blinded to treatment status. All analyses were performed blinded to treatment status.

7.2.3 Intervention

The intervention consisted of a one-academic year (10 months: September 2017 –June 2018) high-intensity training (HIT) style VPA programme delivered by PE teachers during regular Year-8 PE lessons. The intervention required teachers to deliver specific elements of additional VPA, over and above normal VPA, during all Year-8 PE lessons, for the whole school year. Four minutes of VPA (71–85% of HR_{max}) were incorporated into the PE lesson as part of a warm-up and three two-minute infusions per hour of PE were incorporated into the main PE class. The intervention elements were designed to be easy to incorporate into PE lessons, to maximise the training effect, and to minimise the risk of injury with a warm-up that progresses from light to moderate-intensity activities and includes bursts of VPA. Control schools were not given training and were asked to deliver PE as usual.

7.2.4 Intervention fidelity

Intervention fidelity was measured at a school level as part of the main trial of Fit to Study (for details see chapter 2) using a prespecified set of measures, including (1) teacher log books, (2) an objective measure of average VPA per hour of PE using Axivity AX3 accelerometers in a convenience sample of participants per school (at least 50% of the Year group), and (3) pupil-reported compliance with the intervention using a brief three-question survey administered post-intervention asking whether (i) PE lessons started with a warm-up, (ii) PE lessons included bursts of VPA that raised their HR and made them feel out of breath, and (iii) they took part in warm-ups or VPA bursts

if asked by PE teachers. Due to low compliance with the log books, NatCen Social Research additionally collected the teacher-reported percentage of PE lessons in which the intervention was delivered as intended (post-intervention).

In the sub-study, we additionally collected 7-day actigraphy during a regular school week towards the end of the intervention (May–June 2018).

7.2.5 Outcomes

Pupils were assessed before the intervention started (pretest, t_0) and immediately following the intervention (posttest, t_1). Assessments, in order of completion, included tests of cognitive performance, an MRI scan, and a test of CRF.

7.2.5.1 Cardiorespiratory fitness

CRF ($\dot{\mathbf{V}}O_2$peak) was assessed by a maximal graded exercise test on a cycle ergometer (Lode Excalibur, Groningen, The Netherlands), using the *Godfrey* protocol (Godfrey & Mearns, 1971). The test was terminated when the participant reached volitional exhaustion or was unable to maintain a cadence of 60 rpm despite verbal encouragement.

7.2.5.2 Cognitive assessments

The object-location task (OLT) and relational memory paradigm (chapter 5) were used to assess memory performance.

- A modified version of a previously described OLT (Pertzov *et al.*, 2012; Pertzov *et al.*, 2013) was used to assess relational memory performance. The proportion of misbinding / swap errors was used as the outcome measure.
- A modified version of a previously described relational memory task (Chaddock *et al.*, 2010b) was used to assess visual relational memory performance. The proportion of correct responses on valid trials (i.e. accuracy) was the primary outcome measure.

7.2.5.3 MRI acquisition and analysis

MRI acquisition Details of the MRI protocol are described in chapter 2. In brief, all MR data were acquired on a 3 Tesla (Prisma, Erlangen, Germany) MRI scanner at the Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB) using a 32-channel head coil. The protocol took approximately 55 minutes (Figure 7.2) to complete. In this study, only the T_1w, diffusion-weighted, quantitative MRI and arterial spin labelling (ASL) sequences were used and their analysis is discussed below.

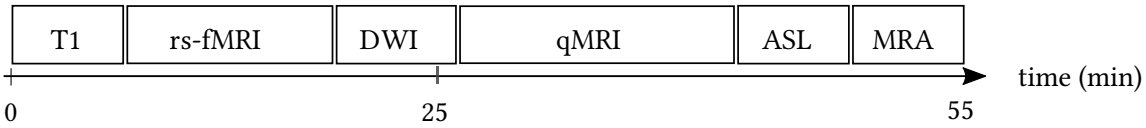


Figure 7.2: MRI protocol. Participants underwent a 55 minute MRI scan, comprising a T_1-weighted scan (T1), a resting-state fMRI (rs-fMRI), a diffusion-weighted scan (DWI), a quantitative MRI sequence (qMRI), an arterial spin labelling (ASL) scan and MR angiography. The size of the scan blocks approximate scan duration.

MRI data analysis Image analysis was performed using tools from the FMRIB Software Library (FSL, version 6; Jenkinson *et al.,* 2012; Smith *et al.,* 2004), Statistical Parametric Mapping (SPM, http://www.fil.ion.ucl.ac.uk/spm/), and FreeSurfer (version 6.0; Dale *et al.,* 1999).

- **Structural T_1-weighted** scans were corrected for gradient distortions and further processed using the FSL_ANAT anatomical processing pipeline available within FSL The volumetric scaling factor from SIENAX (Smith, 2002) was used as a proxy for head size. The hippocampus was segmented using FSL's Multimodal Image Segmentation Tool (MIST, Visser *et al.,* 2016), visually inspected and manually edited (where necessary) by a researcher blind to group allocation. Specifically, images were edited to ensure that (1) voxels into the CSF and neighbouring non-hippocampal structures (e.g. the amygdala and thalamus) were removed, (2) the alveus was included in its entirety, and (3) the fimbria was excluded. The anterior hippocampus was then demarcated following recommendations by

Poppenk *et al.* (2013), and described in detail in chapter 5.2.5.2. Specifically, T_1-weighted images were linearly transformed to MNI152 space using FSL's FLIRT and, moving from posterior to anterior in the coronal plane, we identified the y-coordinate (MNI) where the uncal apex first appeared. Subsequently, a 3D mask was created that started at this coordinate and covered the anterior portion of the brain. The mask was then transformed back to native space (using the inverse of the transformation to MNI) of each participant to extract the anterior portion of the hippocampus.

- **Diffusion-weighted imaging** scans were corrected for eddy currents, EPI distortions, inter-and intra volume subject head motion and gradient distortions using FSL's eddy tool (Andersson & Sotiropoulos, 2015; Andersson & Sotiropoulos, 2016) and bespoke GDC correction scripts (GitHub pipelines). We subsequently fitted the diffusion tensor model and Bingham-Neurite Orientation Dispersion and Density Imaging (NODDI) model (Tariq *et al.*, 2016) to obtain microstructural markers: fractional anisotropy (FA), mean diffusivity (MD), intra-cellular volume fraction (fintra or ICVF), the isotropic (free water) volume fraction (ISOVF) and the orientation dispersion index (ODI).

- **Quantitative MRI** data (i.e. multi-echo PDw and T_1w images) were corrected for gradient distortions and subsequently processed using the Voxel-Based Quantification (VBQ) toolbox available in SPM to obtain R_1 ($1/T_1$) and R_2^* ($1/T_2{}^*$) maps (Callaghan *et al.*, 2014).

- **Pseudo-continuous ASL** scans were corrected for bias-field and gradient distortions and subsequently preprocessed using the BASIL tool available in FSL (Chappell *et al.*, 2009). A perfusion map was derived for each participant.

Diffusion-weighted images, as well as R_1, R_2^* and perfusion maps, were registered to T_1-weighted native space using FLIRT with boundary-based cost function (Greve & Fischl, 2009; Jenkinson *et al.*, 2002).

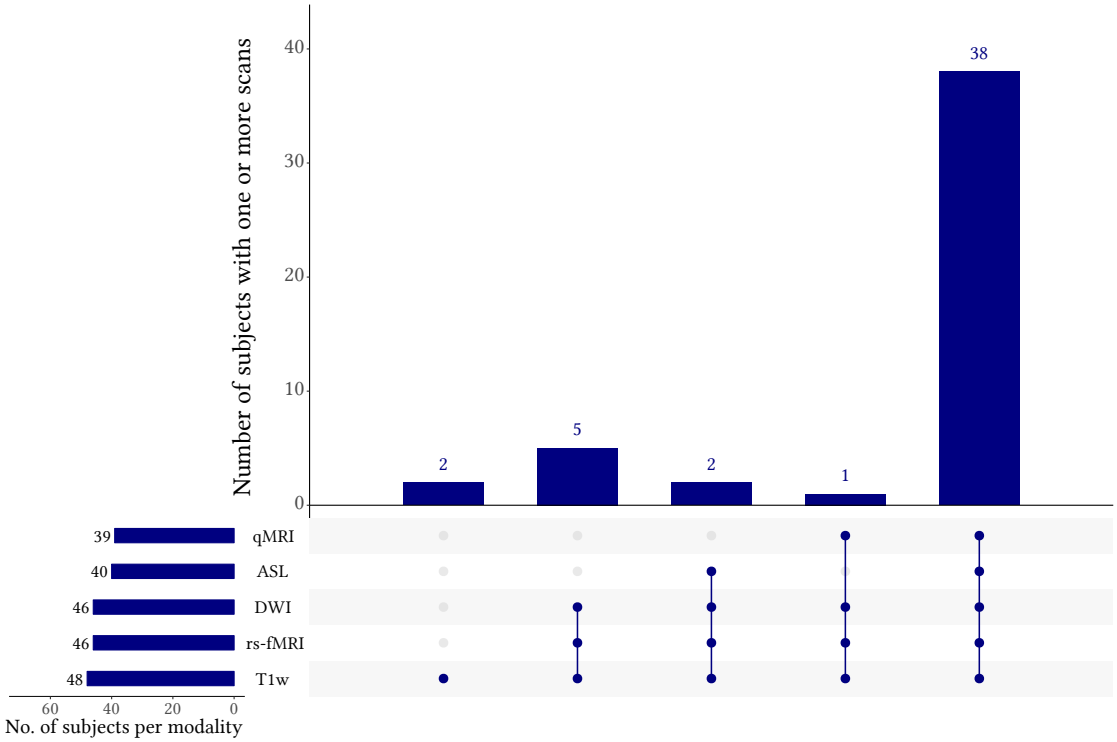


Figure 7.3: Completion of MRI scans across timepoints. *Left bar plot* The number of participants that completed each sequence at both timepoints (baseline and post-intervention). For instance, 48 out of 61 participants had complete T_1-weighted data. *Right bar plot* The number of participants with one or more complete modalities. Linked dots in combination with the bar plot show the sets of modalities (dots) that were present for a subset of participants (bars). For instance, complete MRI data was available for 38 participants.

Quality control and drop-out All raw and preprocessed images were manually checked for gross problems affecting image quality, including motion artifacts and registration errors. At baseline (n = 61), two diffusion-weighted scans, three quantitative MRI scans, and five ASL scans were considered of too low quality –mainly due to motion artifacts –for analysis and were therefore removed from the dataset. Twelve participants did not complete the MRI scan at follow-up (mainly due to drop-out) and of the remaining 49 participants, seven quantitative MRI scans and five ASL scans were considered of too low quality and removed. Forty-eight participants had at least one modality acquired at both timepoints (Figure 7.3), and 38 participants had all five modalities at both timepoints (i.e. the *complete-case* dataset; removal of the rs-fMRI sequence does not change the plot).

7.2.6 Data analysis

7.2.6.1 Univariate analysis

The primary analyses assessed the effect of the intervention anterior hippocampal volume, CRF and hippocampal-dependent memory using multiple imputation assuming that data was missing at random (MAR, details are provided in section 7.2.6.3). Linear mixed models (LMM) fitted with restricted maximum likelihood, with random intercept to capture between-school variability, were used for all primary analyses using *lme4* package (Bates *et al.,* 2015). Analyses followed recommendations for (cluster) randomised controlled trials and clinical trials (EMA, 2015; Twisk *et al.,* 2018): models included baseline values of the outcome measure (i.e. analysis of covariance, ANCOVA), which prevents regression to the mean (Clifton & Clifton, 2019; Twisk *et al.,* 2018; Vickers & Altman, 2001) and is (slightly) more powerful than traditional repeated measures analysis of variance (ANOVA, Van Breukelen, 2006). When the models did not converge due to e.g. a singular fit —usually indicating that the random effects structure does not fit the data —it is recommended to simplify the model while keeping it maximal to allow for a non-singular fit (Barr *et al.,* 2013; Matuschek *et al.,* 2017). Here, the random intercept was removed and the model was refit as a single level linear model using the *stats* package.

In line with analyses of the primary outcome of the trial (Husain *et al.,* 2019) and the secondary outcomes of the main trial chapter 6, we subsequently performed exploratory subgroup analyses to explore whether the effect of the intervention on anterior hippocampal volume was moderated by sex, socioeconomic status or baseline CRF levels. Socioeconomic status (index of multiple deprivation deciles) and baseline CRF levels were treated as continuous covariates.

Finally, we explored whether changes in CRF were related to changes in anterior hippocampal volume, and whether these relationships were moderated by treatment status, adjusting for sex, socioeconomic status and change in pubertal status.

The unstandardized mean differences are provided, as well as the standardized mean differences for completion. The standardized mean difference score was computed using the pooled unconditional standard deviation of the outcome (Husain *et al.,* 2019; White, 2005). For all statistical tests, a two-sided p <0.05 was used. Bonferroni-correction was applied to correct for multiple comparisons where appropriate, but text displays uncorrected p-values unless specified otherwise. An overview of analysis software is provided in Appendix G.1.3.

7.2.6.2 Multivariate analysis

We used permutation-based non-parametric combination (NPC, Pesarin, 1990) with Fisher's combining function (Winkler *et al.,* 2016) to jointly test for an effect of the intervention across MRI markers of tissue microstructure (with each hemisphere being coded as a single modality), while adjusting for baseline values of each modality (i.e. ANCOVA set-up). The MRI markers, or imaging-derived phenotypes (IDPs), that were included matched those used in the cross-sectional analysis (chapter 5): grey matter density (VBM), perfusion (ASL), mean diffusivity (MD), intra-neurite density (fintra, also called intra-cellular volume fraction), R_1 and R_2^*.

Compared to classical multivariate analysis of variance (MANOVA) tests, NPC is more powerful and is based on fewer assumptions (e.g. no normality is required, Winkler *et al.,* 2016). NPC consists of two phases: (1) the null hypothesis of no difference between the groups is tested separately for each metric (i.e. *partial tests*) using permutations that are performed in synchrony, (2) the statistics of the partial tests are converted and combined using one of various combining functions (see Winkler *et al.,* 2016 for an overview) to provide a p-value for the joint test. Here Fisher's combining function was selected based on its favourable properties: it appropriately controls the error rate, it performs well with a large number of tests, and is powerful if an effect is present in multiple tests (Winkler *et al.,* 2016), which is to be expected in the anterior hippocampus (i.e. effects of the intervention may be expected on more than one IDP,

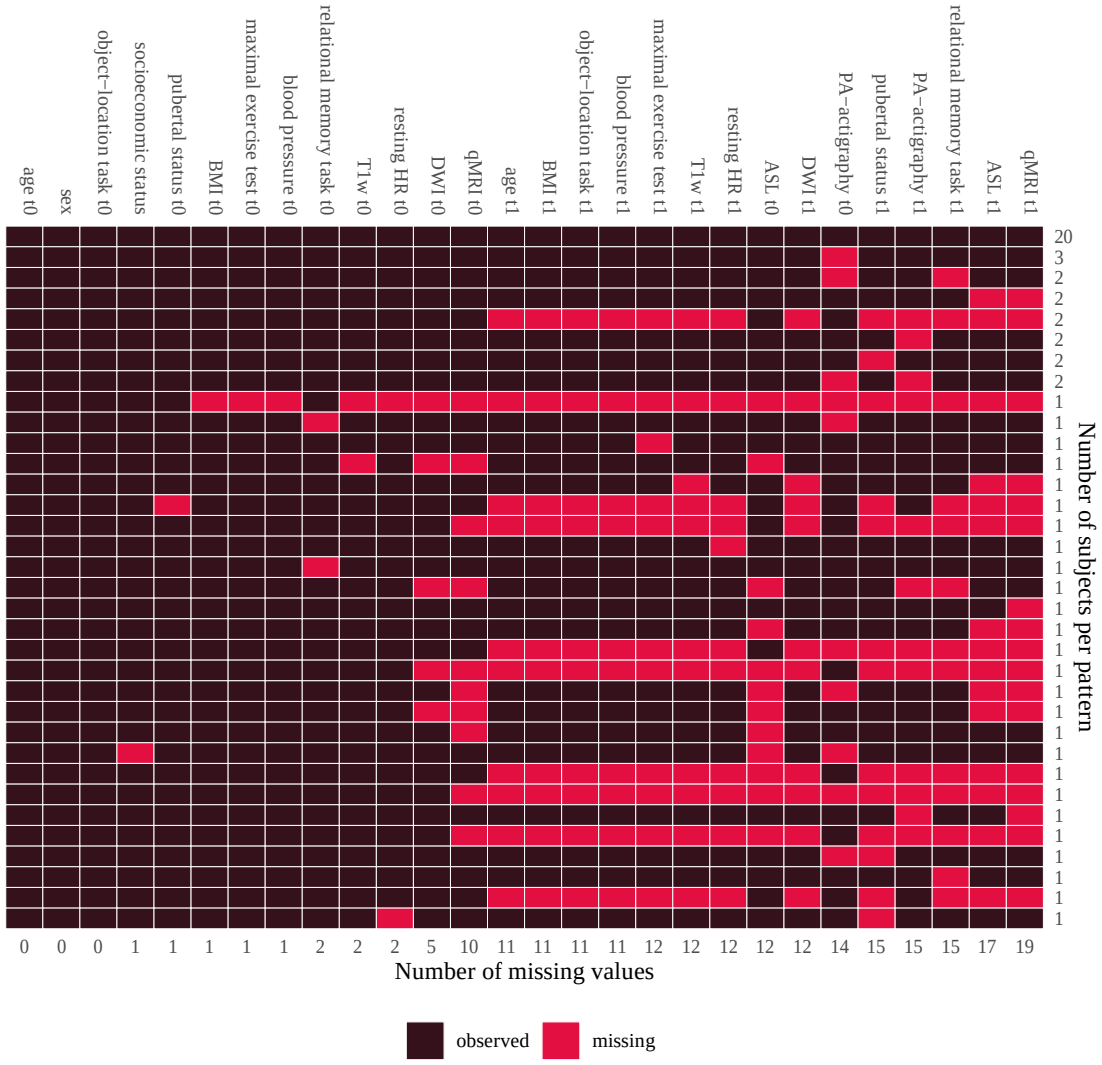
given strong relationships between some IDPs, e.g. MD and intra-neurite density).

Here, we ran NPC across 12 partial tests (six modalities $\times$ two hemispheres), on multiply imputed data, with 1000 permutations and family-wise error rate (FWER) correction across modalities and contrasts. Each partial test examined the difference in MRI marker between intervention and control groups at post-intervention, while controlling for baseline values of the respective MRI marker. It is important to note that this type of permutation test currently does not allow for a multilevel structure (i.e. school clusters) to be modelled. Given that the majority of univariate LMMs indicated that schools accounted for a negligible amount of variance in the hippocampal or behavioural data, we ignored the clustering in the NPC. However, results would need to be interpreted with caution.

In line with the univariate analyses, we secondarily explored whether sex, socioeconomic status or baseline CRF moderated the effect of the intervention on anterior hippocampal MRI markers. We also explored the relationship between changes in CRF and changes in anterior hippocampal MRI markers, and whether this relationship was moderated by treatment status, adjusting for baseline CRF, baseline age and the interaction between changes in age and treatment status of participants. A total of 1000 permutations were run for each analysis and FWER corrected p-values are reported.

7.2.6.3 Missing data

The missing data pattern (Figure 7.4) and relationships between variables with missing data were inspected. Participants had complete baseline data for age, sex and the OLT. Nine out of 61 participants ($\sim$15%, six in intervention) were lost to follow-up. Reasons for the drop-out included fatigue or anxiety during baseline testing (n = 2), physical health reasons (n = 1), non-response to repeated phone calls or lack of interest (n = 6). A further two participants did not attend the follow-up testing session, but completed


age t0
sex
object−location task t0
socioeconomic status
pubertal status t0
BMI t0
maximal exercise test t0
blood pressure t0
relational memory task t0
T1w t0
resting HR t0
DWI t0
qMRI t0
age t1
BMI t1
object−location task t1
blood pressure t1
maximal exercise test t1
T1w t1
resting HR t1
ASL t0
DWI t1
PA−actigraphy t0
pubertal status t1
PA−actigraphy t1
relational memory task t1
ASL t1
qMRI t1
20 3 2 2 2 2 2 2 1 1 1

Univariate analysis The primary univariate intention-to-treat analysis, assessing the effect of the intervention on CRF, anterior hippocampal volume and cognitive performance, assumed that data was missing at random (MAR) and was subsequently multiply imputed using a joint modelling approach for multilevel multiple imputation available in the PAN package (Grund *et al.,* 2016; Schafer & Graham, 2002; Zhao & Schafer, 2018) to create 50 imputed datasets (50,000 burn-in iterations, and 5000 iterations between imputations). We included all variables that were part of the analysis model, as well as any variables that were associated (r > 0.4) with variables in those models or related with missingness in variables in the analysis model (r > 0.3) to make the missing at random (MAR) assumption more plausible (variables are listed in Appendix G.1.1). The imputations were run separately for treatment and control groups (Sullivan *et al.,* 2018). Convergence of the MCMC chains was inspected visually and using the average scale reduction factor across estimates (close to 1). The analysis was then run in each imputed dataset and parameters were combined using Rubin's rules (Rubin, 1976).

Secondary subgroup and change-change analyses were run on complete cases only, due to non-convergence of MCMC chains.

Multivariate analysis Only 36 (57%) participants had complete multimodal MRI data and missing data was multiply imputed using the joint modelling approach described above. We included all data that were part of the analysis model (i.e. all hippocampal MRI metrics and the treatment allocation) as well as any variables that were associated (r >0.5) with variables in those models or related with missingness in variables in the analysis model (r >0.3) to make the missing at random (MAR) assumption more plausible. While it is desirable to impute missing data by intervention group, to allow for group-specific covariance structures, this was not feasible given the large number of variables to be imputed and small number of observations. We subsequently ran the primary multivariate analysis on all imputed data and combined the p-values

using NPC.

Secondary subgroup and change-change analyses were run on complete cases only. Due to the number of variables, it was not possible to impute missing MRI marker data by treatment group, which is required if true relationships are to be preserved.

Sensitivity analysis We repeated all the primary analyses (univariate and multivariate) on complete-case data. In addition, given that the NPC method for multivariate analysis using multiply imputed data is theoretically justified, but not yet empirically assessed, we repeated the primary multivariate analyses on the singly imputed dataset derived using the singular value decomposition imputation method (Husson *et al.,* 2019) described in chapter 5.

7.2.6.4 Supplementary analysis: hippocampus and hippocampal subfields

To date, the majority of exercise studies in humans have examined the entire or anterior hippocampus. Much of the animal literature, however, has taken a more fine-grained approach and examined hippocampal subfields, showing e.g. exercise-induced neurogenesis in the dentate gyrus (DG, Van Praag, 2005). In humans, there is an increasing interest in studying hippocampal subfields (Giuliano *et al.,* 2017), including how they may be affected by exercise. To date, exercise intervention studies have shown exercise-related volumetric increases that are specific to the DG (Frodl *et al.,* 2019; Pereira *et al.,* 2007). A longitudinal study, however, demonstrated that 12 weeks of exercise increased the volume of CA1, but not the DG (Den Ouden *et al.,* 2018).

The primary approach in this chapter has been to examine the anterior hippocampus for two reasons: (1) the majority of previous studies have found exercise-related effects on volume of this region (Erickson *et al.,* 2011; Sagi *et al.,* 2012; Thomas *et al.,* 2016), we therefore decided to take this region as the area of primary interest, (2) among researchers, it is questionable whether hippocampal subfields can be accurately demarcated on standard 3T T_1-weighted (1 mm isotropic) images that lack the necessary contrast and resolution (Cong *et al.,* 2018; Wisse *et al.,* 2020; Yushkevich *et al.,* 2015),

and there is a lack of consensus on how subfields (i.e. borders) are defined (Olsen *et al.,* 2019; Yushkevich *et al.,* 2015).

Bearing in mind these caveats, we *a posteriori* decided to additionally explore the effect of the intervention on hippocampal subfields because no significant intervention effect on the anterior hippocampus was observed (primary analysis) and animal literature suggests that the majority of exercise-related effects occur in the DG. The DG runs along the long-axis of the hippocampus and is only partly included in the anterior hippocampus (DeKraker *et al.,* 2019; Duvernoy, 2005). Hence, by studying the anterior hippocampus we may have been insensitive to subfield-specific effects. Therefore, we segmented the hippocampal subfields using the Multiple Automatically Generated Templates (MAGeT) brain segmentation algorithm (Chakravarty *et al.,* 2013; Pipitone *et al.,* 2014). Details of the pipeline and analysis are provided in Appendix G.1.2.

In addition to the subfields, we also explored the effect of the intervention on the entire hippocampus. Given the exploratory nature of the analyses and because we *a posteriori* decided to undertake these analyses, they were run on complete-cases only.

7.3 Results

Of 104 schools that were randomised to treatment and control, 61 pupils from a subset of 10 schools (5 intervention, 5 control) enrolled in the brain imaging sub-study (Figure 7.5). The 10 schools were representative of schools participating in the main trial sample (details in chapter 2). Higher pupil fitness but not age, sex, eFSM status or self-reported PA, was significantly related with participation in the brain imaging sub-study.

Of 10 schools at baseline, 1 intervention school (1 pupil enrolled) discontinued the intervention for logistical reasons, but its participant completed post-intervention assessments. Of 61 participants at baseline, 9 were lost to follow-up (15%).

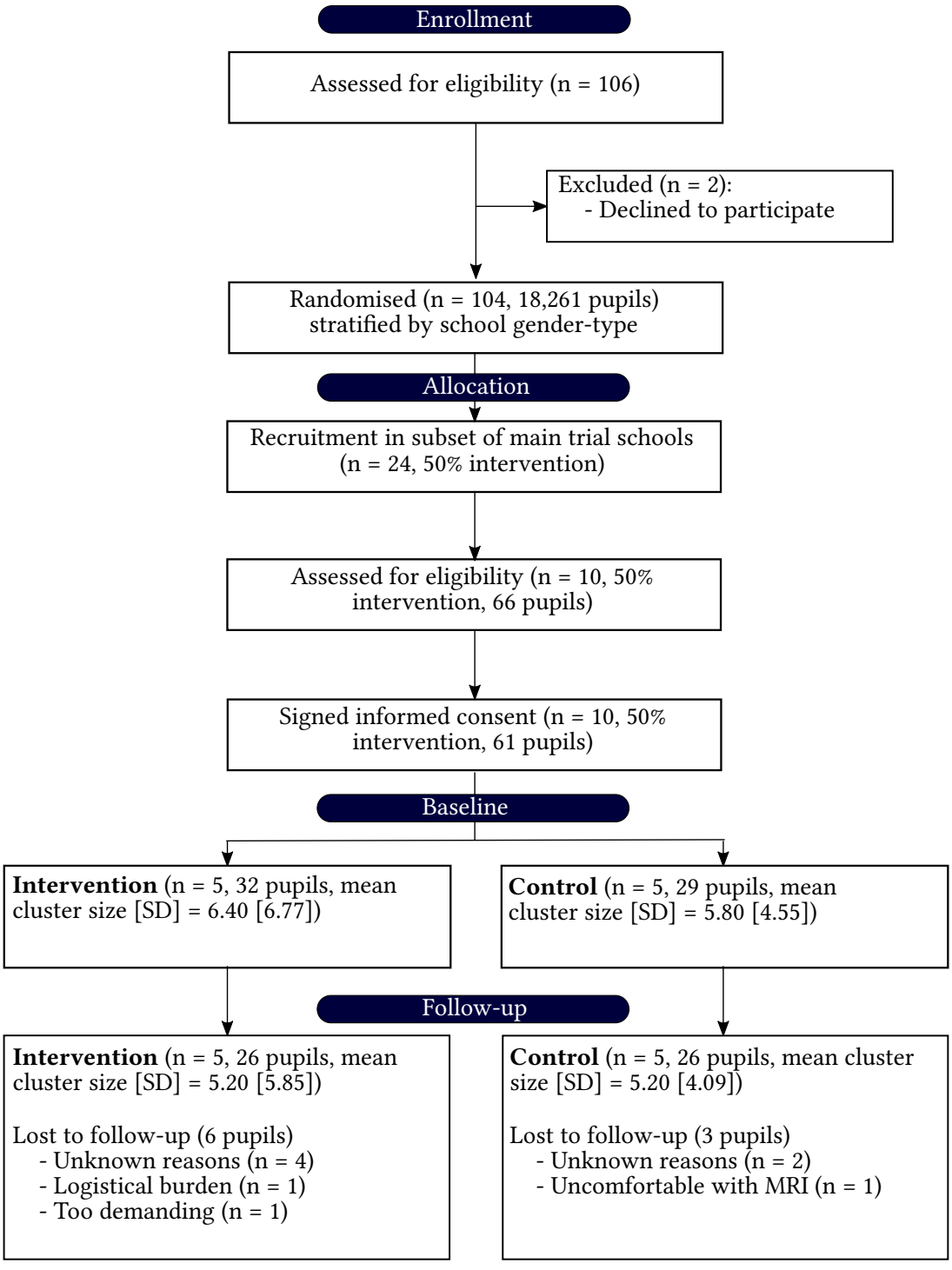


Figure 7.5: CONSORT flow diagram of the brain imaging sub-study. Of 106 main trial schools that were assessed for eligibility, two declined to participate and 104 were randomised into an intervention and control group (n = 52 schools in each arm). Of randomised schools, 24 (50% intervention) were selected for recruitment of brain imaging sub-study participants. A total of 66 participants from 10 schools (50% intervention) were eligible for participation, of which 61 signed an informed consent form. Of 61 participants that were available for baseline assessments, nine dropped out during the course of the study.

7.3.1 Sample characteristics

Descriptive statistics for school- and pupil level characteristics at baseline are presented in Table 7.1. The groups were approximately similar on all measures, although participants in the control group had higher baseline PA levels compared to those in the intervention group. Approximately 34% of the sample reported that they were active for at least 60 min per day for the entire week (>20% of pupils meeting PA recommendations globally). School-level baseline characteristics are provided in Appendix G.2.1.

Table 7.1: Baseline characteristics of schools and participants by treatment group.

	Intervention	Control
Demographics		
No. schools	5	5
No. of individuals	32	29
Age[a], mean (SD), y	12.4 (0.27)	12.5 (0.39)
Female, n (%)	16 (50.0%)	14 (48.3%)
Socioeconomic status		
IMD[b], median (range), decile	8 (2-10)	9 (3-10)
eFSM[c], no. (%), yes	0 (0%)	1 (3.4%)
Pubertal status,mean (SD), a.u.	2.13 (0.72)	2.29 (0.51)
Physical health		
BMI, kg / mm^2,mean (SD)	19.0 (3.62)	19.9 (3.99)
Resting heart rate, mean (SD)	75.4 (12.3)	75.8 (13.4)
Diastolic blood pressure, mean (SD), mmHg	70.8 (7.52)	74.0 (7.03)
Systolic blood pressure, mean (SD), mmHg	104 (10.1)	108 (10.2)
Physical activity		
Daily average MVPA, mean (SD), min	45.4 (21.4)	50.0 (23.2)
Daily average VPA, mean (SD), min	6.90 (4.62)	7.78 (5.51)
Self-reported PA, median (range), days	4.5 (2-7)	6 (2-7)

Abbreviations: BMI = body-mass index, Co-ed = co-education, eFSM = eligible for free school meals, MVPA = moderate to vigorous PA, no. = number, PA= physical activity, SD = standard deviation, VPA = vigorous physical activity

[a] Age at baseline (MRI) assessments

[b] Index of multiple deprivation was collected using the participant's postcode, through Government records: http://imd-by-postcode.opendatacommunities.org/imd/2019

[c] Percent eFSM indicates the percentage of pupils within a school that were enrolled in Fit to Study who are eligible for free school meals.

7.3.2 Intervention fidelity

Of 5 intervention schools, 3 provided information about their adherence to the intervention. One school (n = 4) reported that the intervention was delivered in 75% of PE

classes, whereas two schools (n = 10) reported that the intervention was delivered in over 50% of classes). Objective measures of VPA obtained during a single PE lesson demonstrated that, on average, VPA was higher during the PE lesson in intervention schools compared to control schools (M_{int} [SD] = 6.92 [2.33], M_{ctrl} [SD] = 4.70 [1.80] minutes / hour), though this difference was not significant (p = 0.17). Details on intervention fidelity are provided in Appendix G.2.2.

Pupil reported measures of compliance suggested that PE lessons in both intervention and control schools often started with a warm-up (median = 4, with 1 corresponding to "never" and 5 to "always") and always included infusions (median = 5). Moreover, pupils of both intervention and control schools always took part in intervention elements when asked by the teacher (median = 5). No difference in daily average minutes of VPA during school hours was observed (2.67 [2.06] vs. 2.46 [1.77] min, p = 0.63).

7.3.3 The effect on cardiorespiratory fitness

Outcomes are provided in Table 7.2. The effect of the intervention on CRF was not significantly different between intervention and control at posttest (unstandardized mean difference [UMD] = -2.48, 95% CI: -8.05, 3.10, p = 0.38). Sensitivity analysis using complete-cases did not change these results (UMD = -2.16, 95% CI: -7.02, 2.71 p = 0.44; for details see Appendix G.2.4.1).

Table 7.2: The effect of the intervention on CRF, anterior hippocampal volumes and memory performance.

	Intervention		Control		Adjusted mean difference		*p*
	Baseline M (SD)	Posttest M (SD)	Baseline M (SD)	Posttest M (SD)	Unstandardized M (SD)	Standardized M (SD)	
Cardiorespiratory fitness							
$\dot{V}O_2peak$[b] (ml/kg/min)	35.72 (7.14)	36.92 (6.66)	39.26 (9.74)	41.11 (13.07)	-2.48 (-8.05, 3.10)	-0.24 (-0.78, 0.30)	0.38
Anterior hippocampal volume							
Left (mm^3)	1437.58 (251.12)	1455.65 (255.99)	1430.31 (334.46)	1479.03 (343.84)	-30.62 (-72.17, 10.94)	-0.10, -0.24, 0.04)	0.15
Right (mm^3)	1559.68 (232.17)	1555.53 (266.09)	1543.1 (321.25)	1562.54 (314.03)	-23.63 (-67.99, 20.71)	-0.08, -0.24, 0.07)	0.3
Memory performance							
Object-location task (% errors)	14.22 (8.53)	13.54 (8.7)	13.79 (8.09)	10.7 (7.35)	2.79 (-1.78, 7.37)	0.34 (-0.21, 0.89)	0.23
Relational memory task (accuracy)	69.79 (12.44)	72.93 (15.04)	71.5 (12.79)	70.43 (14)	2.86 (-5.58, 11.40)	0.19 (-0.38, 0.77)	0.51

Abbreviations: M = mean, SD = standard deviation
[a] Standardized mean difference, scaled by the variance of the outcome
[b] Fit using a multilevel model

7.3.4 The effect on the anterior hippocampus

7.3.4.1 Anterior hippocampal volume

The effect of the intervention on left and right anterior hippocampal volume was not significantly different between intervention and control at posttest (UMD_{left} = -30.62, 95% CI: -72.17, 10.94, p = 0.15; UMD_{right} = -23.63, 95% CI: -67.99, 20.73, p = 0.3).

Sensitivity analyses using complete-cases only were in line with the primary analysis for left anterior hippocampal volume (UMD = -27.31, 95% CI: -64.7, 10.09, p = 0.15) and right anterior hippocampal volume (UMD = -1.76, 95% CI: -38.1, 34.57, p = 0.92). Details are provided in Appendix G.2.4.1.

7.3.4.2 MRI markers in the anterior hippocampus

The primary ITT analysis showed a significant effect of the intervention on MRI markers of tissue microstructure within the left or right anterior hippocampus (Fisher's NPC FWE-p = 0.02). Inspection of the uncorrected partial tests showed that this effect was driven by an effect of the intervention on R_1 (left: 80% of tests had uncorrected-p <0.05, right: 90% of tests had uncorrected-p <0.05). Figure 7.6 displays the distributions of change scores for each MRI marker by treatment group. None of the FWER corrected partial tests were significant. Details of these analyses are provided in Appendix G.2.3.1.

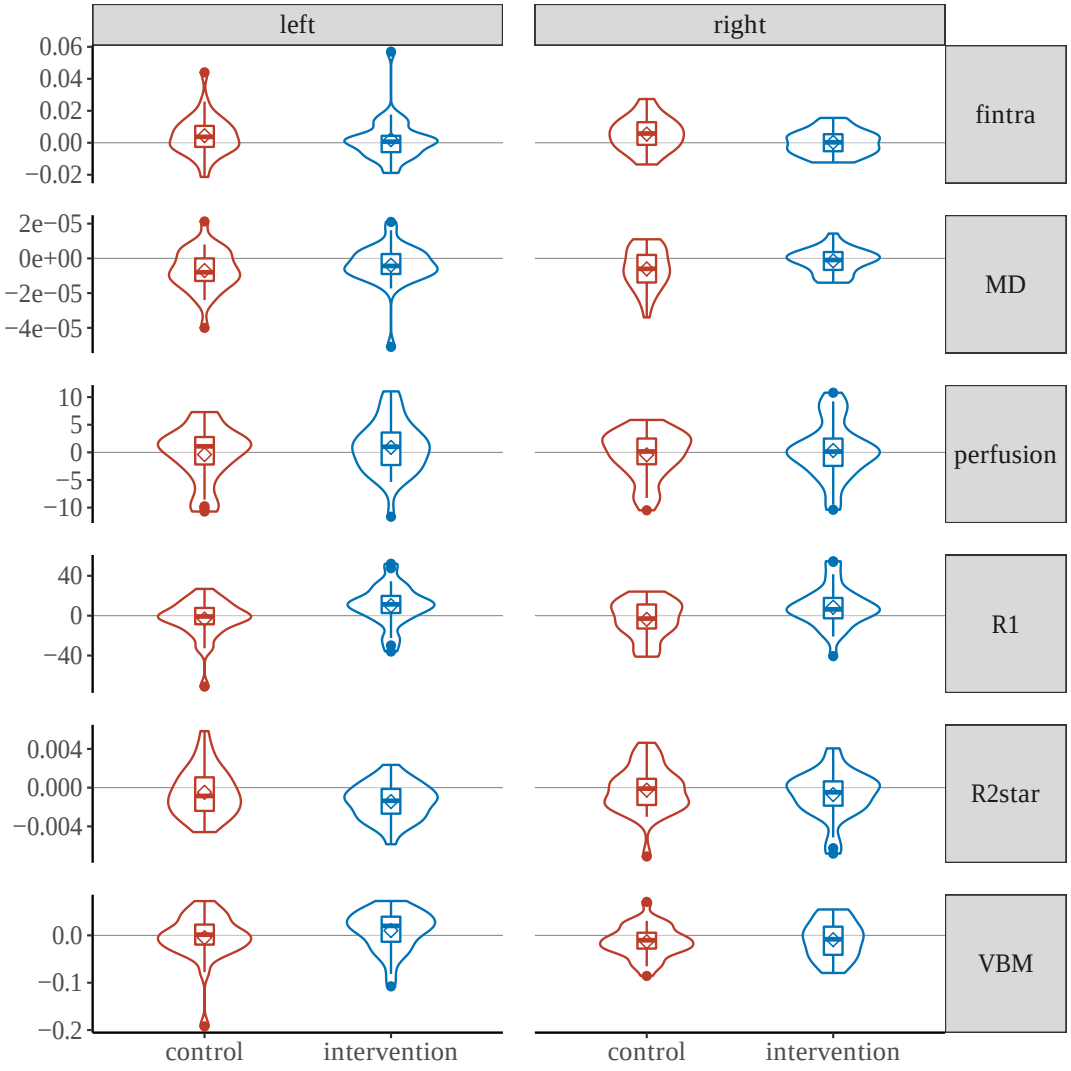


Figure 7.6: Distributions of change scores (post-pre) of MRI markers in the anterior hippocampus by treatment group and hemisphere. Average values for each MRI marker per individual were computed across the 50 multiply imputed datasets. Abbreviations: fintra = intra-cellular volume fraction, MD = mean diffusivity, VBM = voxel-based morphometry

As a sensitivity analysis, we repeated the primary analysis in two ways: (1) using a single (SVD) imputation method that takes into account the hierarchical structure in the data, and (2) using a complete-case (N = 38) dataset. Both methods showed no significant joint effect of the intervention on MRI markers (SVD: Fisher's NPC FWE-p = 0.11; complete-case: Fisher's NPC FWE-p = 0.32). Details are provided in Appendix G.2.3.1.

While the the significant effect of the intervention on R_1 needs to be interpreted with caution in light of the sensitivity analyses, we decided to assess its relevance by assessing the relationship of changes in R_1 with behavioural metrics. In particular, we assessed whether the association between changes in R_1 with CRF or relational memory performance was moderated by treatment status. Given that the multiply imputed data did not contain CRF or relational memory variables (due to the dataset being too small to impute all variables in a single imputation run) nor any interactions with treatment, we ran these analyses in the SVD imputed data only using within-treatment group partial-correlations, followed by r-to-Z transformation and a Z-test to compare r-transformed values. There was no significant difference in associations (r-values) between changes (post-pre) in left or right R_1, and changes in CRF or relational memory performance (all p >0.05, details in Appendix G.2.3.1). Moreover, in the intervention group only, associations between changes in R_1 and changes in CRF or relational memory were not significant either (p >0.05).

7.3.4.3 Exploratory subgroup analysis

Model estimates are presented in Table 7.3. Sex, socioeconomic status or baseline CRF levels did not significantly moderate the effect of the intervention on left and right anterior hippocampal volume (all uncorrected-p >0.05).

Table 7.3: Subgroup analyses assessing whether the effect of the intervention on left and right anterior hippocampal is moderatod by sex, socioeconomic status and baseline CRF.

Moderator	N	β	95% CI	p
Left anterior hippocampus				
sex (reference = female)	47	-42.28	-117.66, 33.1	0.26
SES	47	14.13	-6.99, 35.24	0.18
baseline CRF	47	-0.56	-5.34, 4.22	0.81
Right anterior hippocampus				
sex (reference = female)	47	-13.87	-89.23, 61.49	0.71
SES	47	-2.61	-23.4, 18.17	0.8
baseline CRF	47	-3.96	-8.55, 0.62	0.09

CI = confidence interval, CRF = cardiorespiratory fitness, SES = socioeconomic status (IMD deciles)

We also assessed whether sex, socioeconomic status or baseline CRF moderated the effect of the intervention on MRI markers in the anterior hippocampus. Neither sex (p = 0.77), nor socioeconomic status (p = 0.82) or baseline CRF (p = 0.91) was found to be a significant moderator.

7.3.5 The effect on relational memory

The effect of the intervention on relational memory was not significantly different between the intervention and control group at post test (OLT: UMD = 2.79, 95% CI: -1.78, 7.37, p = 0.23; relational memory paradigm: UMD = 2.86, 95% CI: -5.58, 11.40 p = 0.51).

Sensitivity analysis using complete-cases was in line with these findings, showing no significant effect of the intervention on relational memory performance (OLT: UMD = 2.8, 95% CI: -4.0, 9.61, p = 0.48; relational memory paradigm: UMD = 0.63, 95% CI: -6.81, 8.08, p = 0.86).

7.3.6 Changes in CRF with changes in hippocampal MRI markers

Relationships between a change in CRF and changes in left or right anterior hippocampal volume were not significantly moderated by intervention group (Table 7.4). In

addition, changes in CRF were not significantly associated with either left (β_{CRF} = -0.96, p = 0.63) or right (β_{CRF} = -3.16, p = 0.12) anterior hippocampal volume.

Table 7.4: Relationship between a change in CRF with a change in left and right anterior hippocampal volumes.

Predictor[a]	Left anterior hippocampal volume (N = 42)			Right anterior hippocampal volume (N = 42)		
	β^b	95% CI	p	β^b	95% CI	p
$\dot{\mathbf{V}}O_2$peak × group	0.1	-8.39, 8.59	0.98	0.51	-7.94, 8.97	0.90
$\dot{\mathbf{V}}O_2$peak	-0.96	-4.95, 3.03	0.63	-3.16	-7.13, 0.81	0.12

CI = confidence interval
[a] Predictors assessed in separately models. That is, a single model was fit to assess $\dot{\mathbf{V}}O_2$peak×group. The main effect of change in CRF was assessed in a separate model. All models additionally included sex, socioeconomic status and change in pubertal status (post - pre) as covariates.
[b] Unstandardized values

In addition, we assessed whether a change in CRF was related to a change in MRI markers in the anterior hippocampus, and whether this relationship was moderated by treatment status. In line with the univariate analyses, group allocation did not significantly moderate relationships between changes in CRF and MRI markers (N = 37, Fisher's NPC FWE-p = 0.4), and no significant association between change in CRF and change in MRI markers was observed (Fisher's NPC FWE-p = 0.99).

7.3.7 Supplementary analyses

The majority of studies that have examined the effect of PA interventions on the hippocampus have focused on the entire hippocampus (Firth *et al.,* 2018), and, with the advent of freely available processing pipelines, have started to explore its subfields (Frodl *et al.,* 2019; Maass *et al.,* 2015; Wagner *et al.,* 2015). To allow for a comparison with such studies, we additionally explored these regions of interest, using all available cases.

In line with the findings for the anterior hippocampus, no significant difference in left or right hippocampal volume (left p = 0.56, right p = 0.74), nor in any of the subfields (Bonferroni corrected p >0.05), was observed between the intervention and control group at posttest (p >0.05). Analyses with uncorrected p-values showed a higher

volume of the right CA1 (p = 0.05) and lower volume of CA2-CA3 (p = 0.02) in the intervention group compared to control. Details are provided in Appendix G.2.5.1 and Appendix G.2.5.2.

7.4 Discussion

We assessed the effect of a one-academic year HIT-style VPA intervention delivered during PE on CRF, the anterior hippocampus and hippocampal-dependent memory performance in young adolescents using a combination of univariate and multivariate analyses. In line with findings from the main trial (chapter 6), the intervention did not significantly improve CRF, anterior hippocampal volume or relational memory performance in the intervention group relative to control at posttest. Moreover, the effect of the intervention on anterior hippocampal volume was not significantly moderated by sex, socioeconomic status or baseline CRF, and changes in CRF were not significantly related to changes in anterior hippocampal volumes with the intervention. We observed a significant joint effect of the intervention on MRI markers, which was driven by intervention-related increases in R_1, but not supported by sensitivity analyses in single-imputed and complete-case data.

7.4.1 Anterior hippocampus, CRF and cognitive function

Several experimental studies in young (Wagner *et al.,* 2015), middle-aged (Thomas *et al.,* 2016) and older adults (Erickson *et al.,* 2011, but see Maass *et al.,* 2015; Sexton *et al.,* 2020) have demonstrated PA-related changes in anterior hippocampal volumes. In children and adolescents, however, randomised controlled trials (RCTs) have shown PA-related changes in brain activity, connectivity and white matter microstructure, but no RCTs have assessed the effect of PA interventions on (anterior) hippocampal volume to date. Current evidence in young people is primarily based on cross-sectional studies that have shown inconsistent associations between PA and hippocampal volumes (e.g. chapter 5, and Gorham *et al.,* 2019; Migueles *et al.,* 2020; Ruotsalainen *et al.,* 2019) and a single quasi-experimental study involving 79 young adolescents (n = 40 intervention) that found no effect of doubling the amount of PA (adding two weekly classes of 30–45

min) on left or right hippocampal volume (Bunketorp Käll *et al.,* 2015). Our study adds to the sparse evidence base by showing no significant effect of a one-year HIT-style VPA intervention on anterior hippocampal volumes, and no significant moderation by sex, socioeconomic status or baseline CRF levels.

While the most striking effects of PA interventions in humans have been in the anterior portion of the hippocampus (Erickson *et al.,* 2011; Sagi *et al.,* 2012; Thomas *et al.,* 2016; Wagner *et al.,* 2015, but see Maass *et al.,* 2015; Sexton *et al.,* 2020), there is currently no gold-standard approach for delineating this ill-defined subregion. Previous studies showing positive effects of PA on the anterior hippocampus have used centre-of-gravity- (Erickson *et al.,* 2011) and percentage-based (e.g. anterior 25%, Thomas *et al.,* 2016) approaches. Despite being replicable, they are arbitrary and bear little resemblance to the underlying anatomy. In contrast, we used an internal landmark, the uncal apex, which provides a more accurate delineation of the anterior hippocampus and has therefore been recommended (Poppenk *et al.,* 2013). The results of a previous study using this landmark (albeit in MNI-space) align with the findings presented in this study, showing no significant effect of a PA intervention on anterior hippocampal volume (but in older adults, Sexton *et al.,* 2020). Future studies are encouraged to agree on the anterior hippocampal borders to reduce between-study heterogeneity and to exclude possible segmentation effects on study findings.

An alternative approach, taken by several studies and informed by robust findings of exercise-induced upregulation of neurogenesis in the dentate gyrus (DG) in rodents (Van Praag, 2005; Van Praag *et al.,* 1999; Vivar & Van Praag, 2017), is to examine PA-related changes in volumes of hippocampal subfields. The subregions of the hippocampus have specific cell types, functions, and developmental trajectories (Benoy *et al.,* 2018; Gilbert *et al.,* 2001; Tamnes *et al.,* 2018b; Van Der Meer *et al.,* 2020; Van Strien *et al.,* 2009; Voss *et al.,* 2013). Previous studies have shown mixed evidence that the DG (Frodl *et al.,* 2019; Pereira *et al.,* 2007) or cornu ammonis (CA, Rosano *et al.,* 2017), especially CA1 (Den Ouden *et al.,* 2018), may be sensitive to the effects of PA. We therefore decided,

a posteriori, to additionally examine subfield volumes, showing increased right CA1 and lower right CA2-CA3 volumes in the intervention group compared to control. These findings should, however, be interpreted with caution, given that (1) none of the analyses were significant following appropriate control of multiple comparisons, and (2) it is debated whether $1 \times 1 \times 1$ mm^3 MRI scans provide valid segmentations of hippocampal subregions (Wisse *et al.,* 2020). Future studies are suggested to acquire high-resolution T_1-weighted scans (e.g. 7T) to assess PA effects on hippocampal subfields.

While volumetric measures have shown to be sensitive to PA, they provide little insight into the underlying neurobiological mechanisms. We therefore leveraged the multi-modal dataset acquired in this study and found a significant joint effect of the intervention on MRI (micro)structural markers. Inspection of this finding showed a PA-related increase in R_1 in the intervention group, relative to control. R_1, though poorly validated in grey matter, is affected by the macromolecular content of a voxel, and is suggested to reflect, predominantly, variation in myelin, though iron contributes too (Edwards *et al.,* 2018; Stüber *et al.,* 2014). Given that we only observed an effect on this metric, and not, as previously reported, measures of tissue density (e.g. MD, Kleemeyer *et al.,* 2016) or perfusion (Maass *et al.,* 2015), its increase may reflect enhanced myelination, corroborating findings from a previous study indicating PA-related increases in myelination and not vasculature (Thomas *et al.,* 2016). Crucially, however, we could not replicate this finding in sensitivity analyses and the R_1 increase was only observed when uncorrected p-values of partial tests were inspected. Hence, this result needs to be interpreted with caution and requires replication in future studies.

In line with the findings from the Fit to Study trial (chapter 6), but in contrast to previous studies (Costigan *et al.,* 2015; Costigan *et al.,* 2016; Davis *et al.,* 2011; Jeon & Ha, 2017; Leahy *et al.,* 2020; Logan *et al.,* 2014; Moreau *et al.,* 2017; Owens *et al.,* 2017), we found no significant effect of the HIT-style VPA intervention on CRF or cognitive outcomes. The memory assessments used in this study have previously shown associations with CRF in children (Chaddock *et al.,* 2010b) and hippocampal volumes in children

and older adults carrying a mutation for familial Alzheimer's disease (Chaddock *et al.,* 2010b; Liang *et al.,* 2016), though these assessments had not been used in RCTs before and low sensitivity or test-retest reliability cannot be excluded.

7.4.2 Findings in context

Consistent with the findings of the Fit to Study main trial (chapter 6), we did not observe robust significant effects of the intervention on our primary outcome measures. Major factors impacting on the integrity of the Fit to Study to trial —the majority of which were related to the scale and intervention approach —have been discussed in the previous chapter. These equally apply to the current study, given that the brain imaging sub-study was embedded in the Fit to Study trial, and are briefly touched on below.

First, according to the cardiovascular fitness hypothesis, which posits that changes in CRF may precede (or be necessary for) changes in (neurocognitive) health (Åberg *et al.,* 2009; North *et al.,* 1990; Tuvey *et al.,* 2019; Voss *et al.,* 2016), the lack of significant effects in neurocognitive outcomes may not be surprising in the absence of a PA-related change in CRF. A pre-to-post change in CRF, however, was observed across treatment groups. In previous studies involving older adults, changes in CRF have been related to changes in hippocampal volumes (Maass *et al.,* 2015), as well as changes in hippocampal perfusion (Maass *et al.,* 2015) and measures of tissue density (Kleemeyer *et al.,* 2016). In our study, changes in CRF were not significantly related to pre-to-post differences in anterior hippocampal volumes or MRI markers within the anterior hippocampus.

Second, poor intervention fidelity may offer an alternative explanation for the lack of the trial's effectiveness. Of five intervention schools, only three provided information about intervention adherence, of which only one school reported that they implemented the intervention in over 75% of PE classes. While objective measures of VPA acquired during a single lesson suggest intervention schools spent more PE lesson time in VPA compared to control schools (6.92 vs. 4.70 minutes), this difference

was not statistically significant. Moreover, pupil-level measures suggested that both intervention and control schools delivered the intervention elements, and found no significant different in daily average minutes of VPA during school hours (intervention = 2.67 vs. control = 1.77 min/school day). Without a clear change in exposure level, no changes in outcomes are to be expected.

Third, the Fit to Study intervention consisted of a potential total weekly dose of 20 min of HIT-style VPA, which is lower compared to previous studies (Costigan *et al.,* 2015; Costigan *et al.,* 2016; Leahy *et al.,* 2020). It may be that the actual dose of HIT was too low or fragmented to observe clear effects, though low intervention compliance renders strong statements about dose tentative.

7.4.3 Strengths and limitations

Strengths of this study include its RCT design, allowing for causal conclusions to be drawn, its HIT-style VPA intervention, which is highly scalable and easily incorporated into regular school PE lessons, and its rich set of MRI measures, which are not commonly collected in PA intervention studies and have the potential to inform, albeit indirectly, about neurobiological mechanisms underlying the typically reported PA-related changes in tissue macrostructure (e.g. hippocampal volumes, Erickson *et al.,* 2011).

The study also has various limitations. First, poor measurement of intervention fidelity and large amounts of missing data render conclusions tentative. While seven-day actigraphy was added to the school-level and pupil-level self-report fidelity measures collected as part of Fit to Study (main) trial to obtain participant-specific adherence data, it lacks insight into whether individuals reached the required intensity and was only collected once during the intervention period, providing an unrepresentative estimate of minutes spent in VPA. It is crucial that future studies put in place a variety of measures to capture intervention fidelity. Second, the sample was more fit compared

to the Fit to Study trial population, potentially reducing the generalisability of the findings. Third, due to the size of the sample, it was only possible to appropriately deal with missing data for the primary analyses, hence, secondary analyses ought to be interpreted with caution. Fourth, following the non-significant effects in the anterior hippocampus, we set out to additionally examine the hippocampal subregions. Given that these were not planned for, albeit theoretically informed (Frodl *et al.,* 2019; Pereira *et al.,* 2007), the results should be treated as hypothesis generating rather than hypothesis testing. Finally, no active control group was used in this efficacy trial, and hence, no conclusions regarding the optimal dose can be drawn.

7.5 Conclusion

The one-academic-year HIT-style VPA intervention had no significant effect on CRF, anterior hippocampal volume or memory performance in young adolescents. An intervention-related increase in R_1 was observed, possibly indicating that new tissue, such as myelin, was formed. However, lack of replication in sensitivity analyses, high amounts of missing data and poorly measured intervention fidelity render conclusions tentative. Future well-controlled trials testing for the effects of HIT-style interventions on brain structure and function are warranted.

8

Conclusions and future directions

Contents

8.1 Aims . . . 234
8.2 Summary of findings . . . 235
8.3 Limitations and future directions . . . 236
8.3.1 Physical activity measures . . . 236
8.3.2 Hippocampal measures, hippocampal plasticity and beyond . 237
8.3.3 Cognitive measures . . . 238
8.3.4 Challenges of large-scale RCTs . . . 239
8.3.5 Designing physical activity interventions . . . 241
8.3.6 Moderators and mediators . . . 242
8.3.7 More than physical activity . . . 243
8.4 General conclusion . . . 244

This thesis aimed to investigate relationships between physical activity, cardiorespiratory fitness, the anterior hippocampus and cognitive performance in young adolescents. In this chapter, I will outline and synthesise the major findings presented in the experimental chapters. I will then briefly discuss limitations and present possible directions for future research.

8.1 Aims

The overall aim of this thesis was to examine relationships between measures of physical activity (PA) and cardiorespiratory fitness (CRF) and markers of cognition and anterior hippocampal structure in young adolescents. This work was motivated by a number of factors: (1) across studies in animals and human adults, higher levels of PA and CRF have been related to superior neurocognitive health (Hillman *et al.,* 2008; Vivar & Van Praag, 2017; Voss *et al.,* 2013), yet whether such relationships exist in adolescents is unclear, (2) adolescence has been described as a (neurobiological) critical or sensitive period of development characterised by enhanced plasticity, during which experience or environmental influences strongly affect the brain and may have long-lasting effects on behaviour (Fuhrmann *et al.,* 2015; Larsen & Luna, 2018); this period in life therefore presents an opportunity for interventions to positively impact on short- and long-term health outcomes, (3) the prevalence of physical inactivity is high among adolescents (Guthold *et al.,* 2020; Hallal *et al.,* 2012), which negatively impacts short- and long term health and places an economic burden on society (Ding *et al.,* 2016); even minor changes in PA behaviour may yield significant neurocognitive benefits, and (4) PA is easily modifiable, highly scalable and cost-effective, making it a suitable intervention to positively change public health.

The studies in this thesis used a variety of designs, neuroimaging techniques, behavioural assessments and statistical approaches to assess: (1) whether PA and CRF are associated with cognitive performance in young adolescents (chapter 4), (2) whether PA and CRF are related with MRI markers of anterior hippocampal structure (chapter 5), and (3) whether increasing PA, by means of simple and scalable HIT-style intervention

incorporated into regular school PE lessons, would improve cognitive performance (chapter 6) and benefit hippocampal MRI outcomes (chapter 7).

8.2 Summary of findings

The evidence for the effects of PA on cognitive health in young people, particularly adolescents, has been inconsistent (chapter 1), systematic reviews have been of low quality and there has been a lack of brain imaging studies in this age group. In addition, no randomised controlled trials (RCTs) had assessed the effect of PA interventions on hippocampal volumes in young people, despite converging evidence from research in animals and humans showing PA-induced changes in this region (Erickson *et al.,* 2011; Firth *et al.,* 2018; Van Praag, 2005; Vivar & Van Praag, 2017; Voss *et al.,* 2013). A clear need for high-quality, well-sampled RCTs involving adolescents and brain imaging was identified.

Several of these gaps were addressed using data from the Fit to Study cluster RCT, described in chapter 2, which aimed to improve maths attainment, mental health, CRF, cognitive performance and brain health by means of a vigorous PA intervention delivered during school PE.

In particular, chapter 4 demonstrated that PA and CRF are positively associated with a range of domain-specific cognitive outcomes, but that relationships are small and explain less than 10% of the variance in cognitive performance. Sex moderated the association between CRF and inhibitory control and processing speed, with stronger relationships in girls than boys. We additionally showed that the choice of predictors, outcomes and nuisance variables may impact on observed associations, and that a physically active lifestyle is associated with more favourable cognitive outcomes. These findings lend support for previous cross-sectional studies reporting that higher PA and CRF are associated with more favourable cognitive outcomes, yet do not provide insight into the neural correlates of a physically active lifestyle.

Baseline data from the Fit to Study brain imaging sub-study was used to investigate associations with the anterior hippocampus, a region that has shown to be sensitive to the effects of PA in studies involving animals and humans older adults. Chapter 5 observed no significant relationships between either PA or CRF and anterior hippocampal volume or hippocampal-dependent memory performance. We additionally observed no significant association between a physically active lifestyle and MRI markers of tissue (micro)structure in the anterior hippocampus in this cohort of young adolescents.

The cross-sectional nature of the studies presented in chapter 4 and chapter 5 limits conclusions regarding the directionality of relationships involving PA and CRF. Simply detecting correlates of neurocognitive health will not be sufficient evidence to change policy; to this end, a causal relationship between PA and health outcomes needs to be established. To examine whether increasing the amount of VPA could improve neurocognitive outcomes, the longitudinal data of Fit to Study was used. Chapter 6 demonstrated that the Fit to Study intervention did not significantly improve CRF or cognitive performance, but that these findings should be interpreted with caution given low implementation fidelity and high drop-out. In chapter 7 it was shown that the intervention had no significant effect on anterior hippocampal volume, CRF or memory outcomes, but a significant effect on MRI markers within the anterior hippocampus (p = 0.02) was observed. This effect which was driven by a PA related increase in R_1 in the intervention group. Poor measurement of intervention fidelity and lack of replication in sensitivity analyses, however, suggested that this finding needs to be interpreted with great caution.

8.3 Limitations and future directions

8.3.1 Physical activity measures

The self-report (chapter 4) and objective measures (i.e. daily average MVPA, chapter 5) of PA employed in this thesis are typically used in the field and easily interpretable, but not without their limitations. Self-report measures are easy to acquire, particularly in large-scale studies like Fit to Study, but are subject to social desirability and recall

bias, especially in children (Welk *et al.,* 2000). Actigraphy-derived PA estimates avoid such biases, but are severely impacted by the processing pipeline, e.g. the selection of intensity thresholds (Migueles *et al.,* 2017; Orme *et al.,* 2014; Rowlands *et al.,* 2018b). Moreover, the rich actigraphy data is typically reduced to a single estimate of total daily minutes of MVPA. While MVPA has shown to be relevant to health (Carson *et al.,* 2016; Chinapaw *et al.,* 2018; Grgic *et al.,* 2018), it reflects only a small part of one's PA profile. Recent evidence, for instance, suggests that vigorous PA, rather than MVPA, may yield the greatest benefits to health, though whether it should be accumulated in bouts is unclear (Pindus *et al.,* 2019; Tarp *et al.,* 2018). To better capture the multidimensional nature of PA, future studies would be encouraged to explore novel ways of expressing PA behaviour, such as standardized accelerometer metrics to describe the PA profile (Fairclough *et al.,* 2020; Rowlands *et al.,* 2018a; Rowlands *et al.,* 2018b; Thompson *et al.,* 2015), and use more advanced analyses techniques, such as the canonical correlation analysis (chapter 4 and chapter 5), compositional data analysis (Gupta *et al.,* 2018) and multivariate pattern analysis (Aadland *et al.,* 2018; Aadland *et al.,* 2019).

8.3.2 Hippocampal measures, hippocampal plasticity and beyond

Hippocampal volume, obtained using manual or automated procedures, is the most common outcome measure in studies involving PA (Stillman *et al.,* 2020) and has proven to be an important (bio)marker of health and disease (De Flores *et al.,* 2015; Fotuhi *et al.,* 2012; Jack *et al.,* 2012), though is unspecific to the underlying neurobiology. We therefore leveraged the multi-modal dataset acquired as part of the Fit to Study brain imaging sub-study to investigate micro-scale processes that may underlie volumetric changes. Microstructural MRI measures, as well as perfusion, may provide biologically relevant information, particularly when their information is combined (Cercignani & Bouyagoub, 2018; Tardif *et al.,* 2016), yet this has rarely happened in practice (but see e.g., Den Ouden *et al.,* 2018; Thomas *et al.,* 2016). Despite some studies showing PA- or CRF-related changes in perfusion (Maass *et al.,* 2015), tissue density (Kleemeyer *et al.,* 2016) and myelination (Thomas *et al.,* 2016), this thesis found no robust evidence for changes in MRI markers following the intervention. The mechanism and nature of

the typically observed volumetric changes is therefore still unclear. Future interventional studies that employ a careful selection of multi-modal MRI sequences acquired at high field strength (e.g. 7T) are needed to improve our understanding of exercise-induced neuroplasticity. Developments in myelin (Heath *et al.,* 2018; Mancini *et al.,* 2020; Piredda *et al.,* 2021) and vasculature imaging (Berry *et al.,* 2019; Jezzard *et al.,* 2017; Park *et al.,* 2018; Spallazzi *et al.,* 2019), and the addition of brain metabolite sequences (i.e. MR spectroscopy, e.g. Pajonk *et al.,* 2010; Wagner *et al.,* 2015), may prove crucial in this endeavour.

Rather than focussing on the hippocampus, PA researchers are encouraged to explore the whole brain. In a cross-sectional study involving Fit to Study data, for instance, we showed that a physically active phenotype was associated with measures of brain structure and function across the brain (Salvan *et al.,* 2020). Early intervention studies in children, but particularly in adults, further demonstrate PA-related changes in white matter structure, brain activity and functional connectivity in a wide range of regions, but especially frontal, temporal and parietal areas (e.g. Batouli & Saba, 2017; Erickson *et al.,* 2015; Esteban-Cornejo *et al.,* 2018; Sexton *et al.,* 2016; Stillman *et al.,* 2019; Voelcker-Rehage & Niemann, 2013). This literature is, however, still in its infancy, with large between-study heterogeneity precluding strong conclusions about the role of such brain areas in exercise-mediated cognitive changes (Stillman *et al.,* 2019).

8.3.3 Cognitive measures

A comprehensive battery of computerised cognitive assessments was used to assess a range of cognitive abilities that have shown to be sensitive to PA and CRF: inhibitory control (modified Flanker task, Hillman *et al.,* 2014), cognitive flexibility (colour-shape switching task Hillman *et al.,* 2014), working memory (visual two-back task, Chen *et al.,* 2014b), processing speed (simple reaction time task) and visual relational memory (Chaddock *et al.,* 2010b). The breadth of assessments allowed for an investigation of multiple cognitive domains, which is not typically done in the PA field, and their computerised administration assured precise control over stimuli and accurate recordings

of multiple responses (i.e. speed and accuracy).

Given that these assessments were used in previous PA studies involving children and adolescents, we did not examine their psychometric properties (e.g. reliability and validity) prior to the start of the Fit to Study trial. Despite their widespread usage, their reliability is not typically assessed or reported (Cook & Beckman, 2006; Nyongesa *et al.,* 2019; Wade *et al.,* 2020). What's more, the psychometric properties of these assessments in the population of young adolescents is either unclear or below acceptable values (an extensive discussion is provided in Appendix C.3). This is problematic, given that such properties, but reliability in particular, have implications for the design, analysis and interpretation of studies (Draheim *et al.,* 2019; Hedge *et al.,* 2018a; Paap & Oliver, 2016; Rouder *et al.,* 2019). In correlational research, for instance, the observed correlation value is a function of the reliabilities of the individual measures (Hedge *et al.,* 2018a; Spearman, 1910), with low reliabilities reducing observed correlations. It is crucial that future studies assess and report the psychometric properties in an independent, but demographically comparable sample, prior to the start of the study.

In addition, future studies are encouraged to obtain multiple measures of a single cognitive domain to tackle the task impurity problem (for an example, see Moreau *et al.,* 2017). This was not feasible in Fit to Study due time constraints and may have affected the specificity of the measures used throughout this thesis. Latent variable analysis was used in chapter 6 to create an executive function (EF) construct and (partially) address the task impurity problem, but at the expense of domain-specific information.

8.3.4 Challenges of large-scale RCTs

Most of the limitations mentioned throughout the experimental chapters of this thesis (chapter 4 to chapter 7) are directly or indirectly related to the scale of the Fit to Study trial. The scale was required to adequately power the school-based trial to detect small treatment effects (in the primary outcome) and to obtain a representative sample of young adolescents, yet forced pragmatic decisions regarding the choice, delivery and

collection of outcome measures, intervention design and delivery (e.g. required to be incorporated into school PE, delivered by PE teachers etc.) and measurement of intervention fidelity, and resulted in less control over the intervention (e.g. training, implementation), large amounts of missing data and findings that are complex to be interpreted. That is, the null results presented in chapter 6 and chapter 7 may reflect that (1) the intervention dose was not meeting the required levels, (2) the dose was not received, and/or (3) the specified dose of PA does not improve neurocognitive outcomes.

Despite being conceived as an explanatory, efficacy RCT that aimed to assess whether a vigorous PA intervention could improve academic attainment in a real-world but tightly controlled setting, Fit to Study may be better classified as a pragmatic RCT, which typically aim to maximise external validity (i.e. generalisability of the findings) and more closely resembles usual practice (Ali *et al.*, 2015; Dal-Ré *et al.*, 2018; Ford & Norrie, 2016; Packer & Pitt, 2018). The main trial outcomes may therefore be qualitative in nature: post-intervention surveys suggested that PE teachers would recommend the Fit to Study intervention as a way to promote PA, but struggled to consistently implement the intervention elements in their lessons (Husain *et al.*, 2019). Future trials would therefore need to re-evaluate the intervention design, in collaboration with teachers, to determine how theory-driven elements of vigorous PA can be embedded in PE without disrupting lesson structure.

Moreover, future large-scale exploratory RCTs involving PA interventions in schools are recommended to (1) perform extensive feasibility testing, preferably at different scales to examine the effect of scale-ups, (2) devote time to building relationships with schools and teachers to ensure a high level commitment, and (3) use technological developments, such as smart-phone apps, for the monitoring of intervention fidelity and assessment of outcomes. A thorough evaluation of challenges related to the design and implementation of Fit to Study, as well as recommendations for future trials, is provided in Wheatley *et al.* (2020b).

8.3.5 Designing physical activity interventions

Despite the null results presented in this thesis, there is increasing evidence that high intensity PA interventions, like Fit to Study, have the potential to positively impact on health outcomes in children and adolescents (Costigan *et al.,* 2015; Costigan *et al.,* 2016). It is currently unclear, however, what combination of dose, type of PA (i.e. PA characteristics) and setting yield the largest effects and for whom (e.g. by age, sex, fitness status). Some evidence suggests that running-based interventions with intensities >85%HR_{max}, two-to-three times per week, with longer high intensity intervals (∼4 min) and active recovery periods (of approx. 3 min), lasting more than 7 weeks may provide the optimal stimulus for health improvement (Calverley *et al.,* 2020; Eddolls *et al.,* 2017; Martland *et al.,* 2020; Weston *et al.,* 2014). Other studies suggest that 30 min of moderate intensity PA in children and 10-30 min of moderate-to-high intensity PA in adolescents may benefit cognitive outcomes (Williams *et al.,* 2018). At the same time, there is some evidence suggesting that cognitively engaging exercise interventions (e.g. team games) may yield greater cognitive benefits than those relying solely on physical (aerobic) exertion (Diamond & Ling, 2016; Pesce, 2012; Schmidt *et al.,* 2015; Tomporowski *et al.,* 2015). Moreover, class-room based PA has shown some benefits on academic outcomes (Watson *et al.,* 2017, but see Daly-smith *et al.,* 2018), though most studies have been conducted during school PE, which is where the greatest benefits have been reported (Álvarez-Bueno *et al.,* 2017a; Álvarez-Bueno *et al.,* 2017b). Clearly, a structured approach to designing and assessing intervention characteristics and settings is required, whereby individual elements (e.g. frequency, type, setting) are manipulated to examine their effects on neurocognitive outcomes.

In addition, it is well-known that individuals vary in their responsiveness to PA interventions (Gronwald *et al.,* 2020; Montero & Lundby, 2017; Pickering & Kiely, 2019). To maximise the efficiency of an intervention, the exercise prescription therefore needs to be tailored to an individual's characteristics. In most traditional PA interventions, a comparable external load (e.g. 10 min of running at 10 km/h) is applied across all individuals, which may result in a variable internal load (i.e. the physiological, motor,

biomechanical response, e.g. % of maximum HR or blood lactate concentration) and high inter-individual variability in the outcome (Herold *et al.,* 2019a). To reduce the heterogeneity in exercise responses and increase comparability between individuals, the internal load should be taken into account when prescribing and monitoring PA interventions (Herold *et al.,* 2020). The exact markers of internal load to be used, however, are debated and require further study (Gronwald *et al.,* 2020; Herold *et al.,* 2019a). In Fit to Study, the prescribed intensity involved a marker of internal load (% maximum HR), but was not monitored, hence, whether individuals had a comparable internal load is unclear. It is imperative that future studies closely monitor and report the received dose, to enhance reproducibility and to allow for an investigation of dose-dependent PA effects.

8.3.6 Moderators and mediators

To further our understanding of the relationship between PA and neurocognitive outcomes and maximise the effect of PA interventions, research into moderators and mediators is crucial (Bauman *et al.,* 2002; Erickson *et al.,* 2019; Leckie *et al.,* 2012; Singh *et al.,* 2019; Stenling *et al.,* 2020; Stillman & Erickson, 2018). Knowledge of modifiable (e.g. nutrition, Leckie *et al.,* 2012) and non-modifiable (e.g. genetics or sex, Barha *et al.,* 2017a; Barha *et al.,* 2017b; Prakash *et al.,* 2015) moderators may help explain the high interindividual variability in responsiveness to PA interventions and could be used to design targeted PA interventions. We therefore explored whether sex, socioeconomic status or baseline CRF levels moderated the effects of the intervention on neurocognitive outcomes, but observed no significant evidence. Further exploration of potential moderators (see e.g. Chang *et al.,* 2017; Leckie *et al.,* 2012; Stenling *et al.,* 2020; Stillman *et al.,* 2018), and their interactions (Barha *et al.,* 2017a), is needed.

Studies have also started to examine mediators of (or mechanisms underlying) the PA-cognition relationship, which can be examined at multiple levels (Esteban-Cornejo *et al.,* 2015; Stillman *et al.,* 2016). Most animal research has focussed on cellular and molecular mechanisms, whereas imaging studies in humans have begun to examine

brain (e.g. hippocampal volume, Erickson *et al.,* 2009) and behavioural (e.g. sleep, Wilckens *et al.,* 2018, or mood, Vance *et al.,* 2005) factors that mediate the effect of PA on cognitive outcomes. Fit to Study aimed to interrogate several of these mediators (e.g. CRF, the anterior hippocampus), but no mediation models were applied given the clear null results of the PA intervention. More research on mediators is therefore needed. Structural equation models applied to data from carefully designed RCTs may prove useful in this regard (for an example, see: Eshaghi *et al.,* 2019).

8.3.7 More than physical activity

Most public health reports and research studies, including this thesis, have isolated behaviours, such as PA, to study their effect on neurocognitive health. However, PA —and particularly moderate-to-vigorous PA (MVPA) —only makes up a small proportion of the 24-h day and does not operate in isolation. Other behaviours along the 24-hour movement continuum, such as sleep and sedentary time, interact with PA (Lang *et al.,* 2016) and may independently and jointly affect health outcomes (De Bruin *et al.,* 2017; Saunders *et al.,* 2016; Tremblay *et al.,* 2016a). It is increasingly recognised that a more integrated, holistic approach is required to study how the combination of movement behaviours influence health (Chaput *et al.,* 2014; Tremblay *et al.,* 2016a; Tremblay *et al.,* 2016b).

In response, the 24-hour movement guidelines for children and youth have been established that aim to promote sleep and PA, and reduce sedentary and screen time (Tremblay *et al.,* 2016a). Initial cross-sectional studies have shown that meeting multiple movement recommendations is associated with superior health outcomes (Rollo *et al.,* 2020; Saunders *et al.,* 2016), including higher health related quality of life (Khan *et al.,* 2020; Sampasa-Kanyinga *et al.,* 2017) and better cognitive performance (Walsh *et al.,* 2018), and that the composition of movement behaviours, favouring MVPA, is related with a wide range of health benefits (Carson *et al.,* 2016; Grgic *et al.,* 2018). Moreover, reallocating time from sleep and light PA to MVPA has been related with improvements in EFs in preschoolers (Bezerra *et al.,* 2020), and substitution of sedentary

time with MVPA or sleep has been associated with improvements in EFs in older adults (Fanning *et al.,* 2017; Wei *et al.,* 2020). Future observational and interventional studies may extend this work to assess the joint effect of multiple lifestyle factors (e.g. the movement behaviours, diet; Nyaradi *et al.,* 2013) on brain health in adolescents. Moreover, using multivariate (pattern) analysis (Aadland *et al.,* 2018; Aadland *et al.,* 2019) or compositional and isotemporal substitution analysis (for details see e.g. Bezerra *et al.,* 2020; Dumuid *et al.,* 2019; Grgic *et al.,* 2018; Mota *et al.,* 2020), for instance, they may gain insight into, respectively, the most appropriate movement behaviour or composition of movement behaviours to improve neurocognitive functioning in adolescents.

8.4 General conclusion

In summary, the work in this thesis demonstrates that: (1) the evidence base for the effects of PA on neurocognitive health in adolescents is sparse; (2) higher levels of PA and CRF are consistently associated with superior cognitive outcomes; (3) the relationship between PA and CRF and volume or MRI markers in the anterior hippocampus is unclear; (4) whether a vigorous PA intervention delivered during school PE improves CRF and neurocognitive health requires further investigation.

Appendices

A

Supplementary materials to chapter 1: *Introduction*

Contents

A.1 Terminology 248
A.2 Methods systematic review of reviews 252
A.2.1 Search strategy and selection criteria 253
A.2.2 Review selection and data extraction 254
A.2.3 Quality assessment 254

A.1 Terminology

"*The epidemiologic study of any concept or event requires that the item under investigation be defined and measured.*" Caspersen *et al.* (1985). More than for just epidemiological studies, any study benefits from clear terminology. This is of particular importance in the PA field, where terms such as *physical activity*, *fitness* and *exercise* are often confused and used interchangeably (Budde *et al.,* 2016; Hills *et al.,* 2014).

Physical activity PA is commonly defined as "*any bodily movement produced by skeletal muscles that results in energy expenditure*" (Caspersen *et al.,* 1985). This term is not specific to any type of PA and includes all intensities and domains. Therefore, one typically describes PA by four dimensions (2018 Physical Activity Guidelines Advisory Committee, 2018; Caspersen *et al.,* 1985; Hills *et al.,* 2014; Howley, 2001; Warren *et al.,* 2010):

- **frequency**: the number of episodes / events of PA during a specific period of time
- **duration**: time it takes to complete a single bout of PA
- **intensity**: the physiological effort associated with PA, which can be expressed in absolute or relative units.
  - In *absolute* terms, intensity refers to the rate of energy expenditure, commonly expressed in metabolic equivalents (METs), but also in kilocolories, joules or oxygen consumption. One MET is the resting energy expenditure (REE), or resting oxygen (O_2) consumption —estimated to be approximately 3.5 ml/kg/min in adults —which allows for other activities to be quantified in multiples of METs. Traditionally, METs have been used to classify activities by MET-intensities, with (1) vigorous intensity activities requiring >6 METs, (2) moderate intensity activities requiring between 3–6 METs, (3) light intensity activities requiring between 1.6–3 METs, and (4) sedentary activities

requiring between 1–1.5 METs. However, adult-based METs are not applicable to children, who have higher basal metabolic rates per unit body mass, which question the typical MET-based intensity thresholds (Harrell *et al.,* 2005; Saint-Maurice *et al.,* 2016). For instance, there is some consensus that the lower cut-point for moderate-intensity activities in children and adolescents should be 4 METs, rather than the adult-based 3 METs (Saint-Maurice *et al.,* 2016). Over the years, compendia containing MET values for different physical activities in youth have been developed (Butte *et al.,* 2018; Ridley *et al.,* 2008).

  - In *relative* terms, intensity refers to the (perceived) effort during the activity, which can be expressed in, for instance, a percentage of maximum oxygen consumption or maximum heart rate, or (subjective) perceived exertion. Cut-points for the various intensity categories vary (e.g. Howley, 2001; Norton *et al.,* 2010), but typically —and in this thesis —activities between 55%–70% of the maximum heart rate are considered to be moderate, and >70% to be vigorous (Norton *et al.,* 2010).

- **type or mode**: the type of PA. This can categorised in various, not mutually exclusive, ways, such as (1) by the actual activity under study, e.g. walking or swimming (Lambourne & Tomporowski, 2010), (2) by the metabolic and neuromuscular demands, including aerobic, anaerobic, muscle-strengthening, bone-strengthening, flexibility and balance activities (e.g. Fedewa & Ahn, 2011; Voss *et al.,* 2011b), or (3) by the complexity of the activities, including the coordinative and cognitive demands (e.g. Pesce, 2012; Tomporowski *et al.,* 2015).

The first three dimensions (frequency, duration and intensity) define the *dose* of PA. They are required to estimate energy expenditure, and accumulated over time determine the *volume* of PA. They are commonly referred to as the quantitative PA characteristics, in contrast to the type of PA, which is qualitative in nature (Tomporowski *et al.,* 2015). PA interventions that modify the quantitative characteristics are also referred to

as being *enhanced*, whereas those that (additionally) modify the qualitative characteristics are termed *enriched* or *cognitively demanding* (e.g. Álvarez-Bueno *et al.,* 2017a; Álvarez-Bueno *et al.,* 2017b). The field further distinguishes between *acute* and *chronic* physical activity. The former refers to short, single bouts of PA (typically between 10–40 min), whereas the latter refers to multiple PA sessions over longer periods of time (typically between 6–30 weeks; Rathore & Lom, 2017; Tomporowski *et al.,* 2011; Verburgh *et al.,* 2014). Note that the terms acute and chronic have occasionally been used to refer to the *effects* of physical activity, i.e. immediate versus long-lasting, respectively (Den Heijer *et al.,* 2017). This thesis will, however, refer to PA being acute or chronic and its effects being immediate or long-lasting.

Exercise Exercise is a subset of PA that is planned, structured and often repetitive, and is designed to improve or maintain physical fitness, performance or health (2018 Physical Activity Guidelines Advisory Committee, 2018; Caspersen *et al.,* 1985).

Physical fitness Physical fitness has been defined in various ways, including a "*physiologic attribute determining a person's ability to perform muscle-powered work*" (2018 Physical Activity Guidelines Advisory Committee, 2018), "*a physiological state of well-being that reduces the risk of hypokinetic disease, a basis for participation in sports, and good health, which enables one to complete the tasks of daily living*" (Donnelly *et al.,* 2016), or, perhaps easier to grasp, "*the ability to do moderate to vigorous activity without undue fatigue*" (Bradley *et al.,* 2019). It is a multidimensional construct comprising several domains, such as health-, skill-, or physiological fitness (Caspersen *et al.,* 1985; Corbin, 1991; Pate, 1988). Researchers most commonly refer to health-related domains of fitness, which is often subdivided into five measurable components (Caspersen *et al.,* 1985; Corbin, 1991; Pate, 1988): cardiorespiratory endurance, muscular endurance, muscular strength, flexibility, and body composition, although other subdivisions or components have been used too (e.g. 2018 Physical Activity Guidelines Advisory Committee, 2018; Donnelly *et al.,* 2016; Ruiz-Ariza *et al.,* 2017; Warren *et al.,* 2010). This thesis focuses on cardiorespiratory endurance, also referred to as cardiovascular fitness, aerobic fitness or cardiorespiratory fitness (with minor distinctions, see e.g. Corbin *et al.,* 2014), and

hereafter referred to as cardiorespiratory fitness or CRF, which is defined as: "*ability of the respiratory and circulatory systems to supply fuel during sustained physical activity and to eliminate fatigue products after supplying fuel*" (Caspersen *et al.,* 1985). It is a physiological trait (Lang & Little, 2018), that, unlike PA, does not change on a day-by-day basis (Warren *et al.,* 2010).

Cognition Cognition, from the Latin *cognoscere*, or "to know", refers to the faculty of knowing, but has, despite many attempts, no clear definition (Allen, 2017; Bayne *et al.,* 2019). Textbooks describe cognition in broad terms, often referring to a range of mental processes relating to the acquisition, storage, manipulation, and retrieval of information. For the purpose of this thesis, however, I will treat cognition as an umbrella term that encompasses various more tangible or precisely defined processes, such as memory, learning, and attention, and that is used primarily to define a field of investigation (Allen, 2017). In the exercise-neuroscience field, cognition commonly refers to more specific mental processes such as processing speed, memory and executive functions (Chang *et al.,* 2017; Donnelly *et al.,* 2016), that are different from academic abilities which typically encompass maths, reading and language.

Executive functions Executive functions (EF), also known as cognitive control or executive control, are an umbrella term referring to a set of interrelated cognitive processes that are thought to underlie goal-directed behavior (Best & Miller, 2010; Diamond, 2013; Miyake *et al.,* 2000, but see (Baggetta & Alexander, 2016; Karr *et al.,* 2018) for a discussion of definitions). These processes include, but are not limited to, planning, monitoring, verbal fluency, problem solving, working memory, inhibitory control and cognitive flexibility. To date, there is no consensus regarding the precise components of EF (Baggetta & Alexander, 2016; Packwood *et al.,* 2011), although scientists commonly distinguish three interrelated, but distinct core elements (Diamond, 2013; Miyake & Friedman, 2012; Miyake *et al.,* 2000): (1) working memory, (2) inhibitory control, (3) cognitive flexibility, that give rise to higher-order EFs, such as planning and problem solving. Other components, or models, of EFs have been proposed too, including the central executive system (Baddeley & Hitch, 1974), "hot" (processes driven by emotion)

and "cool" EFs (processes requiring logic and critical analyses, e.g. working memory; Poon, 2018; Zelazo & Carlson, 2012), the executive control system (Anderson, 2002), and various factor models of EFs, that distinguish one or multiple (sub) processes (Baggetta & Alexander, 2016; Friedman & Miyake, 2017; Friedman *et al.,* 2008; Karr *et al.,* 2018). The precise structure of EFs is an area of active investigation and a full account of all models is beyond the scope of this work. For the purpose of this thesis, I treat EFs as an umbrella term, broadly referring to working memory, cognitive flexibility, and inhibitory control.

Adolescence Adolescence, derived from the Latin *adolescere*, or "to grow up", is defined as the period of life between childhood and adulthood (Sawyer *et al.,* 2018). The concept and definition of adolescence has evolved over time, following changes in biological maturation (e.g. puberty starting earlier), as well as cultural and social role transitions (Blakemore, 2019; Fuhrmann *et al.,* 2015; Kinghorn *et al.,* 2018; Larsen & Luna, 2018; Patton *et al.,* 2018; Sawyer *et al.,* 2012; Sawyer *et al.,* 2018). Adolescence starts with puberty and ends with adult independence —the latter of which is particularly hard to define (Ledford, 2018) —and is characterised by dynamic neurobiological, social, and psychological development. Various chronological or age-based definitions of adolescence exist today (Ledford, 2018; Sawyer *et al.,* 2018). The most commonly used boundaries are the ages of 10 and 19 years, but these have recently been challenged (Sawyer *et al.,* 2018) and extended to span the period between 10 and 24 years of age. In this thesis, we will adhere to the updated age-based definition of adolescence (i.e. 10–24 years), and define the period between 10–15 years of age as *young* adolescence (Sawyer *et al.,* 2018).

Back to main text

A.2 Methods systematic review of reviews

The systematic review was registered at PROSPERO (CRD42019124472) and conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-

analysis (PRISMA) guidelines (checklist is provided in Wassenaar *et al.,* 2020).

A.2.1 Search strategy and selection criteria

An information specialist (NR) performed an electronic search of the following databases, originally on June 14th 2018, and updated on February 6th 2019: MedLine, Embase, PsycINFO, Cochrane Library, Web of Science and ERIC. No date restrictions were applied to the initial search. Reference lists of identified articles were examined for additional relevant articles. Data sources and (medical subject headings) search terms are provided in Wassenaar *et al.,* 2020. In line with the a priori defined selection criteria (PROSPERO, ID: CRD42019124472), systematic reviews with or without meta-analysis that were published in English with clearly defined inclusion criteria were included if they reported on the effects of chronic PA interventions (i.e. more than one PA session, over a set period), including randomised controlled trials (RCT), quasi-experimental studies, controlled and pre-post designs, on cognitive, academic and /or brain MRI outcomes in school-aged children or adolescents. Systematic reviews that contained a mixture of observational and interventional evidence, that reported on a single bout of acute PA (i.e. a single session) or cardiovascular fitness only, or included case reports, were excluded.

The in- and exclusion criteria were adapted post-protocol to exclude systematic reviews that also included observational studies because the findings from interventional and observational studies were generally combined in the evidence synthesis. It was deemed too subjective to extract conclusions regarding the effects of intervention studies only. Studies that distinguished between multicomponent (e.g. PA and diet) and PA only interventions were included, but only results from PA only interventions were considered in this systematic review.

Systematic reviews that also included findings of acute PA interventions were considered for inclusion only if they selectively reported on the effects of chronic PA interventions. Following data extraction, it became clear that no single systematic

review had aimed to include MRI studies only. Instead, two out of four systematic reviews included studies that used either MRI or EEG. We therefore decided to adapt our inclusion criteria to also include brain EEG markers as an outcome measure. Because the search criteria included the term MRI, we post-hoc searched the databases (October 1st, 2019) for systematic reviews that included EEG studies. No additional systematic reviews were found.

A.2.2 Review selection and data extraction

Following removal of duplicates, two reviewers (TW, WW) independently screened all titles and abstracts using *Abstrackr* (Rathbone *et al.,* 2015). Short-listed full-text reviews were then independently assessed using the in- and exclusion criteria. Any disagreements between the authors were discussed with two other authors (CS and CF) and resolved by consensus. A data extraction form was piloted independently by two authors (TW, WW) and adjusted to ensure it captured all relevant data, including: year of publication, type of review (systematic review with or without meta-analysis), review methodology (aim, in/exclusion criteria, number of studies included / excluded, database search, quality and bias assessment method), characteristics of included studies (design, number of included participants, type of participants and countries), outcome measures, reporting on intervention fidelity, limitations and recommendations. A single author extracted the data from selected studies using the data extraction template which were verified by a second author and disagreements were resolved by consensus.

A.2.3 Quality assessment

The methodological quality of all reviews was assessed using the updated AMSTAR checklist, AMSTAR-2 (Shea *et al.,* 2017). Two reviewers (TW, WW) independently scored the selected reviews. The quality of each review is reflected by an overall confidence rating, which is determined by an evaluation of non-critical and critical domains

(seven critical items: preregistration, literature search, justification for excluding studies, risk of bias assessment, appropriateness of meta-analytical methods, risk of bias in synthesis of results, publication bias). A lack of addressing one or multiple critical domains resulted in a, respectively, low or critically-low confidence rating. If no critical flaws were present, the presence of non-critical weaknesses determined whether the review received a high (no weaknesses) or moderate (one or more weaknesses) confidence rating. While it is not recommended to combine individual item ratings into an overall score (Shea *et al.,* 2017), we provided the total score to merely acknowledge that a gradient exists between reviews in their study quality.

Back to main text

This page is intentionally left blank.

B

Supplementary materials to chapter 2: *Protocol for a cluster randomised trial of a physical activity intervention in young adolescents: Fit to Study*

Contents

B.1 Sample 258
B.1.1 Drop-out post randomisation main trial schools 258
B.1.2 Representativeness brain imaging sub-sample 261
B.1.3 Recruitment brain imaging cohort 2 268
B.2 Measures 269
B.2.1 Main trial 269
B.2.2 Brain imaging sub-study 279
B.3 Sample size 289

B.1 Sample

B.1.1 Drop-out post randomisation main trial schools

Table B.1 presents characteristics of the schools that dropped-out immediately post-randomisation, but prior to baseline assessments. We examined whether the schools that dropped-out were significantly different from schools that remained part of the sample on key school-level and pupil-level characteristics.

Table B.1: Characteristics of schools and participants throughout the trial.

	Control		Intervention		Overall	
	Included	**Drop-out**	**Included**	**Drop-out**	**Included**	**Drop-out**
School level						
No. schools	47	5	46	6	93	11
Gender status						
Co-ed	38 (80.9%)	4 (80.0%)	37 (80.4%)	5 (83.3%)	75 (80.6%)	9 (81.8%)
Female	8 (17.0%)	0 (0%)	9 (19.6%)	1 (16.7%)	17 (18.3%)	1 (9.1%)
Male	1 (2.1%)	1 (20.0%)	0 (0%)	0 (0%)	1 (1.1%)	1 (9.1%)
Establishment type						
Academy - Sponsor Led	9 (19.1%)	0 (0%)	7 (15.2%)	2 (33.3%)	16 (17.2%)	2 (18.2%)
Academy Converter	1 (2.1%)	0 (0%)	2 (4.3%)	0 (0%)	3 (3.2%)	0 (0%)
Academy Converter - Main	15 (31.9%)	3 (60.0%)	16 (34.8%)	1 (16.7%)	31 (33.3%)	4 (36.4%)
Community School	7 (14.9%)	1 (20.0%)	8 (17.4%)	2 (33.3%)	15 (16.1%)	3 (27.3%)
Foundation School	5 (10.6%)	1 (20.0%)	5 (10.9%)	0 (0%)	10 (10.8%)	1 (9.1%)
Free School - Mainstream	2 (4.3%)	0 (0%)	3 (6.5%)	0 (0%)	5 (5.4%)	0 (0%)
Voluntary aided school	7 (14.9%)	0 (0%)	4 (8.7%)	1 (16.7%)	11 (11.8%)	1 (9.1%)
Voluntary controlled school	1 (2.1%)	0 (0%)	1 (2.2%)	0 (0%)	2 (2.2%)	0 (0%)
School type						
Comprehensive to 16	9 (19.1%)	2 (40.0%)	7 (15.2%)	3 (50.0%)	16 (17.2%)	5 (45.5%)
Comprehensive to 18	33 (70.2%)	3 (60.0%)	34 (73.9%)	3 (50.0%)	67 (72.0%)	6 (54.5%)
Secondary Modern	5 (10.6%)	0 (0%)	5 (10.9%)	0 (0%)	10 (10.8%)	0 (0%)
School size						
Large (>200)	13 (27.7%)	1 (20.0%)	14 (30.4%)	4 (66.7%)	27 (29.0%)	5 (45.5%)
Medium (100-200)	27 (57.4%)	4 (80.0%)	27 (58.7%)	2 (33.3%)	54 (58.1%)	6 (54.5%)
Small (<100)	7 (14.9%)	0 (0%)	5 (10.9%)	0 (0%)	12 (12.9%)	0 (0%)
No. of form groups						
Median (range), no.	7 (3–14)	6 (5–8)	6 (3–12)	8.5 (5–12)	6 (3–14)	7 (5–12)
Missing, no. (%)	1 (2.1%)	0 (0%)	1 (2.2%)	0 (0%)	2 (2.2%)	0 (0%)
Ofsted rating[b]						
Outstanding	14 (29.8%)	2 (40.0%)	8 (17.4%)	2 (33.3%)	22 (23.7%)	4 (36.4%)
Good	24 (51.1%)	1 (20.0%)	29 (63.0%)	2 (33.3%)	53 (57.0%)	3 (27.3%)
Requires Improvement	6 (12.8%)	1 (20.0%)	4 (8.7%)	1 (16.7%)	10 (10.8%)	2 (18.2%)
Serious Weaknesses	0 (0%)	0 (0%)	1 (2.2%)	0 (0%)	1 (1.1%)	0 (0%)
Special Measures	3 (6.4%)	1 (20.0%)	3 (6.5%)	1 (16.7%)	6 (6.5%)	2 (18.2%)
Missing, no. (%)	0 (0%)	0 (0%)	1 (2.2%)	0 (0%)	1 (1.1%)	0 (0%)
Socioeconomic status[a]						
eFSM, mean (SD), %	17.7 (12.9)	17.4 (13.9)	17.0 (9.00)	18.8 (5.04)	17.3 (11.1)	18.2 (9.53)
IMD, median (range), decile	6 (1–10)	8 (1–10)	5 (1–10)	4.5 (3–7)	5 (1–10)	5 (1–10)
School location[c]						
Rural town and fringe	4 (8.5%)	0 (0%)	5 (10.9%)	0 (0%)	9 (9.7%)	0 (0%)
Rural village	0 (0%)	1 (20.0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)
Urban city and town	20 (42.6%)	3 (60.0%)	19 (41.3%)	3 (50.0%)	39 (41.9%)	6 (54.5%)
Urban major conurbation	23 (48.9%)	1 (20.0%)	22 (47.8%)	3 (50.0%)	45 (48.4%)	4 (36.4%)
Pupil level						
No. participants	8157	878	7860	1366	16017	2244
Age[d], mean (SD), year	12.5 (0.29)	12.5 (0.29)	12.5 (0.3)	12.5 (0.23)	12.5 (0.3)	12.5 (0.3)
Female, no. (%)	4495 (55.1%)	340 (38.7%)	4466 (56.8%)	713 (52.2%)	8961 (55.9%)	1053 (46.9%)
eFSM, no. (%), yes	1422 (17.4%)	151 (17.2%)	1243 (15.8%)	252 (18.4%)	2665 (16.6%)	403 (18.0%)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, IMD = index of multiple deprivation, no. = number, SD = standard deviation

[a] Socioeconomic background was expressed for descriptive purposes using two complementary measures (1) average percentage of eFSM pupils within a school, and (2) using the postcode-derived index of multiple deprivation (deciles). A higher percentage of eFSM pupils reflects a lower socioeconomic background, whereas higher IMD scores reflect a higher socioeconomic background.

[b] Ofsted rating from schools was derived from the UK government website https://reports.ofsted.gov.uk

[c] School location was derived using: https://www.gov.uk/government/collections/rural-urban-classification

[d] Age at baseline (1 September 2017)

At a school level, we used a drop-out model to assess whether any of the school-level demographics were associated with drop-out. That is, a bias-reduced generalised linear model (GLM) with logit link function (available in *brglm2*, Firth, 1993; Kosmidis, 2020) that contained all school-level demographics was fit to predict school drop-out. None of the school-level demographical variables were significant predictors of school drop-out (Table B.2).

Table B.2: School level drop-out model: Analysis of Deviance (Type II Wald χ^2 test).

Predictor	df	χ^2	p
Gender type	2	0.73	0.69
School type	2	2.45	0.29
Establishment type	7	2.32	0.94
School size	2	0.97	0.62
Number of form groups	1	0.00	0.98
Ofsted rating	3	3.42	0.33
Percent eFSM	1	3.42	0.33
IMD decile	1	0.00	0.95
School location	3	3.72	0.29

Abbreviations: eFSM = eligible for free school meals, IMD = index of multiple deprivation

Pupil level variables (sex, age, eFSM) were assessed using a Generalized linear mixed model (GLMM) with logit link function and random intercepts fitted for schools (available in *lme4* in R). Age, sex or eFSM did not significantly predict school drop-out (Table B.3), despite loosing a comparatively higher proportion of eFSM pupils (Table B.1). While these results do not provide evidence in favour of the null hypothesis (of no relationship), they at least suggest that none of the predictors significantly predicts school drop-out (i.e. we cannot reject the null hypothesis). Based on these findings, we assume that the remaining sample is comparable to the original sample and will therefore be treated as the main trial sample in the remainder of the thesis.

Table B.3: Pupil level drop-out model.

Predictor	OR	95% CI	p
Sex (reference = female)	1.18	0.001, 1043.13	0.96
Age	0.99	0.002, 452.8	1.00
eFSM (reference = no)	1.06	0.0002, 5870	0.99

Abbreviations: CI = confidence interval, eFSM = eligible for free school meals, OR = odds ratio
Note: intercept of model not displayed (all 0)

Back to main text

B.1.2 Representativeness brain imaging sub-sample

The representativeness of the brain imaging sub-sample was assessed on a school- and pupil level.

B.1.2.1 School level characteristics

Table B.4 shows a comparison of school level demographics between schools (n = 93) that are part of the main trial only (n = 83) and those that are part of the brain imaging sub-study (n = 10). Compared to the main trial sample (and total sample), a greater proportion of sub-study schools were co-educational (i.e. mixture of male and female), located in rural areas, were of medium school size and were of higher socioeconomic status (IMD and percentage eFSM). All of the intervention schools in the sub-study sample were trained face-to-face, compared to only 21% of the remaining main trial schools.

We assessed whether any of the school level characteristics are associated with inclusion / selection for the brain imaging sub-study, and thereby whether any of these variables is different between the two samples. To this end, we fit a bias-reduced generalized linear model with logit link function (available in R package *brglm2*, Firth, 1993; Kosmidis, 2020). Table B.5 presents the analysis of deviance results (Type II Wald χ^2 test), showing that none of the school level characteristics significantly predicts inclusion in the brain imaging sub-sample.

Table B.4: School level demographics for the main trial and sub-sample

	Main trial only		Sub-study		Overall	
	Schools (n = 83)	**Pupils (n = 15,956)**	**Schools† (n = 10)**	**Pupils (n = 61)**	**Schools (n = 93)**	**Pupils (n = 16,017)**
Gender type, no. (%)						
Co-ed	66 (79.5)	13227 (82.9)	9 (90.0)	60 (98.4)	75 (80.6)	13287 (83.0)
Female	16 (19.3)	2666 (16.7)	1 (10.0)	1 (1.6)	17 (18.3)	2667 (16.7)
Male	1 (1.2)	63 (0.4)	0 (0)	0 (0)	1 (1.1)	63 (0.4)
School type, no. (%)						
Comprehensive to 16	16 (19.3)	2893 (18.1)	0 (0)	0 (0)	16 (17.2)	2893 (18.1)
Comprehensive to 18	60 (72.3)	11451 (71.8)	7 (70.0)	41 (67.2)	67 (72.0)	11492 (71.7)
Secondary Modern	7 (8.4)	1612 (10.1)	3 (30.0)	20 (32.8)	10 (10.8)	1632 (10.2)
Establishment type, no. (%)						
Academy - Sponsor Led	16 (19.3)	2435 (15.3)	0 (0)	0 (0)	16 (17.2)	2435 (15.2)
Academy Converter	2 (2.4)	599 (3.8)	1 (10.0)	1 (1.6)	3 (3.2)	600 (3.7)
Academy Converter - Main	26 (31.3)	5562 (34.9)	5 (50.0)	43 (70.5)	31 (33.3)	5605 (35.0)
Community School	13 (15.7)	2947 (18.5)	2 (20.0)	5 (8.2)	15 (16.1)	2952 (18.4)
Foundation School	9 (10.8)	1886 (11.8)	1 (10.0)	1 (1.6)	10 (10.8)	1887 (11.8)
Free School - Mainstream	4 (4.8)	513 (3.2)	1 (10.0)	11 (18.0)	5 (5.4)	524 (3.3)
Voluntary aided school	11 (13.3)	1715 (10.7)	0 (0)	0 (0)	11 (11.8)	1715 (10.7)
Voluntary controlled school	2 (2.4)	299 (1.9)	0 (0)	0 (0)	2 (2.2)	299 (1.9)
School size, no. (%)						
Large (>200)	26 (31.3)	6835 (42.8)	1 (10.0)	1 (1.6)	27 (29.0)	6836 (42.7)
Medium (100-200)	46 (55.4)	8262 (51.8)	8 (80.0)	49 (80.3)	54 (58.1)	8311 (51.9)
Small (<100)	11 (13.3)	859 (5.4)	1 (10.0)	11 (18.0)	12 (12.9)	870 (5.4)
No. of form groups						
Median (range), n	7 (3–14)	8 (3–14)	6 (4–11)	6 (4–11)	6 (3–14)	8 (3–14)
Missing, no. (%)	2 (2.4)	251 (1.6)	0 (0)	0 (0)	2 (2.2)	251 (1.6)
Socioeconomic status[a]						
Percent eFSM, mean (SD), %	18.4 (11.0)	16.7 (10.8)	8.35 (6.75)	6.39 (4.59)	17.3 (11.1)	16.6 (10.8)
IMD, median (range), decile	5 (1–10)	6 (1–10)	8.50 (3–10)	8 (3–10)	5 (1–10)	6 (1–10)
Ofsted rating[b], no. (%)						
Outstanding	21 (25.3)	8680 (54.4)	1 (10.0)	33 (54.1)	22 (23.7)	8713 (54.4)
Good	49 (59.0)	4458 (27.9)	4 (40.0)	1 (1.6)	53 (57.0)	4459 (27.8)
Requires Improvement	7 (8.4)	1774 (11.1)	3 (30.0)	18 (29.5)	10 (10.8)	1792 (11.2)
Serious Weaknesses	1 (1.2)	99 (0.6)	0 (0)	0 (0)	1 (1.1)	99 (0.6)
Special Measures	4 (4.8)	765 (4.8)	2 (20.0)	9 (14.8)	6 (6.5)	774 (4.8)
Missing	1 (1.2)	180 (1.1)	0 (0)	0 (0)	1 (1.1)	180 (1.1)
Location[c], no. (%)						
Rural town and fringe	4 (4.8)	1414 (8.9)	5 (50.0)	38 (62.3)	9 (9.7)	1452 (9.1)
Urban city and town	34 (41.0)	6408 (40.2)	5 (50.0)	23 (37.7)	39 (41.9)	6431 (40.2)
Urban major conurbation	45 (54.2)	8134 (51.0)	0 (0)	0 (0)	45 (48.4)	8134 (50.8)
Teacher training, no. (%)						
Face-to-face	10 (12.0)	2893 (18.1)	5 (50.0)	32 (52.5)	15 (16.1)	2925 (18.3)
Online	26 (31.3)	4228 (26.5)	0 (0)	0 (0)	26 (28.0)	4228 (26.4)
None	47 (56.6)	8835 (55.4)	5 (50.0)	29 (47.5)	52 (55.9)	8864 (55.3)

Abbreviations: eFSM = eligible for free school meals, SD = standard deviation

† Values are based on those from the entire school where brain imaging sub-study participants are based

[a] Socioeconomic background was expressed for descriptive purposes using two complementary measures (1) average percentage of eFSM pupils within a school, and (2) using the postcode-derived index of multiple deprivation (deciles). A higher percentage of eFSM pupils reflects a lower socioeconomic background, whereas higher IMD scores reflect a higher socioeconomic background.

[b] Ofsted rating from schools was derived from the UK government website https://reports.ofsted.gov.uk

[c] School location was derived using: https://www.gov.uk/government/collections/rural-urban-classification

Table B.5: School level comparison between main trial sample and subsample: Analysis of Deviance (Type II Wald χ^2 test.

Predictor	df	χ^2	p
Gender type	2	1.07	0.59
School type	2	0.11	0.95
Establishment type	7	2.38	0.94
School size	2	2.19	0.33
Number of form groups	1	0.022	0.88
Ofsted rating	3	4.37	0.22
Percent eFSM	1	0.07	0.80
IMD decile	1	0.50	0.48
School location	1	2.64	0.10

Abbreviations: eFSM = eligible for free school meals, IMD = index of multiple deprivation

Back to main text

B.1.2.2 Pupil level characteristics

We subsequently assessed whether any pupil level characteristics predict inclusion/selection for the brain imaging sub-study, in two ways: (1) considering all pupils in all schools that participate in the main trial, (2) considering only pupils in schools (n = 1,641; 10 schools) of which pupils enrolled in the brain imaging sub-study.

All pupils in all schools Table B.6 presents a comparison of pupil level characteristics of those pupils who participated in the main trial only and those who also participated in the sub-study. The sub-study sample is comparable to the main trial sample on average age (Table B.6), but includes proportionally fewer females (49 % versus 56%) and participants that are eFSM (1.6% versus 16.7% , i.e. higher socioeconomic status). None of the pupil characteristics is significantly associated with enrolment in the sub-study when taking into account the hierarchical structure of the data (using a Generalized Linear Mixed Model (GLMM) with logit link function: all p >0.05, Table B.7). However, pupils from a lower socioeconomic background are less likely to enrol in the sub-study (OR = 0.08, 95% CI: 0.005—0.37, p = 0.01) when the structure in the data is ignored (i.e. a single level model, controlling for sex and age, Table B.8).

Table B.6: Pupil level demographics for the main trial and sub-sample

	Main trial only (n = 15,956)	**Sub-study** (n = 61)	**Overall** (n=16,017)
Age, mean (SD), years	12.5 (0.29)	12.5 (0.32)	12.5 (0.3)
Female, no (%),	8931 (56.0%)	30 (49.2%)	8961 (55.9%)
eFSM, no (%), yes	2664 (16.7%)	1 (1.6%)	2665 (16.6%)

Abbreviations: eFSM = eligible for free school meals, SD = standard deviation

Table B.7: Multilevel logistic regression (logit link) of pupil level characteristics on inclusion in the brain imaging sub-study.

term	OR	95% CI	p
Intercept	0	0, 0	0
Age, years (mean-centered)	0.84	0.34, 2.04	0.69
Sex (reference = female)	1.1	0.65, 1.88	0.72
eFSM (reference = not eFSM)	0.22	0.03, 1.6	0.13

Abbreviations: eFSM = eligible for free school meals, OR = odds ratio

Table B.8: Single level logistic regression (logit link) of pupil level characteristics on inclusion in the brain imaging sub-study.

term	OR	95% CI	p
Intercept	0	0, 0.01	0
Age, years (mean-centered)	0.81	0.34, 1.89	0.63
Sex (reference = female)	1.25	0.75, 2.07	0.39
eFSM (reference = not eFSM)	0.08	0, 0.37	0.01

Abbreviations: eFSM = eligible for free school meals, OR = odds ratio

Back to main text

Within sub-study schools Table B.9 presents a comparison of pupil level characteristics of all pupils in the 10 schools (n = 1,651) that participated in the brain imaging sub-study, showing similar findings to the above. That is, the pupils who enrolled in the sub-study are comparable to their classmates in age, but are proportionally less often female (49.1% versus 53.9%) and eFSM (1.6% versus 9.1%). None of the pupil level characteristics is significantly associated with enrolment in the sub-study (all $p > 0.05$). (Table B.10). These results remained unchanged if the school-level random intercept was removed (Table B.11).

Table B.9: Pupil level characteristics of pupils in brain imaging sub-study schools

	Main trial only (n = 1,590)	Sub-study (n = 61)	Overall (n = 1,651)
Age, mean (SD), years	12.5 (0.29)	12.5 (0.32)	12.5 (0.29)
Female, no. (%)	857 (53.9%)	30 (49.2%)	887 (53.7%)
eFSM, no. (%), yes	145 (9.1%)	1 (1.6%)	146 (8.8%)

Abbreviations: eFSM = eligible for free school meals

Table B.10: Multilevel logistic regression (logit link) of pupil level characteristics on inclusion in the brain imaging sub-study (within sub-study schools).

term	OR	95% CI	p
Intercept	0.03	0.01, 0.06	0
Age, years (mean-centered)	0.84	0.35, 2	0.7
Sex (reference = female)	1.13	0.67, 1.89	0.65
eFSM (reference = not eFSM)	0.21	0.03, 1.49	0.12

Abbreviations: eFSM = eligible for free school meals, OR = odds ratio

Table B.11: Single-level logistic regression (logit link) of pupil level characteristics on inclusion in the brain imaging sub-study (within sub-study schools).

term	OR	95% CI	p
Intercept	0.04	0.03, 0.05	0
Age, years (mean-centered)	0.85	0.35, 2.02	0.71
Sex (reference = female)	1.23	0.73, 2.05	0.43
eFSM (reference = not eFSM)	0.16	0.01, 0.75	0.07

Abbreviations: eFSM = eligible for free school meals, OR = odds ratio

In addition to the pupil level characteristics, we also examined whether baseline cardiorespiratory fitness (CRF) and self-reported physical activity were related with inclusion in the brain imaging sub-study. Descriptives are provided in Table B.12 and a plot of the distribution is shown in Figure B.1. Pupils in the brain imaging sub-study were on average more fit (CRF: M [SD] = 48.9 [24.3] laps) and active (M [SD] = 5.43 [1.83] days), compared to those in the main trial (CRF: M [SD] = 37.4 [20.6] laps, and active: M [SD] = 4.42 [1.92] days). However, only higher levels of CRF predicted enrolment in the brain imaging sub-study over and above physical activity levels (OR = 1.03, 95% CI: 1.002—1.06, p = 0.036; Table B.13). Higher levels of physical activity were not significantly related to enrolment in the sub-study (OR = 1.08, 95% CI: 0.77—1.53, p = 0.65). These relationships remained unchanged within the sub-sample of schools from which brain imaging sub-study pupils were recruited (i.e. OR_{CRF} = 1.03, p = 0.04, OR_{PA} = 1.06, p = 0.73). It is important to note that a substantial portion of participants (33%) did not complete the baseline fitness assessment or the baseline questionnaire (39%). Those who completed the assessments were preferentially males from a higher

socioeconomic background.

Table B.12: Aerobic fitness and self-reported physical activity of pupils participating in the main trial only and those who additionally participated in the brain imaging sub-study.

	Main trial only	Sub-study	Overall
All schools			
Pupils, no.	15956	61	16017
CRFs (20MSR)			
Mean (SD), laps	37.4 (20.6)	48.9 (24.3)	37.5 (20.6)
Missing, no. (%)	5298 (33.2%)	23 (37.7%)	5321 (33.2%)
Self-reported physical activity			
Mean (SD)	4.42 (1.92)	5.43 (1.83)	4.42 (1.92)
Median (range)	5 (0–7)	6 (2–7)	5 (0–7)
Missing, no. (%)	6258 (39.2%)	31 (50.8%)	6289 (39.3%)
brain imaging sub-study schools			
Pupils, no.	1590	61	1651
CRF (20MSR)			
Mean (SD), laps	38.7 (20.2)	48.9 (24.3)	39.0 (20.4)
Missing, no. (%)	405 (25.5%)	23 (37.7%)	428 (25.9%)
Self-reported physical activity			
Mean (SD)	4.83 (1.81)	5.43 (1.83)	4.85 (1.81)
Median (range)	5 (0–7)	6 (2–7)	5 (0–7)
Missing, no. (%)	579 (36.4%)	31 (50.8%)	610 (36.9%)

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, no. = number, SD = standard deviation

Table B.13: Multilevel logistic regression (logit link) of pupil level fitness and physical activity on inclusion in the brain imaging sub-study.

	All schools			Within sub-study schools		
	OR	**95% CI**	***p***	**OR**	**95% CI**	***p***
Intercept	0	0, 0	0	0	0, 0.02	0
CRF, laps	1.03	1, 1.06	0.04	1.03	1, 1.06	0.04
Self-reported PA, days	1.08	0.77, 1.53	0.65	1.06	0.75, 1.51	0.73

Abbreviations: CI = confidence interval, OR = odds ratio

Back to main text

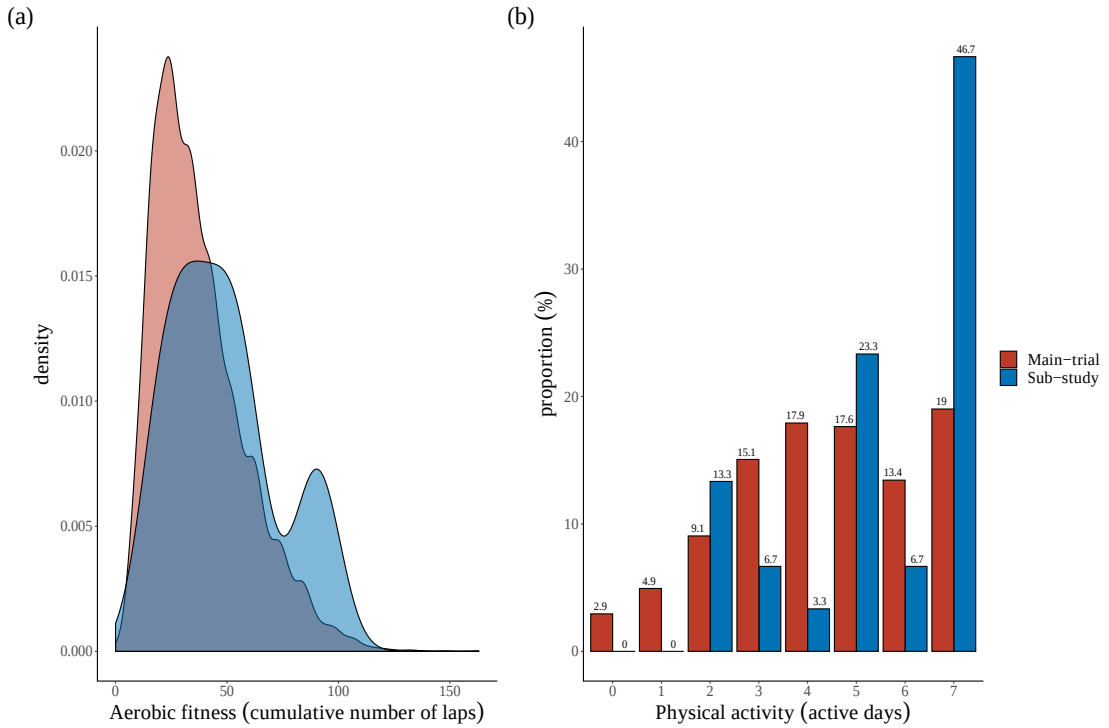


Figure B.1: A comparison of CRF and self-reported PA levels in main trial and sub-sample. Pupils in the sub-sample are more fit (*a*) and physically active (*b*) compared to the remainder of the main trial sample. CRF was assessed using the 20MSR, whereas self-reported days of physical activity was assessed using a single item questionnaire asking on how many days pupils have done a total of 60 minutes or more of moderate-to-vigorous physical activity.

B.1.3 Recruitment brain imaging cohort 2

A second sample of main trial schools was identified mid-intervention (in January 2018) based on them (1) having high completion rates of baseline cognitive, questionnaire and fitness data relative to other schools, (2) being responsive to messages from the Fit to Study research team, (3) having a lower socioeconomic background, as indicated by the percentage of FSM-eligible Year-7 pupils or postcode-derived index of multiple deprivation, and, for intervention schools, and (4) adhering to the intervention. From those schools that met criteria, a sub-sample of 10 schools was visited by a researcher blinded to treatment status. The researcher aimed to collect expressions of interest (EOI) from a sub-sample of 30 low-fit pupils per school selected by the PE teacher. If a sub-sample could not be selected, EOI were collected from the entire year group, and selection for the sub-study was based on baseline fitness data (i.e. lowest scores on

the 20MSR or Cooper Run test). Selected schools that were not interested in the brain imaging sub-study were replaced by other schools from the Fit to Study sample that were within reasonable travel distance (max 1.5h) of the brain imaging centre (Oxford, UK). Inclusion criteria for individual pupils was identical to the criteria for the first cohort, apart from the maximal exercise test ($\dot{V}O_2max$) screening, which was not part of the assessments for the second cohort. The trial manager ensured a balance between subjects from intervention and control schools.

Back to main text

B.2 Measures

Back to main text: outcome measures

B.2.1 Main trial

B.2.1.1 Cognitive assessments

Processing speed Processing speed was assessed with a simple reaction time task (Figure B.2). The task comprised fixation trials and a maximum of 40 target trials. During fixation trials, a black fixation cross was presented centrally for either 1,000 ms, 2,000 ms, 3,000 ms, or 4,000 ms. The stimulus duration was determined at random, with equal occurrence of each duration. If the participant responded while the cross was on the screen, they received feedback telling them to wait until they saw the target stimulus before pressing, followed by another fixation trial. If the participant did not respond while the cross was on the screen, a target trial was run. During the target trial, a red triangle was presented centrally on a white screen. The triangle disappeared once the participant responded by pressing the spacebar on the keyboard. Responses between 100–3,000 ms were considered valid. The task terminated after 240,000 ms, irrespective of whether all 40 target trials had been completed. Participants completed a practice block of four target trials and as many fixation trials as needed. The average reaction time across valid trials was the primary outcome measure.

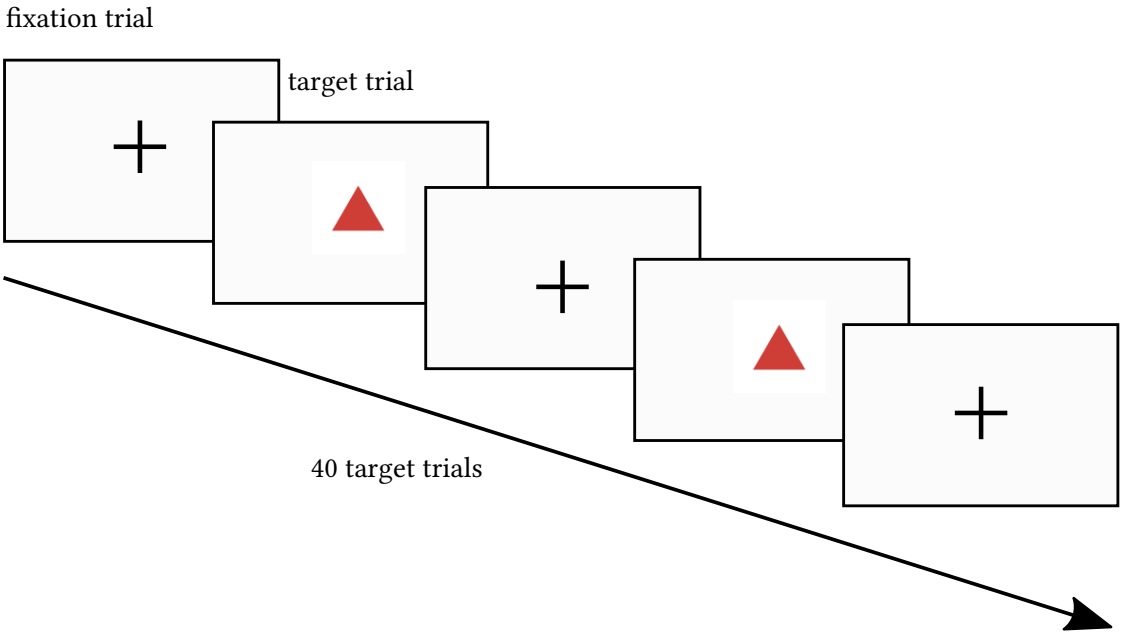


Figure B.2: Reaction time task paradigm. Participants had to respond as quickly as possible to the appearance of a red cross on the screen.

Back to main text

Relational memory Relational memory performance was assessed with a modified version of the relational memory paradigm described in Chaddock *et al.* (2010a). The task consisted of an encoding phase and recognition phase (Figure B.3). During the encoding phase, participants were presented with two blocks of 18 study trials (i.e. events) containing scene, face and object stimuli. Within a trial, each type of stimulus was shown individually and sequentially for 1,300 ms in the centre of a white screen. Trials were separated by a centrally presented black fixation cross for 1,000 ms. Participants were instructed to memorise the stimuli as belonging to a single event. During the recognition phase, (short-term) memory was tested for associations between stimuli within a trial. A single block of 36 test trials was presented. Half of the test trials were intact, i.e. "target test trials": the place, face and object all appeared in the same study trial, and the other half of the test trials were rearranged, i.e. "foil test trials": one of the items was from a different study trial. Stimuli were categorised and split evenly between target and foil test trials to prevent any systematic differences. During a test

trial, a place, object and face stimulus were presented simultaneously on a white screen (place on the left, face in the centre and object on the right). Participants were asked to indicate whether all items were from the same event, or whether one of the items did not belong to the event, by pressing a button on the keyboard. Only responses within 200–15,000 ms were considered valid. A trial was followed by a blank screen for 250 ms. Prior to the task, participants completed a short practice encoding phase and recognition phase. The proportion of correct responses on valid trials to the total number of valid trials (i.e. accuracy) was considered an index of relational memory.

The trials were kept the same for all participants, but the order of the study- and test trials was randomised. Moreover, there were two different versions of the task, each with its own set of stimuli. Within a school, half of students were randomly assigned version 1, the other half version 2. Participants who completed version 1 at baseline, were presented with version 2 post-intervention, and vice versa.

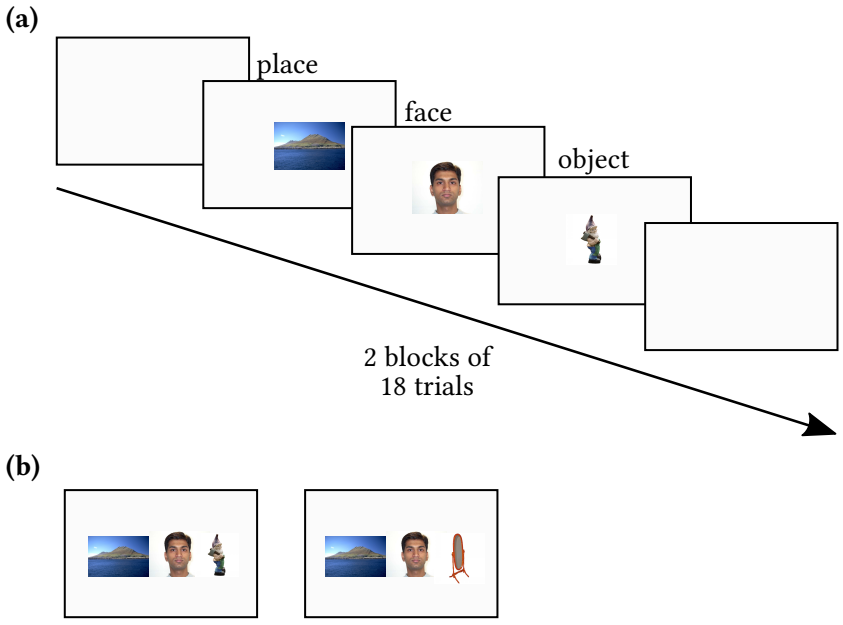


Figure B.3: Relational memory paradigm. (*a*) In the *encoding phase*, participants were presented with two blocks of 18 study trials containing scene, face and object stimuli. They were instructed to memorise the stimuli as belonging to a single event. (*b*) In the *recognition phase* participants had to indicate whether all items were from the same event or whether one of the items did not belong to the event.

Back to main text

Cognitive flexibility Cognitive flexibility was assessed using a modified version of the colour-shape switch task (Hillman *et al.,* 2014). In the first part of the task, participants were required to learn a set of response actions related to particular attributes of a centrally presented character: (1) its shape (circle or square) and (2) its colour (blue or green) (Figure B.4). Then, participants were instructed to use a rule-based cue – the direction of the character's arms – that governed what attribute to pay attention to, requiring them to flexibly shift between attributes and corresponding response actions.

In the task, participants completed two blocks of 32 homogeneous trials each; one block of shape only (Figure 2.5A), and one block of colour only (Figure 2.5B). The order of these blocks was randomised between participants. Participants were instructed to pay attention to the centrally presented character and respond as quickly and as accurately as possible by pressing the "F" key on the keyboard with their left index finger if the the character was blue or a circle, or the"J" key with their right index finger if the character was green or a square.

During the heterogeneous condition (the mixing block, Figure 2.5C), participants had to flexibly shift their attention towards the correct attribute (colour or shape) based on the direction of the character's arms. If the character's arms were up, the participant had to make a shape-based response; if the arms were down, a color-based response was required. Following a practice block of 32 trials (16 trials per attribute), participants completed two heterogeneous blocks of 64 trials each. Within each block, the direction of the character's arms switched every two trials, in line with Rogers & Monsell (1995) alternating-runs paradigm. The stimulus presented on a trial was chosen randomly from the set of stimuli appropriate for that trial type. The different stimuli were presented equally often throughout the task. The first trial of each heterogeneous block was not considered to be a switch or non-switch trial. Within a trial, the stimulus was presented centrally on a black background for 10,000 ms or until a response was given.

Trials were preceded by a 50 ms black screen and followed by a red cross for 50 ms in case of an erroneous response. Only responses within 200–6,000 ms were considered valid. Outcome variables were the average reaction time and accuracy on valid switch and non-switch trials, as well as the difference between the average reaction time on correct, valid non-switch and switch trials within the heterogeneous block, i.e. the "traditional" *switch cost* (Monsell, 2003; Wylie & Allport, 2000), and its equivalent for accuracy. The switch cost was computed as follows:

$$Switchcost_{\Delta X} = X_{switch} - X_{non-switch} \tag{B.1}$$

whereby X can be replaced with reaction time or accuracy on valid and correct trials.

Back to main text

Working memory A modified version of the visual two-back task (Chen *et al.*, 2016; Chen *et al.*, 2014b; Guiney & Machado, 2013) was used to assess working memory (Figure B.5). Stimuli were pseudo-random sequences of weather symbols (i.e. sun, moon, rainbow, snowflake and star) presented centrally on a white screen. Participants were instructed to monitor each stimulus and to identify as quickly and accurately as possible whether the stimulus was identical to the one that appeared two trials ago. When the current stimulus matched the previously observed stimulus, the participant was instructed to press the "F" key with their left index finger to indicate a positive response; in case of a mismatch, the participant was asked to press the "J" key. The task consisted of two blocks of 26 trials each. Of the last 24 trials in each block, eight (33%) were target trials and 16 were foil trials. Whether the current trial was a target or foil was randomly decided at the start of each trial. For target trials, the stimulus from two trials ago was presented on the screen. For foil trials, one of the other stimuli, with the exception of stimuli that were presented one or two trials ago, was selected at random. A trial started with a blank screen for 500 ms, followed by the stimulus for 1,500 ms and another blank screen for 1,000 ms. Responses within 150–2,500 ms after stimulus onset were considered valid. At the beginning of the task, participants were given a

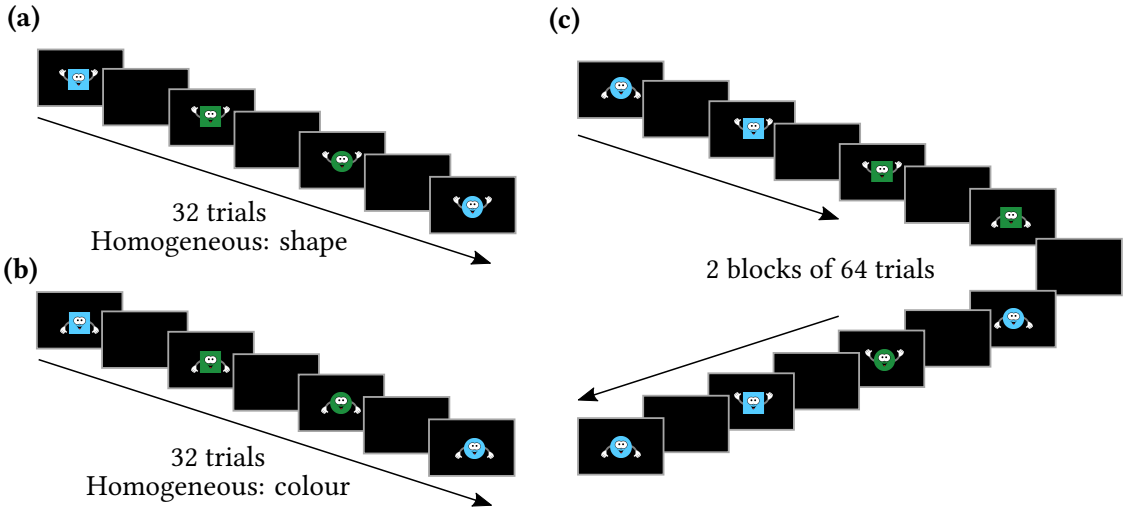


Figure B.4: Colour-shape switch task paradigm. (*a*) In the homogeneous *shape* condition (here: arms up), participants had to pay attention to the character's shape and press the F key if the character was a circle or the "J" key if the character was a square, (*b*) in the homogeneous *colour* condition (arms down), participants had to focus on the character's colour and press the F key if the character was blue, or the "J" key if the character was green, (*c*) in the heterogeneous *colour-shape switch* condition, participants had to shift their attention to the colour or shape based on direction of the character's arms: shape if the arms were up, colour if the arms were down. The direction of the characters arms *switched* every two trials. On non-switch trials, the participant had to keep their attention towards the same attribute, whereas on switch trials, the participant had to *switch* between attributes. It is assumed that individuals use top-down voluntary control when switching between competing tasks, requiring them to respond more slowly (reaction time) or less accurate (see e.g. Kiesel *et al.*, 2011; Monsell, 2003). Figure is adapted from Hillman *et al.* (2014).

demonstration of the task to help their understanding. To test their understanding, participants were asked to complete a hypothetical sequence of stimuli by indicating whether the current stimulus would be a target or a foil. If the participant passed the understanding test (<2 errors), they completed a practice block of 12 trials. Practice included feedback about their accuracy and response time, e.g. a prompt was shown for too slow responses. The average response time and accuracy on valid trials were used as indices of working memory performance.

Back to main text

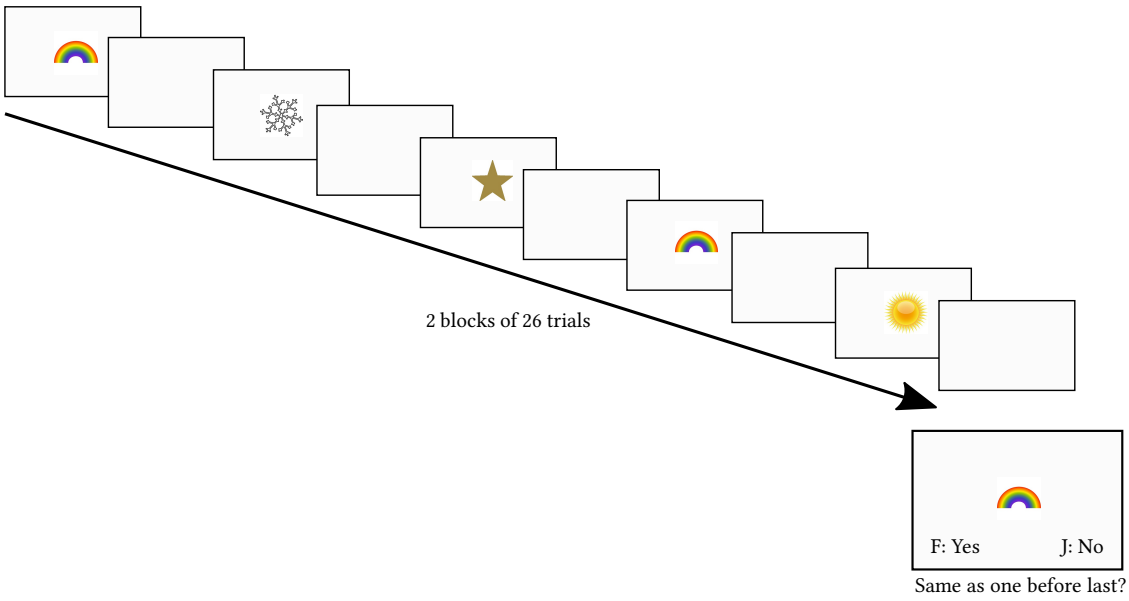


Figure B.5: Two-back working memory paradigm. Participants were shown pseudo-random sequences of weather symbols and were instructed to identify as *quickly* and *accurately* as possible whether the current stimulus was identical ("F" key) or not ("J" key) to the one that appeared two trials ago.

Inhibitory control A modified version of the Flanker task was used to assess inhibitory control (Hillman *et al.*, 2014). In the task, participants were required to respond as quickly and accurately as possible to the direction of a centrally presented target arrow which was flanked by an array of arrows that were either congruent (>>>>>) or incongruent (<<><<) to the central target stimulus (Figure B.6). Participants were instructed to press the "F" key if the target stimulus pointed to the left, and the "J" key if the stimulus pointed to the right. The task consisted of two blocks of 52 trials each. Each block contained four trial types: congruent left, congruent right, and incongruent left and incongruent right. Within a block, a total of 13 trials of each type was presented in random order. A trial started with a blank screen for 500 ms, followed by the stimulus array for 333 ms and another blank screen for 1,067 ms. Responses within 200–1,400 ms after stimulus onset were considered valid. At the beginning of the task, participants were given three practice blocks of eight trials each: (1) target stimuli only, with feedback, (2) target and flanked stimuli, with feedback, and (3) target and flanked stimuli, without feedback. Outcome variables were accuracy and reaction time for the congruent and incongruent condition, as well as a reaction time and accuracy

interference score, defined as the difference in performance on valid incongruent and congruent trials ($X_{inter} = X_{incon} - X_{con}$), where X could be accuracy or reaction time.

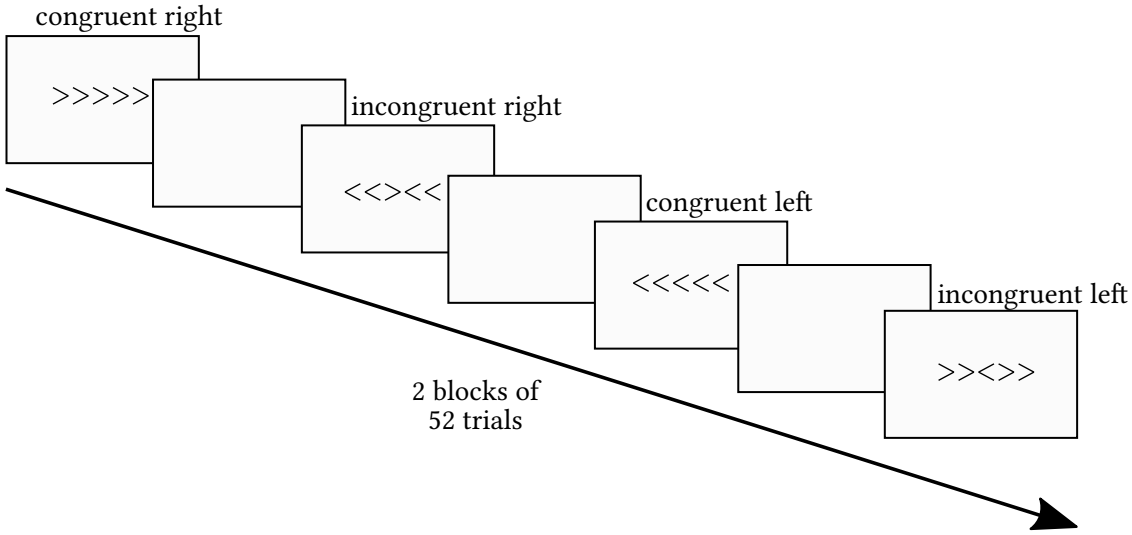


Figure B.6: Modified flanker task The task contained four trial types: congruent left, congruent right, and incongruent left and incongruent right. Participants were instructed to press, as quickly and accurately as possible, the "F" key if the centrally located stimulus pointed to the left, and the "J" key if the stimulus pointed to the right.

Back to main text

Randomisation of cognitive assessments The order of cognitive assessments was pseudo-randomised across participants within schools to reduce the probability of task-order effects. All participants started with the reaction time task (i.e. processing speed), followed by the remaining four tasks. This arrangement was because pilot testing suggested that compliance may be higher following a relatively simple task. Randomisation was performed by the database developer who was not involved in the analysis of the data. Randomisation was performed per school to ensure a balance in possible task sequences (a total of 24, given that all participants started with the reaction time task) across participants, both within and between schools. For each school, participants were sorted randomly and then assigned a number between 1 and 24 (in order), corresponding to each of the possible task sequences, until every

participant was allocated a task sequence. This process was performed separately for baseline and post-intervention assessments in Microsoft Excel.

Back to main text

B.2.1.2 Psychological variables linked to daily MVPA

Descriptions of questionnaire-based psychological variables linked to daily MVPA, adapted from Wassenaar *et al.* (2019a), were included in this thesis for completion and given that they have been used in various missing data imputation approaches presented in chapter 4 to chapter 7. Dr C. Wheatley led the selection of these measures.

Government health guidelines suggest that young people accumulate 60 minutes of MVPA per day (2018 Physical Activity Guidelines Advisory Committee, 2018). To investigate attitudes and beliefs about this volume of activity, we used a description of MVPA drawn from previous research (Inchley *et al.*, 2016; Scott *et al.*, 2015). Participants were instructed to think of PA as *"any activity that raises your heart rate and makes you breathe harder and feel warmer. Fast cycling to and from school, running around during school breaks, doing PE and taking part in after-school sports clubs are all examples of physical activit"*. Participants were instructed that *"An hour of physical activity can either be in one block, such as football or netball practice, or several shorter activities that add up to an hour, such as cycling to and from school and running during break and at lunchtime."*

Attitudes were measured using a semantic differential scale, anchored with three items drawn from previous research (Rivis & Sheeran, 2003; Wheatley *et al.*, 2017): *"For me, being physically active for an hour every day in a typical week during term would be stressful/relaxing; boring/fun; dissatisfying/satisfying."*.

Subjective norms Two items (1 = "strongly disagree" to 7 = "strongly agree") assessed injunctive norms, or perceived approval of PA, among people in my family and most of my friends, e.g.: *"People in my family think I should be physically active for an hour every day in a typical week during term."* A further two items captured descriptive norms, or prevalence of the behaviour among people in my family and most of my friends, e.g.: *"Most of my friends will be physically active for an hour every day in a typical week during term."* .

Perceived behavioural control was measured with two items (1 = "strongly disagree" to 7 = "strongly agree"). The first taps self-efficacy: *"If I wanted to, I am confident that I could be physically active for an hour every day in a typical week during term."* The second captures controllability: *"Whether or not I am physically active for an hour every day in a typical week during term is completely up to me."* .

Intention was measured with a single item (1 = "definitely no" to 7 = "definitely yes"): *"I intend to be physically active for an hour every day in a typical week during term."*.

Past behaviour Habitual physical activity over the past six months was measured using a single item (1 = "never" to 7 = "always"): *"Thinking about the past six months, how often have you been physically active for an hour every day during a typical week in term?"* (Hagger *et al.,* 2001).

Prototype favourability In line with theoretical guidelines (Gibbons *et al.,* 1995) participants read a definition of a prototype. Then they were asked to consider active prototypes: *"Think of someone your age who is physically active for an hour every day in a typical week during school term" and indicate "how far the following words describe this type of person"* (1 = "not at all" to 7 = "extremely") using four adjectives (confident, popular, determined, attractive) drawn from previous research (Wheatley *et al.,* 2017).

Prototype similarity In line with previous research (Rivis & Sheeran, 2003), perceived similarity to active and inactive prototypes was measured with single items (1 = "very dissimilar" to 7 = "very similar"): *"In general, how similar are you to the type of person who is physically active for an hour every day in a typical week during term?"*

Behavioural willingness In line with previous research (Davies *et al.*, 2016; Gibbons *et al.*, 1998), willingness was measured by presenting participants with imaginary situations in school where they could choose to be physically active or inactive in the presence of peers. For each scenario, two items (1 reversed) assessed whether participants would be willing to be or avoid being physically active (1 = "extremely unlikely" to 7 = "extremely likely").

Back to main text

B.2.2 Brain imaging sub-study

B.2.2.1 Psychological variables linked to daily MVPA

The study also included exploratory explicit and implicit measures of psychological variables linked to daily average PA, using the frameworks set out in the Theory of Planned Behaviour (Ajzen, 1985; Ajzen, 1991) and the Prototype Willingness Model. Like the psychological measures described above, they have been included in this thesis for completion and given their role in various missing data imputation approaches presented in chapter 4 to chapter 7. Dr C. Wheatley led the selection of these measures.

Explicit measures The explicit measures set out in B.2.1.2 above were taken from the main trial and repeated in the brain imaging sub-study to (1) obtain a well-controlled, lab-based, measure of the task for cross-validation, and (2) ensure that all brain-study participants had a score on this task, independent from their main trial completion.

Implicit Association Test (IAT) Individuals are thought to have cognitive associations and biases that exist outside conscious awareness (Greenwald *et al.,* 2009). Furthermore, implicit cognitions can influence behaviour (Strack & Deutsch, 2004) and can also, potentially, lend additional predictive power over and above explicit measures (Greenwald *et al.,* 2009). The Implicit Association Test (IAT, Greenwald *et al.,* 1998) is a computer-based sorting task that measures biases (Strack & Deutsch, 2004). This study used two IATs developed according to theoretical guidelines (Greenwald *et al.,* 1998; Greenwald *et al.,* 2003; Greenwald *et al.,* 2009; Nosek *et al.,* 2005) to test the predictive power of implicit prototype evaluations as specified by the Prototype Willingness Model (Gerrard *et al.,* 2008). The Prototype Favourability IAT measured adolescents' implicit attitudes towards physically active and inactive images. The Prototype Similarity IAT, which also draws from the Self-Concept IAT (Greenwald & Farnham, 2000) and the Drinker Identity IAT (Gray *et al.,* 2011) measured the extent to which participants associated active/inactive characteristics with their own identities. Each IAT assessed the relative strengths of associations between a target concept (active/inactive) and a target attribute: positive/negative in the favourability IAT; and me/not me in the similarity IAT. Participants classified stimuli according to whether they were active/inactive or positive/negative (or me/not me) as quickly as possible. The task rested on the assumption that the more strongly participants associate two concepts (e.g. "active" and "positive" or "active" and "me") the faster and more accurately they perform.

Target concepts The words "active" and "inactive" represented the concept of the physically active prototype and the physically inactive prototype. To ensure participants understood these concepts, and that they did not confound "physically active (inactive)" with "physical activity", the on-screen instructions also described "the sort of person your age who is physically active (inactive)".

Attribute concepts The favourability IAT used "positive" and "negative" as evaluative attributes, while the similarity IAT used "me" and "not me".

Stimuli The number of stimuli were limited to four in each category of target concept and attribute (i.e. 16 overall), which is sufficient to maintain validity without sacrificing brevity.

- **Physically active (inactive) stimuli.** In both IATs, the stimuli used to represent physically active (inactive) prototypes were images of stick-people so that the activity level changed while all other individual characteristics were held constant.

- **Positive (negative) stimuli.** These were words describing social characteristics. Positive attributes —confident, popular, determined, attractive —and negative attributes —annoying, bossy, judgmental, grumpy —were drawn from previous research with adolescents (Wheatley *et al.,* 2017).

- **Me (not me) stimuli.** In line with previous research (Greenwald & Farnham, 2000), these were synonyms for me/not me: I, my, mine, self and they, them, their, others.

Both IATs were programmed in IATGEN, a free, validated tool that runs in Qualtrics and allows researchers to combine IATs and explicit self-report measures in a single survey (Carpenter *et al.,* 2017). IATGEN runs a standard seven-block IAT (Greenwald *et al.,* 1998), which takes approximately five minutes to complete. Following the recommended procedure (Greenwald *et al.,* 2003), IATGEN removes participants who record <300 ms for more than 10% of responses, and deletes individual trials over 10,000 ms. Where participants have made an error, the time spent correcting the mistake was included in latency score. A positive IAT- or D score indicated that participants had relatively stronger associations between the congruent target and attribute ("active" and "positive") than between the incongruent target and attribute ("active" and "negative").

Back to main text

B.2.2.2 Cognitive assessments

Each cohort completed a battery of cognitive assessments (Table B.14). *Cohort 1* completed the object-location task (OLT; main text: Object-location task), IAT, relational memory task and colour-shape switch task. *Cohort 2* completed the OLT and the Tower of London task (described below). The OLT was included to allow for across-cohort analyses or cross-validation of primary outcome findings in an independent cohort. The Tower of London task was included to obtain a measure of planning skills, a higher-level executive function that is hypothesised to improve with physical activity (Álvarez-Bueno *et al.*, 2017b).

Table B.14: Cognitive assessments completed by cohort 1 and cohort 2

Cohort 1	Cohort 2
• Object-location memory task • Implicit association task • Relational memory task • Colour-shape switching task	• Object-location memory task • Tower of London task

Tower of London task

Planning was assessed using the Tower of London task (Anderson *et al.*, 1996; Shallice, 1982; Van Der Niet *et al.*, 2015). In the task, a participant was instructed to move three coloured balls arranged on three pegs of differing heights from their initial position into another pre-specified target configuration. Only one ball could be moved at a time from one peg to another with certain additional restrictions: (1) each of the three pegs could only hold three, two or one balls respectively, and (2) a ball could not be moved when another ball was lying on top of it. A total of 12 puzzles/problems had to be solved, that differed in difficulty by the number of required moves, ranging from two to five moves. A problem was solved correctly when the target configuration was achieved in the set number of moves, with a total of three attempts per problem. A participant received

three points for solving the problem in one attempt, two points if two attempts were needed, one point if three attempts were required and zero points if more than three attempts were needed. The sum of the points for all 12 problems constituted their total score (maximum 36 points). Further, the time it took between problem presentation and the first move was recorded ("planning time"), as well as the time between the first move and the completion of the problem ("execution time"). The task was programmed in Inquisit 4 and available through the Inquisit website.

Back to main text

B.2.2.3 Health behaviour in school-aged children questionnaire

A selection of items from the Health Behaviour in School-Aged Children (HBSC) questionnaire was included (Table B.15) to capture health-related behaviours of participants. Items were selected based on relevance for the study and research aims; items that were for >15 year old individuals only were excluded.

Table B.15: Health behaviour in school-aged children questionnaire

Category	Item	Response
Demographics		
Sex	Are you a boy or a girl?	B/G
Age	What month were you born?	
Age	What year were you born?	
Eating habits		
Breakfast	How often do you usually have breakfast during **weekdays** (more than a glass of milk or fruit juice)?	1 = never, 6 = 5 days
Breakfast	How often do you usually have breakfast during the **weekend** (more than a glass of milk or fruit juice)?	1 = never, 3 = both days
Food consumption freq.	How many times a week do you usually eat or drink .... ? [Fruits]	1 = never, 7 = every day (>1)
Food consumption freq.	How many times a week do you usually eat or drink .... ? [Vegetables]	1 = never, 7 = every day (>1)
Food consumption freq.	How many times a week do you usually eat or drink .... ? [Sweets (candy or chocolate)]	1 = never, 7 = every day (>1)
Food consumption freq.	How many times a week do you usually eat or drink .... ? [Coke or other soft drinks that contain sugar]	1 = never, 7 = every day (>1)
Oral health		
Oral health	How often do you brush your teeth?	1 = >once a day, 5 = never
Weight control & body image		
Weight control	At present are you on a diet or doing something else to lose weight?	1 = no, 4 = yes
Body image	Do you think your body is.....? [thin - fat]	1 = too thin, 5 = too fat
Physical activity		
MVPA	For this next question, add up all the time you spent in physical activity each day. Over the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?	1 = 0 days, 8 = 7 days
VPA	OUTSIDE SCHOOL HOURS: How OFTEN do you usually exercise in your free time so much that you get out of breath or sweat?	1 = every day, 7 = never
VPA	OUTSIDE SCHOOL HOURS: How many HOURS a week do you usually exercise in your free time so much that you get out of breath or sweat?	1 = none, 6 = 7 hours or more
Leisure time activity		
Television use	About how many hours a day do you usually watch television (including DVDs and on-line videos/streaming) in your free time during **weekdays**?	1 = none, 9 = 7 hours or more a day
Television use	About how many hours a day do you usually watch television (including DVDs and on-line videos/streaming) in your free time during the **weekend**?	1 = none, 9 = 7 hours or more a day

Continued on next page

Table B.15 – continued from previous page

Category	Item	Response
Computer use	About how many hours a day do you usually play games on a computer or games console (Playstation, Xbox, GameCube etc.) in your free time during **weekdays**?	1 = none, 9 = 7 hours or more a day
Computer use	About how many hours a day do you usually play games on a computer or games console (Playstation, Xbox, GameCube etc.) in your free time during the **weekend**?	1 = none, 9 = 7 hours or more a day
Computer use	About how many hours a day do you usually use a computer for chatting on-line, internet, emailing, homework etc. in your free time during **weekdays**?	1 = none, 9 = 7 hours or more a day
Computer use	About how many hours a day do you usually use a computer for chatting on-line, internet, emailing, homework etc. in your free time during the **weekend**?	1 = none, 9 = 7 hours or more a day
Risk behaviour		
Tobacco use	Have you ever smoked tobacco? (At least one cigarette, cigar or pipe)	Y/N
Tobacco use	How often do you smoke tobacco at present?	1 = every day, 4 = not smoke
Alcohol consumption	***** At present, how often do you drink anything alcoholic, such as beer, wine or spirits? Try to include even those times when you only drink a small amount. [Beer]	1 = every day, 5 = never
Alcohol consumption	***** At present, how often do you drink anything alcoholic, such as beer, wine or spirits? Try to include even those times when you only drink a small amount. [Wine]	1 = every day, 5 = never
Alcohol consumption	***** At present, how often do you drink anything alcoholic, such as beer, wine or spirits? Try to include even those times when you only drink a small amount. [Spirits / Liquor]	1 = every day, 5 = never
Alcohol consumption	***** At present, how often do you drink anything alcoholic, such as beer, wine or spirits? Try to include even those times when you only drink a small amount. [Alcopops (i.e. Smirnoff Ice, Bacardi Breezer, Mike's Hard,Lemonade)]	1 = every day, 5 = never
Alcohol consumption	***** At present, how often do you drink anything alcoholic, such as beer, wine or spirits? Try to include even those times when you only drink a small amount. [Any other drink that contains alcohol]	1 = every day, 5 = never
Drunkenness	Have you ever had so much alcohol that you were really drunk?	1 = never, 5 = yes, more than 10 times
Injuries, fighting, bullying		
Injuries	Many young people get hurt or injured from activities such as playing sports or fighting with others at different places such as the street or home. Injuries can include being poisoned or burned. Injuries do not include illnesses such as Measles or the Flu. The following questions are about injures you may have had during the past 12 months. During the past 12 months, how many times were you injured and had to be treated by a doctor or nurse?	1 = not injured in past 12 months, 5 = 4 times or more
Fighting	During the past 12 months, how many times were you in a physical fight?	1 = no physical fight in past 12 months, 5 = 4 times or more
Bullying	How often have you been bullied at school in the past couple of months?	1 = not been bullied at school the past couple of months, 5 = several times a week

Continued on next page

Table B.15 – continued from previous page

Category	Item	Response
Bullying others	How often have you taken part in bullying another student(s) at school in the past couple of months?	1 = not bullied others at school the past couple of months, 5 = several times a week
Family		
Family structure	All families are different (for example, not everyone lives with both their parents, sometimes people live with just one parent, or they have two homes or live with two families) and we would like to know about yours. Please answer this first question for the home where you live all or most of the time and tick the people who live there.	
Family structure	Please say how many brothers live here (including half, step or foster brothers). Please write in the number or write 0 (zero) if there are none.	
Family structure	Please say how many sisters live here (including half, step or foster sisters). Please write in the number or write 0 (zero) if there are none.	
Family structure	Do you have another home or another family, such as the case when your parents are separated or divorced?	
Family structure	How often do you stay there?	
Family structure	Please tick the people (adults) who live there:	
Family structure	Please say how many brothers live here (including half, step or foster brothers). Please write in the number or write 0 (zero) if there are none.	
Family structure	Please say how many sisters live here (including half, step or foster sisters). Please write in the number or write 0 (zero) if there are none.	
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Father]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Stepfather (or mother's boyfriend)]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Mother]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Stepmother (or father's girlfriend)]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Elder brother (s)]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Elder sister (s)]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Best friend]	1 = very easy, 5 = don't have or see this person
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Friends of the same sex]	1 = very easy, 5 = don't have or see this person

Continued on next page

Table B.15 – continued from previous page

Category	Item	Response
Family communication	How easy is it for you to talk to the following persons about things that really bother you? Please tick one box for each line [Friends of the opposite sex]	1 = very easy, 5 = don't have or see this person
Peers		
Number of close friends	At present, how many close male and female friends do you have? [Males]	1 = none, 4 = three or more
Number of close friends	At present, how many close male and female friends do you have? [Females]	1 = none, 4 = three or more
Peer contact frequency	How many days a week do you usually spend time with friends right after school?	1 = 0 days, 7 = 6 days
Peer contact frequency	How many evenings per week do you usually spend out with your friends?	1 = 0 days, 8 = 7 days
Peer contact frequency	How often do you talk to your friend(s) on the phone or send them text messages or have contact through the internet?	1 = rarely or never, 5 = every day
Positive health		
Subj. health complaints	In the last 6 months: how often have you had the following....? [Headache]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Stomach-ache]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Back ache]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Feeling low]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Irritability or bad temper]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Feeling nervous]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Difficulties in getting to sleep]	1 = about every day, 5 = rarely or never
Subj. health complaints	In the last 6 months: how often have you had the following....? [Feeling dizzy]	1 = about every day, 5 = rarely or never
Self-rated health	Would you say your health is.....?	1 = excellent, 4 = poor
Life satisfaction	Here is a line from 0 to 10. The right corner is the best possible life for you; whilst the left corner is the worst possible life for you. In general, where on the line do you feel you stand at the moment?	0 = worst possible life, 10 = best possible life
School setting		
Academic achievement	In your opinion, what does your class teacher(s) think about your school performance compared to your classmates?	1 = very good, 4 = below average
Satisfaction with school	How do you feel about school at present?	1 = like it a lot, 4 = don't like it at all
Student relations	Here are some statements about the students in your class(es). Please show how much you agree or disagree with each one. [The students in my class(es) enjoy being together]	1 = strongly agree, 5 = strongly disagree
Student relations	Here are some statements about the students in your class(es). Please show how much you agree or disagree with each one. [Most of the students in my class(es) are kind and helpful]	1 = strongly agree, 5 = strongly disagree

Continued on next page

287

Table B.15 – continued from previous page

Category	Item	Response
Student relations	Here are some statements about the students in your class(es). Please show how much you agree or disagree with each one. [Other students accept me as I am]	1 = strongly agree, 5 = strongly disagree
School-related stress	How pressured do you feel by the schoolwork you have to do?	1 = not at all, 4 = a lot
Social inequality		
Parental occupation	Does your father have a job?	1 = yes, 4 = don't have or see father
Parental occupation	Please say in what place he works (for example: hospital, bank, restaurant)	
Parental occupation	Please write down exactly what job he does there (for example: teacher, bus driver)	
Parental occupation	Why does your father not have a job? (Please tick the box that best describes the situation)	1 = he is sick, retired, or student, 4 = I don't know
Parental occupation	Does your mother have a job?	1 = yes, 4 = don't have or see mother
Parental occupation	Please say in what place she works (for example: hospital, bank, restaurant)	
Parental occupation	Please write down exactly what job she does there (for example: teacher, bus driver)	
Parental occupation	Why does your mother not have a job? (Please tick the box that best describes the situation)	1 = she is sick, retired, or student, 4 = I don't know
Family affluence scale	Does your family own a car, van or truck?	1 = no, 3 = yes, two or more
Family affluence scale	Do you have your own bedroom for yourself?	Y/N
Family affluence scale	During the past 12 months, how many times did you travel away on holiday / vacation with your family?	1 = not at all, 4 = more than twice
Family affluence scale	How many computers does your family own?	1 = none, 4 = more than two
Perceived fam. wealth	How well off do you think your family is?	1 = very well off, 5 = not at all well off
Hunger	Some young people go to school or to bed hungry because there is not enough food at home. How often does this happen to you?	1 = always, 4 = never

Abbreviations: B/G = boy / girl, Y/N = yes / no

Back to main text

B.3 Sample size

Sample size calculations were performed by NatCen for the primary intention-to-treat analysis, which compared maths test performance between intervention and control schools (Husain, 2017). The initial sample size calculation demonstrated a range of minimum detectable effect sizes (MDES) varying between 0.18–0.26 for various achieved sample sizes (80—100 schools), assuming 80% power, $\alpha = 0.05$ (two-tailed), 90 pupils per school and baseline covariates explaining 50% of variance at pupil and school level (Husain, 2017). These estimates were updated (in February–March 2018) to reflect school drop-out (n = 18 schools) and a revised set of assumptions (Bartasevicius & Jabin, 2018). In particular, the new MDES were computed based on: 88 participating schools (47% intervention), on average 78 pupils per school, baseline covariates explaining 40% of pupil level variance and 10% of school level variance, and an intraclass correlation coefficient (ICC) of 0.16. The newly computed MDES was equal to 0.24 standard deviations. The MDES would have been 0.21 standard deviations if all 106 recruited schools had participated in the trial. At the final analysis stage, the MDES was calculated to be 0.22 SD (Husain *et al.,* 2019), mainly due to a lower ICC (0.11) than was originally assumed. Power calculations were performed in PowerUp! (Dong & Peng, 2013).

The MDES for the cognitive tests was calculated for the randomisation sample using the following set of assumptions: (1) 104 schools, 50% of which in the intervention group, (2) an average of 176 pupils per school (n_{tot} = 18,261), (3) baseline covariates explaining 40% of pupil level variance and 10% of school level variance (based on NatCen's adjusted estimates), (4) an ICC of 0.02, following a previous study measuring executive functions in a comparable sample (Aadland *et al.,* 2019), and (5) 80% power with $\alpha = 0.05$ (two-tailed). Under these assumptions MDES was 0.09 SD (95% CI: 0.027, 0.16; it would be 0.15 SD, 95% CI: 0.045, 0.26, if the ICC was 0.06, as suggested by Aadland *et al.,* 2017b). With a similar set of assumptions and an ICC = 0.25 (NB. teacher-rated; Berry *et al.,* 2016), the MDES for the SDQ was 0.23 SD. For CRF, the

MDES was 0.08 SD with a similar set of assumptions and an ICC = 0.03 (Kriemler *et al.*, 2010). Notably, the trial was not designed to measure or affect cognitive, CRF or mental health scores (being secondary measures), hence these MDES' solely serve as *a-posteriori* indications of possible detectable effect sizes under these particular sets of assumptions.

Back to main text

C

Supplementary materials to chapter 3: *Methodological considerations when measuring and assessing physical activity, the brain and cognition*

Contents

C.1 Physical activity: alternative objective methods for assessing physical activity . . . 292
C.2 Brain imaging: confounds . . . 292
C.2.1 Scanner-related confounds . . . 292
C.2.2 Subject-related confounds . . . 293
C.3 Cognitive testing: psychometrics . . . 296
C.4 Data analysis . . . 300
C.4.1 Cluster randomised trials . . . 300
C.4.2 Missing data . . . 306

C.1 Physical activity: alternative objective methods for assessing physical activity

Besides actigraphy, a range of other methods may be used to objectively assess PA. In brief, doubly labelled water and indirect calorimetry measure energy expenditure —with the former being the "gold standard" —and have been used extensively for validation of other PA measures, but are expensive, time-intensive, cumbersome and cannot capture qualitative data (e.g. activity type). Direct observation can provide such qualitative information alongside quantitative measurements (e.g. duration and frequency) and is flexible, making it suitable for studies involving young children, yet is time-intensive, requires well-trained observers, has the potential for reactivity (of the observed) and is impractical for population studies. Pedometers are motion sensors that count steps (acceleration in one direction) at low cost, can be used for population studies, but provide no information about intensity, duration or frequency of PA. Heart rate (HR) monitors indirectly measure energy expenditure and PA (including duration, frequency and intensity), are unobtrusive, can be worn for long periods of time by a large number of participants, but require calibration (to account for inter-individual differences in the relation between heart rate and energy expenditure) and measurements are affected by factors other than PA, such as e.g. body position, stress and level of fitness. Some of the limitations of HR monitoring, however, can be overcome if they are combined with accelerometers, yet this may be cumbersome and too expensive for large-scale studies.

Back to main text

C.2 Brain imaging: confounds

C.2.1 Scanner-related confounds

A large number of scanner related factors impact on the MR signal and, in turn, influence MRI-derived measures. Variations in scanner hardware and software, including scanner manufacturer (Noble *et al.,* 2017), strength of the magnetic field (Chen *et al.,*

2014b; Han *et al.,* 2006; Tardif *et al.,* 2010), radio-frequency (RF) coils (Chen *et al.,* 2014b), imaging protocols (Han *et al.,* 2006; Krueger *et al.,* 2012; Tardif *et al.,* 2009), as well as scanner upgrades (Keenan *et al.,* 2019, but see Han *et al.,* 2006), table position (Alfaro-Almagro *et al.,* 2020) and processing protocols (Han *et al.,* 2006; Lerch *et al.,* 2017), could bias MRI-derived estimates. In typical MRI studies that are conducted at a single site, these factors are kept constant throughout the study and are of little practical importance. In the current era of multi-site studies and publicly available MRI data, however, such factors need to be considered during data collection and (statistical) modelling (Chen *et al.,* 2014b; Focke *et al.,* 2011; Fortin *et al.,* 2017; Noble *et al.,* 2017).

Image distortions due to magnetic field ($\mathbf{B}_0$ and $\mathbf{B}_1$) inhomogeneities and eddy currents are more complex to deal with, are particularly troublesome for weighted sequences (e.g. T1-weighted and diffusion), and require additional online or offline processing. For instance, shimming —the adjustment for variation in the magnetic field by applying additional magnetic fields —could be used to correct for inhomogeneities in the main magnetic field, whereas additional sequences (e.g. a $\mathbf{B}_1$ map) may be required to correct for RF ($\mathbf{B}_1$) field inhomogeneities (Glasser *et al.,* 2013; Vovk *et al.,* 2007). In addition, a host of pre-processing techniques exist that use image features to adjust for scanner-related intensity variations and artefacts (Vovk *et al.,* 2007).

C.2.2 Subject-related confounds

One of the most notorious subject-related confounds in structural and functional neuroimaging is subject motion. For instance, motion affects estimates of functional connectivity (Satterthwaite *et al.,* 2012) and has been related to lower estimates of GM volume and cortical thickness (Reuter *et al.,* 2015). Motion is of particular concern in developmental studies (Barkovich *et al.,* 2019; Galván *et al.,* 2012; Lerch *et al.,* 2017), because of its relationship with age (children tend to move more than adults) and sex (boys move more than girls; Alexander-Bloch *et al.,* 2016; Dosenbach *et al.,* 2017; Pardoe *et al.,* 2016; Satterthwaite *et al.,* 2012). Moreover, the motion-induced artefacts are anatomically heterogeneous, especially affecting brain regions (e.g. frontal areas) that

undergo changes during development (Alexander-Bloch *et al.,* 2016), showing biologically plausible effects. Techniques have been developed to adjust for motion during image acquisition (Van Der Kouwe *et al.,* 2006; White & Carlin, 2010; Zaitsev *et al.,* 2017), as well as during image processing (i.e. *post hoc*, Alexander-Bloch *et al.,* 2016). Among the latter, researchers tend to exclude cases with excessive motion, which may be inefficient and could bias findings (e.g. Ducharme *et al.,* 2016), or attempt to reduce the impact of subject motion, either using bespoke algorithms or through the inclusion of motion parameters in the statistical analysis.

Subject-related confounds that are less frequently considered but have shown to affect MRI estimates include hydration levels (Streitbürger *et al.,* 2012), seasonality (Miller *et al.,* 2015), and diurnal fluctuations (i.e. time-of-day, Jiang *et al.,* 2014; Karch *et al.,* 2019; Nakamura *et al.,* 2015; Thomas *et al.,* 2018; Trefler *et al.,* 2016, but also see Karch *et al.,* 2019 for a complete overview of factors). For instance, dehydration and a shorter photoperiod have been linked to, respectively, lower GM and WM volume (Streitbürger *et al.,* 2012), and lower hippocampal volume (Miller *et al.,* 2015). In addition, there is an apparent decrease in GM and WM volume (Trefler *et al.,* 2016), increase in CSF and MD (Jiang *et al.,* 2014; Thomas *et al.,* 2018) and decrease in functional connectivity and cerebral blood flow (Hodkinson *et al.,* 2014; Orban *et al.,* 2020) from morning to afternoon. A study assessing whether such subject-related confounds are predictive of within-subject variance in structural imaging measures, however, emphasised the importance of controlling for time-of-day in longitudinal analyses but found no noteworthy effect of a wide range of other physiological (e.g. hydration, caffeine intake), behavioural (e.g. stress) and scan-day-related (e.g. temperature) variables (Karch *et al.,* 2019).

MRI-derived estimates are also influenced by factors that describe subjects, such as their age (Barnes *et al.,* 2010), sex (Herting *et al.,* 2018; Paus *et al.,* 2017; Persson *et al.,* 2014), pubertal status (Goddings *et al.,* 2014; Herting *et al.,* 2017; Vijayakumar *et al.,* 2018b) or socioeconomic status (Farah, 2018; Hackman *et al.,* 2010; Yu *et al.,*

2018). In contrast to contextual confounds such as time-of-day or a subject's hydration level, which are hardly ever of interest, these demographic variables are of interest in their own right, yet are treated as confounds in studies examining brain-behaviour relationships. Head or brain size is related to regional variations in brain size and shape, but is seldom of interest and commonly treated as a confounding variable in volumetric analyses (Barnes *et al.,* 2010; De Jong *et al.,* 2017; Miller *et al.,* 2016; Nordenskjöld *et al.,* 2015; O'Brien *et al.,* 2011; Voevodskaya *et al.,* 2014).

A number of subject-related confounds can be controlled for by means of the experimental design (Jager *et al.,* 2008; Kahlert *et al.,* 2017). In randomised experiments, for instance, comparability of the groups (e.g. in terms of confounding variables such as sex) is assured by randomisation, and contextual confounds (e.g. time-of-day) are typically under control of the experimenter. Alternative methods to control for confounding are *matching*, whereby subject-groups are matched on important confounding variables, and *restriction*, where inclusion may be restricted to a certain group (e.g. <15 year olds) to remove the effect of confounding variables (e.g. age). In instances where the experimental design cannot control for confounding variables, such as in most inter-individual differences research, a researcher can use stratification by or statistical adjustment for the confounding variable (Jager *et al.,* 2008). The latter is particularly common in the neuroimaging field, and either involves regressing out confounds from the data (Dinga *et al.,* 2020; Snoek *et al.,* 2019), or including them as additional regressors in the analysis, e.g. using multiple regression or an analysis of covariance (ANCOVA). Attention should be paid, however, to the selection and treatment of confounding variables, given that erroneous adjustment for such variables may not remove but introduce bias (Dinga *et al.,* 2020; Janszky *et al.,* 2010).

Back to main text

C.3 Cognitive testing: psychometrics

When selecting a cognitive task, one typically considers whether it measures the construct of interest, i.e. its *validity*, and whether it measures this construct consistently, i.e. its *reliability*. Different types of validity (e.g. face, convergent, criterion-related validity) and reliability (e.g. test-retest, inter-rater reliability) exist (Cook & Beckman, 2006; Heale & Twycross, 2015), and testing for these psychometric properties is essential to ensure the quality of the assessment. Test-retest reliability, for instance, can be assessed by computing the Pearson correlation or intra-class correlation (ICC) between scores obtained at different times, whereas convergent validity can be assessed by computing correlations between the test score and scores on tests of theoretically similar measures.

In PA research, as well as more generally in cognitive neuroscience, the reliability and validity of cognitive tasks are rarely reported (Cook & Beckman, 2006; Nyongesa *et al.,* 2019; Wade *et al.,* 2020). This is problematic, given that such properties, but reliability in particular, have implications for the design, analysis and interpretation of studies (Draheim *et al.,* 2019; Hedge *et al.,* 2018a; Paap & Oliver, 2016; Rouder *et al.,* 2019). For instance, tasks that have reliably produced robust experimental effects (i.e. they replicated across studies or labs), may have low reliability in inter-individual differences (i.e. cross-sectional) research, where reliability refers to the extent to which a measure consistently *ranks* individuals across measurements (Draheim *et al.,* 2019; Hedge *et al.,* 2018a; Paap & Oliver, 2016; Rouder *et al.,* 2019); a phenomenon that Hedge *et al.* (2018a) termed the *reliability paradox*. Whereas experimental research minimises between-subjects variance (or maximises within-subjects variance), cross-sectional (or differential) research requires high between-subjects variability. With low between-subjects variability (and constant error variance), one is unable to distinguish between subjects, resulting in low reliability. Measures with low reliability are particularly problematic for correlational research, because the observed (sample) correlation between two measures is a function of the reliabilities of the individual measures, shown in Equation C.1 (Hedge *et al.,* 2018a; Spearman, 1910): the true correlation (r between two measures x and y is attenuated when one or both measures have non-perfect reliability

(<1).

$$r(x,y)_{sample} = r(x,y)_{true}\sqrt{reliability(x)reliability(y)} \quad \text{(C.1)}$$

Hence, researchers ought to be careful when using seemingly reliable experimental measures in cross-sectional studies, and be aware of the impact of low reliability on observed correlations.

Difference scores, in which a score on one condition is subtracted from a score on another condition, have notoriously low reliability (Draheim *et al.,* 2019; Hedge *et al.,* 2018a), despite their component scores being fairly reliable (albeit not perfectly). Difference score are commonly reported outcome measures of various EF tasks, including the Flanker task (interference effect) and tasks-switching paradigms (switch cost). For the Flanker task, for instance, a test-retest reliability of 0.48 was found for the interference score (reaction time based) in young people, while reliability estimates of approximately 0.6 have been reported for the component scores (Van Leeuwen *et al.,* 2007). Such estimates, however, vary widely, with reliability estimates as high as 0.91 and low as 0.4 being reported for interference scores in adults (Hedge *et al.,* 2018a; Wöstmann *et al.,* 2013). Importantly, reliability is a characteristic of the score, and not of the cognitive task itself (Downing, 2004). A cognitive task may yield vastly different reliability estimates, depending on the setting or subjects (Cook & Beckman, 2006). Variability in reliability estimates could also be due to factors such as homogeneity of the sample, task difficulty, trial-to-trail variation, fatigue, or motivation (Cook & Beckman, 2006; Paap & Oliver, 2016).

The cognitive assessments included in the main trial of Fit to Study were used in previous studies (Chaddock *et al.,* 2010a; Chen *et al.,* 2014b; Hillman *et al.,* 2014), and their reliability was not assessed prior to the start of the trial. Studies assessing their reliability, however, have done so primarily in adult samples and have shown mixed

results for the two-back task (varying between 0.16–0.91, see Friedman *et al.,* 2008; Jaeggi *et al.,* 2010; Van Leeuwen *et al.,* 2007), colour-shape switch task (0.57–0.78 for RT switch costs, Paap & Oliver, 2016; Paap & Sawi, 2014; Timmer *et al.,* 2018), Flanker task (0.40–0.91, for RT interference, and 0.14–0.72 for accuracy interference, Hedge *et al.,* 2018a; Paap & Oliver, 2016; Paap & Sawi, 2014; Rey-Mermet *et al.,* 2018; Tarp *et al.,* 2016; Van Leeuwen *et al.,* 2007; Wöstmann *et al.,* 2013), visual paired associates tasks (0.45–0.8, Fawns-Ritchie & Deary, 2020; Youngjohn *et al.,* 1992). To provide an *indication* of the reliability of the cognitive assessments used in Fit to Study, reliability estimates were computed using baseline and follow-up data. These are presented in Table C.1, and are in line with values reported in the literature, but should be interpreted with caution, given that they are not obtained in a separate sample and age may have affected the between-session estimates.

Table C.1: Reliability statistics of cognitive assessments. Test-retest reliability is displayed within a session (at baseline), and between sessions (baseline and follow-up). The between-session statistics are based on participants in the control group only.

	Within session[a]			Between session		
	N[b]	Pearson *r*	ICC[c]	N	Pearson *r*	ICC[c]
Flanker task						
Accuracy congruent	5400	0.74	0.73	1489	0.50	0.56
Accuracy incongruent	5399	0.74	0.73	1489	0.56	0.56
Accuracy interference	5392	0.54	0.54	1489	0.43	0.45
Reaction time congruent	5384	0.76	0.72	1489	0.52	0.50
Reaction time incongruent	5367	0.75	0.72	1489	0.49	0.49
Reaction time interference	5349	0.45	0.35	1489	0.25	0.30
Two-back task						
Accuracy	5012	0.60	0.59	1468	0.47	0.49
Reaction time	4986	0.60	0.57	1468	0.41	0.42
Colour-shape task switching						
Accuracy switch	4696	0.68	0.62	1221	0.54	0.56
Accuracy non-switch	4701	0.72	0.66	1221	0.58	0.59
Accuracy switch cost	4689	0.04	0.02	1221	0.08	0.09
Reaction time switch	4656	0.70	0.66	1221	0.43	0.45
Reaction time non-switch	4658	0.59	0.55	1221	0.39	0.38
Reaction time switch cost	4626	0.55	0.52	1221	0.41	0.43
Relational memory task						
Accuracy	—	—	—	1445	0.33	0.33
Reaction Time task						
Reaction time	—	—	—	2626	0.34	0.37

[a] The test-retest statistics were computed between blocks within the baseline session. Note, only the Flanker task, Two-back task and Colour-shape switch task consisted of two blocks.
[b] Total number of participants with data on this measure across blocks within a session. Numbers vary due to, e.g. participants having anticipatory responses on a single reaction-time condition only (e.g. congruent or incongruent), which does not allow for an average reaction time to be computed.
[c] Intra-class correlation, $ICC_{2,1}$, using a two-way random effects model, available in the *psych* package in R. The ICC accounts for consistency of performance over time (within subject) and change in average performance over time (i.e. mean change). In the commonly used Shrout & Fleiss (1979) nomenclature, this corresponds to $ICC_{2,1}$.

Besides being reliable, tasks are required to be valid. Established cognitive assessments such as the Flanker task, two-back task, colour-shape switching task and associative memory paradigms tend to have high *face validity*: they seem to measure what they are supposed to measure. The two-back task, for instance, involves processes that are typically associated with working memory, including maintenance, updating and processing of information. Despite face validity, however, there is variable evidence for other forms of validity, in line with findings for the reliability of these tasks. For instance, the colour-shape task-switching task seems to load consistently on a shifting latent variable (Friedman & Miyake, 2017; Friedman *et al.*, 2008), but low

convergent validity was reported elsewhere (Paap & Sawi, 2014). Similarly variable validity estimates have been reported for the two-back task (Jaeggi *et al.*, 2010), and typically low convergent validity for the Flanker task (Paap *et al.*, 2020; Paap & Sawi, 2014). Such low validity may be explained, at least in part, by the low reliability of interference scores (Paap *et al.*, 2020), or perhaps, as recently proposed, by the absence of a unified construct of inhibition (Rey-Mermet *et al.*, 2018; Rouder *et al.*, 2019).

Back to main text

C.4 Data analysis

C.4.1 Cluster randomised trials

In cluster randomised controlled trials (cRCT, also called group or community RCTs), intact groups or clusters of individuals are randomised (Cornfield, 1978). The unit of randomisation can be diverse, including for instance families, classrooms, schools, hospitals, study sites or communities. In Fit to Study, schools are randomised to intervention and control, and pupils are nested within schools (Figure C.1). Additional clustering arises from repeated measurements over time.

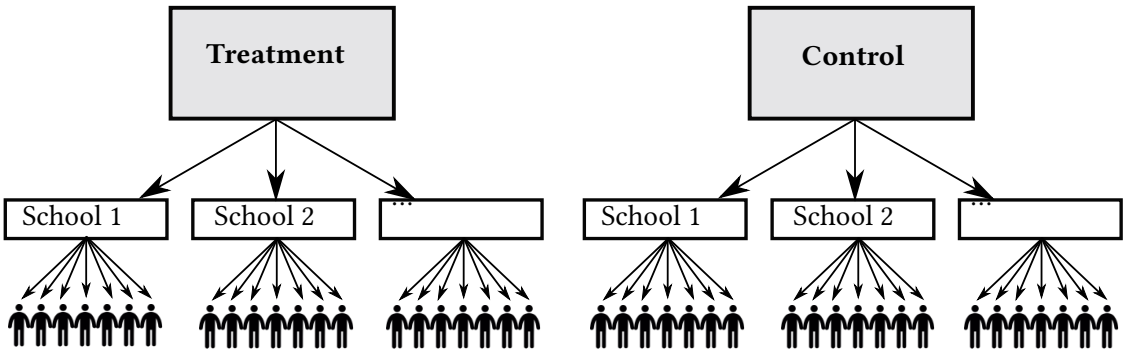


Figure C.1: Illustration of clustering in Fit to Study. Schools were nested within study arms and pupils were nested within schools The school level is commonly denoted level 2, whereas level 1 often refers to pupils. However, if repeated measurements are collected, time may be nested within pupils who are nested within schools (i.e. three levels, not shown)

Cluster RCTs have been used extensively in education and healthcare settings (Ukoumunne *et al.,* 1999), and they are the method of choice if (1) the intervention is designed to operate on a group level, as is the case for Fit to Study (increasing ecological validity), (2) the intervention cannot be delivered to individuals without the risk of experimental contamination (e.g. changes to the physical or social environment), or (3) practical considerations prevent the intervention to be delivered to individuals (Donner & Klar, 2004; Murray *et al.,* 2004; Turner *et al.,* 2017a; Turner *et al.,* 2017b). The delivery of an intervention to groups of individuals increases administrative efficiency, may reduce implementation costs and is thought to enhance compliance (Donner & Klar, 2004).

However, compared to individually randomised trials, cRCTs are often more complex to design and analyse (Donner & Klar, 2004; Murray *et al.,* 2004). Design-related challenge include obtaining consent (i.e. ethics, Taljaard *et al.,* 2013), recruitment bias (Eldridge *et al.,* 2009; Hahn *et al.,* 2005), randomisation (Ivers *et al.,* 2012; Li *et al.,* 2017a; Turner *et al.,* 2017a), and sample size calculations (Donner *et al.,* 1981; Rutterford *et al.,* 2015), whereas analysis-related obstacles include modelling of the hierarchical structure (Turner *et al.,* 2017b), as well as dealing with non-adherence and missing data (Díaz-Ordaz *et al.,* 2014; Taljaard *et al.,* 2008). These methodological challenges have been widely discussed elsewhere (Donner & Klar, 2004; Hayes & Moulton, 2017) and a thorough discussion is outside the scope of this thesis. Here, we briefly discuss one factor inherent to cRCTs that affects both the design and analysis: outcomes on members of the same group tend to be correlated (i.e. non-independent), which needs to be accounted for in the design to prevent underpowered studies and in the analysis to avoid inflated type I errors (i.e. false-positives). The degree to which participants' scores within cluster are correlated is commonly expressed as the intraclass correlation coefficient (ICC), denoted by ρ. Formally, the ICC is the proportion of total variation in the outcome that can be explained by variation between clusters (Eldridge *et al.,* 2009), which, for continuous outcomes, is provided in Equation C.2:

$$\rho = \frac{{\sigma^2}_B}{{\sigma^2}_B + {\sigma^2}_W} \tag{C.2}$$

with between cluster variance ${\sigma^2}_B$, within cluster variance ${\sigma^2}_W$, and the intracluster correlation coefficient ρ. The ICC is bound between 0 and 1, with higher values indicating that scores of individuals within a cluster tend to be more similar. While ICC estimates differ by outcome measure, they commonly vary between 0.1-0.25 for academic scores when the randomisation was performed at the school level (Hedges & Hedberg, 2007, EEF 2015).

The non-independence that comes with clustering reduces the effective sample size to be less than the total number of individual participants, resulting in lower statistical efficiency (i.e. power and precision). The reduced efficiency is a function of, amongst others, the number of clusters and the design effect —also known as the variance inflation (VI) factor —which is determined by the average cluster size and ICC (Equation C.3):

$$D = 1 + (\tilde{m} - 1)\rho \tag{C.3}$$

where D is the design effect, $\tilde{m}$ is the average cluster size and ρ is the ICC. To retain power comparable to that for individually randomised trials, the sample size should be multiplied by the VI (with $N_{effective} = \frac{N_{total}}{D}$). Hence, a greater sample size is required when the average cluster size $\tilde{m}$ is high or when the clusters are more homogeneous (i.e. higher ρ).

While researchers generally have little control over the ICC, the number (k) and size (m) of clusters are to be decided when designing the trial. Figure C.2 demonstrates that, for a given ICC, increasing the k is a more efficient way of increasing power than increasing m (Brown *et al.*, 2015; Snijders & Bosker, 2012; Vanhove, 2015). Moreover, with increasing ICC, power drops more rapidly for low than for high k (with constant cluster size). Although sample-size choices are often constrained by cost and feasibility,

such simulations indicate that cRCTs should prioritise higher k over m to attain sufficient power. Power is not only affected by the number of clusters, but also by factors such as imbalanced cluster sizes, unbalanced allocation of clusters to treatment groups and imbalances in baseline covariates (Eldridge *et al.*, 2006; Turner *et al.*, 2017a), the majority of which researchers can control for when designing a study (Turner *et al.*, 2017a).

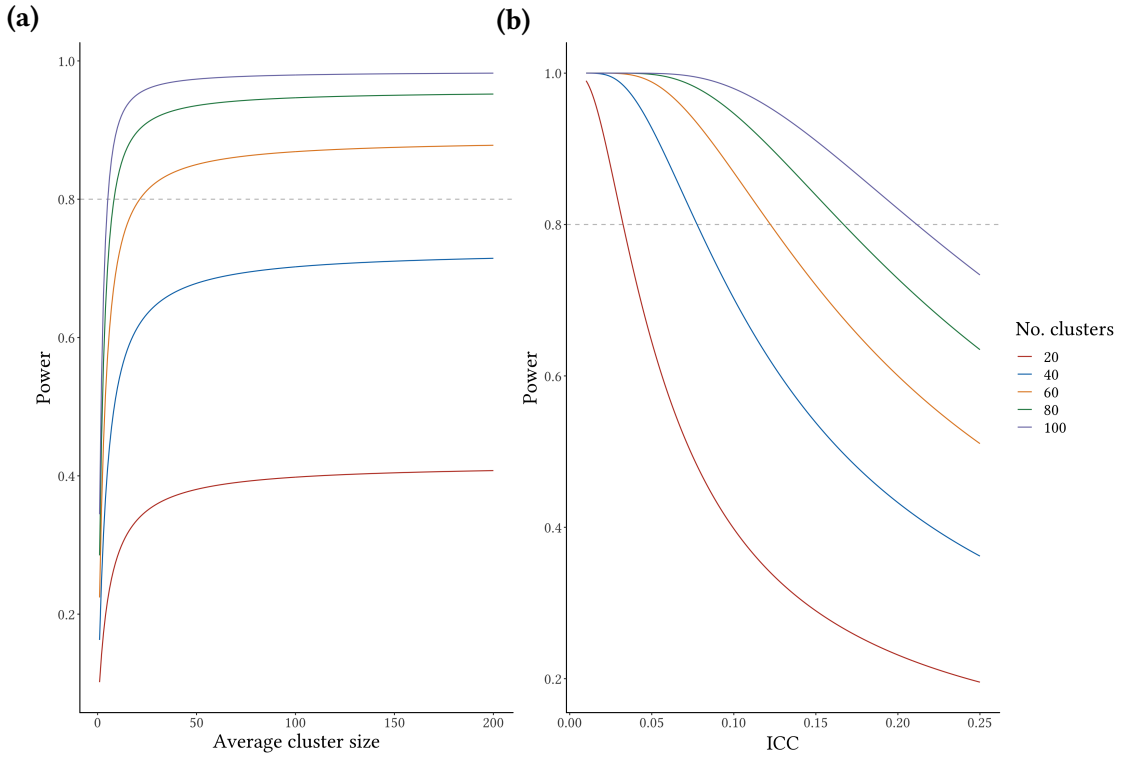


Figure C.2: Simulated effects of clustering on statistical power. Simulations of effect of cluster size, ICC and average number of clusters is displayed, assuming all clusters contain an equal number of individuals, an effect size of 0.25, two-tailed α=0.05, equal (50%) split of treatment groups, and baseline covariates explaining 40% of level 1 variance and 10% of level 2 variance, in line with power calculations of NatCen Social Research (Bartasevicius & Jabin, 2018). In (*a*) the number of clusters varied with the ICC while average cluster size was held constant at 100 participants. Power is higher with a greater number of clusters, and drops with greater ICC. In (*b*) the number of clusters varied with the average number of individuals in each cluster while the ICC was held constant at 0.1, which is a reasonable number for school-level interventions. Power is predominately affected by the total number of clusters, not cluster size. Simulations were performed in R using the *PowerUpR* package.

Besides its effect on power, the non-independence of observations due to clustering

also impact on the analysis of cRCTs. In particular, as can be observed in Figure C.3, simulations have shown that tests that ignore clustering will result in an inflated type I error (i.e. increased false positives; for further reading: Austin & Leckie, 2018; Hayes & Moulton, 2017; Kahan *et al.,* 2016; Zucker, 1990). Moreover this effect worsens with increasing ICC. Various analysis techniques have been developed that can take clustering into account (McNeish *et al.,* 2016; Molenberghs & Verbeke, 2000; Snijders & Bosker, 2012; Turner *et al.,* 2017b), including parametric mixed- or random-effects models (Laird & Ware, 1982), semi-parametric Generalised Estimating Equations (GEE, Zeger *et al.,* 1988), or non-parametric methods such as permutation tests (Gail *et al.,* 1996).

Systematic reviews that evaluated the reporting of cRCTs demonstrated large variability in whether such trials accounted for clustering (ranging from around 30–80%, Díaz-Ordaz *et al.,* 2014; Fiero *et al.,* 2016; Richardson *et al.,* 2016). To improve reporting of cRCTs, the The Consolidated Standards of Reporting Trials (CONSORT) statement has been extended for cRCTs (Campbell *et al.,* 2012) and recommendations on the reporting of ICCs have been developed (Campbell *et al.,* 2004). Moreover, many resources exist to assist researchers with the design and analysis of cRCTs, including various papers (Hemming *et al.,* 2017; Turner *et al.,* 2017a; Turner *et al.,* 2017b), books (Eldridge & Kerry, 2012; Gelman & Hill, 2006; Hayes & Moulton, 2017; Moerbeek & Teerenstra, 2015; Murray, 1998; Snijders & Bosker, 2012) and web-pages (e.g. https://clusterrandomisedtrials.qmul.ac.uk). Moreover, freely available software packages in R (R Core Team, 2020) have been developed to assist with the randomisation (e.g. Carter & Hood, 2008), sample size and/or power calculation (e.g. *powerUp!*, Dong & Peng, 2013; *powerUpR*, Bulus *et al.,* 2019, *simR*, Green & Macleod, 2016; *clusterPower*, Reich *et al.,* 2012), and analysis (e.g. *cvcrand*, Yu *et al.,* 2020; and packages that support multilevel models, see Mixed Models CRAN). Similar implementations are available for other statistical packages (see e.g. McNeish *et al.,* 2016; Snijders & Bosker, 2012, https://clusterrandomisedtrials.qmul.ac.uk).

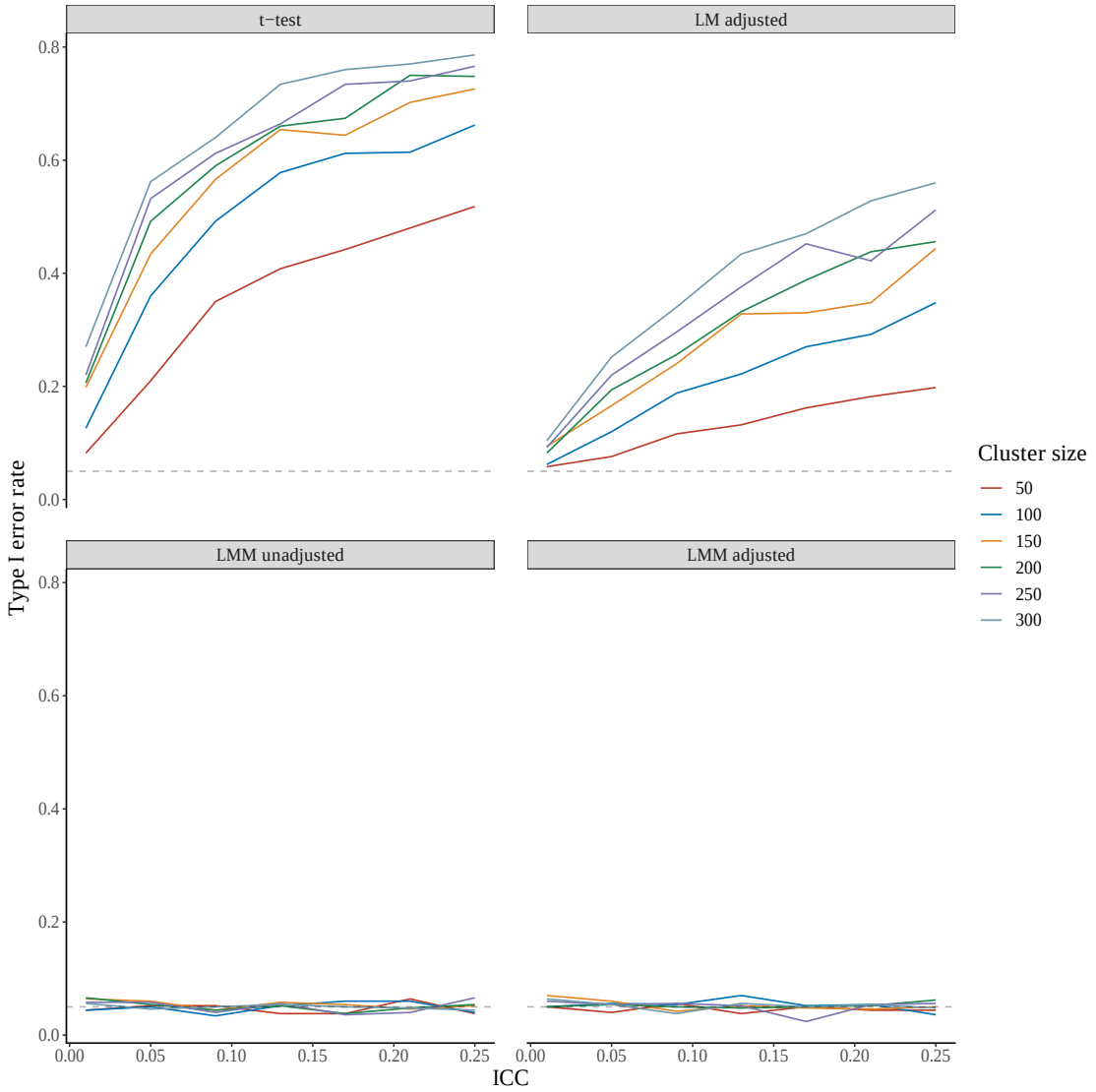


Figure C.3: Simulated effects of clustering on type I error rates with increasing ICC and different cluster sizes across various parametric models. Type I errors rates were computed for simulated cluster randomised trials (500 samples), analysed by means of (*top left*) a t-test on post-intervention scores, (*top right*) a linear model adjusted for baseline scores, (*bottom left*) a linear mixed model on post-intervention scores that takes into account the clustered nature of the data, and (*bottom right*) a linear mixed model that also adjusts for baseline scores and takes the clustered data into account. Type I error rates are inflated with increasing ICC if the model does not appropriately account for clustering (i.e. the t-test and LM). Moreover, a higher proportion of type I errors is observed with increasing cluster sizes. The type I error rate is only slightly different for different numbers of clusters (data not shown). All simulations were based on 60 clusters (50% intervention), and assumed no effect was present and the correlation between pre- and post-test scores was 0.7. Simulations were run in R, using the *cannonball* and *lme4* packages.

Back to main text

C.4.2 Missing data

C.4.2.1 Introduction to missing data

Missing data are common in nearly all fields of research, and may arise for various reasons. In some cases, the researcher plans for data to be missing (i.e. missingness is *intentional*, Van Buuren, 2018). For instance, a questionnaire is only given to a subsample of participants, or subjects are directed to skip an item following a particular response. In most cases, however, missing data is not planned for (i.e. *unintentional*). This is the case, for example, if a respondent decided to skip a question on a questionnaire or refused to complete the survey. In a multilevel context, variables may be missing for all individuals in a cluster because they were not measured (i.e. *systematically missing*), or they may be missing for some individuals in a cluster (i.e. *sporadically missing*, Resche-Rigon & White, 2018; Resche-Rigon *et al.,* 2013). Often, multiple types of missing data occur in a dataset. For example, a questionnaire may be administered to a sub-set of schools only (i.e. intentional and systematically missing data), and some pupils within a school may skip certain questions (i.e. unintentional and sporadically missing data). As a researcher, it is important to be aware of these types of missingness, and consider missing data at each stage of the research process in order to come up with strategies to minimize the impact of non-response.

While researchers are often aware of missing data in their datasets, the missing values are usually excluded or ad-hoc methods (section C.4.2.3) are used. Misconceptions about the missing data methods (Van Ginkel *et al.,* 2020), or a lack of understanding of the potential impact of missing data on analysis results may help explain why only few studies deal with missing data appropriately. Reviews on the reporting of missing data in various fields, including clinical trials (Díaz-Ordaz *et al.,* 2014; Powney *et al.,* 2014) and intervention studies (Rioux & Little, 2019), epidemiological research (Eekhout *et al.,* 2012) and quality of life outcomes (Fielding *et al.,* 2008; Rombach *et al.,* 2016), have indicated that studies often (1) lack an adequate description of the missing data, (2)

use complete-case analysis without mention, (3) do not report in sufficient detail on the missing data method, (4) use ad-hoc methods relative to more advanced missing data methods, and (5) do not perform sensitivity analysis. These findings are both surprising and worrying: inadequate handling of missing data can lead to loss of power, less precision, and biased inferences (Rubin & Little, 2002; White *et al.,* 2011).

Approaches to address missing data have developed since the 1950's (see Molenberghs *et al.,* 2014) and are now widely available to researchers. To improve missing data practices, guidelines now incorporate reporting on missing data. For instance, the *Strengthening the Reporting of Observational Studies in Epidemiology* (STROBE) guidelines (Von Elm *et al.,* 2007), the *Standard Protocol Items: Recommendations for Interventional Trials* (SPIRIT) guidelines (Chan *et al.,* 2013), and the guidelines proposed by Sterne *et al.* (2009), indicate how to report missing data in research papers. In addition, an increasing number of resources is now available to assist researchers wishing to deal with missing data in their own research. These include various books (e.g. Carpenter & Kenward, 2012; Daniels & Hogan, 2008; Enders, 2010; Molenberghs *et al.,* 2014; O'Kelly & Ratitch, 2014; Rubin & Little, 2002; Schafer, 1997; Van Buuren, 2018), research papers (e.g. Grund *et al.,* 2018b; Lang & Little, 2018; Nguyen *et al.,* 2017; Perkins *et al.,* 2018; Sterne *et al.,* 2009), and online materials and courses (e.g. http://missingdata.org.uk, https://rmisstastic.netlify.com).

Back to main text

C.4.2.2 Missing data mechanisms

In 1976, Rubin (1976) described three types, or mechanisms, of missing data which are still being used today (Graham, 2009) and have important implications for the handling and analysis of missing data.

To aid the interpretation of these mechanisms, imagine that we have collected a dataset, represented as an $n \times p$ matrix (with *n* units and *p* variables) so that $Y = (Y_{ij})$.

The complete dataset Y_{ij}^{com} can be partitioned into an observed part Y_{ij}^{obs} and a missing part Y_{ij}^{mis}. Then, suppose that R is a missing indicator matrix $R = (R_{ij})$, that indicates whether data is missing ($R_{ij} = 0$) or whether data is present ($R_{ij} = 1$). Given this notation, the three missingness mechanisms can be defined.

1. *Missing completely at random (MCAR)*: Data are missing completely at random and none of the variables are related to or predict missingness. That is, the distribution of missingness does not depend on Y_{ij}^{obs} or Y_{ij}^{mis} (equation C.4), and the missing data are thus a random subset of the dataset. MCAR is the case, for example, if the computer breaks down on the day of testing or if a questionnaire gets lost in the post.

$$P(R|Y_{ij}^{com}) = P(R) \tag{C.4}$$

2. *Missing at random (MAR)*: Data are said to be missing at random if the probability of missingness is unrelated to unobserved data conditional on the observed data (equation C.5). That is, once one has controlled for all the observed data, any remaining missingness is completely at random (Graham *et al.*, 2007). This is the case, for instance, if females skip a question on the questionnaire more often than males, and sex is observed for all participants, or if drop-out of participants in a longitudinal study is related to their baseline characteristics, but not future unobserved variables.

$$P(R|Y_{ij}^{com}) = P(R|Y^{obs}) \tag{C.5}$$

3. *Missing not at random (MNAR)*: Data are said to be missing not at random if the probability of missingness depends on unobserved and observed data (equation C.6). This is the case, for example, if participants from a lower socio-economic background skip questions about income or status, and no information about socioeconomic background has been collected. Some strategies exist to deal with data that are MNAR. For example, one can try to collect more data on the cause(s) of the missingness. Given that this may not be feasible in many occasions, one can attempt to model the missingness mechanism and perform sensitivity analyses to examine how robust the results are to various scenarios (Kenward & Carpenter, 2007).

$$P(R|Y_{ij}^{com}) = P(R|Y_{ij}^{obs}, Y_{ij}^{mis}) \tag{C.6}$$

MCAR and MAR are *ignorable* mechanisms. That is, there is no need to model their non-response mechanisms for inferences to be unbiased and valid. MNAR on the other hand, is *nonignorable*, and if not modelled correctly, yields biased parameter estimates (Lang & Little, 2018).

It is important to note that one can only test for MCAR. That is, one can demonstrate that data are not MCAR if participants with missing data in some variables show mean differences with other variables in the dataset. The absence of mean differences, however, does not provide evidence for the data to be MCAR (Raykov, 2011). Testing whether data are MAR or MNAR is impossible, given that the data that are needed for such a test are missing. The MAR assumption can be made more plausible, however, by collecting explanatory variables that may be related to missingness (White *et al.*, 2011).

C.4.2.3 Methods to deal with missing data

Many methods have been developed to deal with missing data. A distinction is typically made between somewhat older, *ad-hoc* methods (also termed *unprincipled* methods, Kenward & Carpenter, 2007) that provide a simple and quick fix of the missing data

issue but only give valid inferences under certain assumptions and settings, and those that are more advanced and commonly based on statistical models (*principled* methods), including multiple imputation, likelihood and bayesian methods, and weighting approaches, as well as non-parametric alternatives such as those based on random Forests and single value decomposition (SVD). While all of these methods have been described in detail in several textbooks (e.g., Enders, 2010; Molenberghs *et al.,* 2014; Rubin & Little, 2002; Van Buuren, 2018), we briefly review the most common approaches here.

Quick-and-dirty, antiquated or ad hoc approaches

1. **Listwise- and pairwise deletion** In *list-wise deletion*, also known as complete-case analysis, all cases with one or more missing values on the analysis variables are excluded from statistical tests. This is the default setting in many statistical analysis packages and has historically been the most popular method to deal with missing data. While complete-case analysis is easy to apply, one should carefully consider the missingness mechanism before using this approach. If the data are MCAR, complete-case analysis will result in unbiased estimates (Graham, 2009). There will, however, be loss of power given that only a subset of the entire sample is included in the analysis. If the data are not MCAR, complete-case analysis can lead to biased estimates because the subsample may not be representative of the entire population (Dziura *et al.,* 2013; Rubin & Little, 2002; Schafer & Graham, 2002; Van Buuren, 2018). Only in special cases can complete-case analysis be used and provide unbiased estimates (Hughes *et al.,* 2019; Jakobsen *et al.,* 2017; Van Buuren, 2018; Van Ginkel *et al.,* 2020; White & Carlin, 2010). One such case is if missingness occurs in either the outcome, one or more predictors, or both, and the probability of being a complete case is independent of the outcome, conditional on the predictors.

   In an attempt to resolve the wastefulness of complete-case analysis, one can decide to calculate the statistics of interest on all the observed data. This method is known as *pair-wise deletion* or *available-case analysis*. Pair-wise deletion can be used to compute means and variances, but is most commonly used in the

context of creating correlation or covariance matrices. In this case, correlation or covariance values are computed from cases with observed data on both variables. Consequently, fewer cases are discarded from computations which results in slightly higher power. Moreover, if the data are MCAR, estimates of the mean and variance may be unbiased (Van Buuren, 2018). While pair-wise deletion includes all available data, it does not come without its own issues. For instance, statistical estimates may be biased if the data is not MCAR. In addition, it is unclear what sample size to use for computing standard errors: each pair of correlation values can be based on a different number of cases. Pair-wise deletion is therefore not recommended as the primary solution to deal with missing data (but see: Graham, 2009).

2. **Single imputation** An alternative to exclusion of missing data is to "fill in", that is to *impute*, the gaps in the data. Data imputation has several desirable properties, including higher power (no data is removed) and the possibility of using other available information to replace missing data with synthetic values (Schafer & Graham, 2002). However, not all imputation methods give unbiased results and are easy to implement. Several imputation methods have been developed that result in missing values being replaced by a single synthetic value. A brief overview of some of these methods is presented here.

   (a) **Mean imputation** In mean imputation (also known as *unconditional mean substitution*) the missing data of a variable is replaced by the mean value of the observed cases of that variable. This method provides a "quick and dirty" solution for missing data, but is generally not recommended. While the estimate of the mean is unbiased under MCAR, it does not preserve conditional relationships between variables, leads to an underestimation of the variances and covariances, and will lead to biased estimates in most cases (Schafer & Graham, 2002; Van Buuren, 2018).

   (b) **Regression imputation** In regression imputation (also known as *conditional mean substitution*) the observed data is used in a regression model

to predict values for cases with missing data. While this method incorporates information from other observed variables and is therefore expected to make better predictions of the missing data values, it is only valid in certain conditions. If the data are MCAR or MAR and the model includes variables that predict missingness, the regression imputation method results in unbiased estimates of the regression coefficients. The imputed values, however, all fall on the regression line, which is unrealistic in most cases and leads to an overestimation of correlation coefficients and underestimation of the variance (Enders, 2010; Rubin & Little, 2002). An extension of simple regression imputation is *stochastic regression imputation*, in which a random error (such as the residual) is added to the predicted values. Like simple regression imputation, this method yields unbiased regression coefficients if the data are MCAR or MAR. Furthermore, the addition of the random error prevents perfect prediction of imputed values and ensures that correlation coefficients are unbiased under MAR too.

(c) **Last observation carried forward** Last observation carried forward (LOCF) is a way to deal with missing data in a longitudinal framework by taking the last (i.e. previous) value to replace the missing value. While this method has been popular in clinical trials, LOCF is generally not recommended for missing data imputation (for details see Molenberghs *et al.*, 2014). LOCF can result in biased estimates, even under MCAR, and typically leads to underestimation of the variance (Dziura *et al.*, 2013).

While the single imputation approaches are easy to implement, they only lead to valid inference under specific conditions that are often difficult to establish in practice (for an overview of the bias under different approaches, see Dziura *et al.*, 2013; Schafer & Graham, 2002; Van Buuren, 2018). Moreover, the majority of these approaches understate the level of uncertainty (i.e. too low standard errors) because they do not distinguish between imputed and observed variables (Schafer & Graham, 2002). Multiple imputation incorporates adjustment for the uncertainty in the imputations themselves and is therefore the method of choice for most researchers.

Multiple imputation

Multiple imputation consists of three phases: an imputation phase, an analysis phase and a pooling phase (Figure C.4). In the imputation phase, several copies (M) of the data are created in which missing entries are replaced by imputed values. The imputation procedure typically uses a regression model (i.e. the imputation model) to specify a distribution from which plausible synthetic values are drawn. That is, M random draws are made from the posterior (predictive) distribution of the missing data, given the observed data, to generate M imputed datasets (Enders *et al.,* 2018b). Various multiple imputation methods have been developed, that differ in their imputation model and are typically based on one of two frameworks: joint model imputation (JM, Rubin, 1987; Schafer, 1997) or fully condition specification (FCS) imputation (Raghunathan *et al.,* 2001; Van Buuren *et al.,* 2006, but other approaches have been proposed too, e.g. Audigier *et al.,* 2017). JM draws synthetic values for all incomplete variables simultaneously from a multivariate distribution, whereas FCS, also known as Multiple Imputation by Chained Equations (MICE) or sequential regression multivariate imputation, imputes multivariate missing data on a variable-by-variable basis using a series of univariate regression models (Raghunathan *et al.,* 2001; Resche-Rigon & White, 2018; Van Buuren *et al.,* 2006; Van Buuren & Groothuis-oudshoorn, 2011). JM has stronger theoretical underpinnings and is computationally less intensive, but is less flexible compared to FCS. In terms of performance, FCS and JM perform equally well with continuous and binary variables, but FCS outperforms JM when the data are categorical (Kropko *et al.,* 2014; Lee & Carlin, 2010).

Following the imputation phase, each of the multiple imputed datasets are analysed using standard statistical techniques (e.g. a mixed effects model), resulting in m sets of parameter estimates. These estimates will likely differ between imputed datasets, and reflect the uncertainty in the imputations. In the pooling phase, the estimates are combined, or pooled, across imputations, according to Rubin's rules (Rubin, 1987). Point estimates, Q (e.g. β), are averaged across imputed datasets (Equation C.7):

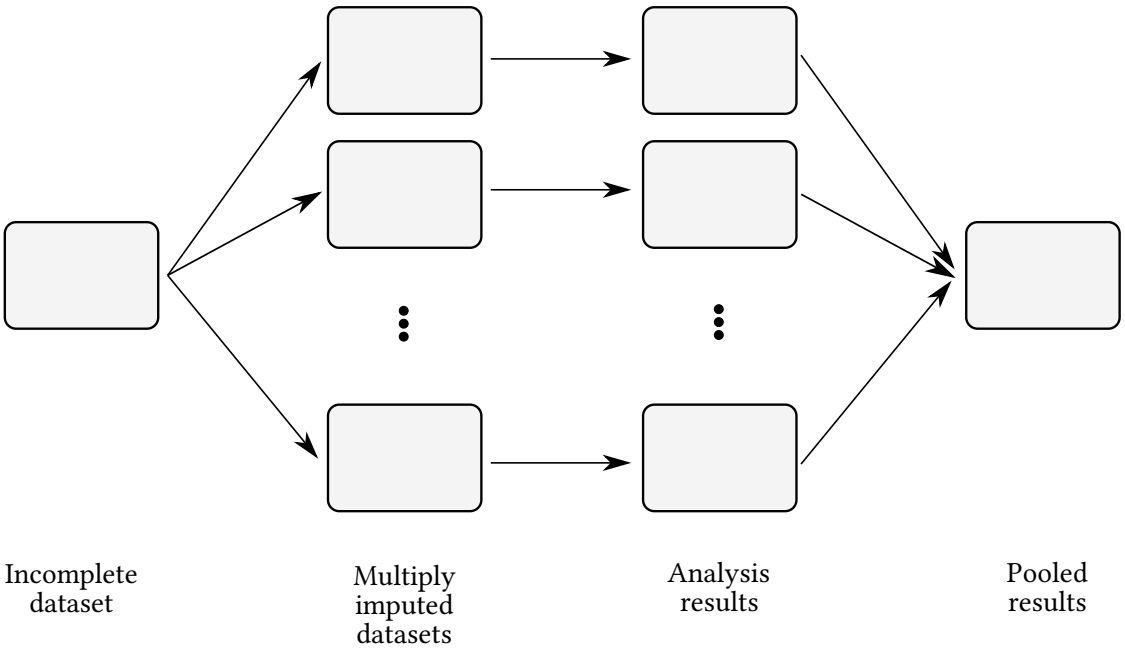


Figure C.4: Principle of multiple imputation. From left to right, from a single incomplete dataset several multiply imputed datasets are created. Each is analysed individually and the results are pooled to obtain unbiased estimates with appropriately adjusted variance. Adapted from Van Buuren (2018).

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\hat{Q}_l \tag{C.7}$$

where, l corresponds to the l-th imputation.

To obtain the variance parameters, U ($var(\beta) = se(\beta)^2$), the *within-imputation* (Equation C.8) variance and the *between-imputation* variance (Equation C.9) are combined into a total variance, T, (Equation C.10).

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m}\hat{U}_l \tag{C.8}$$

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})^T(\hat{Q}_l - \bar{Q}) \tag{C.9}$$

$$T = \bar{U} + B + \frac{B}{m} \tag{C.10}$$

Multiple imputation is available in many software packages, and with various books and online resources providing insight into its procedure, has become increasingly accessible to non-statisticians. However, researchers using multiple imputation are likely to encounter various practical challenges and are required to make several (non-trivial) modelling choices (e.g. see Enders *et al.,* 2018b; Molenberghs *et al.,* 2014; Van Buuren, 2018; Von Hippel, 2009; Von Hippel, 2020). For instance, researchers should assess the plausibility of the MAR assumption, decide on which variables to include in the imputation model (e.g. Van Buuren *et al.,* 1999) and whether to transform such variables (e.g.Lee & Carlin, 2017; Von Hippel, 2013), establish the form of the imputation model, and determine the number of iterations (of the algorithm) and imputations (e.g. Von Hippel, 2020).

An important notion in multiple imputation literature is that the imputation model should more general than the (substantive) analysis model, often referred to as *congeniality* (Meng, 1994, or *compatibility*, albeit its definition is grounded in mathematics, for details see e.g. Enders *et al.,* 2020; Van Buuren, 2018). If the analysis model contains particular relations between variables, such as power terms or interactions, they should be included in the imputation model to obtain unbiased results (Von Hippel, 2009). Failing to include such associations biases the parameters estimates towards zero. Variables that are not part of the analysis model, but are either related to missingness or explain a large amount of variance in variables of interest, should be included as auxiliary variables. Such variables aid imputations and make the MAR assumption more plausible (Graham, 2009; Hughes *et al.,* 2019; Lang & Little, 2018).

Following the concept of congeniality, a multilevel imputation model should be used if the data is clustered and analysed with a multilevel model. Multilevel missing data may yield complexities beyond those described for single-level imputation. For instance, missing data can now be present in predictors at multiple levels, as well as the class variable and the outcome. Moreover, the nature of the data, such as a small number of clusters, small cluster sizes, or systematically missing data could pose issues

for multiple imputations. It has been suggested that dummy-indicator variables could be used to represent the multilevel structure when imputing missing data (Drechsler, 2015). However, such an approach may distort the true variability between clusters and could result in biased parameters estimates (Andridge, 2011; Lüdtke *et al.,* 2017). In the past years, the multilevel multiple imputation field has evolved rapidly (e.g. Audigier *et al.,* 2018; Goldstein *et al.,* 2009; Grund *et al.,* 2016; Grund *et al.,* 2018a; Grund *et al.,* 2018b; Huque *et al.,* 2018; Mistler & Enders, 2017; Van Buuren, 2018; Yucel, 2008) and various implementations of multilevel multiple imputation now exist (e.g. JOMO Quartagno & Carpenter, 2016; Quartagno & Carpenter, 2018, PAN Grund *et al.,* 2016; Schafer & Yucel, 2002, MICE Van Buuren & Groothuis-oudshoorn, 2011 in R, BLIMP Enders *et al.,* 2018a, MPlus Muthén & Muthén, n.d.) that appropriately account for the multilevel structure in the data and deal with many (but not all) of the challenges inherent to hierarchical data.

Alternative approaches While multiple imputation is frequently used to deal with missing data, other methods have been developed too and may be equally appropriate. It is outside the scope of this thesis to discuss such methods at length; a very general overview is provided below.

- **Parametric methods: Likelihood-based and (fully) Bayesian approaches**. These methods explicitly model the observed data, without the need to impute the missing data. Full information maximum likelihood (FIML) is one such approach, whereby parameters estimates are obtained that maximise the fit to the observed data. FIML is available in various structural equation modelling packages, e.g. MPlus, and is easy to use, but tends to be less flexible than multiple imputation (Grund *et al.,* 2018b; Lang & Little, 2018) and may not be preferred for multilevel models that contain interactions (Enders *et al.,* 2020).

- **Semi-parametric and non-parametric approaches** Many other approaches have been developed to deal with missing data that are not or less model-based compared to likelihood-based and Bayesian approaches. Weighting methods, for

instance, were originally developed for survey data, where responses of respondents are weighted to account for unequal selection probabilities. The inverse of the selection probability was used as a weight. One can similarly re-weight complete-cases to account for nonreponse (under MAR), however, it uses listwise deletion and estimation of variance parameters requires specialist software (Dziura *et al.,* 2013). Imputation methods based on classification and regression trees (e.g. missForrest in R by Stekhoven & Bühlmann, 2012) and principal component analysis (PCA) or singular value decomposition (SVD, e.g. Husson *et al.,* 2019; Josse *et al.,* 2011) have shown great promise to replace missing values with credible synthetic values, even in the context of multilevel data structures, an example of which is presented in chapter 5. Most of these methods, however, only impute the data once, and therefore do not account for the uncertainty caused by the missing data, resulting in too small confidence intervals. Other tree- and PCA-based methods are available that do multiply impute the missing data (e.g. Burgette & Reiter, 2010; Doove *et al.,* 2014; Josse *et al.,* 2011), and the machine learning community is rapidly developing new techniques to impute missing data (for an overview see Lin & Tsai, 2020).

In this thesis, we use a combination of missing data approaches, which assume that data is missing at random (MAR). In particular, for the studies involving the Fit to Study main trial data (chapter 4 and chapter 6), we use multiple imputation. The cross-sectional analyses in chapter 4 contain complex three-way interactions, for which model-based imputation is used (based on the JM), whereas the multiple imputations for chapter 6 are split by treatment group to allow the covariance structures to vary within each group (FCS based). The sample size of the brain-imaging sub-study, combined with the rich set of measures, renders large-scale multiple imputations complex and in some cases unfeasible (e.g. due to non-convergence). We therefore used a pragmatic approach to deal with missing cases in those chapters. In particular, the primary cross-sectional analyses presented in chapter 5 contain many variables, which resulted in non-converging imputation models. SVD-based imputation was therefore used in this chapter. The primary intervention analyses presented in chapter 7, however, used

multiple imputation (based on JM, split by treatment group), specific to each research question. To assess the robustness of the imputation results, complete-case analyses are presented in all chapters.

Back to main text

D

Supplementary materials to chapter 4: *Associations between cardiorespiratory fitness, physical activity and cognitive function in a large sample of young British adolescents*

Contents

D.1 Methods 320
D.1.1 Outcome measures 320
D.1.2 Specification curve analysis 322
D.1.3 Missing data 324
D.1.4 Software 328
D.2 Results 328
D.2.1 Data cleaning 328
D.2.2 Descriptives 329
D.2.3 Univariate analysis 334
D.2.4 Specification curve analysis 338
D.2.5 Canonical correlation analysis 360

D.1 Methods

D.1.1 Outcome measures

D.1.1.1 PA questionnaires

Table D.1 details the PA questionnaire items that were used in this study.

Table D.1: Physical activity questionnaire items

Measures	Item	Reference
Past week physical activity	*In the past week, on how many days have you done a total of 60 minutes or more of physical activity, which was enough to raise your breathing rate? This may include sport, exercise, and brisk walking or cycling for fun, or to get to and from places.* (1 = never to 7 = every day).	Scott *et al.* (2015)
Habitual physical activity	*Thinking about the past six months, how often have you been physically active for an hour every day during a typical week in school term?* (1 = never to 7 = always).	Hagger *et al.* (2001)
Attitudes towards physical activity	*For me, being physically active for an hour every day in a typical week during term would be:* *1. Boring - - - - - - - Fun* *2. Dissatisfying - - - - - - - Satisfying* *3. Stressful - - - - - - - Relaxing* (7 points, incl. extremely/very/quite/neutral)	Michie *et al.* (2008)

Back to main text

D.1.1.2 CRF assessment

Schools that did not complete the 20 meter shuttle run, completed the 12 min Cooper run test instead. The Cooper Run test has a similar validity compared to the 20MSR (r = 0.78, Mayorga-Vega *et al.*, 2016) and requires pupils to run or walk as far as possible in 12 minutes around a measured track (between 80 m and 400 m). The total distance run is typically used as an outcome measure.

Of all schools, only four completed the Cooper Run test (n = 561 of n = 692 in total) at baseline, and 63 completed the 20MSR (n = 10,697). Given the difference in scale and units of the outcome measures (20MSR: laps, versus Cooper: total distance), the scores of these tests cannot easily be combined. Classically, researchers have used estimating equations to convert 20MSR (Batista *et al.*, 2013; Léger *et al.*, 1988; Mahar

et al., 2011; Mahar *et al.,* 2018; Matsuzaka *et al.,* 2004; Melo *et al.,* 2011; Paradisis *et al.,* 2014; Tomkinson *et al.,* 2017) and Cooper Test performance (Cooper, 1968; Grant *et al.,* 1995) to estimated $\dot{V}O_2$max values. To this end, they related field test performance to the gold-standard $\dot{V}O_2$max obtained with a maximal graded exercise test. While the conversion to a common metric ($\dot{V}O_2$max) is attractive, we decided not to pursue this for several reasons.

1. A large number of estimating equations exist, particularly for 20MSR performance (>17 equations, see Tomkinson *et al.,* 2017), but selecting one that is appropriate for the Fit to Study sample is not straightforward. Depending on the fitness protocol used, these equations vary in their predictors, most commonly including one or more of sex, age, speed or body weight, and more generally in their accuracy of predicting $\dot{V}O_2$max (Batista *et al.,* 2013; Melo *et al.,* 2011). Moreover, these equations have been established in samples that may not generalise to the population.

2. Estimating equations may underestimate (Grant *et al.,* 1995; Penry *et al.,* 2011) or overestimate (Scott *et al.,* 2019) an individual's real $\dot{V}O_2$max. This is particularly problematic if the degree of under / over estimation differs between the Cooper test and 20MSR. A comparison of estimated $\dot{V}O_2$max values from 20MSR and the Cooper test in healthy adults demonstrated that, although 20MSR underestimated an individual's level of CRF, the converted Cooper test results were biased: underestimation of $\dot{V}O_2$max values in low fit individuals (i.e. low real $\dot{V}O_2$max) and overestimation of $\dot{V}O_2$max in high fit individuals (Penry *et al.,* 2011).

3. The $\dot{V}O_2$max data that has been collected as part of the brain imaging sub-study (details in chapter 2) in a sub-sample of 61 participants is not appropriate for the creation of conversion equations. Despite high correlations between the 20MSR and $\dot{V}O_2$peak (Figure D.1, Spearman's ρ = 0.77, $p \leq 0.01$), no such relationships were found in the small sample (n = 9) of participants who completed the Cooper test (Spearman's ρ = 0.23, p = 0.56). Moreover, participants of the sub-study are generally more fit compared to the main trial sample. Consequently, an

equation that is based on the sub-sample may not be valid in the main trial sample. Nonetheless, the strong relationship between 20MSR and $\dot{V}O_2$peak suggests that 20MSR is a valid measure of CRF. Hence, the raw 20MSR results were treated as the main measure of CRF.

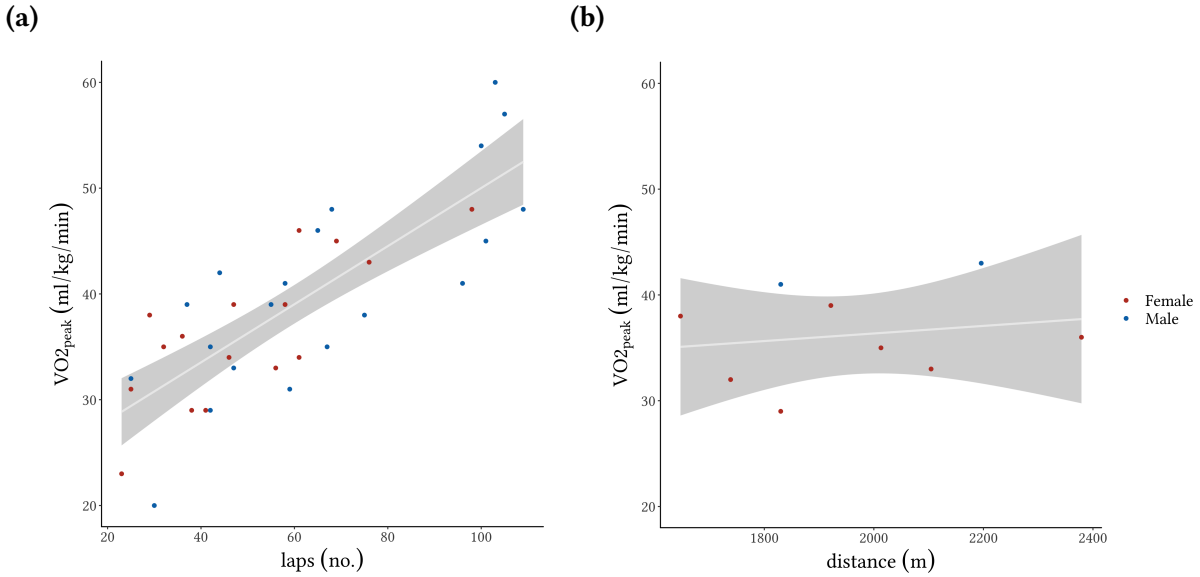


Figure D.1: Validation of 20 MSR as an estimate of CRF. *(a)* Higher 20MSR scores are related to higher levels of $\dot{V}O_2$peak (Spearman's ρ = 0.77) —the gold standard measure of CRF —in a sub-sample of participants (n = 36, from 7 schools) who additionally completed a maximal exercise test as part of their enrolment in the brain imaging sub-study. *(b)* In sub-study pupils who completed the Cooper test (n = 9, from 1 school), no such relationship was observed. Grey shaded area around linear fit represent 95% confidence interval of predicted values.

Back to main text

D.1.2 Specification curve analysis

To inspect the robustness of the relationships between cardiorespiratory fitness or physical activity and cognitive outcomes to variations in (1) the cleaning procedure, and (2) the selected predictor, outcome and nuisance variables, we used specification curve analysis (SCA). SCA is an analytical approach, proposed by Simonsohn *et al.* (2015), that deals with the flexibility in data analysis (i.e. researchers degrees of freedom) by running all theoretically justifiable, statistically valid and non-redundant models (Simonsohn *et al.*, 2020). Rather than reporting the results of a single analysis, SCA

involves presenting the outcomes of all justifiable models, also termed specifications. SCA typically consists of three steps: (1) identify all defensible and valid specifications, (2) implement and display the results of all specifications, (3) perform statistical inference across all specifications.

Here, we limit the analyses to steps 1 and 2, given that (1) our main aim was to provide insight into the variability in outcomes when considering other defensible analytical pathways, not performing statistical inference across the specifications, (2) the various cleaning procedures yield different sub-samples of the data, resulting in statistical testing across possibly non-comparable sub-groups (e.g. it is known that boys and those from lower socioeconomic status (i.e. eFSM) were more likely to be removed during cleaning, see Appendix D.2.1). Moreover, by showing the results for a different set of predictors (i.e. dichotomising continuous variables) and outcome variables (e.g. difference scores and/or pure reaction time measures) that are not recommended (Draheim *et al.,* 2019), yet justifiable, and have been used in previous studies, we hoped to provide a useful resource for future studies investigating such relationships.

To identify all justifiable specifications, we made four key analytical decisions: (1) how to perform the cleaning of the cognitive assessments, (2) how to measure PA and CRF, (3) how to measure cognition, and (4) how to correct for nuisance variables. For the primary analyses, we used a conservative approach to cleaning whereby participants with <30% of valid trails were excluded, under the assumption that a minimum number of trials was required to obtain a reliable estimate of performance. With SCA, we inspected the robustness of the relationships to variation in such exclusion thresholds (i.e. 0, 30% or 90% valid trials) and accuracy-based cleaning (i.e. 0 or at chance performance). We further inspected the robustness of the relationships to dichotomising predictor variables, variation in outcome measures (e.g. accuracy, RT, difference scores) or nuisance variables (i.e. no nuisance, task-specific nuisance variables, demographic variables, or task and demographic variables). Further details on the analytic choices

are provided in Appendix D.2.4.

Prior to running the SCA, we excluded participants with data collection errors from the dataset. We provide the median and range of standardized estimates of each predictor across cleaning procedures and outcomes, as well as the proportion of specifications that yielded significant results.

Back to main text

D.1.3 Missing data

Table D.2 presents the proportion of missing data of task related variables. We used model-based multilevel imputation available in Blimp (Enders *et al.*, 2020) to multiply impute the missing data in the presence of interaction terms. The variables included in each imputation model (i.e. a model per cognitive task and predictor, resulting in 10 imputation models) are shown in Table D.3. Each imputation model contained all variables that were part of the analysis model, as well as variables that were correlated (r >0.4) with variables in the analysis model or related with missingness in variables in the analysis model, and the treatment status indicator.

Table D.2: Missing data (proportion) of task related variables

	No missing	% missing
age	0	0
sex	0	0
eFSM	0	0
20MSR pretest	5320	33.21
Self-reported PA pretest	6318	39.45
20MSR posttest	7903	49.34
Reaction time task pretest	8077	50.43
Self-reported PA posttest	9619	60.05
Reaction time task posttest	9878	61.67
Flanker task pretest	10569	65.99
Relational memory task pretest	10699	66.8
Two-back pretest	11005	68.71
Colour-shape task switch pretest	11295	70.52
Flanker task posttest	11500	71.8
Two-back posttest	11516	71.9
Relational memory task posttest	11525	71.95
Colour-shape task switch posttest	12081	75.43

Abbreviations: 20MSR = 20meter shuttle run, eFSM = eligible for free school meals, PA = physical activity

Table D.3: Variables used in the imputation models for each of the primary univariate analysis models

	Flanker task	Flanker task	task-switching task	Relational memory task	Reaction time task
CRF as predictor models					
Analysis model	school	school	school	school	school
Analysis model	Incongruent accuracy (tp_1)	Accuracy (tp_1)	Accuracy switch (tp_1)	Accuracy (tp_1)	RT (tp_1) - transformed
Analysis model	CRF (tp_1)	CRF (tp_1)	CRF (tp_1)	CRF (tp_1)	CRF (tp_1)
Analysis model	sex	sex	sex	sex	sex
Analysis model	eFSM	eFSM	eFSM	eFSM	eFSM
Analysis model	age	age	age	age	age
Auxiliary	Treatment indicator	Treatment indicator	Treatment indicator	Treatment indicator	Treatment indicator
Auxiliary	CRF (tp_2)	CRF (tp_2)	CRF (tp_2)	CRF (tp_2)	CRF (tp_2)
Auxiliary	RT congruent (tp_1)	RT (tp_1)	RT heterogeneous (tp_1)	Accuracy (tp_2)	RT (tp_2)
Auxiliary	RT incongruent (tp_1)	Accuracy (tp_2)	Accuracy non-switch (tp_1)		
Auxiliary	Accuracy congruent (tp_1)		Accuracy switch (tp_2)		
Auxiliary	Accuracy incongruent (tp_2)		Accuracy non-switch (tp_2)		
PA as predictor models					
Analysis model	school	school	school	school	school
Analysis model	Incongruent accuracy (tp_1)	Accuracy (tp_1)	Accuracy switch (tp_1)	Accuracy (tp_1)	RT (tp_1) - transformed
Analysis model	PA (tp_1)	PA (tp_1)	PA (tp_1)	PA (tp_1)	PA (tp_1)
Analysis model	sex	sex	sex	sex	sex
Analysis model	eFSM	eFSM	eFSM	eFSM	eFSM
Analysis model	age	age	age	age	age
Auxiliary	Treatment indicator	Treatment indicator	Treatment indicator	Treatment indicator	Treatment indicator
Auxiliary	RT congruent (tp_1)	RT (tp_1)	RT heterogeneous (tp_1)	Accuracy (tp_2)	RT (tp_2)
Auxiliary	RT incongruent (tp_1)	Accuracy (tp_2)	Accuracy non-switch (tp_1)	Q intent to be active (tp_1)	Q intent to be active (tp_1)
Auxiliary	Accuracy congruent (tp_1)	Q intent to be active (tp_1)	Accuracy switch (tp_2)	Q habitual PA (tp_1)	Q habitual PA (tp_1)
Auxiliary	Accuracy incongruent (tp_2)	Q habitual PA (tp_1)	Accuracy non-switch (tp_2)	Q attitudes to PA (tp_1)	Q attitudes to PA (tp_1)
Auxiliary	Q intent to be active (tp_1)	Q attitudes to PA (tp_1)	Q intent to be active (tp_1)	Q willingness (tp_1)	Q willingness (tp_1)
Auxiliary	Q habitual PA (tp_1)	Q willingness (tp_1)	Q habitual PA (tp_1)	Q similar to active (tp_1)	Q similar to active (tp_1)
Auxiliary	Q attitudes to PA (tp_1)	Q similar to active (tp_1)	Q attitudes to PA (tp_1)	Q habitual PA (tp_2)	Q habitual PA (tp_2)
Auxiliary	Q willingness (tp_1)	Q habitual PA (tp_2)	Q willingness (tp_1)	Q PA (tp_2)	Q PA (tp_2)
Auxiliary	Q similar to active (tp_1)	Q PA (tp_2)	Q similar to active (tp_1)	Q intent to be active (tp_2)	Q intent to be active (tp_2)

Continued on next page

Table D.3 – continued from previous page

	Flanker task	Flanker task	task-switching task	Relational memory task	Reaction time task
Auxiliary	Q habitual PA (tp_2)	Q intent to be active (tp_2)	Q habitual PA (tp_2)		
Auxiliary	Q PA (tp_2)		Q PA (tp_2)		
Auxiliary	Q intent to be active (tp_2)		Q intent to be active (tp_2)		

CRF = cardiorespiratory fitness, eFSM = eligible for free school meals, PA= physical activity, RT = reaction time, tp_1 = timepoint 1, tp_2 = timepoint 2, Q = questionnaire

Back to main text

D.1.4 Software

Univariate analyses were performed in R (version 6.0) using multilevel models available in *lme4* and *lmerTest*. Interactions were inspected using *emtrends* available in *emmeans*. The data of the primary analyses were multiply imputed using Blimp and subsequently analysed in R using the *mitml* package. The SCA was performed using code available in the *specr* package. The CCA testing procedure was performed in MATLAB (version 6.0), using code available on GitHub (link: CCA script).

Back to main text

D.2 Results

D.2.1 Data cleaning

For the primary analyses, we used a conservative approach to data cleaning that aimed to remove careless responders and data collection errors only. Table D.4 shows the number of participants that completed each assessment pre and post data cleaning.

Table D.4: Number (proportion) of participants that completed each assessment pre and post data cleaning

	Unclean		Clean	
	No.	% of sample	No.	% of sample
CRF	10697	66.79	10697	66.79
Flanker task	5713	35.67	5448	34.01
Questionnaire	9805	61.22	9775	61.03
Reaction time task	8059	50.32	7940	49.57
Relational memory task	5518	34.45	5318	33.20
task-switching task	5191	32.41	4722	29.48
Two-back task	5323	33.23	5012	31.29

Abbreviations: CRF = cardiorespiratory fitness

We assessed whether participant characteristics could predict exclusion during data cleaning using a linear mixed effects model with logit link function available in *lme4*. Results of the models are presented in Table D.5. Being male was predictive of being

excluded during data cleaning for the majority of cognitive assessments (relational memory task, Flanker task, two-back task, colour-shape switch task, PA questionnaire), as was being eligible for free school meals (relational memory task, Flanker task, two-back task, and colour-shape switch task) and older age (PA questionnaire).

Table D.5: Models testing whether baseline characteristics predict exclusion during cleaning

	OR	95% CI	*p*
Relational memory			
sex (reference = female)	3.13	2.28, 4.3	0
eFSM (reference = non eligible)	2.42	1.71, 3.42	0
age	0.79	0.48, 1.29	0.35
Processing speed (Reaction time task)			
sex (reference = female)		0.87, 1.86	0.22
eFSM (reference = non eligible)	1.13	0.69, 1.87	0.62
age	0.91	0.49, 1.68	0.76
Inhibitory control (Flanker task)			
sex (reference = female)	3.14	2.37, 4.17	0
eFSM (reference = non eligible)	1.79	1.3, 2.46	0
age	0.91	0.59, 1.4	0.68
Working memory (two-back task)			
sex (reference = female)	1.64	1.28, 2.1	0
eFSM (reference = non eligible)	1.9	1.4, 2.56	0
age	1.16	0.78, 1.72	0.47
Cognitive flexibility (colour-shape task switching)			
sex (reference = female)	2.24	1.81, 2.77	0
eFSM (reference = non eligible)	1.84	1.43, 2.38	0
age	1.01	0.72, 1.4	0.97
Questionnaire			
sex (reference = female)	3.03	1.39, 6.59	0.01
eFSM (reference = non eligible)	1.54	0.62, 3.83	0.36
age	7.55	2.06, 27.61	0

Abbreviations: eFSM = eligible for free school meals, OR = odds ratio

Back to main text

D.2.2 Descriptives

School-level and pupil-level characteristics for the total sample, the cognitive sample and the complete-case sample are displayed in Table D.6. The cognitive sample includes participants that completed at least one cognitive assessment (at baseline), whereas the complete-case sample includes participants who completed all cognitive assessments, the 20MSR and the PA questionnaire (at baseline). Compared to the total sample, a

smaller proportion of males and eFSM pupils are part of the cognitive- and complete-case sample.

Descriptive statistics of predictor and outcome variables of are provided in Table D.7. Please note, these statistics are computed on the complete-case dataset for each predictor or outcome variable.

Table D.6: Total-, cognitive- and complete-case sample demographics

	Total sample	Cognitive sample	Complete-case[a] sample
School level			
No. of schools	93	83	54
Gender status, no. (%)			
Co-ed	75 (80.6%)	66 (79.5%)	40 (74.1%)
Female	17 (18.3%)	16 (19.3%)	13 (24.1%)
Male	1 (1.1%)	1 (1.2%)	1 (1.9%)
School type, no. (%)			
Comprehensive to 16	16 (17.2%)	16 (19.3%)	12 (22.2%)
Comprehensive to 18	67 (72.0%)	57 (68.7%)	36 (66.7%)
Secondary Modern	10 (10.8%)	10 (12.0%)	6 (11.1%)
School establishment, no. (%)			
Academy - Sponsor Led	16 (17.2%)	14 (16.9%)	11 (20.4%)
Academy Converter	3 (3.2%)	3 (3.6%)	1 (1.9%)
Academy Convertor - Mainstream	31 (33.3%)	27 (32.5%)	16 (29.6%)
Community School	15 (16.1%)	11 (13.3%)	7 (13.0%)
Foundation School	10 (10.8%)	10 (12.0%)	7 (13.0%)
Free School - Mainstream	5 (5.4%)	5 (6.0%)	3 (5.6%)
Voluntary aided school	11 (11.8%)	11 (13.3%)	8 (14.8%)
Voluntary controlled school	2 (2.2%)	2 (2.4%)	1 (1.9%)
School size, no. (%)			
Large (>200)	27 (29.0%)	24 (28.9%)	16 (29.6%)
Medium (100–200)	54 (58.1%)	48 (57.8%)	30 (55.6%)
Small (<100)	12 (12.9%)	11 (13.3%)	8 (14.8%)
Number of form groups			
Median (range)	6 (3-14)	7 (3-14)	6 (3-14)
Missing, no. (%)	2 (2.2%)	2 (2.4%)	2 (3.7%)
Ofsted rating, no. (%)			
Outstanding	22 (23.7%)	18 (21.7%)	13 (24.1%)
Good	53 (57.0%)	50 (60.2%)	33 (61.1%)
Requires Improvement	10 (10.8%)	8 (9.6%)	3 (5.6%)
Serious Weaknesses	1 (1.1%)	1 (1.2%)	1 (1.9%)
Special Measures	6 (6.5%)	5 (6.0%)	4 (7.4%)
Missing, no. (%)	1 (1.1%)	1 (1.2%)	0
Socioeconomic status			
Percent eFSM, mean (SD)	17.3 (11.1)	17.6 (11.4)	17.8 (12.4)
IMD, median (range)	5 (1-10)	5 (1-10)	5 (1-10)
School location, no. (%)			
Rural town and fringe	9 (9.7%)	8 (9.6%)	4 (7.4%)
Rural village	0 (0%)	0 (0%)	0 (0%)

Continued on next page

Table D.6 – continued from previous page

	Total sample	Cognitive sample	Complete-case[a] sample
Urban city and town	39 (41.9%)	34 (41.0%)	25 (46.3%)
Urban major conurbation	45 (48.4%)	41 (49.4%)	25 (46.3%)
Pupil level			
No. pupils	16.017	8433	1791
Age, mean (SD), yr	12.5 (0.295)	12.5 (0.295)	12.5 (0.294)
Females, no. (%)	8961 (55.9%)	5208 (61.8%)	1285 (71.7%)
eFSM, no.(%), yes	2665 (16.6%)	1199 (14.2%)	150 (8.4%)

Abbreviations: eFSM = eligible for free school meals, IMD = index of multiple deprivation, SD = standard deviation
[a] Complete-case dataset based on participants having data on all outcome measures.

Table D.7: Mean participant scores on all predictor and outcome variables, split by eFSM status and sex

	Range	eFSM		Sex		Overall (n = 16,017)
		No (n = 13,352)	Yes (n = 2,665)	Female (n = 8,961)	Male (n = 7,056)	
20MSR						
Mean (SD)	0–163	38.5 (20.9)	32.0 (18.4)	32.7 (16.7)	43.8 (23.5)	37.5 (20.7)
Missing, no. (%)		4302 (32.2%)	1018 (38.2%)	2883 (32.2%)	2437 (34.5%)	5320 (33.2%)
PA (past week)						
Mean (SD)	0–7	4.47 (1.90)	4.13 (1.99)	4.25 (1.88)	4.67 (1.95)	4.42 (1.92)
Missing, no. (%)		5044 (37.8%)	1274 (47.8%)	3125 (34.9%)	3193 (45.3%)	6318 (39.4%)
Habitual PA						
Mean (SD)	1–7	5.18 (1.39)	4.87 (1.47)	5.06 (1.38)	5.25 (1.43)	5.14 (1.41)
Missing, no. (%)		4996 (37.4%)	1247 (46.8%)	3086 (34.4%)	3157 (44.7%)	6243 (39.0%)
Attitude to PA						
Mean (SD)	1–7	5.02 (1.25)	4.73 (1.34)	4.87 (1.25)	5.15 (1.28)	4.98 (1.27)
Missing, no. (%)		4996 (37.4%)	1247 (46.8%)	3086 (34.4%)	3157 (44.7%)	6243 (39.0%)
Relational memory						
Accuracy, mean (SD)	8.33–9.72	61.0 (11.8)	57.7 (11.9)	60.9 (11.8)	60.0 (11.9)	60.6 (11.9)
Missing, no. (%)		8695 (65.1%)	2004 (75.2%)	5477 (61.1%)	5222 (74.0%)	10699 (66.8%)
Reaction time task						
RT, mean (SD)	245–1630	374 (81.7)	395 (116)	375 (86.2)	380 (89.6)	377 (87.6)
Missing, no. (%)		6519 (48.8%)	1558 (58.5%)	4071 (45.4%)	4006 (56.8%)	8077 (50.4%)
Two-back						
Accuracy, mean (SD)	2.08–100	56.6 (18.0)	49.3 (18.9)	56.1 (17.4)	55.0 (19.8)	55.7 (18.3)
RT, mean (SD)	217–1270	868 (162)	817 (174)	875 (158)	839 (172)	862 (164)
Missing, no. (%)		8920 (66.8%)	2085 (78.2%)	5702 (63.6%)	5303 (75.2%)	11005 (68.7%)
Flanker						
RT incongruent, mean (SD)	254–1160	558 (122)	546 (136)	575 (121)	522 (122)	557 (124)
RT congruent, mean (SD)	275–1080	492 (89.3)	495 (101)	506 (90.9)	466 (84.7)	492 (90.9)
Accuracy incongruent, mean (SD)	4.26–100	63.0 (19.8)	58.5 (20.2)	65.8 (19.5)	56.0 (19.1)	62.4 (19.9)
Accuracy congruent, mean (SD)	1.96–100	84.6 (16.3)	79.7 (17.8)	84.9 (16.5)	82.3 (16.6)	84.0 (16.6)
Missing, no. (%)		8579 (64.3%)	1990 (74.7%)	5404 (60.3%)	5165 (73.2%)	10569 (66.0%)

Continued on next page

Table D.7 – continued from previous page

	Range	eFSM		Sex		Overall (n = 16,017)
		No (n = 13,352)	Yes (n = 2,665)	Female (n = 8,961)	Male (n = 7,056)	
Colour-shape task switch						
RT switch, mean (SD)	250–3970	1500 (572)	1320 (572)	1510 (586)	1400 (546)	1470 (575)
RT non-switch, mean (SD)	270–3120	1080 (335)	1020 (341)	1090 (342)	1040 (322)	1070 (336)
Accuracy switch, mean (SD)	26.7–100	72.3 (16.8)	66.1 (16.5)	72.5 (17.0)	69.7 (16.5)	71.6 (16.9)
Accuracy non-switch, mean (SD)	20.6–100	76.3 (16.9)	69.3 (16.7)	76.3 (17.1)	73.8 (16.8)	75.5 (17.1)
Missing, no. (%)		9186 (68.8%)	2109 (79.1%)	5815 (64.9%)	5480 (77.7%)	11295 (70.5%)

Abbreviations: eFSM = eligible for free school meals, RT = reaction time

Back to main text

D.2.3 Univariate analysis

D.2.3.1 Primary analysis

The primary analysis assessed the relationship between CRF or PA and cognitive outcomes was moderated by sex or eFSM status, adjusting for age, using multilevel models on multiply imputed data under the assumption that data was missing at random (MAR). Table D.8 shows the unstandardized model estimates for each outcome and predictor (CRF or PA, or interactions with sex and/or eFSM), 95% confidence intervals, p-values and explained variance (marginal and conditional) for all models.

Table D.8: Unstandardized regression coefficients of multiple imputed data

Outcome	Predictor	β	95% CI	p	$R^2_{(m,c)}$ [c]
Relational memory (% correct)	CRF	0.05	0.03, 0.07	<0.001	0.01, 0.04
Relational memory (% correct)	PA	0.18	-0.02, 0.4	0.08	0.01, 0.04
Processing speed (RT, trans)[a]	CRF×sex[b]	0.004	0.002, 0.007	0.001	0.03, 0.09
Processing speed (RT, trans)[a]	PA	-0.04	-0.05, -0.02	<0.001	0.01, 0.08
IC (Flanker, % correct)	CRF×sex[b]	-0.07	-0.13, -0.02	0.005	0.07, 0.12
IC (Flanker, % correct)	PA	0.56	0.25, 0.87	0.001	0.05, 0.10
CF (Task switch, % correct)	CRF	0.004	0.002, 0.006	<0.001	0.02, 0.07
CF (Task switch, % correct)	PA	0.53	0.28, 0.78	<0.001	0.02, 0.07
WM (two-back, % correct)	CRF	0.08	0.05, 0.11	<0.001	0.02, 0.08
WM (two-back, % correct)	PA	0.48	0.18, 0.78	0.002	0.01, 0.08

Abbreviations: CRF = cardiorespiratory fitness, CF = cognitive flexibility, IC = inhibitory control, RT = reaction time, PA = physical activity, WM = working memory
[a] Reaction times were transformed using an inverse normal transformation prior to imputations. To get an approximation of the relationship on the original scale, the regression coefficient could be multiplied by the standard deviation of the outcome pre-transformation (SD = 87.56ms)
[b] Reference level is female
[c] Marginal (m) and conditional (c) explained variance

Back to main text

D.2.3.2 Sensitivity analysis

As a sensitivity analysis, we repeated the primary analyses in the complete-case (i.e. available case) dataset. This set of analyses was conducted first adjusting for age only, and subsequently additionally adjusting for where and when assessments were conducted. Table D.9 and Table D.10 provide the, respectively, standardized and unstan-

dardized estimates for the age-adjusted analyses, whereas Table D.11 and Table D.12 provides the estimates for the fully adjusted models.

We inspected the significant interactions (Table D.9), and computed regression slopes for males and females. The relationship between CRF and cognitive flexibility was stronger in females (standardized β_{fit} = 0.16, p <0.001) compared to males (β_{fit} = 0.02, p = 0.44). Moreover, females had a stronger relationship between CRF and processing speed (β_{fit} = -0.2, p <0.001) compared to males (β_{fit} = -0.12, p <0.001).

Regression coefficients of fully adjusted models (Table D.11 and Table D.12) are consistent with those of the partly adjusted models, albeit with minor variations in their size.

Table D.9: Standardized regression coefficients of complete-case models

Outcome	Predictor	$N_{clusters}$	β	95% CI	p	$R^2_{(m,c)}$ [c]
Relational memory (% correct)	CRF	4205 (70)	0.08	0.05, 0.11	<0.001	0.01, 0.04
Relational memory (% correct)	PA	4093 (76)	0.03	0, 0.06	0.034	0.01, 0.04
Processing speed (RT, trans)[a]	CRF×sex[b]	6158 (70)	0.08	0.03, 0.13	0.003	0.03, 0.1
Processing speed (RT, trans)[a]	PA	5806 (76)	-0.07	-0.09, -0.04	<0.001	0.01, 0.08
IC (Flanker, % correct)	CRF	4339 (71)	0.12	0.09, 0.15	<0.001	0.06, 0.12
IC (Flanker, % correct)	PA	4185 (75)	0.06	0.03, 0.09	<0.001	0.05, 0.11
CF (Task switch, % correct)	CRF×sex[b]	3741 (70)	-0.14	-0.2, -0.07	<0.001	0.02, 0.08
CF (Task switch, % correct)	PA	3650 (76)	0.05	0.02, 0.08	0.002	0.01, 0.07
WM (two-back, % correct)	CRF	3964 (70)	0.09	0.06, 0.13	<0.001	0.02, 0.09
WM (two-back, % correct)	PA	3799 (76)	0.05	0.02, 0.08	0.002	0.02, 0.09

Abbreviations: CRF = cardiorespiratory fitness, CF = cognitive flexibility, IC = inhibitory control, RT = reaction time, PA = physical activity, WM = working memory
[a] Reaction times were transformed using an inverse normal transformation prior to imputations
[b] Reference level is female
[c] Marginal (m) and conditional (c) explained variance

Table D.10: Unstandardized regression coefficients of complete-case models

Outcome	Predictor	$N_{clusters}$	β	95% CI	p	$R^2_{(m,c)}$ [c]
Relational memory (% correct)	CRF	4205 (70)	0.05	0.03, 0.07	<0.001	0.01, 0.04
Relational memory (% correct)	PA	4093 (76)	0.21	0.02, 0.4	0.034	0.01, 0.04
Processing speed (RT, trans)[a]	CRF×sex[b]	6158 (70)	0.004	0.001, 0.006	0.003	0.03, 0.1
Processing speed (RT, trans)[a]	PA	5806 (76)	-0.04	-0.05, -0.02	<0.001	0.01, 0.08
IC (Flanker, % correct)	CRF	4339 (71)	0.12	0.09, 0.15	<0.001	0.06, 0.12
IC (Flanker, % correct)	PA	4185 (75)	0.62	0.31, 0.92	<0.001	0.05, 0.11
CF (Task switch, % correct)	CRF×sex[b]	3741 (70)	-0.12	-0.18, -0.07	<0.001	0.02, 0.08
CF (Task switch, % correct)	PA	3650 (76)	0.46	0.17, 0.74	0.002	0.01, 0.07
WM (two-back, % correct)	CRF	3964 (70)	0.09	0.06, 0.12	<0.001	0.02, 0.09
WM (two-back, % correct)	PA	3799 (76)	0.46	0.17, 0.76	0.002	0.02, 0.09

Abbreviations: CRF = cardiorespiratory fitness, CF = cognitive flexibility, IC = inhibitory control, RT = reaction time, PA = physical activity, WM = working memory

[a] Reaction times were transformed using an inverse normal transformation (mean = 0, SD = 1). To get an approximation of the relationship on the original scale, the regression coefficient could be multiplied by the standard deviation of the outcome pre-transformation, which is for models with CRF as predictor: 83.57ms, and PA as predictor: 85.56ms.

[b] Reference level is female

[c] Marginal (m) and conditional (c) explained variance

Table D.11: Standardized regression coefficients of fully-adjusted complete-case models

Outcome	Predictor	$N_{clusters}$	β	95% CI	p	$R^2_{(m,c)}$ [c]
Relational memory (% correct)	CRF	4205 (70)	0.08	0.05, 0.11	<0.001	0.03, 0.04
Relational memory (% correct)	PA	4083 (76)	0.03	0, 0.06	0.050	0.03, 0.05
Processing speed (RT, trans)[a]	CRF×sex[b]	6158 (70)	0.08	0.03, 0.13	0.003	0.04, 0.09
Processing speed (RT, trans)[a]	PA	5787 (76)	-0.07	-0.09, -0.04	<0.001	0.02, 0.08
IC (Flanker, % correct)	CRF	4339 (71)	0.12	0.09, 0.15	<0.001	0.09, 0.11
IC (Flanker, % correct)	PA	4171 (75)	0.06	0.03, 0.09	<0.001	0.08, 0.11
CF (Task switch, % correct)	CRF×sex[b]	3741 (70)	-0.13	-0.2, -0.07	<0.001	0.06, 0.09
CF (Task switch, % correct)	PA	3638 (76)	0.05	0.02, 0.09	0.001	0.04, 0.08
WM (two-back, % correct)	CRF	3964 (70)	0.09	0.06, 0.12	<0.001	0.06, 0.09
WM (two-back, % correct)	PA	3788 (76)	0.05	0.02, 0.08	0.001	0.04, 0.08

Abbreviations: CRF = cardiorespiratory fitness, CF = cognitive flexibility, IC = inhibitory control, RT = reaction time, PA = physical activity, WM = working memory

[a] Reaction times were transformed using an inverse normal transformation (mean = 0, SD = 1)

[b] Reference level is female

[c] Marginal (m) and conditional (c) explained variance

Table D.12: Unstandardized regression coefficients of fully-adjusted complete-case models

Outcome	Predictor	$N_{clusters}$	β	95% CI	p	$R^2_{(m,c)}$ [c]
Relational memory (% correct)	CRF	4205 (70)	0.07	0.04, 0.1	<0.001	0.03, 0.04
Relational memory (% correct)	PA	4083 (76)	0.19	0, 0.38	0.050	0.03, 0.05
Processing speed (RT task, trans)[a]	CRF×sex[b]	6158 (70)	0.004	0.001, 0.006	0.003	0.04, 0.09
Processing speed (RT task, trans)[a]	PA	5787 (76)	-0.04	-0.05, -0.02	<0.001	0.02, 0.08
IC (Flanker, % correct)	CRF	4339 (71)	0.12	0.09, 0.15	<0.001	0.09, 0.11
IC (Flanker, % correct)	PA	4171 (75)	0.61	0.3, 0.91	<0.001	0.08, 0.11
CF (Task switch, % correct)	CRF×sex[b]	3741 (70)	-0.12	-0.18, -0.06	<0.001	0.06, 0.09
CF (Task switch, % correct)	PA	3638 (76)	0.48	0.2, 077	0.001	0.04, 0.08
WM (two-back, % correct)	CRF	3964 (70)	0.09	0.06, 0.12	<0.001	0.06, 0.09
WM (two-back, % correct)	PA	3788 (76)	0.48	0.19, 0.78	0.001	0.04, 0.08

Abbreviations: CRF = cardiorespiratory fitness, CF = cognitive flexibility, IC = inhibitory control, RT = reaction time, PA = physical activity, WM = working memory

[a] Reaction times were transformed using an inverse normal transformation (mean = 0, SD = 1). To get an approximation of the relationship on the original scale, the regression coefficient could be multiplied by the standard deviation of the outcome pre-transformation, which is for models with CRF as predictor: 83.57ms, and PA as predictor: 85.56ms.

[b] Reference level is female

[c] Marginal (m) and conditional (c) explained variance

Back to main text

D.2.4 Specification curve analysis

To assess the robustness of the relationships between PA or CRF and cognitive outcomes to variations in (1) data cleaning procedures, (2) the selected predictor, outcome and nuisance variables, we used specification curve analysis (Orben & Przybylski, 2019; Simonsohn *et al.,* 2015; Simonsohn *et al.,* 2020). Outcomes and continuous predictors were scaled by 2 standard deviations, such that regression estimates of continuous and binary predictors can be more readily compared (Gelman, 2008). No diagnostic checks (e.g. normality of residuals) were performed for individual specifications. Multilevel linear models were used throughout using *lme4* (Bates *et al.,* 2015) in R (version 6.0.0)

D.2.4.1 Robustness to data cleaning

To assess the robustness of the relationships to justifiable alterations in data cleaning of cognitive outcomes, we varied the cleaning procedure. In particular, we altered the following cleaning parameters:

1. Varying the minimum or % of valid trials (defined by a response within predefined RT bounds); here we chose for the following three thresholds: 0 (i.e. no exclusion based on the number of valid trials), 30% (i.e. a participant with fewer than 30% of trials is excluded), 90% (i.e. a participant with data on fewer than 90% of trials is excluded)

2. Accuracy-based cleaning, with the following thresholds: 0 (i.e. no exclusion based on accuracy), or chance (i.e. participants are excluded if they scored <chance performance, under the assumption that the task was not understood correctly)

Justifiable adaptations in the cleaning approach resulted in small variations in the observed standardized regression coefficients (β) for CRF and PA (Table D.13), but generally did not change conclusions regarding directionality or significance of the relationships. Figures D.2 to D.6 display the specification curves for each cognitive outcome, showing the variability in standardized regression estimates for the CRF or

PA predictor to variations in cleaning of the cognitive outcomes. CRF has consistently stronger relationships (i.e. a higher median standardized regression coefficient) compared with PA. Varying the cleaning approach resulted in variability in the number of tests that reached statistical significance (at p <0.05) when assessing relationships between PA and working memory (Figure D.2, 2 of 6 specifications were significant) or relational memory (Figure D.3, 4 of 6 specifications were significant), whereas all other relationships were robust to variations in data cleaning.

Table D.13: The robustness of relationships between PA or CRF and cognitive outcomes to variations in cleaning of cognitive assessments

	N[a]	**No. of tests**[b]	**Standardized β** Median (range)	**Significant** No. (%)
Relational memory (accuracy)				
CRF	3693	6	0.07 (0.06, 0.08)	6 (100)
PA	3616	6	0.03 (0.02, 0.04)	2 (33)
All	3665	12	0.05 (0.02, 0.08)	8 (67)
Processing speed (reaction time task; RT)				
CRF	6160	3	-0.16 (-0.16, -0.16)	3 (100)
PA	5819	3	-0.07 (-0.07, -0.07)	3 (100)
All	5917	6	-0.11 (-0.16, 0.07)	6 (100)
Inhibitory control (Flanker; accuracy)				
CRF	4069	6	0.12 (0.11, 0.13)	6 (100)
PA	3933	6	0.06 (0.06, 0.07)	6 (100)
All	3997	12	0.09 (0.06, 0.13)	12 (100)
Cognitive flexibility (task switching; accuracy)				
CRF	3552	6	0.1 (0.1, 0.12)	6 (100)
PA	3494	6	0.05 (0.04, 0.07)	6 (100)
All	3519	12	0.09 (0.04, 0.12)	12 (100)
Working memory (two-back; accuracy)				
CRF	3452	6	0.10 (0.09, 0.1)	6 (100)
PA	3367	6	0.05 (0.03, 0.05)	4 (67)
All	3410	12	0.07 (0.03, 0.10)	10 (83)

Abbreviations: CRF = cardiorespiratory fitness, PA = physical activity
[a] The median sample size across tests (rounded to nearest integer)
[b] The number of statistical tests (i.e. specifications) conducted for a set of predictor and outcome variables

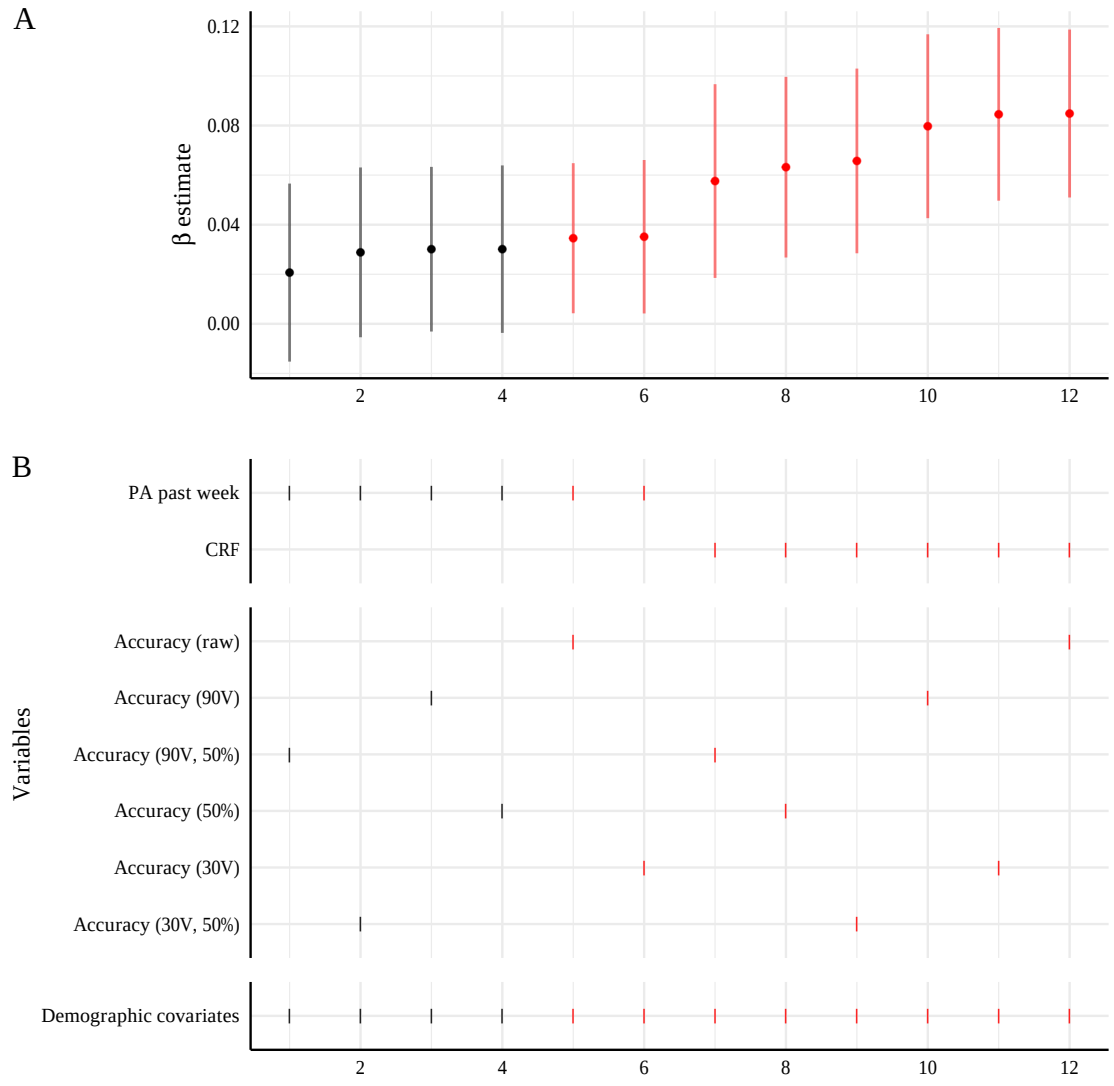


Figure D.2: Results of the specification curve analysis for relational memory outcomes. The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and relational memory outcomes that varied by cleaning procedure. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimate represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top)* block contains the predictors (PA past week or CRF), the second (*middle*) block presents the outcomes and the third (*bottom*) block shows the nuisance variables (demographics: age, sex, eFSM). Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by a small vertical bar. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning.

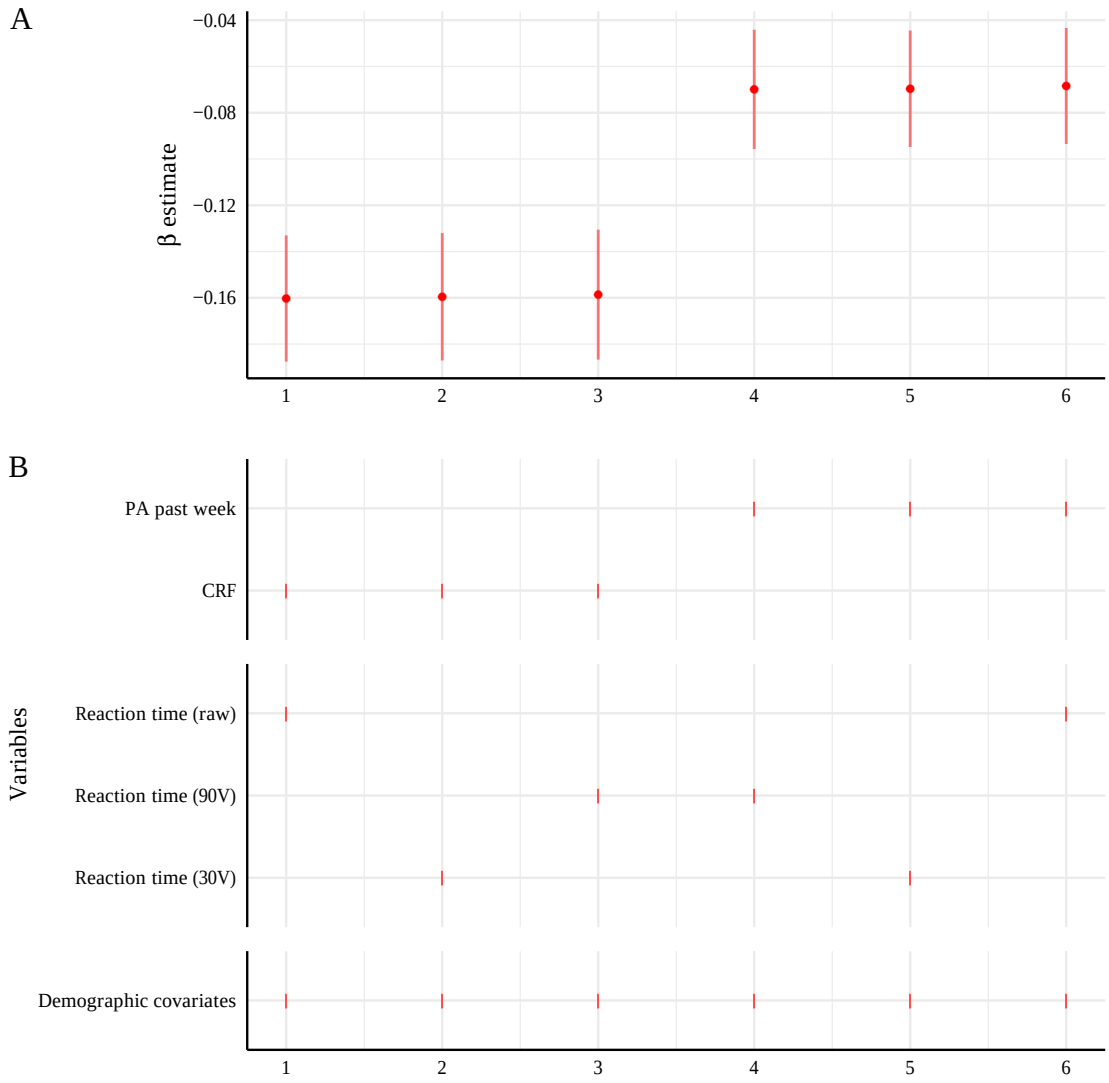


Figure D.3: Results of the specification curve analysis for processing speed outcomes. The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and RT outcomes that varied by cleaning procedure. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimate represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top)* block contains the predictors (PA past week or CRF), the second (*middle*) block presents the outcomes and the third (*bottom*) block shows the nuisance variables (demographics: age, sex, eFSM). Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by a small vertical bar. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning.

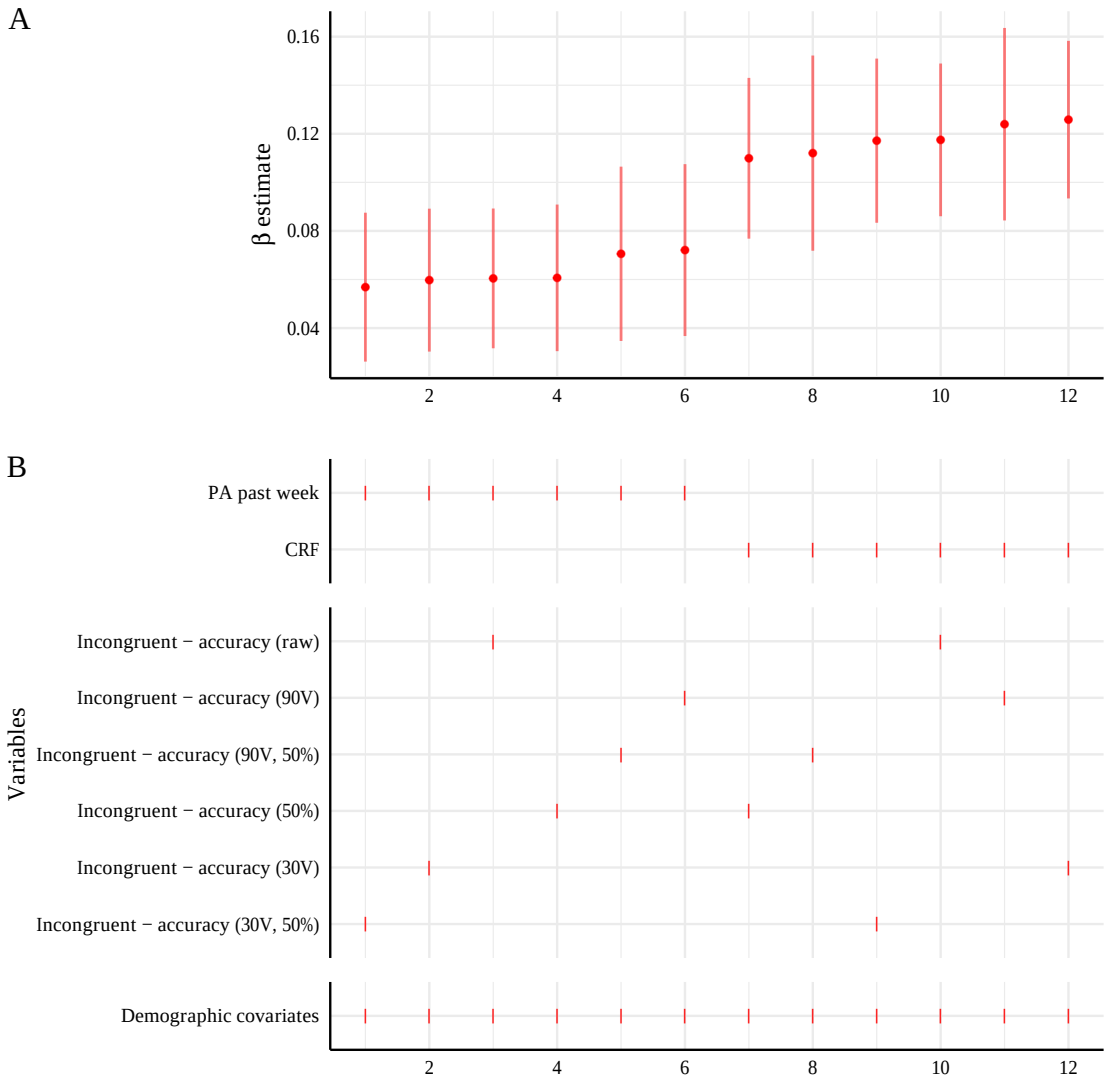


Figure D.4: Results of the specification curve analysis for inhibitory control (Flanker task). The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and Flanker task outcomes that varied by cleaning procedure. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimate represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top)* block contains the predictors (PA past week or CRF), the second (*middle*) block presents the outcomes and the third (*bottom*) block shows the nuisance variables (demographics: age, sex, eFSM). Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by a small vertical bar. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning.

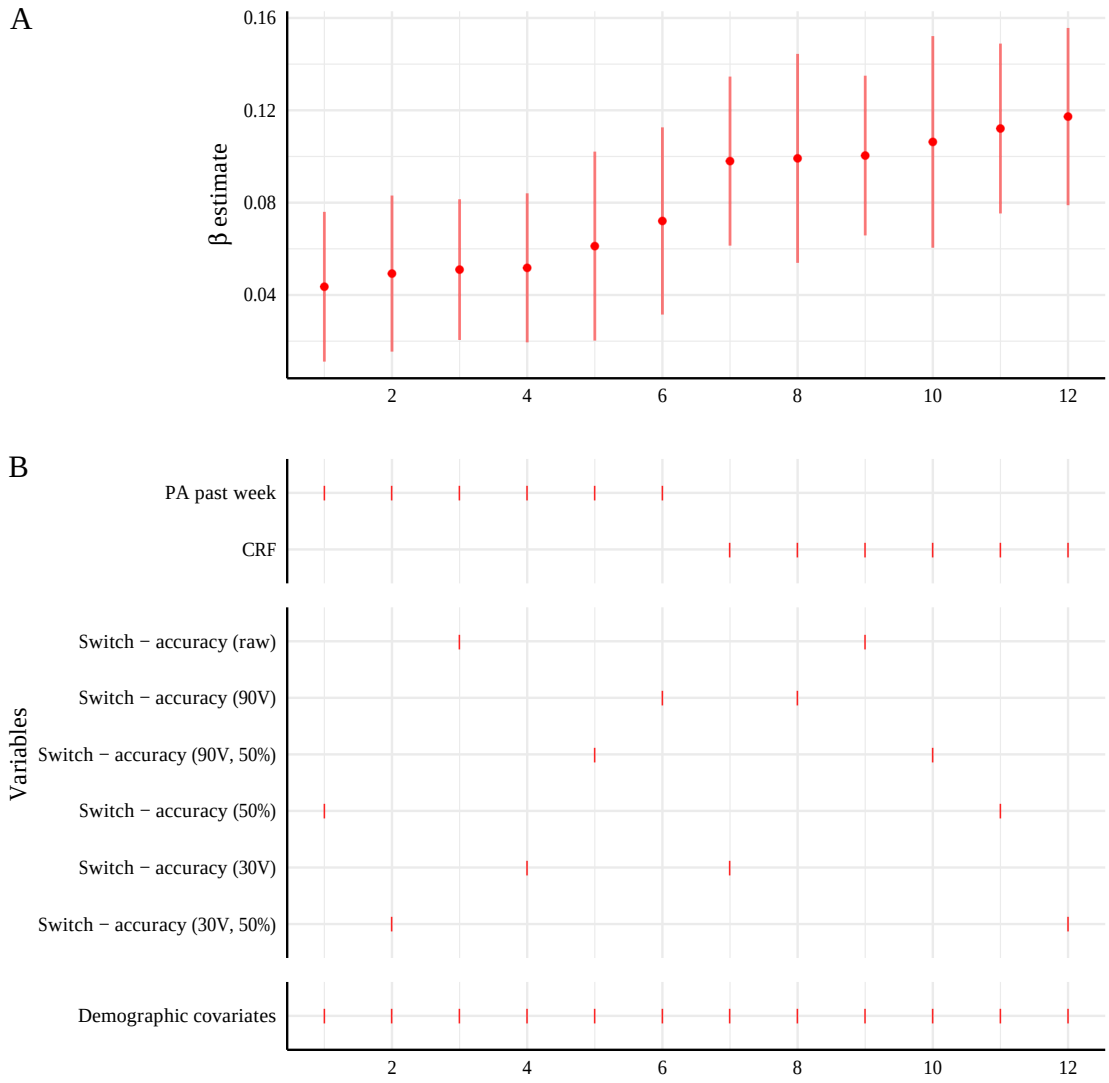


Figure D.5: Results of the specification curve analysis for cognitive flexibility (task-switching task). The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and task switching outcomes that varied by cleaning procedure. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimate represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top)* block contains the predictors (PA past week or CRF), the second (*middle*) block presents the outcomes and the third (*bottom*) block shows the nuisance variables (demographics: age, sex, eFSM). Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by a small vertical bar. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning.

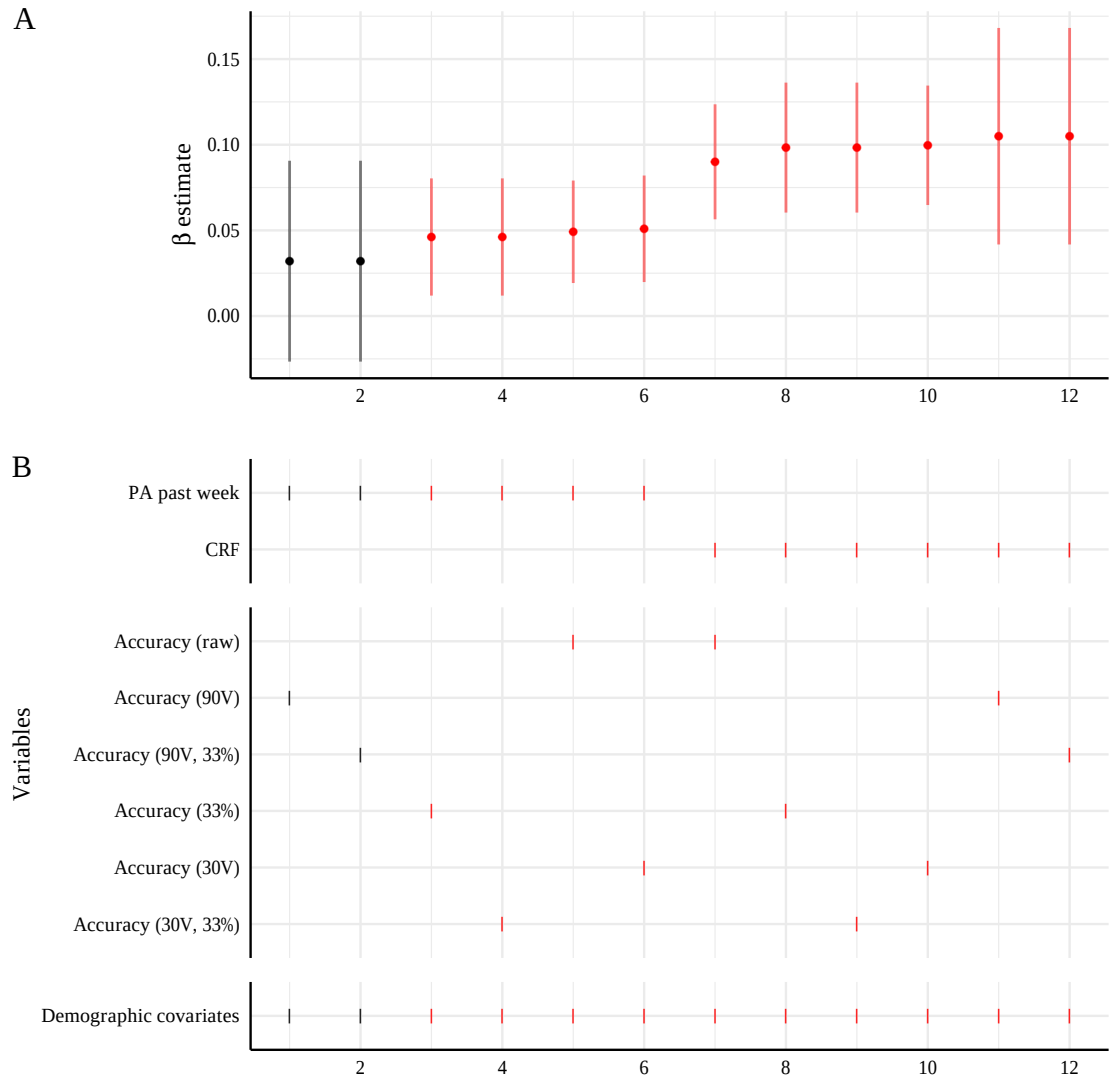


Figure D.6: Results of the specification curve analysis working memory (two-back task). The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and two-back outcomes that varied by cleaning procedure. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimate represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top)* block contains the predictors (PA past week or CRF), the second (*middle*) block presents the outcomes and the third (*bottom*) block shows the nuisance variables (demographics: age, sex, eFSM). Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by a small vertical bar. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 33% = chance-level (accuracy) cleaning.

Back to main text

D.2.4.2 Robustness to predictors, outcomes, nuisance and data cleaning

In addition to assessing the robustness of CRF or PA —cognition relationships to data cleaning procedures, we inspected the influence of other justifiable analytic choices on the standardized regression coefficients.

We made four key analytical decisions (Table D.14):

1. Cleaning of cognitive assessments, using two justifiable cleaning approaches:

   (a) Varying the minimum or % of valid trials (defined by a response within pre-defined reaction time bounds); here we chose for the following three thresholds: 0 (i.e. no exclusion based on the number of valid trials), 30% (i.e. a participant with fewer than 30% of trials is excluded), 90% (i.e. a participant with data on fewer than 90% of trials is excluded).

   (b) Accuracy-based cleaning, with the following thresholds: 0 (i.e. no exclusion based on accuracy), or chance (i.e. participants are excluded if they scored <chance performance, under the assumption that the task was not understood correctly).

2. Predictor variables, chosen to be PA and CRF in line with the primary analyses, have been treated as dichotomous or continuous variables in previous studies. We dichotomised the continuous variables using existing cut-offs. In particular, participants were classified as high PA if they were active for at least 60 min a day on 7 consecutive days (i.e. in line with the international PA guidelines, 2018 Physical Activity Guidelines Advisory Committee, 2018; World Health Organization, 2018), and as high fit if they scored above sex specific norm values (50 percentile; males >39 laps, females >28 laps, Tomkinson *et al.*, 2017).

3. Outcome variables, chosen to be the most commonly reported outcome measures per cognitive assessment:

(a) Relational memory assessment: accuracy (on valid trials). While d-prime (*d'*) has been used in some studies, we decided not to include this measure here because a low and variable number of trials across participants may introduce bias (Miller, 1996; Verde, 2006). Participants were not instructed to be as quick as possible, and hence reaction time was not considered as an outcome measure.

(b) Processing speed (simple reaction time task): mean reaction time (valid trials).

(c) Inhibitory control (Flanker task): incongruent accuracy or reaction time, interference accuracy (i.e. $Accuracy_{congruent} - Accuracy_{incongruent}$, with lower scores indicating better performance) or interference reaction time (i.e. $RT_{incongruent} - RT_{congruent}$, with lower scores indicating better performance).

(d) Cognitive flexibility (colour-shape task-switching task): switch and non-switch accuracy or reaction time, interference accuracy (i.e. $Accuracy_{non-switch} - Accuracy_{switch}$, with lower scores indicating better performance) or interference reaction time (i.e. $RT_{switch} - RT_{non-switch}$, with lower scores indicating better performance).

(e) Working memory (two-back task): accuracy and reaction time. While d-prime (*d'*) has been used in some studies, we decided not to include this measure here because a low and variable number of trials across participants may introduce bias (Miller, 1996; Verde, 2006).

4. Nuisance variables: no nuisance, demographic variables (age, sex, eFSM), task-specific variables (location and date of task completion), demographic and task-specific variables. Rather than considering each nuisance variable individually —which is not common practice if nuisance variables have been identified —we grouped nuisance variables into nuisance sets and in/excluded sets from specifications.

Table D.14: Specifications (analytical choices) that are tested for each cognitive assessment

Analytic domains	Analysis choices
Cleaning of cognitive assessments	% of valid trials (0, 30, 90) Accuracy-based cleaning (e.g. chance level)
Operationalising cognition	Accuracy / RT Difference scores (if applicable)
Operationalising PA and CRF	Continuous vs binary
Operationalising confounds	No confounds Task-specific confounds Demographic confounds Task + demographic confounds

Abbreviations: CRF = cardiorespiratory fitness, PA = physical activity, RT = reaction time

D.2.4.3 Key observations

Table D.15 shows the median standardized regression coefficients for each predictor, across tests that varied outcomes, cleaning procedures and nuisance variables. Notably, median regression coefficients were generally higher for relationships involving CRF as predictor, whereas dichotomising the predictor had little effect. The choice of outcome variables had a strong effect on median standardized regression coefficients (details below), resulting in estimates that were close to zero for tasks with multiple distinct outcome measures, such as those measuring inhibitory control and cognitive flexibility. The specification curves for each of the cognitive assessments are shown in Figures D.7 to D.11. The median standardized regression coefficients for the inhibitory control task, cognitive flexibility task and working memory task, per outcome and predictor, are shown in, respectively, Table D.16, Table D.17 and Table D.18. Outcomes and continuous predictors were scaled by 2 standard deviations.

Key observations of SCA results (across Figures and Tables):

- Across cognitive domains, higher CRF and PA are generally associated with more favourable cognitive outcomes.

- Cognitive assessments with multiple outcomes, such as the working memory, inhibitory control and cognitive flexibility tasks, demonstrate consistently positive relationships of CRF and PA with accuracy outcomes (of component scores), but no clear relationships with reaction time outcomes (in line with several previous studies, e.g. Chaddock *et al.,* 2010a; Raine *et al.,* 2018).

- RT difference scores (interference scores) of the Flanker task (inhibitory control) and colour-shape switching task (cognitive flexibility) were positively related with CRF and PA, suggesting that higher CRF and PA are related to larger difference scores.

- Higher CRF and PA were also related (in addition to accuracy), albeit less consistently, with RT outcomes on the switching (switch condition) and working memory task, possibly indicating that higher fit and active pupils use a different strategy compared to their lower fit/active counterparts whereby accuracy is favoured over RT.

Table D.15: Robustness of relationships between PA or CRF and cognitive outcomes to variations in predictors, outcomes, nuisance and cleaning of cognitive assessments

	N[a]	No. of tests[b]	No. of y [c]	Standardized β Median (range)	Significant No. (%)
Relational memory (accuracy)					
CRF (binary)	3693	24	1	0.08 (0.07 ,0.09)	24 (100)
CRF (continuous)	3693	24	1	0.07 (0.05 ,0.08)	24 (100)
PA (continuous)	3612	24	1	0.03 (0.02 ,0.04)	8 (33)
PA (binary)	3612	24	1	0 (-0.02 ,0.01)	0 (0)
All	3661	96	1	0.04 (-0.02, 0.09)	56 (58)
Processing speed (reaction time task; RT)					
CRF (binary)	6160	12	1	-0.12 (-0.12, -0.11)	12 (100)
CRF (continuous)	6160	12	1	-0.16 (-0.16, -0.15)	12 (100)
PA (continuous)	5810	12	1	-0.07 (-0.07, -0.07)	12 (100)
PA (binary)	5810	12	1	-0.06 (-0.06, -0.06)	12 (100)
All	5917	48	1	-0.09 (-0.16, -0.06)	48 (100)
Inhibitory control (Flanker; accuracy)					
CRF (binary)	4066	144	6	0.01 (-0.05, 0.1)	81 (56)
CRF (continuous)	4066	144	6	0.02 (-0.08, 0.13)	95 (66)
PA (continuous)	3923	144	6	0.03 (-0.04, 0.07)	66 (46)
PA (binary)	3923	144	6	0.01 (-0.05, 0.06)	14 (10)
All	3997	576	6	0.02 (-0.08, 0.13)	256 (44)
Cognitive flexibility (task switching; accuracy)					
CRF (binary)	3545	144	6	0.03 (-0.02, 0.10)	71 (49)
CRF (continuous)	3545	144	6	0.03 (-0.05, 0.12)	68 (47)
PA (continuous)	3482	144	6	0.02 (-0.04, 0.09)	48 (33)
PA (binary)	3482	144	6	-0.01 (-0.03, 0.07)	14 (10)
All	3506	576	6	0.02 (-0.05, 0.12)	201 (35)
Working memory (two-back; accuracy)					
CRF (binary)	3452	48	2	0.08 (0.04, 0.11)	40 (83)
CRF (continuous)	3452	48	2	0.08 (0.03, 0.12)	40 (83)
PA (continuous)	3362	48	2	0.05 (0.02, 0.06)	32 (67)
PA (binary)	3362	48	2	0 (-0.02, 0.03)	0 (0)
All	3410	192	2	0.05 (-0.02, 0.12)	112 (58)

Abbreviations: CRF = cardiorespiratory fitness, PA = physical activity, RT = reaction time

[a] Continuous predictors are scaled by 2 standard deviations to improve comparisons with binary predictors (Gelman, 2008). Outcomes were scaled by 2SD to allow for a comparison of regression coefficients with those presented in Table D.13. Regression coefficients of specifications involving RT or interference (i.e. difference) score outcomes were multiplied by -1 to allow for meaningful medians to be computed.

[b] The number of statistical tests (i.e. specifications) conducted for a set of predictor and outcome variables.

[c] This corresponds to the number of distinct outcome measures, excluding variations in cleaning procedures. For inhibitory control, for instance, distinct outcomes include: incongruent accuracy, congruent accuracy, interference accuracy, incongruent reaction time, congruent reaction time, interference reaction time.

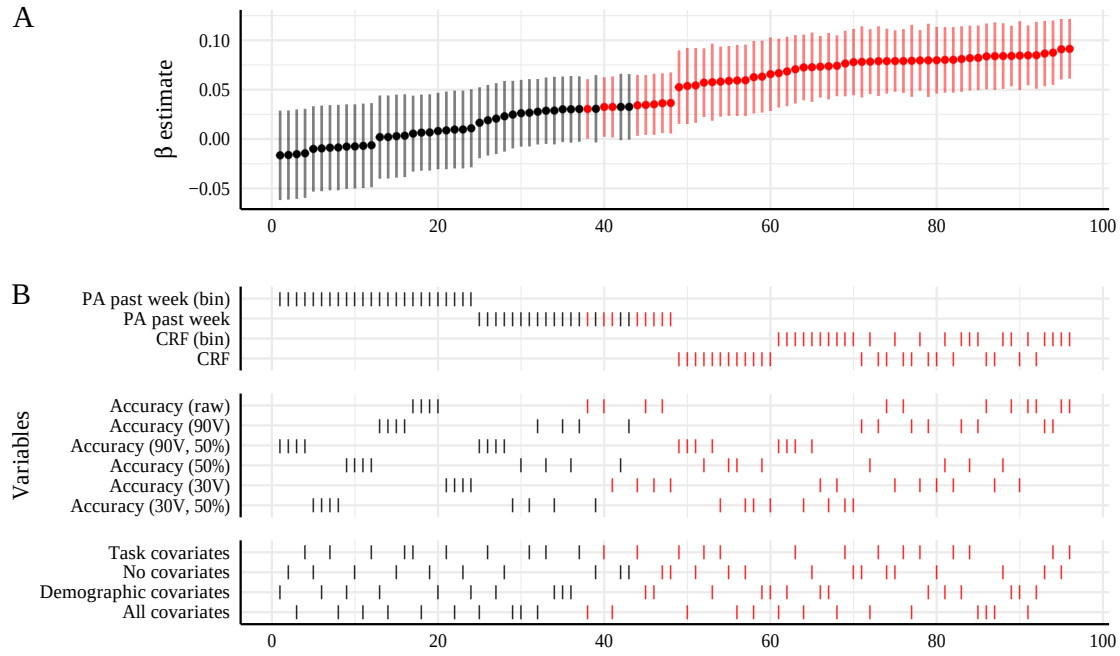


Figure D.7: Results of the specification curve for relational memory. The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and relational memory. Each specification tests a different combination of cleaning procedures, outcomes, predictors and nuisance variables. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimates represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top*) block contains the predictors, the second (*middle*) block presents the outcomes and the third (*bottom*)block shows the nuisance variables. Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by small vertical bars. makes up a single specification (i.e. model). Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning. Outcomes and continuous predictors were standardized by 2 standard deviations.

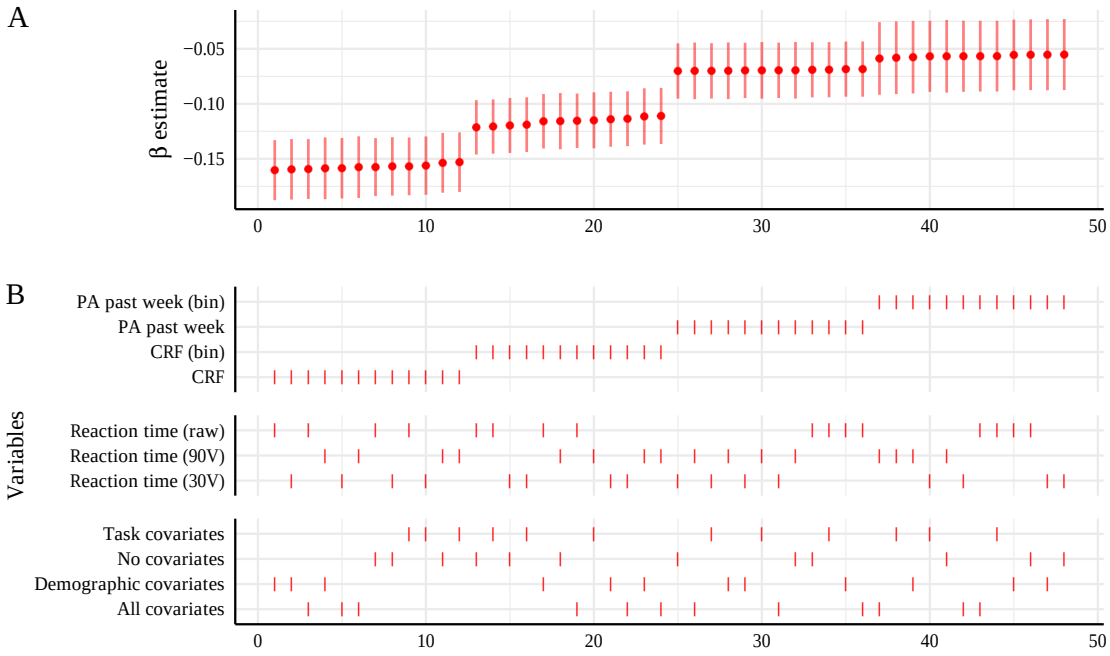


Figure D.8: Results of the specification curve for processing speed (simple reaction time task). The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and processing speed. Each specification tests a different combination of cleaning procedures, outcomes, predictors and nuisance variables. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimates represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top*) block contains the predictors, the second (*middle*) block presents the outcomes and the third (*bottom*)block shows the nuisance variables. Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by small vertical bars. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V. Outcomes and continuous predictors were standardized by 2 standard deviations.

Figure D.9: ***(page 353)*** **Results of the specification curve for inhibitory control (Flanker task).** The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and inhibitory control. Each specification tests a different combination of cleaning procedures, outcomes, predictors and nuisance variables. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimates represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top*) block contains the predictors, the second (*middle*) block presents the outcomes and the third (*bottom*)block shows the nuisance variables. Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by small vertical bars. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning. Outcomes and continuous predictors were standardized by 2 standard deviations.

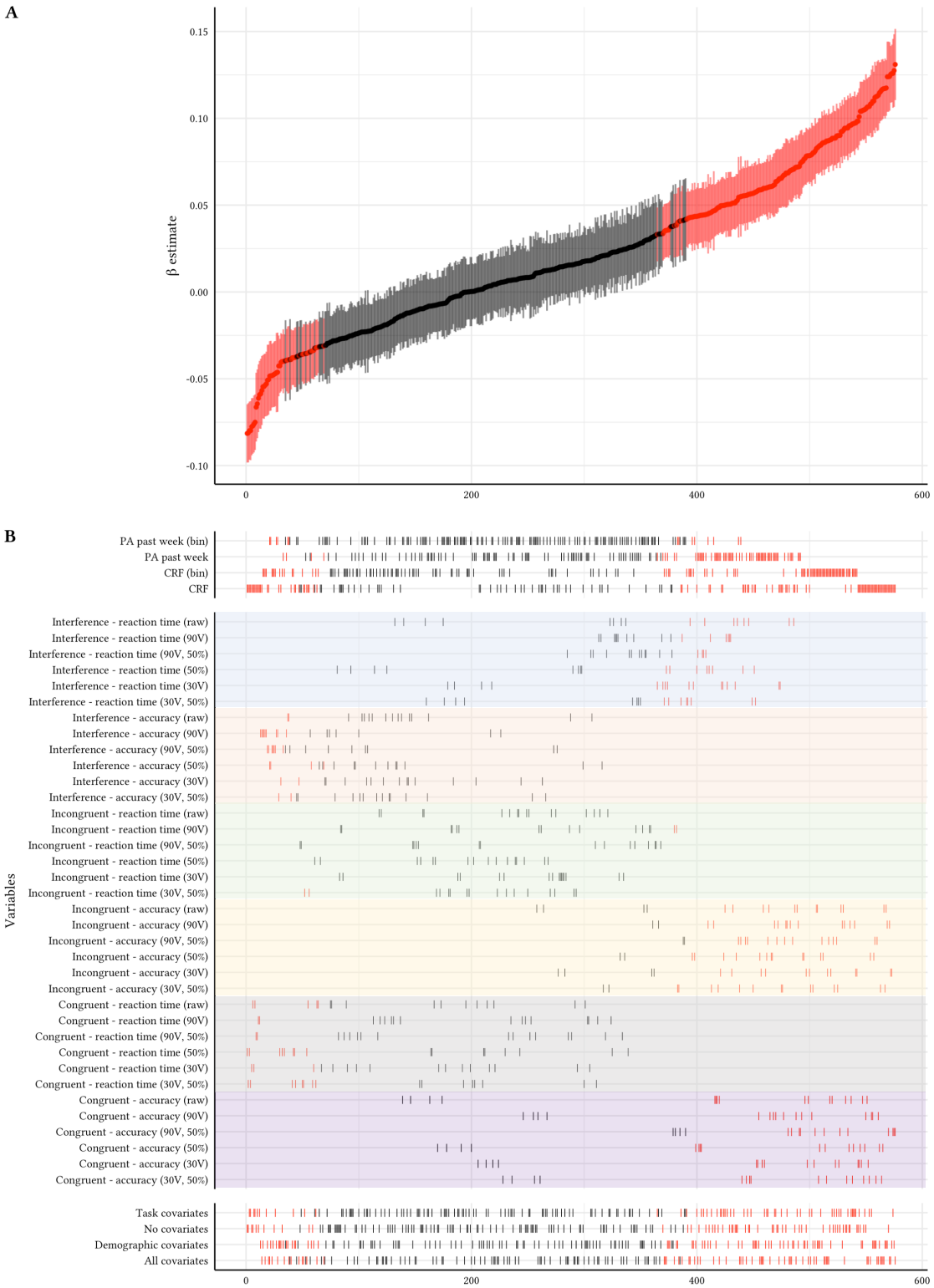

A
0.15
0.10
0.05
0.00
-0.05
-0.10
β estimate
0
200
400
600
B
PA past week (bin)
PA past week
CRF (bin)
CRF
Interference - reaction time (raw)
Interference - reaction time (90V)
Interference - reaction time (90V, 50%)
Interference - reaction time (50%)
Interference - reaction time (30V)
Interference - reaction time (30V, 50%)
Interference - accuracy (raw)
Interference - accuracy (90V)
Interference - accuracy (90V, 50%)
Interference - accuracy (50%)
Interference - accuracy (30V)
Interference - accuracy (30V, 50%)
Incongruent - reaction time (raw)
Incongruent - reaction time (90V)
Incongruent - reaction time (90V, 50%)
Incongruent - reaction time (50%)
Incongruent - reaction time (30V)
Incongruent - reaction time (30V, 50%)
Incongruent - accuracy (raw)
Incongruent - accuracy (90V)
Incongruent - accuracy (90V, 50%)
Incongruent - accuracy (50%)
Incongruent - accuracy (30V)
Incongruent - accuracy (30V, 50%)
Congruent - reaction time (raw)
Congruent - reaction time (90V)
Congruent - reaction time (90V, 50%)
Congruent - reaction time (50%)
Congruent - reaction time (30V)
Congruent - reaction time (30V, 50%)
Congruent - accuracy (raw)
Congruent - accuracy (90V)
Congruent - accuracy (90V, 50%)
Congruent - accuracy (50%)
Congruent - accuracy (30V)
Congruent - accuracy (30V, 50%)
Variables
Task covariates
No covariates
Demographic covariates
All covariates
0
200
400
600

Table D.16: Results of the specification curve analysis of the Flanker task by outcome measure

	Incongruent			Congruent			Interference		
	N	Standardized β Median (range)	Significant No. (%)[a]	N	Standardized β Median (range)	Significant No. (%)[a]	N	Standardized β Median (range)	Significant No. (%)[a]
Accuracy									
CRF (binary)	4069	0.09 (0.08, 0.1)	24 (100)	4069	0.09 (0.07, 0.1)	24 (100)	4068	-0.02 (-0.05, -0.01)	8 (33)
CRF (continuous)	4069	0.09 (0.06, 0.13)	24 (100)	4069	0.11 (0.09, 0.13)	24 (100)	4068	-0.01 (-0.06, 0.02)	8 (33)
PA (continuous)	3926	0.06 (0.05, 0.07)	24 (100)	3926	0.06 (0.04, 0.07)	24 (100)	3925	-0.02 (-0.04, 0)	4 (17)
PA (binary)	3926	0.03 (0.01, 0.06)	8 (33)	3926	0.01 (-0.01, 0.04)	0 (0)	3925	-0.03 (-0.05, -0.01)	6 (25)
Reaction time									
CRF (binary)	4067	0 (0.01, -0.01)	0 (0)	4068	-0.03 (-0.04, -0.02)	9 (38)	4066	0.04 (0.02, 0.05)	16 (67)
CRF (continuous)	4067	-0.01 (-0.04, 0.02)	2 (8)	4068	-0.05 (-0.08, -0.01)	19 (79)	4066	0.05 (0.03, 0.07)	18 (75)
PA (continuous)	3923	0.02 (0, 0.04)	2 (8)	3925	0 (-0.01, 0.02)	0 (0)	3923	0.03 (0.02, 0.05)	12 (50)
PA (binary)	3923	0.01 (-0.01, 0.03)	0 (0)	3925	0.01 (0, 0.03)	0 (0)	3923	0 (-0.03, 0.03)	0 (0)

Abbreviations: CRF = cardiorespiratory fitness, N = number of participants, PA = physical activity
[a] Number of significant tests and % of total number of tests run for that predictor and outcome

Figure D.10: *(page 356)* **Results of the specification curve for cognitive flexibility (Colour-shape switching task).** The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and cognitive flexibility. Each specification tests a different combination of cleaning procedures, outcomes, predictors and nuisance variables. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimates represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top*) block contains the predictors, the second (*middle*) block presents the outcomes and the third (*bottom*)block shows the nuisance variables. Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by small vertical bars. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 50% = chance-level (accuracy) cleaning. Outcomes and continuous predictors were standardized by 2 standard deviations.

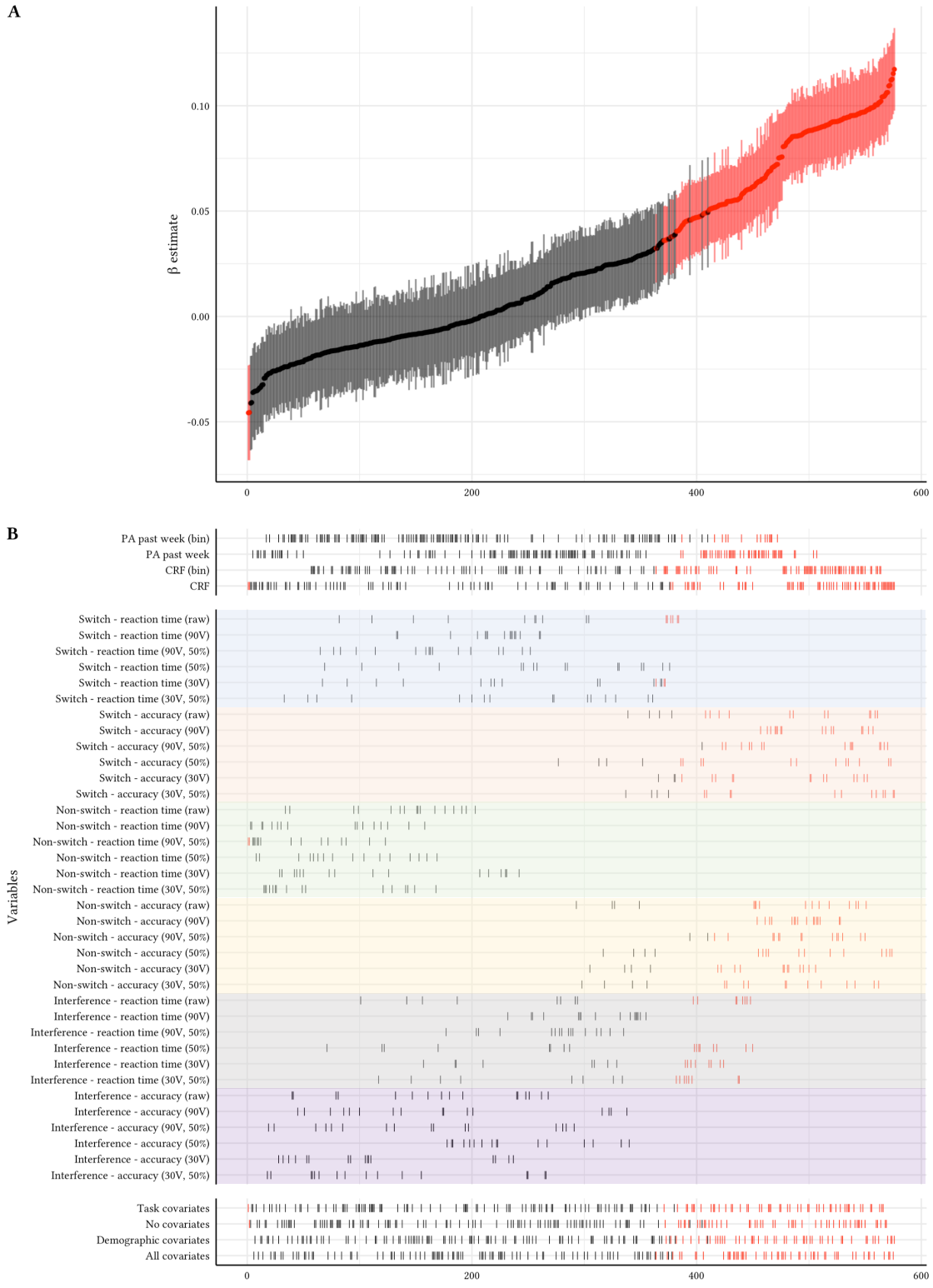

A
β estimate
0.10
0.05
0.00
-0.05
0
200
400
600
B
PA past week (bin)
PA past week
CRF (bin)
CRF
Variables
Switch - reaction time (raw)
Switch - reaction time (90V)
Switch - reaction time (90V, 50%)
Switch - reaction time (50%)
Switch - reaction time (30V)
Switch - reaction time (30V, 50%)
Switch - accuracy (raw)
Switch - accuracy (90V)
Switch - accuracy (90V, 50%)
Switch - accuracy (50%)
Switch - accuracy (30V)
Switch - accuracy (30V, 50%)
Non-switch - reaction time (raw)
Non-switch - reaction time (90V)
Non-switch - reaction time (90V, 50%)
Non-switch - reaction time (50%)
Non-switch - reaction time (30V)
Non-switch - reaction time (30V, 50%)
Non-switch - accuracy (raw)
Non-switch - accuracy (90V)
Non-switch - accuracy (90V, 50%)
Non-switch - accuracy (50%)
Non-switch - accuracy (30V)
Non-switch - accuracy (30V, 50%)
Interference - reaction time (raw)
Interference - reaction time (90V)
Interference - reaction time (90V, 50%)
Interference - reaction time (50%)
Interference - reaction time (30V)
Interference - reaction time (30V, 50%)
Interference - accuracy (raw)
Interference - accuracy (90V)
Interference - accuracy (90V, 50%)
Interference - accuracy (50%)
Interference - accuracy (30V)
Interference - accuracy (30V, 50%)
Task covariates
No covariates
Demographic covariates
All covariates
0
200
400
600

Table D.17: Results of the specification curve analysis of the task-switching task by outcome measure

	Switch			Non-switch			Switch cost		
	N	Standardized β Median (range)	Significant No. (%)[a]	N	Standardized β Median (range)	Significant No. (%)[a]	N	Standardized β Median (range)	Significant No. (%)[a]
Accuracy									
CRF (binary)	3552	0.09 (0.08, 0.1)	24 (100)	3555	0.09 (0.08, 0.1)	24 (100)	3549	-0.01 (-0.02, 0)	0 (0)
CRF (continuous)	3552	0.1 (0.08, 0.12)	24 (100)	3555	0.09 (0.08, 0.11)	24 (100)	3549	-0.02 (-0.03, 0.01)	0 (0)
PA (continuous)	3488	0.05 (0.04, 0.08)	24 (100)	3490	0.06 (0.05, 0.09)	24 (100)	3486	0.02 (0, 0.03)	0 (0)
PA (binary)	3488	0.04 (0.02, 0.07)	8 (33)	3490	0.03 (0.02, 0.07)	6 (25)	3486	-0.01 (-0.03, 0)	0 (0)
Reaction time									
CRF (binary)	3548	0.03 (0, 0.04)	7 (29)	3552	-0.01 (-0.02, 0.01)	0 (0)	3549	0.05 (0.02, 0.06)	16 (67)
CRF (continuous)	3548	0.02 (-0.02, 0.04)	2 (8)	3552	-0.02 (-0.05, 0)	2 (8)	3549	0.05 (0.02, 0.06)	16 (67)
PA (continuous)	3484	0 (-0.01, 0.01)	0 (0)	3488	-0.02 (-0.04, -0.01)	0 (0)	3486	0.02 (0.01, 0.03)	0 (0)
PA (binary)	3484	-0.01 (-0.02, 0)	0 (0)	3488	-0.02 (-0.03, -0.01)	0 (0)	3486	0 (-0.02, 0.01)	0 (0)

Abbreviations: CRF = cardiorespiratory fitness, N = number of participants, PA= physical activity
[a] Number of significant tests and % of total number of tests run for that predictor and outcome

357

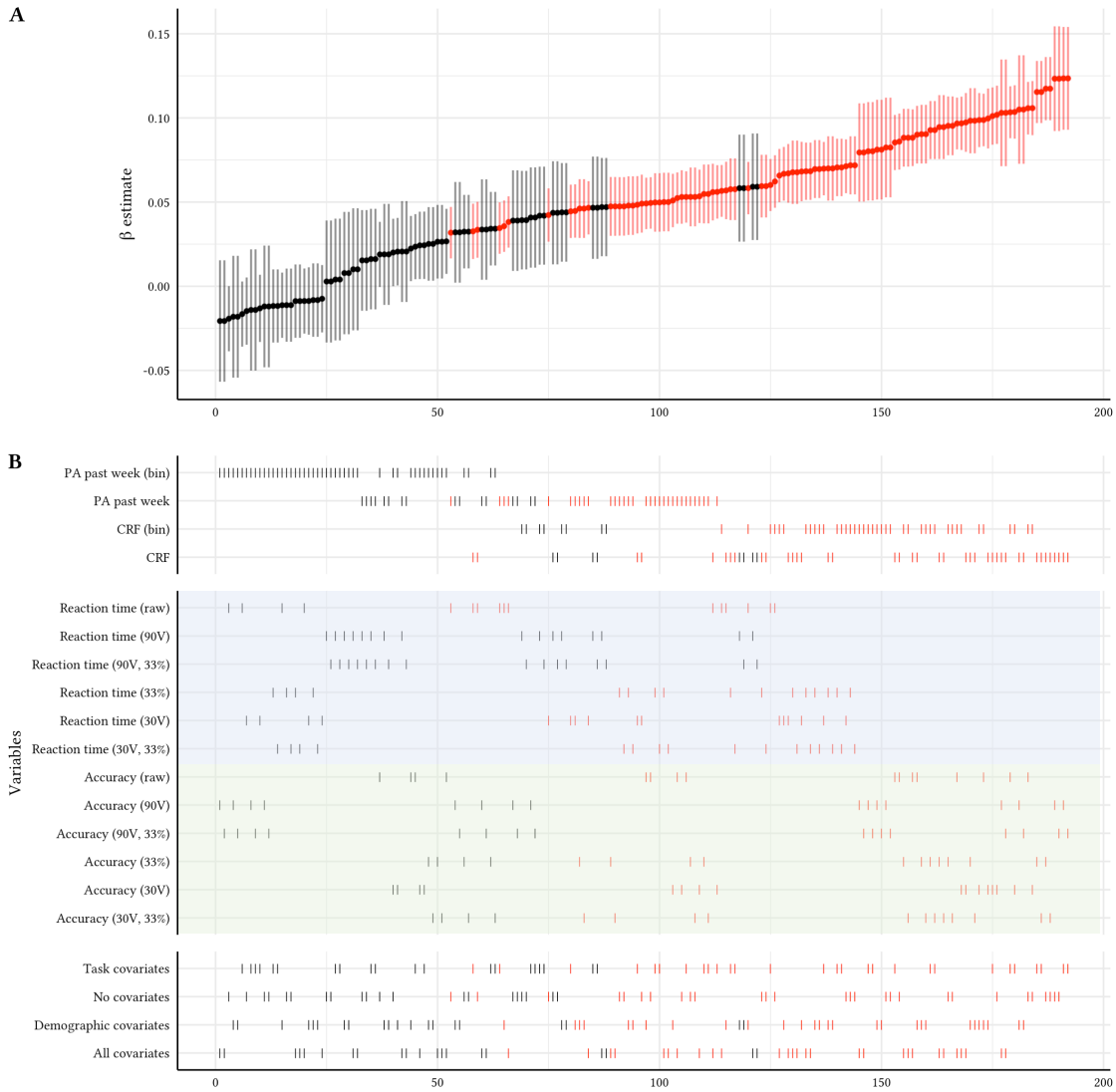


Figure D.11: Results of the specification curve working memory (two-back task). The specification curve analysis shows the range of possible results for multiple different specifications of a cross-sectional multilevel linear model assessing the relationship between PA or CRF and working memory. Each specification tests a different combination of cleaning procedures, outcomes, predictors and nuisance variables. In (*a*) the standardized regression estimates (y-axis) are displayed for each specification (x-axis), ordered by size of the estimate. Bars around estimates represent 95% confidence intervals. In (*b*) details of the specifications are provided, presented in three blocks that are supposed to be read vertically. The first (*top*) block contains the predictors, the second (*middle*) block presents the outcomes and the third (*bottom*)block shows the nuisance variables. Each column in (*b*) aligns with a datapoint in (*a*) and corresponds to a single specification (i.e. model tested), consisting of a predictor, outcome and nuisance variable, indicated by small vertical bars. Bars coloured in red denote specifications that resulted in a significant (p <0.05) relationship between the predictor and outcome. Outcomes contain details of the cleaning procedure: raw = no cleaning, 30V or 90V denotes a percentage of valid trials, e.g. 30% valid trials = 30V, 33% = chance-level (accuracy) cleaning. Outcomes and continuous predictors were standardized by 2 standard deviations.

The variability in size of the confidence intervals around β estimates in Figure D.11 may be explained by the variability in sample sizes. That is, when the 90V cleaning criteria are used, a test is performed on scores from approximately 1,100 participants, whereas other cleaning parameters leave a sample of N >3,000.

Table D.18: Results of the specification curve analysis of the two-back task by outcome measure

	Accuracy			Reaction time		
	N	Standardized β Median (range)	Significant No. (%)[a]	N	Standardized β Median (range)	Significant No. (%)[a]
CRF (binary)	3452	0.09 (0.08, 0.11)	24 (100)	3452	-0.06 (-0.07, -0.04)	16 (67)
CRF (continuous)	3452	0.1 (0.09, 0.12)	24 (100)	3452	-0.06 (-0.07, -0.03)	16 (67)
PA (continuous)	3362	0.05 (0.03, 0.06)	16 (67)	3362	-0.04 (-0.05, -0.02)	16 (67)
PA (binary)	3362	0.02 (-0.02, 0.03)	0 (0)	3362	0.01 (-0.01, 0.02)	0 (0)

Abbreviations: CRF = cardiorespiratory fitness, N = number of participants, PA = physical activity
[a] Number of significant tests and % of total number of tests run for that predictor and outcome

Back to main text

D.2.5 Canonical correlation analysis

D.2.5.1 Primary CCA

Canonical correlation analysis (CCA) was used to assess the link between a physically active lifestyle and cognitive performance. Table D.19 displays the canonical correlations and corresponding permutation-based p-values for each of the modes. The first two modes were statistically significant and therefore interrogated further. Table D.20 shows the cross-loadings for the first mode and Table D.21 shows the cross-loadings for the second mode.

Table D.19: Canonical correlations (ρ) and p-values for each mode

Modes	ρ	p
Mode 1	0.31	0.001
Mode 2	0.24	0.001
Mode 3	0.15	0.14
Mode 4	0.09	0.78
Mode 5	0.09	0.81
Mode 6	0.07	0.91
Mode 7	0.05	0.95
Mode 8	0.04	0.96
Mode 9	0.03	0.96

Abbreviations: ρ = canonical correlation, p = p-value

Table D.20: Cross-loadings mode 1

PA lifestyle	r	Cognition	r
CRF	0.19	Two-back (acc)	0.23
Habitual PA	0.14	Flanker task (congruent acc)	0.16
PA past week	0.13	Switch task (non-switch acc)	0.15
Attitude PA	0.09	Switch task (switch acc)	0.14
Sex	0.07	Two-back (RT)	0.13
Lifestyle × eFSM	0.02	Relational memory task (acc)	0.10
Lifestyle × sex × eFSM	-0.02	Flanker task (incongruent acc)	0.08
Lifestyle × sex	-0.06	Switch task (switch RT)	0.02
eFSM	-0.20	Flanker task (incongruent RT)	-0.02
		Switch task (non-switch RT)	-0.05
		Flanker task (congruent RT)	-0.09
		Reaction time task	-0.14

Abbreviations: acc = accuracy, eFSM = eligible for free school meals, PA = physical activity, r = Pearson correlation, RT = reaction time

Table D.21: Cross-loadings mode 2

PA lifestyle	r	Cognition	r
CRF	0.08	Flanker task (incongruent acc)	0.21
Attitude PA	0.07	Flanker task (incongruent RT)	0.17
Habitual PA	0.04	Flanker task (congruent RT)	0.16
PA past week	0.04	Flanker task (congruent acc)	0.08
eFSM	0.02	Switch task (switch acc)	0.06
Lifestyle × sex × eFSM	-0.01	Switch task (non-switch acc)	0.04
Lifestyle × sex	-0.04	Relational memory task (acc)	0.04
Lifestyle v eFSM	-0.06	Two-back (RT)	0.03
Sex	-0.22	Switch task (switch RT)	0.01
		Reaction time task	0.00
		Switch task (non-switch RT)	-0.02
		Two-back (acc)	-0.03

Abbreviations: acc = accuracy, eFSM = eligible for free school meals, PA = physical activity, r = Pearson correlation, RT = reaction time

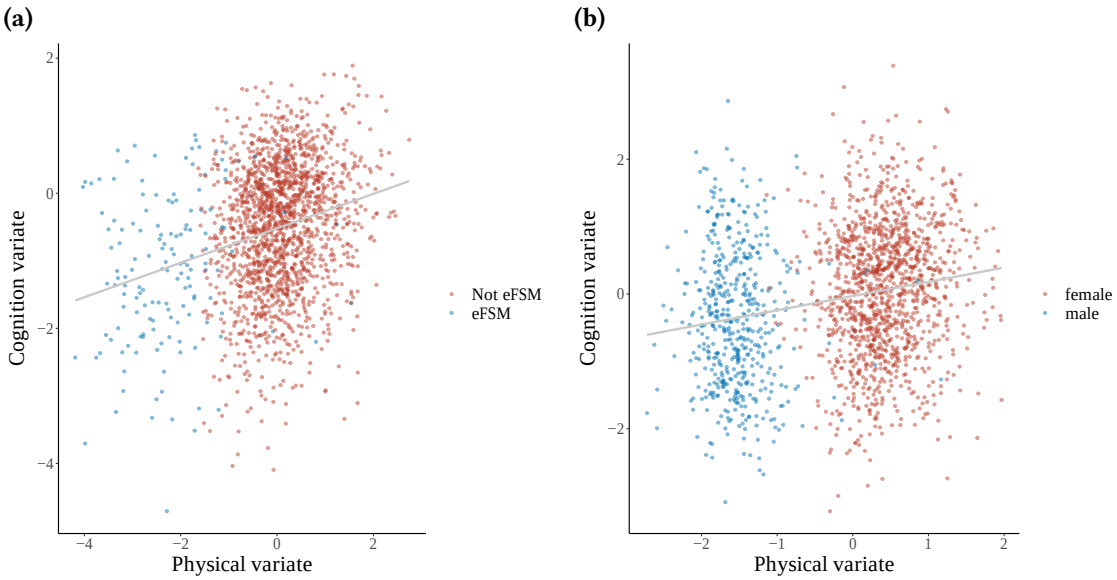


Figure D.12: Cross-validated physical and cognitive variates. (*a*) displays the cross-validated correlation between the first physical- and cognitive variate (mode 1). Participants on eFSM (blue) have generally lower physical and cognitive scores compared to those not eFSM (red). (*b*) shows the cross-validated correlation between the second physical- and cognitive variate (mode 2). Females (red) have generally higher scores on the physical- and cognitive variate compared to males (blue).

Cross-validated CCA

We used leave-on-school-out cross-validation to assess the generalisability of the CCA solutions. Figure D.12 displays the cross-validated canonical correlations for the first (ρ_{CV} = 0.26; Figure D.12a) and second mode (ρ_{CV} = 0.21; Figure D.12b).

Back to main text

D.2.5.2 Secondary analysis

In the secondary analysis, the CCA was repeated, but the PA lifestyle phenotype was adjusted for sex and eFSM. Table D.22 displays the canonical correlations and permutation-based p-values. Only the first canonical correlation was found to be statistically significant and further inspected. Figure D.13 shows a scatterplot of the first physical and cognitive variate, as well as the cross-loadings corresponding to the first mode of covariation. Table D.23 shows the values of the cross-loadings for the variates of the first CCA mode.

Table D.22: Canonical correlations (ρ) and p-values for each mode

Modes	ρ	p
Mode 1	0.24	0.001
Mode 2	0.09	0.77
Mode 3	0.07	0.78
Mode 4	0.05	0.95

Abbreviations: ρ = Pearson correlation coefficient, p = p-value

Table D.23: Cross-loadings mode 1

PA lifestyle	r	Cognition	r
CRF	0.22	Flanker task (congruent acc)	0.16
Habitual PA	0.16	Two-back (acc)	0.14
PA past week	0.15	Flanker task (incongruent acc)	0.13
Attitude PA	0.12	Switch task (non-switch acc)	0.11
		Switch task (switch acc)	0.10
		Relational memory task (acc)	0.09
		Two-back (RT)	0.07
		Flanker task (incongruent RT)	0.03
		Switch task (switch RT)	0.00
		Flanker task (congruent RT)	-0.01
		Switch task (non-switch RT)	-0.06
		Reaction time task	-0.14

Abbreviations: acc = accuracy, eFSM = eligible for free school meals, PA = physical activity, r = Pearson correlation, RT = reaction time

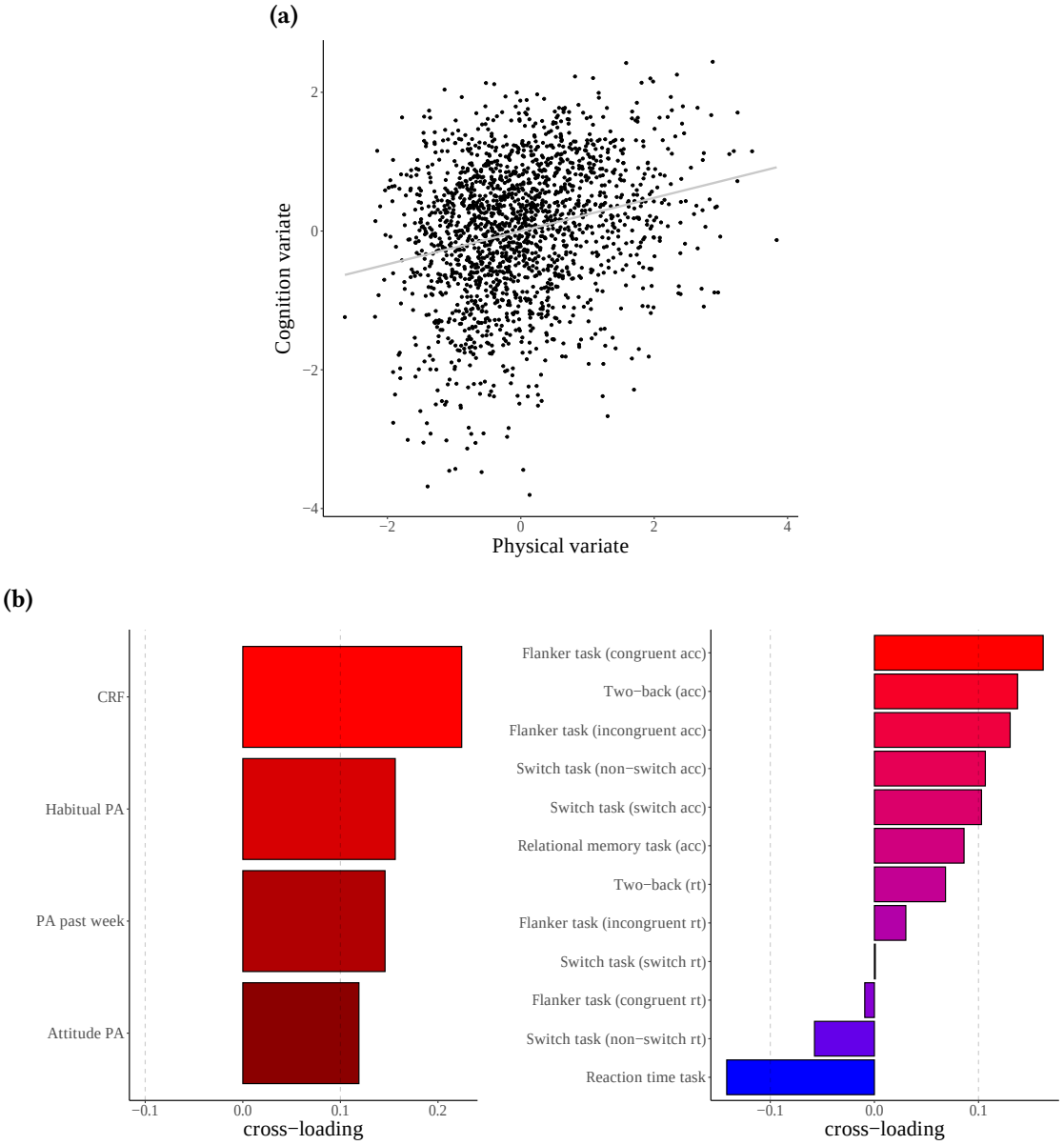


Figure D.13: Canonical correlation and cross-loadings of the first mode of covariation. (*a*) shows the ubject scores on the first physical and cognition variate, (*b*) presents the cross-loadings between the physically active lifestyle variables and the cognition variate (*left*) and between the cognitive variables and the PA lifestyle variate (*right*). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest.

Cross-validated CCA

We used leave-on-school-out cross-validation to assess the generalisability of the CCA solution. Figure D.14 displays the cross-validated correlation for the first mode (ρ_{CV} = 0.21).

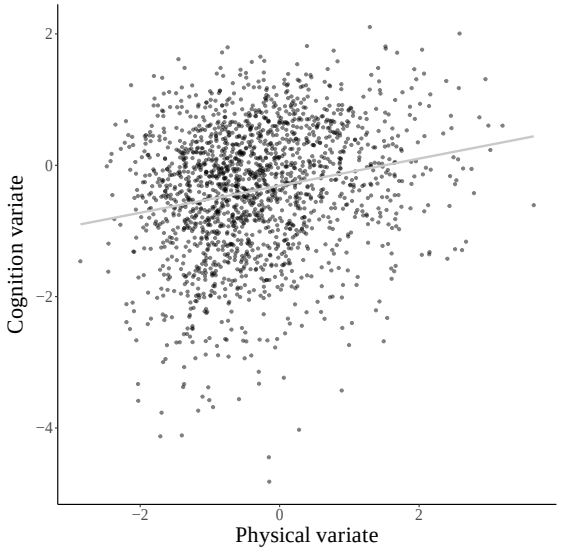


Figure D.14: Cross-validated physical and cognitive variates

Back to main text

E

Supplementary materials to chapter 5: *Associations between cardiorespiratory fitness, physical activity, the anterior hippocampus and memory*

Contents

E.1 Methods 366
E.1.1 Cognitive assessments 366
E.1.2 MRI acquisition 367
E.1.3 Hippocampal segmentation 369
E.1.4 Time-of-day effects 375
E.1.5 BLUS residuals 378
E.1.6 Missing data 379
E.1.7 Software 382
E.2 Results 382
E.2.1 Primary analyses 382
E.2.2 Entire hippocampus 386
E.2.3 Multivariate relationship between fitness, physical activity and brain MR markers 391

E.1 Methods

E.1.1 Cognitive assessments

1. **Object-location task** In the OLT (Pertzov *et al.,* 2012; Pertzov *et al.,* 2013), participants were shown three fractal objects, randomly located on an interactive touch screen. The locations of the fractals were determined by a MATLAB script (MathWorks, Inc). Participants were instructed to remember the identity of objects and their locations. Following a delay of 1 or 8 s, two fractals were displayed on the vertical meridian. One of these fractal objects was part of the original memory array (i.e. the target), whereas the other was not (i.e. the foils). Participants were instructed to touch the remembered object and "drag" it to its remembered location. Memory for object identity (accuracy) and object location were computed, as well as the number of *swap* or misbinding errors (i.e. the correct object was selected, but located near the location of one of the other fractals in the memory array). Swap errors are a proxy of relational memory and the proportion of swap errors was the primary outcome measure.

2. **Relational memory paradigm** A modified version of a previously described relational memory task (Chaddock *et al.,* 2010b) was used to assess visual relational memory performance. Participants were presented with two blocks of 18 study trials (i.e. events) containing scene, face and object stimuli. Within each trial, each type of stimulus was presented individually and sequentially (1,300 ms). Participants were instructed to memorise the stimuli within each trial as belonging to a single event. During the recognition phase, participants were shown 36 test trials, each containing a place, object and face stimulus, half of which were intact (i.e. the stimuli appeared in the same study trial). Participants were instructed to indicate whether all stimuli were from the same event, or whether one of the stimuli did not belong to the event. Responses within 200–15,000 ms were considered valid. The proportion of correct responses was the primary outcome measure.

Back to main text

E.1.2 MRI acquisition

Table E.1 shows the scan parameters of the main sequences collected as part of the Fit to Study brain imaging sub-study.

Table E.1: MRI scan parameters.

	Structural						Functional	
	T1-weighted	**DWI**	**qMRI**			**Angiography**	**pcASL**	**resting-fMRI**
Sequence	3D MPRAGE	multi-shell, multi-band EPI	3D multi-echo FLASH (PDw)	3D multi-echo FLASH (T1w)	Single-echo FLASH	3D time-of-flight	single shot EPI	multi-band EPI
TR	1900	2483	25	25	4	21	4100	933
TE (ms)	3.97	78.2	2.34	2.34	2	3.58	14	33.4
Flip angle (°)	8	80	6	21	5	18	90	60
Voxel size (mm^3)	1×1×1	1.75×.175×1.75	1×1×1	1×1×1	4×4×4	0.6×0.6×0.6	3.4×3.4×3.5	2×2×2
FoV (mm)	192	214	256	256	256	192	220	192
Acceleration factor	-	4	2	2	-	3	-	6
b-value (s/mm^2)	-	0, 1250, 2500	-	-	-	-	-	-
No. of volumes	-	-	-	-	-	-	96	644
No. of directions	-	11, 60, 60	-	-	-	-	-	-
Labelling delays (s)	-	-	-	-	-		0.25, 0.5, 0.75, 1, 1.25, 1.5	-
No. of slabs/slices	-	-	-	-	-	4/32	-	-
Scan time (min)	5.31	5.4	5.11	5.11	0.13	6	5.34	10.1

Abbreviations: DWI = diffusion-weighted imaging, pcASL = pseudo-continuous Arterial Spin Labelling, qMRI = quantitative MRI

368

Back to main text

E.1.3 Hippocampal segmentation

E.1.3.1 Quantitative and qualitative comparison of algorithms

We inspected the performance of the frequently used, freely available hippocampal segmentation algorithms FSL FIRST and FreeSurfer, and the recently developed tool MIST (Visser *et al.*, 2016). While it was outside the scope of this thesis to perform an in-depth comparison of algorithms —which would involve a formal comparison against the gold standard manual segmentations, see e.g. Mulder *et al.*, 2014; Schoemaker *et al.*, 2016 —we qualitatively (visually) and quantitatively explored their performance to help decide which algorithm to use for hippocampal segmentations.

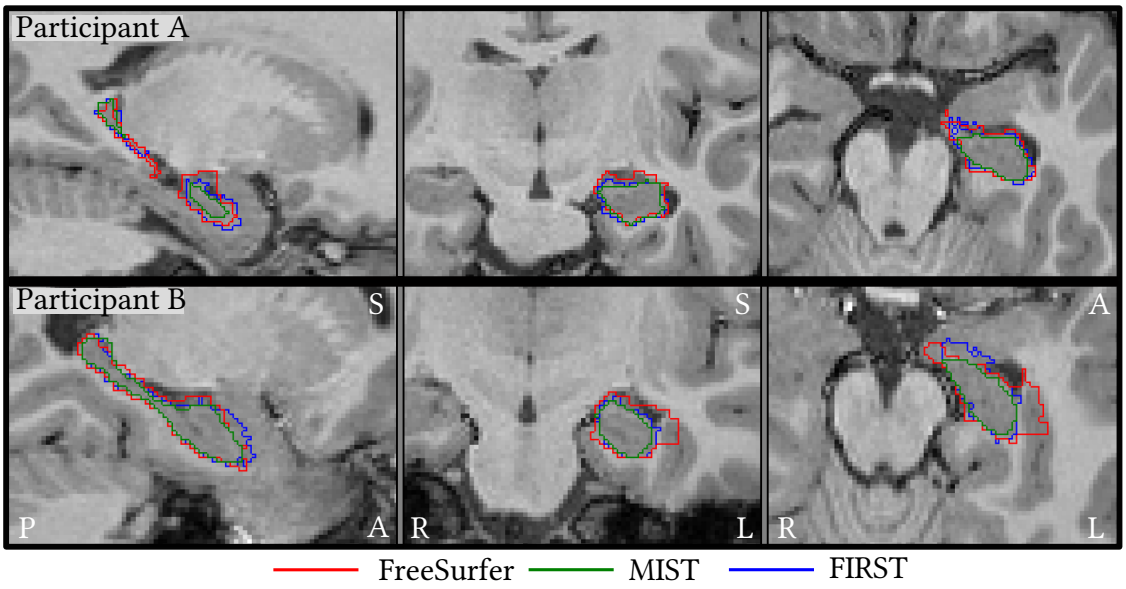


Figure E.1: Qualitative comparison of automated hippocampal segmentation algorithms. Outlines of hippocampal segmentations from FreeSurfer (red), MIST (green) and FIRST (blue) in the sagittal plane (left column), coronal plane (middle column) and transverse plane (right column) are presented for participants A and B. Abbreviations: A = anterior, L = left, P = posterior, R = right, S = superior.

To that end, we visually inspected hippocampal segmentations from each algorithm. Figure E.1 illustrates segmentations of the left hippocampus from FreeSurfer (red), MIST (green) and FIRST (blue) for two participants. FreeSurfer and FIRST seem to overestimate the hippocampal borders, whereas MIST may slightly underestimate the hippocampus. Figure E.2 demonstrates a qualitative comparison of left and right

hippocampal volumes for the entire cohort (at baseline), showing higher estimated hippocampal volumes from FreeSurfer and FIRST compared to MIST. The volumes estimated by MIST are in the range of volumes derived from manual segmentations in a comparable cohort of preadolescents (average volumes between 2,700—2,800 mm^3, Schoemaker *et al.*, 2016). Moreover, the overestimation of hippocampal volumes by FIRST and FreeSurfer is in line with previous literature (Cherbuin *et al.*, 2009; Morey *et al.*, 2009; Schoemaker *et al.*, 2016). Based on these observations, we decided to use MIST for hippocampal segmentations.

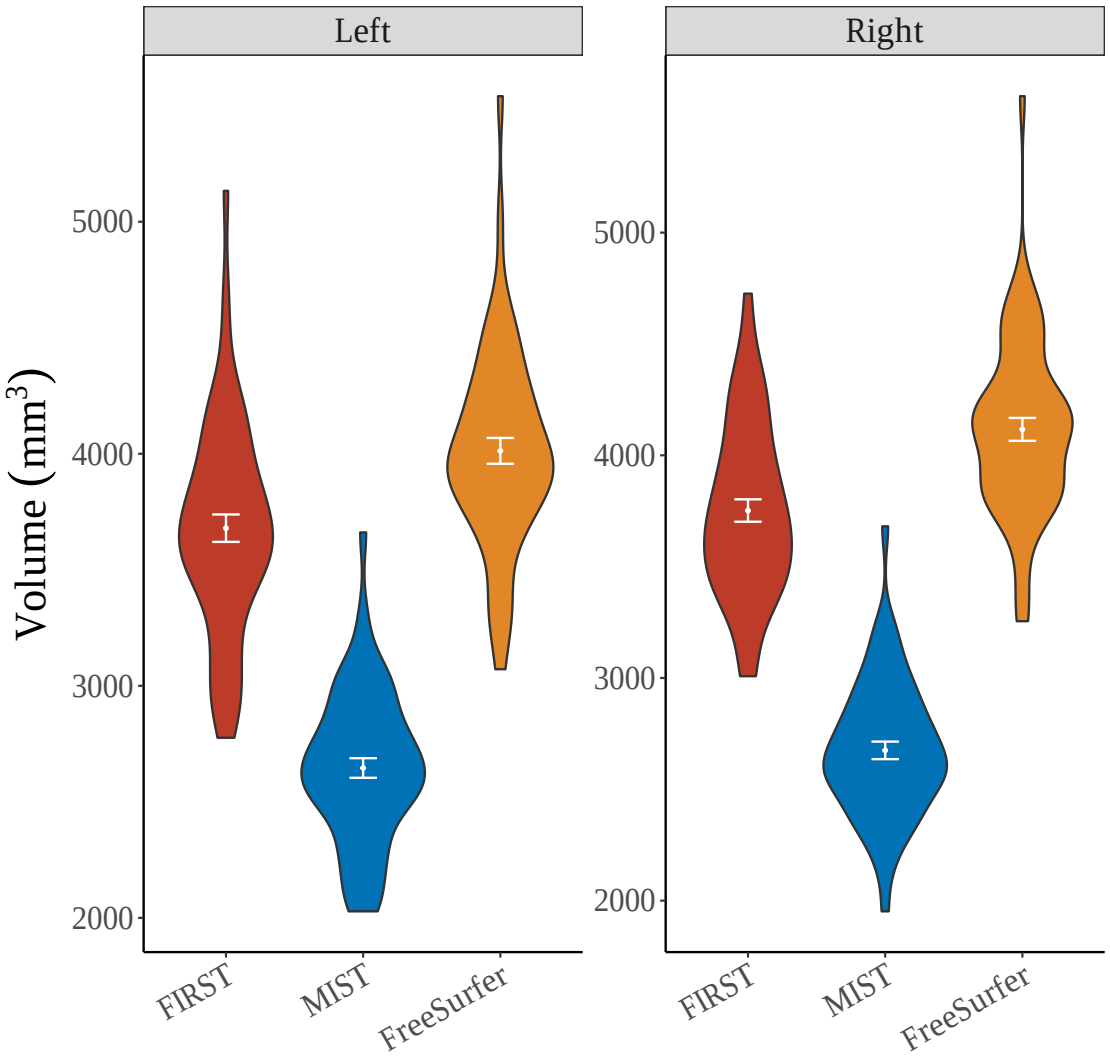


Figure E.2: Quantitative comparison of automated hippocampal segmentation algorithms. Large differences are observed between the estimated volumes from FreeSurfer, MIST and FIRST. Error bars reflect standard errors of the mean.

Back to main text

E.1.3.2 Manual adjustments

Visual inspection of the MIST-derived hippocampal segmentations indicated a considerable degree of variability in the accuracy of segmentations of the hippocampus across subjects (Figure E.3). All hippocampi where therefore manually edited. The main corrections were made in regions that are among the major sources of variation between segmentation protocols (Konrad *et al.*, 2009):

- Posteriorly: at the level of the tail of the hippocampus, voxels extending into the CSF and thalamus
- Supero-medially: voxels extending into the fimbria
- Anterior-inferiorly: voxels extending into the amygdala
- Inferolaterally: voxels extending into the temporal horn of the lateral ventricle, or voxels missing at the infero-lateral border of the hippocampus with the surrounding white matter

Voxels extending into the CSF and neighbouring non-hippocampal structures (e.g. the amygdala and thalamus) were removed. The alveus is a clear internal landmark that is commonly used to demarcate the hippocampal-amygdala border. In line with Geuze *et al.* (2005), the alveus was included in its entirety. The fimbria is clearly visible on T1-weighted MR scans, does not belong to the hippocampal formation and was therefore removed.

To assess inter-rater and intra-rater reliability of the manual adjustments, the MIST-derived hippocampi of 10 randomly chosen subjects were manually adjusted twice (two weeks apart) by two independent raters (ND and TW). The intraclass correlation coefficient (ICC) was computed for the left and right hippocampus separately, using a two-way mixed effects model (using *R* package *psych*, Revelle, 2018) examining the agreement between raters and within raters over time (Koo & Li, 2016).

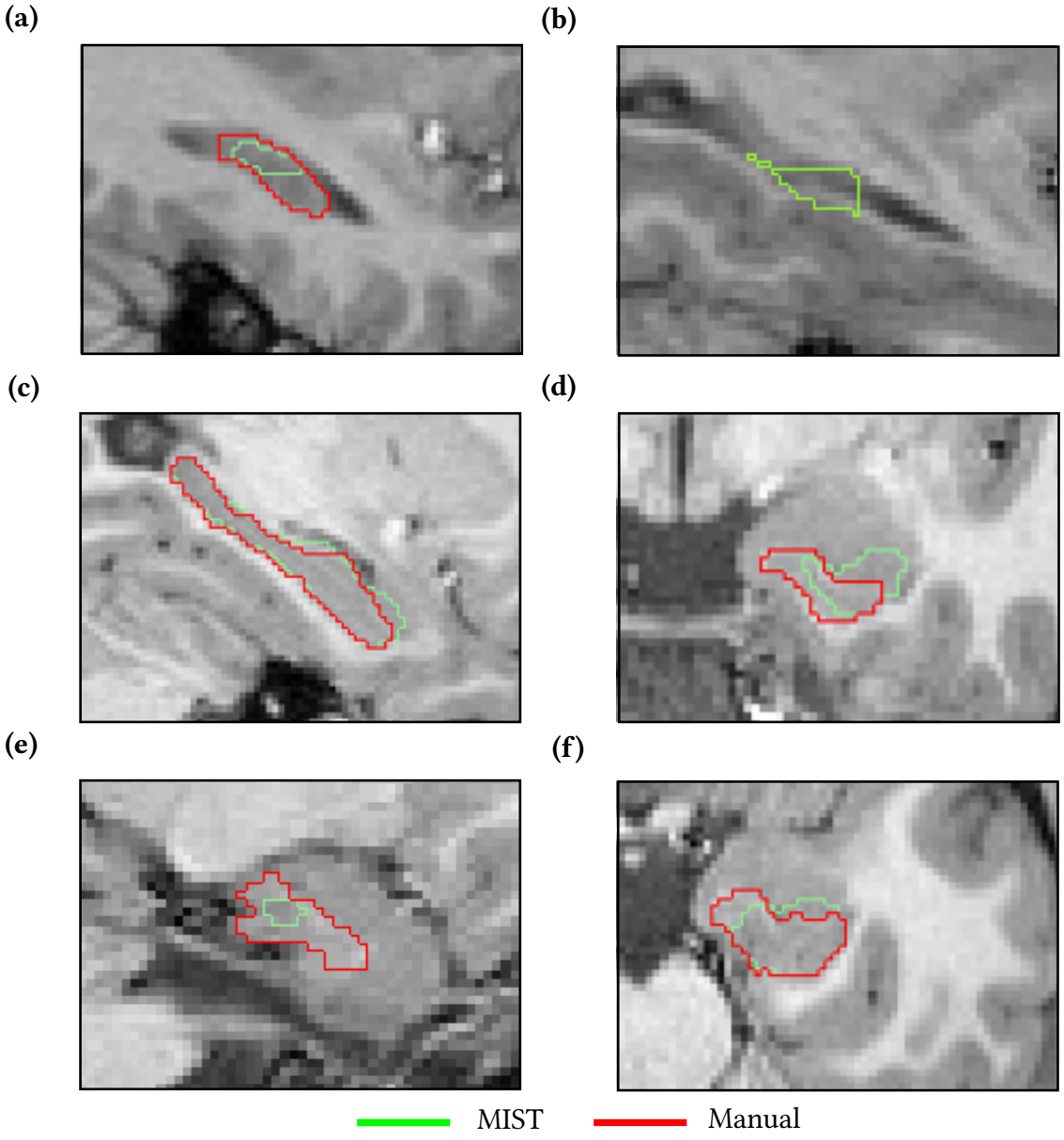


Figure E.3: Illustration of manual edits and MIST errors. A sample of segmentations was selected to illustrate the variability and inaccuracy in MIST segmentations that are particularly relevant if one is interested in the anterior hippocampus. MIST segmentations are presented in green, and manual edits in red, depicting: (*a*) a cross-section of the brain in the sagittal plane, showing that MIST misses out on a substantial portion of the inferolateral hippocampus, (*b*) a cross-section of the brain in the sagittal plane in which MIST included non-hippocampal tissue, which has been removed manually, (*c*) a cross-section of the brain in the sagittal plane, depicting an overshoot of the MIST segmentation into the amygdala (anterior-inferiorly), (*d*) a cross-section of the brain in the coronal plane showing that MIST included amygdala tissue and excluded hippocampal tissue (inferomedially), (*e*) a cross-section of the brain in the sagittal plane, showing that MIST excluded a large portion of the hippocampus inferomedially, (*f*) a cross-section of the brain in the axial plane, showing that MIST excluded hippocampal voxels (inferiomedially) and included amygdala tissue.

The ICC between raters was, for respectively left and right hippocampus, 0.9 (95% CI: 0.67–0.97) and 0.9 (95% CI: 0.21–0.98). The ICCs reflecting intra-rater reliability (for rater TW) was 0.9 (95% CI: 0.56–0.976) and 0.91 (95% CI: 0.49–0.98). These reliability statistics are generally considered good to excellent, and are comparable to reliability statistics found for the majority of manual segmentation protocols (Boccardi *et al.*, 2011; Konrad *et al.*, 2009).

While the ICC is a useful metric to assess agreement, it does not take the spatial domain into account. Therefore, the Dice similarity coefficient (DSC, Dice, 1945) was computed to assess the degree of spatial overlap between the manually adjusted segmentations of the two raters.

$$DSC = \frac{2|N_A \cap N_B|}{N_A + N_B} \tag{E.1}$$

where N_A and N_B are the number of voxels in masks A and B respectively, and N_{AB} is the number of voxels of the region where the masks overlap. The DICE coefficients were 0.91 and 0.92 for left and right hippocampus respectively. The intra-rater (rater TW) DICE coefficient (i.e. across time) was 0.94 for both hippocampi. To put these numbers into perspective: DICE coefficents for many automated methods (compared to manual) generally vary between 0.8–0.9 (Dill *et al.*, 2015).

Back to main text

E.1.3.3 Anterior hippocampal segmentation

In line with recommendations by Poppenk *et al.* (2013), we used the uncal apex as internal anatomical landmark for delineation. The uncus is particularly developed in humans and primates (Duvernoy, 2005), and is clearly visible on T1-weighted scans (Poppenk *et al.*, 2013; Zeidman & Maguire, 2016). We aligned the T1-weighted scan with the MNI-152 template using FSLs FLIRT and then defined the anterior hippocampus as the region anterior to the most posterior coronal slice in which the uncal apex was visible (Figure E.4).

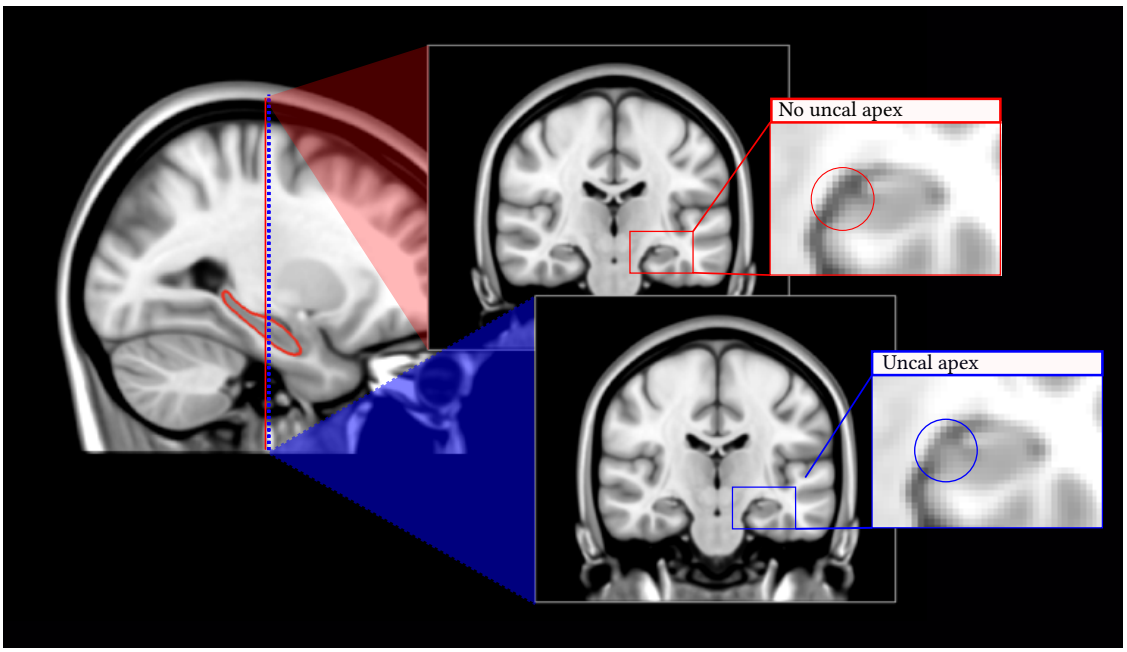


Figure E.4: Demarcation of the anterior hippocampus. The hippocampus was divided into an anterior and posterior portion based on the location of the uncal apex. That is, the most posterior coronal slice that included the uncal apex was considered the posterior border of the anterior hippocampus (in blue). The Y-coordinate of the uncal apex was approximately located at Y = -20 in MNI space, yet there was considerable inter-individual variability in its exact location (see Figure E.5)

Visual inspection of the results of this method suggested that segmentations were more accurate (i.e. better aligned with the uncal apex) compared to MNI-coordinate based segmentation due to inter-individual variability of the location of the uncal apex in MNI space. Figure E.5 shows the distribution of Y-coordinates (MNI space) where the uncal apex became visible across subjects (for illustration purposes, data from both timepoints were included in the plot) for the left (median = -20, IQR = 2.0) and right (median = -20, IQR = 2.25) hemisphere.

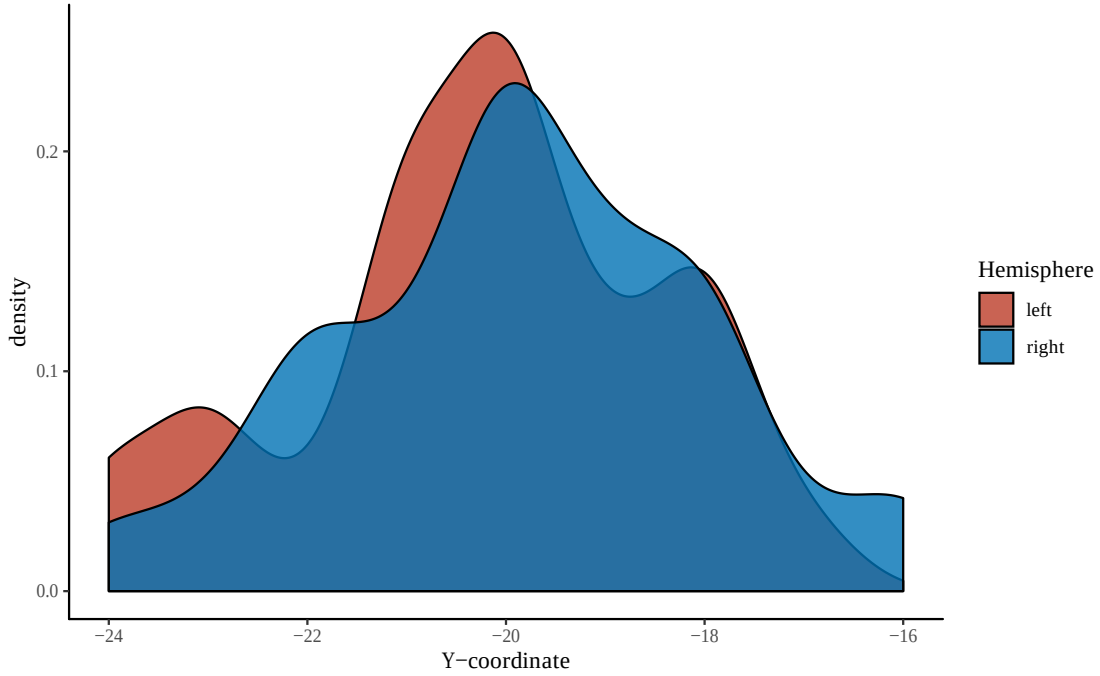


Figure E.5: Variability of the Y-coordinate of the uncal apex in MNI152-standard space across participants.

The demarcation was performed by a single rater, blind to group allocation and in random order. We assessed inter- and intra-rater reliability of the detection of the uncal apex (i.e. the Y-coordinate in MNI-space). The ICC between raters was, for respectively left and right hippocampus, 0.96 (95% CI: 0.81–0.99) and 0.987 (95% CI: 0.95–0.996). The ICCs reflecting intra-rater reliability (for rater TW) was 0.97 (left, 95% CI: 0.90–0.99) and 0.993 (right, 95% CI: 0.98–0.998).

Back to main text

E.1.4 Time-of-day effects

Previous literature has observed robust time-of-day effects on various brain MRI metrics (Jiang *et al.*, 2014; Karch *et al.*, 2019; Nakamura *et al.*, 2015; Thomas *et al.*, 2018; Trefler *et al.*, 2016), and has suggested to adjust for this nuisance variable in the analysis. Here, we explored the relationship between acquisition time of the MRI sequence (expressed in hours, following Karch *et al.*, 2019) and hippocampal MRI markers. In particular, we found no significant (zero-order) relationships between time

of day and (anterior) hippocampal volume (Figure E.6). We additionally corrected for differences in head size (using the scaling factor of SIENAX, Smith, 2002) in univariate linear models —due a lack of school-level variability —but no significant relationships were observed (Table E.2). However, in line with findings from previous studies (Karch *et al.*, 2019; Trefler *et al.*, 2016), a reduction in (anterior) hippocampal volume was observed from morning to afternoon.

Table E.2: The relationship between scan acquisition time and hippocampal volume. The relationship between each volumetric marker and time-of-day was assessed using a univariate linear model (adjusted for head size). Statistics (β, 95% CI, and p) are presented for the time-of-day variable

Dependent variable	$\beta^{\dagger}$	95% CI	p	p-adjusted‡
Left hippocampus	-0.12	-0.33, 0.1	0.28	0.37
Right hippocampus	-0.19	-0.4, 0.03	0.08	0.26
Left anterior hippocampus	-0.11	-0.35, 0.14	0.38	0.38
Right anterior hippocampus	-0.19	-0.43,0.06	0.13	0.26

† Standardized estimates, i.e. the variables were standardized (mean = 0, SD = 0) prior to model fitting
‡ FDR-adjusted p-values.

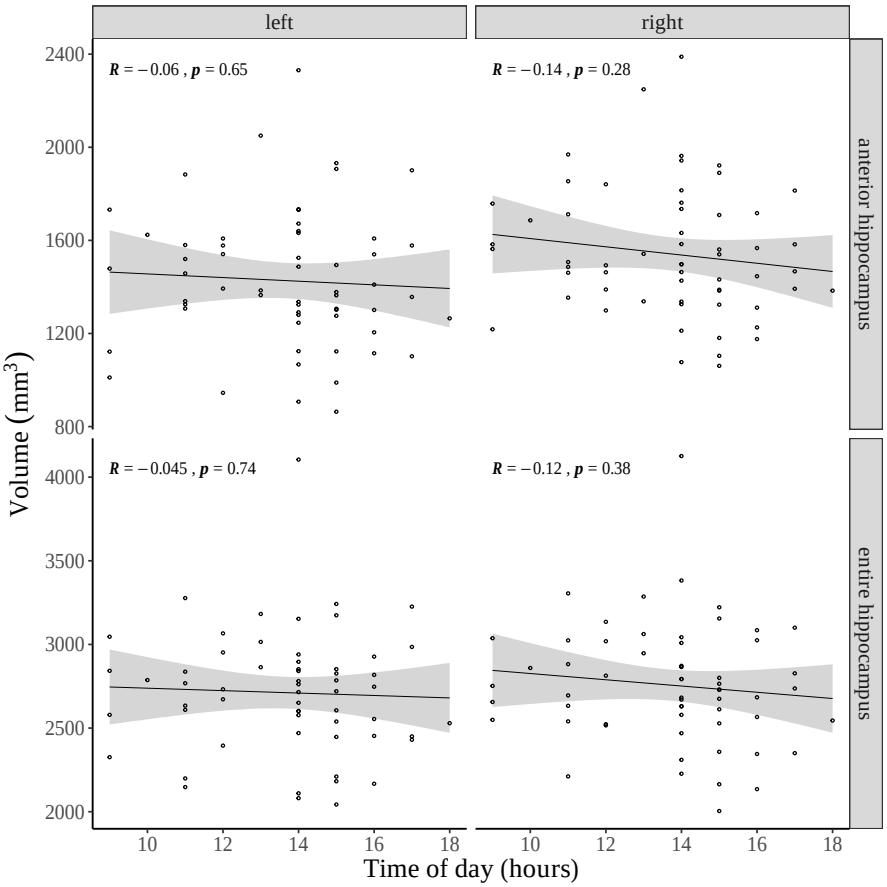


Figure E.6: The relationship between scan acquisition time and left or right (anterior) hippocampal volume. *Left column* shows the relationship between time-of-day (x-axis) and *left* anterior hippocampal volume (*top row*), and left hippocampal volume (*bottom row*). *Right column* shows the relationship between time-of-day (x-axis) and *right* anterior hippocampal volume (*top row*), and right hippocampal volume (*bottom row*). Models were fit using single-level linear models due to singular fits.

We also explored the relationship between time of day and MRI markers of tissue microstructure within the anterior hippocampus, adjusted for head size (Table E.3). While the relationship was significant for R_1 and perfusion in the right hippocampus, false-disovery rate (FDR) correction for multiple comparisons yielded no significant relationships. The regression coefficient is negative for the majority of metrics, suggesting a reduction in these metrics from morning to afternoon.

Table E.3: The relationship between scan acquisition time and hippocampal MRI markers of tissue (micro)structure. The relationship between each MRI marker and time-of-day was assessed using a univariate linear model (adjusted for head size). Statistics (β, 95% CI, and p) are presented for the time-of-day variable.

Dependent variable	$\beta^{\dagger}$	95% CI	p	p-adjusted‡
Left fintra	-0.09	-0.29, 0.1	0.36	0.58
Right fintra	-0.15	-0.29, 0.04	0.12	0.30
Left MD	0.09	-0.34, 0.28	0.38	0.58
Right MD	0.15	-0.11, 0.34	0.12	0.30
Left perfusion	-0.18	-0.04, 0.02	0.08	0.30
Right perfusion	-0.23	-0.38, -0.03	0.02	0.29
Left R_1	-0.15	-0.43, 0.06	0.16	0.32
Right R_1	-0.21	-0.42,0.002	0.05	0.29
Left R_2^*	0.03	-0.42, 0.24	0.80	0.80
Right R_2^*	0.07	-0.19, 0.28	0.49	0.59
Left VBM	-0.07	-0.13, 0.12	0.45	0.59
Right VBM	0.05	-0.27, 0.24	0.59	0.64

Abbreviations: fintra = intra-axonal compartment of NODDI, MD = mean diffusivity, VBM = voxel-based morphometry
† Standardized estimates, i.e. the variables were standardized (mean = 0, SD = 0) prior to model fitting
‡ FDR-adjusted p-values.

Back to main text

E.1.5 BLUS residuals

The BLUS residuals were computed in R using the function defined below (adapted from TE Nichols, Oxford).

```
BLUSres <- function(Y,X,...){

  #This script was adapted from TE Nichols, Sept 2019
  #Coded in R: TWassenaar Sept 2019

  require(expm)
  require(dplyr)
  require(lintools)
  require(matlib)
  #Start script:
  argss = list(...)

  #Make all factors numeric
  X <- X %>% mutate_if(is.factor, as.numeric)
  Y <- Y %>% mutate_if(is.factor, as.numeric)

  demean <- function(x, na.rm = FALSE) (x - mean(x, na.rm = na.rm))
  #Mean center all variables
```

```
  X <- X %>% mutate_if(is.numeric, demean, na.rm = TRUE)
  Y <- Y %>% mutate_if(is.numeric, demean, na.rm = TRUE)

  Y <- as.matrix(Y)
  X <- as.matrix(X)
  N <- nrow(Y)
  P <- ncol(X)
  if (!(nrow(X)==N)){
    print("error rows and and X and Y don't match")}

  stopifnot(N>0)
  if (length(argss) > 0){
    I1 <- argss[[1]]
    I0 <- setdiff(1:N,I1)
    if (qr(X[I0,])$rank < P){
      print("error omitted subset not full rank")}}
  else {
    Done <- TRUE
    while (Done)
    {
      I1 = sample(N,N-P)
      I0 = setdiff(1:N,I1)
      stopifnot(I0 > 0)
      if (qr(X[I0,])$rank == P){
        Done <- FALSE
      }
    }
  }

  M <- diag(N) - X %*% solve(t(X)%*%X) %*% Conj(t(X))
  S <- diag(N)
  S <- S[,I1]
  A <- M %*% S %*% sqrtm(solve(Conj(t(S)) %*% M %*% S))
  u <- Conj(t(A)) %*% Y
  return(list(u = u, I1=I1, I0 = I0))
}
```

Back to main text

E.1.6 Missing data

Missing data was imputed using a single multilevel singular value decomposition (SVD, Husson *et al.*, 2019). The variables included in the imputation procedure were the main outcome variables of the assessments, displayed in Table E.4. An intervention ("group") dummy was included to capture any intervention effects.

Table E.4: Overview of variables included in the imputations. Variables ending with ".0" were collected at baseline, whereas those ending with ".1" were collected at follow-up.

Variable names		
School	inactive_similar.0	hours_weekendGaming.0
sex	willingness.0	hours_weekPC.0
age.0	paDaysPastWeek.0	hours_weekendPC.0
height.0	id1s.0	breakfastWeek.0
weight.0	id8s.0	breakfastWeekend.0
bmi.0	error1s.0	timesFruit.0
pubertalDevelopmentScale.0	error8s.0	timesVegetables.0
highestDegreeParent.0	ni1s.0	timesCandy.0
yearsEduParent.0	ni8s.0	timesSoftdrinks.0
imdDecile.0	swap1s.0	group
inattentionSubPercentile.0	swap8s.0	L_Hipp.0
hyperactivitySubPercentile.0	percentCorrectHeterogeneousValid.0	R_Hipp.0
protocol.0	percentCorrectHomogeneousValid.0	L_Hipp_ant.0
sysBP.0	percentCorrectNonswitchValid.0	R_Hipp_ant.0
diaBP.0	percentCorrectSwitchValid.0	timeOfDay.0
restHR.0	reactionTimeHeterogeneousCorrectMean.0	Vscaling.0
vo2KgPeak.0	reactionTimeHomogeneousCorrectMean.0	timeEndMins.0
hrMax.0	reactionTimeNonswitchCorrectMean.0	VPA_schoolhours.0
workloadMax.0	reactionTimeSwitchCorrectMean.0	age.1
rerMax.0	version.0	height.1
rpeEndTest.0	countValid.0	weight.1
totWeekVPA1s.0	countValidCorrect.0	bmi.1
totWeekMPA1s.0	elapsedtime.0	pubertalDevelopmentScale.1
totWeekMVPA1s.0	correctRejectionRateValid.0	protocol.1
totWeekSed1s.0	falseAlarmRateValid.0	sysBP.1
avgWeekDayVPA1s.0	hitRateValid.0	diaBP.1
avgWeekendDayVPA1s.0	missRateValid.0	restHR.1
avgWeekDayMVPA1s.0	percentCorrect.0	vo2KgPeak.1
avgWeekendDayMVPA1s.0	percentCorrectValid.0	hrMax.1
dailyAvgVPA1s.0	reactionTimeCorrectMean.0	workloadMax.1
avgDailyMVPA1s.0	reactionTimeCorrectRejectionMean.0	rerMax.1
avgWeekDay60s.0	reactionTimeHitMean.0	rpeEndTest.1
avgWeekendDay60s.0	sleepConditionIndicator.0	totWeekVPA1s.1
dailyAvgMVPA_60s.0	adolescentSleepQuest.0	totWeekMPA1s.1
totWeekVPA60s.0	totMoodDisturbance.0	totWeekMVPA1s.1
totWeekMPA60s.0	prosocialScale.0	totWeekSed1s.1
totWeekSedentary60s.0	emotionalScale.0	avgWeekDayVPA1s.1
dScoleSimilar.0	conductScale.0	avgWeekendDayVPA1s.1
dscoreFavourable.0	hyperactivityScale.0	avgWeekDayMVPA1s.1
paPastSixMonths.0	peerScale.0	avgWeekendDayMVPA1s.1
intentToBeActive.0	selfRatedQoL.0	dailyAvgVPA1s.1
attitude.0	selfRatedHealth.0	avgDailyMVPA1s.1
subjNorms.0	subjHealthIndex.0	avgWeekDay60s.1
perceivedControl.0	timesPA.0	avgWeekendDay60s.1
active_favourable.0	hours_weekPA.0	dailyAvgMVPA_60s.1
active_similar.0	hours_weekTV.0	totWeekVPA60s.1
inactive_favourable.0	hours_weekendTV.0	totWeekMPA60s.1
totWeekSedentary60s.1	hours_weekGaming.0	R1_right.1
dScoleSimilar.1	hyperactivityScale.1	FA_right.1
dscoreFavourable.1	peerScale.1	MD_right.1

Continued on next page

Table E.4 – continued from previous page

Variable names		
paPastSixMonths.1	selfRatedQoL.1	Kurtosis_right.1
intentToBeActive.1	selfRatedHealth.1	fintra_right.1
attitude.1	subjHealthIndex.1	ODI_right.1
subjNorms.1	timesPA.1	VBMs_right.1
perceivedControl.1	hours_weekPA.1	Jacobian_left.0
active_favourable.1	hours_weekTV.1	perfusion_left.0
active_similar.1	hours_weekendTV.1	perfusion_pvcorr_left.0
inactive_favourable.1	hours_weekGaming.1	arr_tim_left.0
inactive_similar.1	hours_weekendGaming.1	R1_left.0
willingness.1	hours_weekPC.1	R2star_left.0
paDaysPastWeek.1	hours_weekendPC.1	FA_left.0
id1s.1	breakfastWeek.1	MD_left.0
id8s.1	breakfastWeekend.1	Kurtosis_left.0
error1s.1	timesFruit.1	fintra_left.0
error8s.1	timesVegetables.1	ODI_left.0
ni1s.1	timesCandy.1	VBMs_left.0
ni8s.1	timesSoftdrinks.1	Jacobian_right.0
swap1s.1	warmUp.1	perfusion_right.0
swap8s.1	infusion.1	perfusion_pvcorr_right.0
percentCorrectHeterogeneousValid.1	participation.1	arr_tim_right.0
percentCorrectHomogeneousValid.1	maths.1	R1_right.0
percentCorrectNonswitchValid.1	L_Hipp.1	R2star_right.0
percentCorrectSwitchValid.1	R_Hipp.1	FA_right.0
reactionTimeHeterogeneousCorrectMean.1	L_Hipp_ant.1	MD_right.0
reactionTimeHomogeneousCorrectMean.1	R_Hipp_ant.1	Kurtosis_right.0
reactionTimeNonswitchCorrectMean.1	timeOfDay.1	fintra_right.0
reactionTimeSwitchCorrectMean.1	Vscaling.1	ODI_right.0
version.1	timeEndMins.1	VBMs_right.0
countValid.1	VPA_schoolhours.1	MTbeforeST.0
countValidCorrect.1	Jacobian_left.1	
elapsedtime.1	perfusion_left.1	
correctRejectionRateValid.1	perfusion_pvcorr_left.1	
falseAlarmRateValid.1	arr_tim_left.1	
hitRateValid.1	R1_left.1	
missRateValid.1	R2star_left.1	
percentCorrect.1	FA_left.1	
percentCorrectValid.1	MD_left.1	
reactionTimeCorrectMean.1	Kurtosis_left.1	
reactionTimeCorrectRejectionMean.1	fintra_left.1	
reactionTimeHitMean.1	ODI_left.1	
sleepConditionIndicator.1	VBMs_left.1	
adolescentSleepQuest.1	Jacobian_right.1	
totMoodDisturbance.1	perfusion_right.1	
prosocialScale.1	perfusion_pvcorr_right.1	
emotionalScale.1	arr_tim_right.1	
conductScale.1	R2star_right.1	

Back to main text

E.1.7 Software

All analyses were performed in R (version 6.0). Multilevel models were fit using *lme4* (Bates *et al.*, 2015) and *lmerTest* (Kuznetsova *et al.*, 2017) packages. The CCA was performed using functions available in the *candisc* package (Friendly & Fox, 2020).

Back to main text

E.2 Results

E.2.1 Primary analyses

E.2.1.1 Available-case analysis

Table E.5 presents the unstandardized results of the models assessing associations between daily average MVPA or CRF and left or right anterior hippocampal volumes.

Table E.5: Analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and left or right anterior hippocampal volume (unstandardized).

	Left anterior hippocampus					Right anterior hippocampus				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
$\dot{V}O_2$peak	56	7.73	-3.6, 19.05	0.18	0.26	56	4.66	-6.36, 15.68	0.40	0.26
MVPA[a]	44	0.41	-3.33, 4.16	0.83	0.14, 0.2	44	-0.55	-4.29, 3.2	0.77	0.2
Model 2										
$\dot{V}O_2$peak	56	6.39	-7.6, 20.39	0.43	0.25, 0.29	56	5.06	-8.2, 18.31	0.45	0.31
MVPA[a]	44	0.26	-3.66, 4.18	0.90	0.17, 0.29	44	-0.3	-4.35, 3.75	0.88	0.25

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.6 presents the unstandardized results of the models assessing associations between daily average MVPA or CRF and memory outcomes.

Table E.6: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and memory outcomes (unstandardized).

	Object-location task (swap errors, %)					Relational memory paradigm (accuracy, %)				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
$\dot{V}O_2$peak	57	-0.28	-0.59, 0.04	0.12	0.17, 0.22	55	0.22	-0.25, 0.69	0.36	0.1
MVPA[a]	45	0.06	-0.06, 0.18	0.35	0.14, 0.29	44	-0.01	-0.19, 0.18	0.95	0.11
Model 2										
$\dot{V}O_2$peak	57	-0.13	-0.55, 0.30	0.59	0.18, 0.29	55	0.02	-0.58, 0.62	0.95	0.2
MVPA[a]	45	0.06	-0.07, 0.19	0.37	0.14, 0.28	44	-0.01	-0.19, 0.18	0.95	0.25

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.7 presents the unstandardized results of the models assessing associations between daily left and right anterior hippocampal volumes and memory outcomes.

Table E.7: Univariate analyses assessing relationships between left and right anterior hippocampal volumes and memory outcomes (unstandardized).

	Object-location task (swap errors, %)					Relational memory paradigm (accuracy, %)				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
Left ant. hip.	56	0.003	-0.005, 0.012	0.45	0.15, 0.24	54	0.006	-0.007, 0.019	0.35	0.1
Right ant. hip.	56	0.001	-0.007, 0.01	0.77	0.15, 0.23	54	0.002	-0.012, 0.016	0.75	0.08
Model 2										
Left ant. hip.	56	0.003	-0.006, 0.012	0.55	0.18, 0.31	54	0.006	-0.007, 0.02	0.34	0.2
Right ant. hip.	56	0	-0.009, 0.01	0.90	0.18, 0.30	54	0.004	-0.011, 0.018	0.61	0.19

Abbreviations: CI = confidence interval, ant. hip. = anterior hippocampus
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Back to main text

E.2.1.2 Single imputed analysis

Table E.8 and Table E.9 present the, respectively, standardized and unstandardized, results of the models assessing associations between daily average MVPA or CRF and left or right anterior hippocampal volumes.

Table E.8: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and left or right anterior hippocampal volume (standardized, N = 61).

	Left anterior hippocampus				Right anterior hippocampus			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
$\dot{V}O_2$peak	0.07	-0.24, 0.37	0.70	0.23, 0.26	0.05	-0.25, 0.35	0.75	0.25
MVPA[a]	-0.08	-0.33, 0.17	0.55	0.23, 0.26	-0.12	-0.38, 0.15	0.38	0.26
Model 2								
$\dot{V}O_2$peak	0.15	-0.26, 0.56	0.53	0.25, 0.3	0.13	-0.27, 0.52	0.52	0.3
MVPA[a]	−0.07	-0.33, 0.19	0.61	0.24, 0.3	-0.1	-0.38, 0.17	0.45	0.3

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting HR and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.9: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and left or right anterior hippocampal volume (unstandardized, N = 61).

	Left anterior hippocampus				Right anterior hippocampus			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
$\dot{V}O_2$peak	2.28	-8.04, 12.59	0.70	0.23, 0.26	1.53	-8.13, 11.19	0.75	0.25
MVPA[a]	-1.07	-4.53, 2.4	0.55	0.23, 0.26	-1.49	-4.87, 1.88	0.38	0.26
Model 2								
$\dot{V}O_2$peak	5	-8.98, 18.98	0.53	0.24, 0.3	4.09	-8.7, 16.89	0.52	0.3
MVPA[a]	-0.93	-4.51, 2.64	0.61	0.24, 0.3	-1.33	-4.87, 2.2	0.45	0.3

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting HR and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

We secondarily assessed whether daily average minutes of MVPA and CRF were associated with performance on tasks of relational memory. Table E.10 and Table E.11

present the, respectively, standardized and unstandardized, estimates of those models. None of the results were statistically significant.

Table E.10: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and memory outcomes (standardized, N = 61).

	Object-location task				Relational memory paradigm			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
$\dot{V}O_2$peak	-0.24	-0.5, 0.02	0.10	0.19, 0.22	0.01	-0.28, 0.3	0.95	0.07
MVPA[a]	0.04	-0.21, 0.3	0.75	0.15, 0.22	-0.18	-0.46, 0.1	0.20	0.1
Model 2								
$\dot{V}O_2$peak	-0.16	-0.57, 0.25	0.49	0.2, 0.26	0.02	-0.4, 0.44	0.93	0.17
MVPA[a]	0.09	-0.18, 0.35	0.53	0.2, 0.28	-0.16	-0.45, 0.13	0.28	0.19

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.11: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and the left or right anterior hippocampal volume (unstandardized, N = 61).

	Object-location task (swap errors, %)				Relational memory paradigm (accuracy, %)			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
$\dot{V}O_2$peak	-0.24	-0.5, 0.02	0.10	0.19, 0.22	0.01	-0.39, 0.42	0.95	0.07
MVPA[a]	0.02	-0.09, 0.12	0.75	0.15, 0.22	-0.1	-0.26, 0.06	0.20	0.1
Model 2								
$\dot{V}O_2$peak	-0.16	-0.56, 0.25	0.49	0.2, 0.26	0.03	-0.55, 0.61	0.93	0.17
MVPA[a]	0.03	-0.07, 0.14	0.53	0.20, 0.28	-0.09	-0.25, 0.08	0.28	0.19

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

We also assessed whether left or right anterior hippocampal volumes were associated with performance on the object-location and relational memory tasks. Table E.12 and Table E.13, present, respectively, the standardized and unstandardized, estimates of those models. Neither left or right anterior hippocampal volume was significantly associated with memory outcomes across models.

Table E.12: Univariate analyses assessing relationships between left and right anterior hippocampal volumes and memory outcomes (standardized, N = 61).

	Object-location task (swap errors)				Relational memory paradigm (accuracy)			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
Left ant. hip.	0.09	-0.19, 0.36	0.55	0.20, 0.25	0.20	-0.1, 0.50	0.19	0.11
Right ant. hip.	0.02	-0.25, 0.29	0.88	0.19, 0.24	0.10	-0.20, 0.41	0.51	0.09
Model 2								
Left ant. hip.	0.05	-0.24, 0.33	0.76	0.22, 0.29	0.22	-0.09, 0.52	0.15	0.21
Right ant. hip.	-0.02	-0.30, 0.26	0.88	0.22, 0.29	0.15	-0.16, 0.46	0.33	0.19

Abbreviations: CI = confidence interval, ant. hip. = anterior hippocampus
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Table E.13: Univariate analyses assessing relationships between left and right anterior hippocampal volumes and memory outcomes (unstandardized, N = 61).

	Object-location task (swap errors, %)				Relational memory paradigm (accuracy, %)			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
Left ant. hip.	0.003	-0.005, 0.01	0.55	0.19, 0.25	0.008	-0.004, 0.02	0.19	0.11
Right ant. hip.	0.001	-0.008, 0.009	0.88	0.19, 0.24	0.004	-0.009, 0.018	0.51	0.09
Model 2								
Left ant. hip.	0.001	-0.007, 0.01	0.76	0.21, 0.29	0.009	-0.003, 0.021	0.15	0.21
Right ant. hip.	-0.001	-0.009, 0.008	0.88	0.21, 0.28	0.007	-0.007, 0.02	0.33	0.19

Abbreviations: CI = confidence interval, ant. hip. = anterior hippocampus
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Back to main text

E.2.2 Entire hippocampus

E.2.2.1 Available-case analysis

We repeated the primary univariate analyses for the entire hippocampus. Table E.14 and Table E.15 display the results of the multilevel models assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and volume of the entire hippocampus (left and right), showing respectively, standardardized and unstandardized

regression coefficients. In model 1, analyses are controlled for age, sex, socioeconomic status, pubertal status and head size; in model 2, analyses are additionally controlled for BMI, blood pressure and resting HR.

Table E.14: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and volume of the entire hippocampus (standardized).

	Left hippocampus					Right hippocampus				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
$\dot{V}O_2$peak	56	0.12	-0.17, 0.41	0.48	0.33, 0.34	56	0.11	-0.18, 0.39	0.45	0.39
MVPA[a]	44	-0.08	-0.38, 0.22	0.61	0.28, 0.32	44	-0.15	-0.46, 0.16	0.33	0.34
Model 2										
$\dot{V}O_2$peak	56	0.12	-0.22, 0.47	0.55	0.36, 0.37	56	0.09	-0.26, 0.43	0.61	0.42
MVPA[a]	44	-0.11	-0.42, 0.19	0.46	0.35, 0.41	44	-0.18	-0.51, 0.15	0.27	0.39

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.15: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and volume of the entire hippocampus (unstandardized).

	Left hippocampus					Right hippocampus				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
$\dot{V}O_2$peak	56	5.39	-7.70, 18.49	0.48	0.33, 0.34	56	4.99	-8.16, 18.13	0.45	0.39
MVPA[a]	44	-1.1	-4.92, 2.83	0.61	0.28, 0.32	44	-1.92	-5.85, 2.01	0.33	0.34
Model 2										
$\dot{V}O_2$peak	56	5.56	-10.2, 21.32	0.55	0.36, 0.37	56	4.12	-11.84, 20.07	0.61	0.42
MVPA[a]	44	-1.42	-5.35, 2.51	0.49	0.35, 0.41	44	-2.3	-6.52, 1.91	0.27	0.39

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.16 and Table E.17 display the results of the multilevel models assessing relationships between the left and right hippocampus and relational memory performance, showing respectively, standardardized and unstandardized regression coeffi-

cients. In model 1, analyses are controlled for age, sex, socioeconomic status, pubertal status and head size; in model 2, analyses are additionally controlled for BMI, blood pressure and resting HR.

Table E.16: Univariate analyses assessing relationships between left or right hippocampal volume and memory performance (standardized).

	Object-location task					Relational memory paradigm				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
Left hip.	56	0.02	-0.29, 0.34	0.88	0.14, 0.23	54	0.24	-0.10, 0.59	0.16	0.12
Right hip.	56	-0.04	-0.36, 0.28	0.82	0.15, 0.22	54	0.17	-0.19, 0.53	0.35	0.1
Model 2										
Left hip.	56	-0.02	-0.35, 0.31	0.92	0.18, 0.30	54	0.28	-0.08, 0.63	0.12	0.23
Right hip.	56	-0.06	-0.39, 0.27	0.72	0.18, 0.30	54	0.9	-0.18, 0.55	0.30	0.21

Abbreviations: CI = confidence interval, hip. = hippocampus
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Table E.17: Univariate analyses assessing relationships between left or right hippocampal volume and memory performance (unstandardized).

	Object-location task (swap errors, %)					Relational memory paradigm (accuracy, %)				
	N	β	95% CI	p	$R^2_{m,c}$	N	β	95% CI	p	$R^2_{m,c}$
Model 1										
Left hip.	56	0.001	-0.007, 0.008	0.88	0.14, 0.23	54	0.008	-0.003, 0.019	0.16	0.12
Right hip.	56	-0.001	-0.008, 0.007	0.82	0.15, 0.22	54	0.005	-0.006, 0.017	0.35	0.1
Model 2										
Left hip.	56	0	-0.008, 0.007	0.92	0.18, 0.30	54	0.009	-0.002, 0.02	0.12	0.23
Right hip.	56	-0.001	-0.009, 0.006	0.72	0.18, 0.30	54	0.006	-0.006, 0.018	0.30	0.21

Abbreviations: CI = confidence interval, hip. = hippocampus
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Back to main text

E.2.2.2 Imputed analysis

We repeated the primary univariate analyses for the entire hippocampus (single imputed data). Table E.18 and Table E.19 display the results of the multilevel models

assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and volume of the entire hippocampus (left and right), showing respectively, standardardized and unstandardized regression coefficients. In model 1, analyses are controlled for age, sex, socioeconomic status, pubertal status and head size; in model 2, analyses are additionally controlled for BMI, blood pressure and resting HR.

Table E.18: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and volume of the entire hippocampus (standardized, N = 61).

	Left hippocampus				Right hippocampus			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
$\dot{V}O_2$peak	0.03	-0.24, 0.3	0.86	0.36, 0.38	0.09	-0.17, 0.36	0.48	0.4
MVPA[a]	-0.16	-0.39, 0.06	0.17	0.38, 0.4	-0.18	-0.41, 0.05	0.12	0.42
Model 2								
$\dot{V}O_2$peak	0.13	-0.22, 0.48	0.53	0.4, 0.4	0.08	-0.28, 0.44	0.65	0.43
MVPA[a]	-0.16	-0.39, 0.07	0.19	0.41, 0.43	-0.2	-0.44, 0.04	0.09	0.46

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.19: Univariate analyses assessing relationships between daily average minutes of MVPA or $\dot{V}O_2$peak and volume of the entire hippocampus (unstandardized, N = 61).

	Left hippocampus				Right hippocampus			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
$\dot{V}O_2$peak	1.15	-10.43, 12.73	0.86	0.36, 0.38	3.97	-7.34, 15.28	0.48	0.4
MVPA[a]	-2.77	-6.64, 1.1	0.17	0.38, 0.4	-3.05	-6.96, 0.86	0.12	0.42
Model 2								
$\dot{V}O_2$peak	5.46	-9.42, 20.35	0.53	0.4, 0.4	3.41	-11.65, 18.48	0.65	0.43
MVPA[a]	-2.73	-6.72, 1.25	0.19	0.41, 0.43	-3.45	-7.51, 0.6	0.09	0.46

Abbreviations: CI = confidence interval, MVPA = moderate-to-vigorous physical activity
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.
[a] Corresponds to daily average minutes of MVPA

Table E.20 and Table E.21 display the results of the multilevel models assessing relationships between the left and right hippocampal volume and relational memory

performance, showing respectively, standardardized and unstandardized regression coefficients. In model 1, analyses are controlled for age, sex, socioeconomic status, pubertal status and head size; in model 2, analyses are additionally controlled for BMI, blood pressure and resting HR.

Table E.20: Univariate analyses assessing relationships between left or right hippocampal volume and memory performance (standardized, N = 61).

	Object-location task				Relational memory paradigm			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
Left hippocampus	0.005	-0.3, 0.31	0.97	0.19, 0.24	0.27	-0.06, 0.6	0.10	0.12
Right hippocampus	-0.05	-0.35, 0.25	0.74	0.20, 0.23	0.16	-0.18, 0.49	0.35	0.09
Model 2								
Left hippocampus	-0.05	-0.37, 0.27	0.76	0.22, 0.29	0.31	-0.03, 0.65	0.07	0.23
Right hippocampus	-0.08	-0.39, 0.23	0.60	0.22, 0.28	0.19	-0.15, 0.53	0.26	0.19

Abbreviations: CI = confidence interval
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Table E.21: Univariate analyses assessing relationships between left or right hippocampal volume and memory performance(unstandardized, N = 61).

	Object-location task (swap errors, %)				Relational memory paradigm (accuracy, %)			
	β	95% CI	p	$R^2_{m,c}$	β	95% CI	p	$R^2_{m,c}$
Model 1								
Left hippocampus	0	-0.007, 0.007	0.97	0.19, 0.24	0.01	0, 0.02	0.10	0.12
Right hippocampus	-0.001	-0.008, 0.006	0.74	0.20, 0.23	0.01	-0.01, 0.02	0.35	0.09
Model 2								
Left hippocampus	-0.001	-0.009, 0.006	0.76	0.22, 0.29	0.01	0, 0.02	0.07	0.23
Right hippocampus	-0.002	-0.009, 0.005	0.60	0.22, 0.28	0.01	0, 0.02	0.26	0.19

Abbreviations: CI = confidence interval
R^2 statistics are provided for each model. If a hierarchical model was fit, the marginal R^2 (R^2_m) is presented followed by the conditional R^2 (R^2_c).
Model 1 includes as covariates: age, sex, socioeconomic status, pubertal status and head size. *Model 2* additionally includes BMI, resting heart rate and systolic- and diastolic blood pressure.

Back to main text

E.2.3 Multivariate relationship between fitness, physical activity and brain MR markers

E.2.3.1 CCA on imputed data

Physical activity, fitness and brain MR markers

Table E.22 displays the cross-loadings for the first mode of covariation, involving physical activity and fitness markers.

Table E.22: Cross-loadings mode 1

PA lifestyle	r	Brain	r
Self-reported PA	0.20	R_2^*_right	0.32
Max workload	0.08	R_2^*_left	0.21
Total weekly VPA	-0.07	fintra_right	0.06
Total weekly MPA	-0.10	fintra_left	0.05
Total weekly sedentary	-0.13	perfusion_right	-0.02
$\dot{\mathbf{V}}O_2$peak	-0.44	MD_left	-0.06
		VBM_left	-0.08
		R_1_left	-0.08
		R_1_right	-0.09
		MD_right	-0.14
		perfusion_left	-0.21
		VBM_right	-0.36

Abbreviations: fintra = intra-axonal compartment, MD = mean diffusivity, MPA = moderate physical activity, PA = physical activity, r = Pearson correlation, VBM = voxel-based morphometry, VPA = vigorous physical activity

In Figure E.7 the leave-one-school-out cross-validated canonical correlation, between a physical variate and brain variate, is displayed (ρ_{CV} = 0.11).

Back to main text

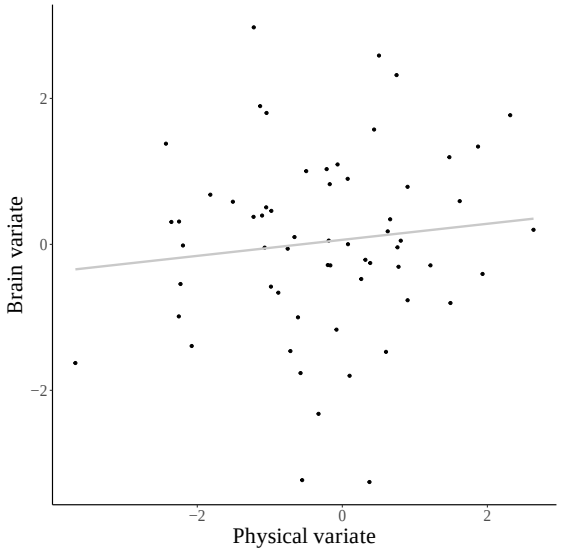


Figure E.7: Cross-validated canonical correlation between the first physical variate and brain variate.

Physical health and brain MR markers

Table E.23 displays the cross-loadings for the first mode of covariation for the CCA involving all physical health markers (i.e. BMI, blood pressure and resting heart rate).

Table E.23: Cross-loadings mode 1

PA health	r	Brain	r
BMI	0.45	fintra_right	0.23
Diastolic BP	0.14	R_2^*_left	0.2
Self-reported PA	0.14	R_2^*_right	0.19
Total weekly VPA	0.02	fintra_left	0.14
Max workload	0.01	R_1_right	-0.02
Systolic BP	-0.03	R_1_left	-0.09
Total weekly MPA	-0.06	perfusion_right	-0.11
Total weekly sedentary	-0.17	MD_left	-0.12
Resting HR	-0.29	perfusion_left	-0.21
$\dot{\mathbf{V}}O_2$peak	-0.36	VBM_left	-0.3
		MD_right	-0.33
		VBM_right	-0.48

Abbreviations: bmi = body mass index, BP = blood pressure, fintra = intra-axonal compartment, HR = heart rate, MD = mean diffusivity, MPA = moderate physical activity, PA = physical activity, r = Pearson correlation, VBM = voxel-based morphometry, VPA = vigorous physical activity

In Figure E.8 the leave-one-school-out cross-validated canonical correlation, between a physical health variate and brain variate, is displayed (ρ_{CV} = 0.21).

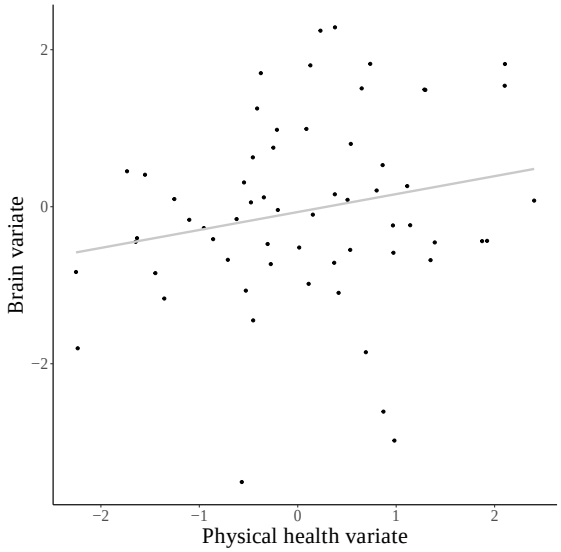


Figure E.8: Cross-validated canonical correlation between the first physical health variate and brain variate.

Back to main text

E.2.3.2 CCA on complete-case data

We additionally examined the multivariate relationship between physical activity behaviour and hippocampal MRI markers of tissue microstructure in the complete-case dataset (n = 36). We ran the CCA twice, first with the original physical phenotype and then with the phenotype that also included markers of physical health (i.e. BMI, blood pressure and resting heart rate), hereafter named the "physical health" phenotype. In line with findings from the imputed dataset, no significant mode of covariation was detected for the physical phenotype (ρ = 0.71, p = 0.98, ρ_{CV} = -0.12). Figure E.9 shows the canonical correlation and cross-loadings of the first mode of covariation.

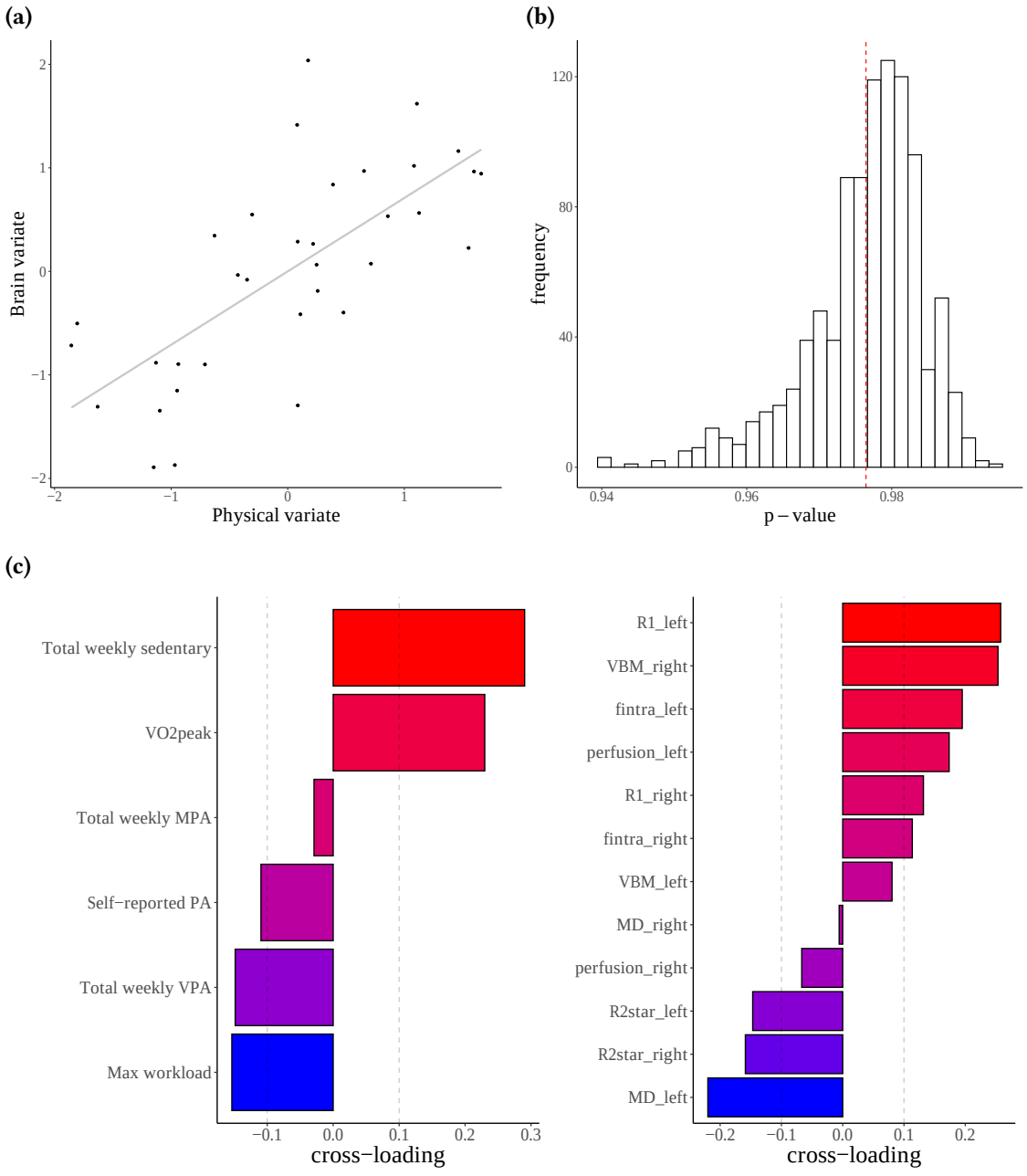


Figure E.9: Canonical correlation and cross-loadings of the first mode of covariation (complete-case). (*a*) Subject scores on the first physical and brain variate. (*b*) Distribution of p-values. Red dashed line corresponds to median p-value. (*c*) Cross-loadings between the physically active lifestyle variables and the brain variate (left) and between the brain variables and the PA lifestyle variate (right). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest.

Table E.24 displays the cross-loadings for the first mode of covariation for the CCA involving PA and fitness markers only (adjusted for age, sex, socioeconomic status, pubertal status and head size).

Table E.24: Cross-loadings mode 1 (complete-case)

PA lifestyle	***r***	**Brain**	***r***
Total weekly sedentary	0.29	R_1_left	0.26
$\dot{\mathbf{V}}O_2$peak	0.23	VBM_right	0.25
Total weekly MPA	-0.03	fintra_left	0.2
Self-reported PA	-0.11	perfusion_left	0.17
Total weekly VPA	-0.15	R_1_right	0.13
Max workload	-0.15	fintra_right	0.11
		VBM_left	0.08
		MD_right	-0.01
		perfusion_right	-0.07
		R_2^*_left	-0.15
		R_2^*_right	-0.16
		MD_left	-0.22

Abbreviations: fintra = intra-axonal compartment, MD = mean diffusivity, MPA = moderate physical activity, PA = physical activity, r = Pearson correlation, VBM = voxel-based morphometry, VPA = vigorous physical activity

The leave-one-school-out cross-validated correlation (ρ_{CV} = -0.12) is presented in Figure E.10.

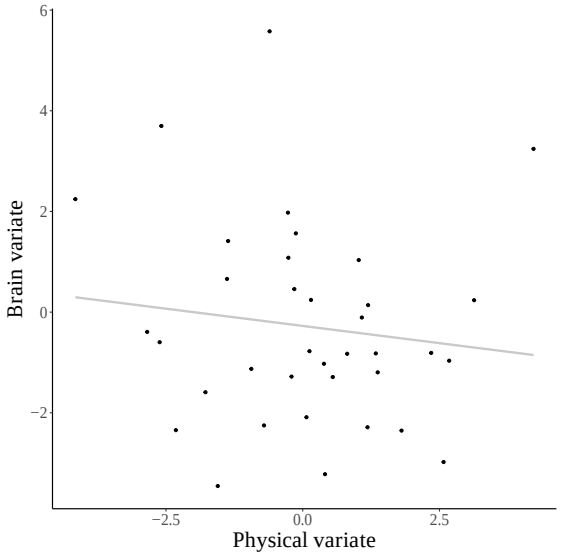


Figure E.10: Cross-validated canonical correlation between the first physical variate and brain variate (complete-case).

Back to main text

No significant mode of covariation was detected for the physical health phenotype (ρ = 0.88, p = 0.77, ρ_{CV} = -0.21). Figure E.11 shows the canonical correlation and cross-loadings of the first mode of covariation.

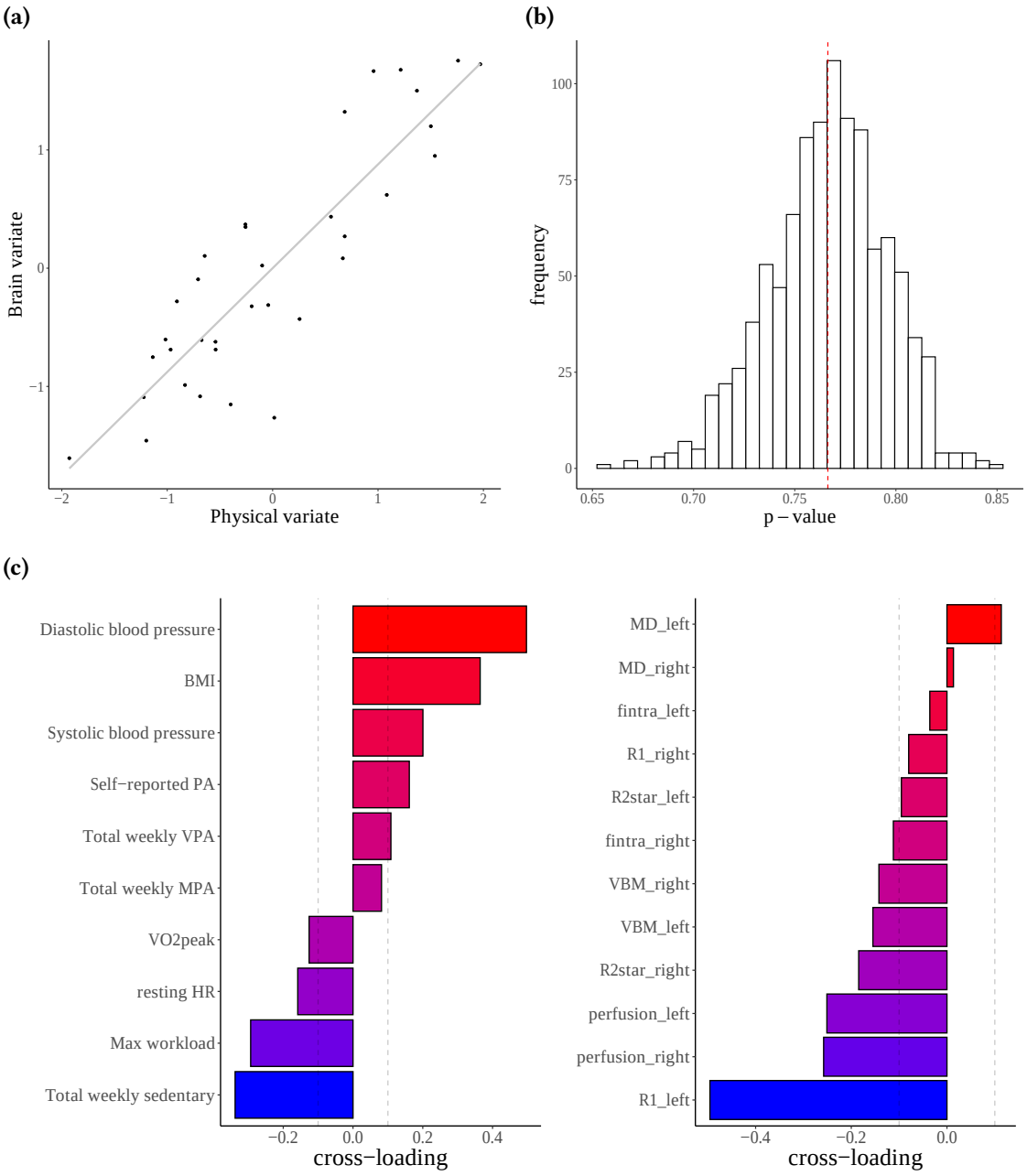


Figure E.11: Canonical correlation and cross-loadings of the first mode of covariation (complete-case). (*a*) Subject scores on the first physical health and brain variate. (*b*) Distribution of p-values. Red dashed line corresponds to median p-value. (*c*) Cross-loadings between the physical health variables and the brain variate (left) and between the brain variables and the physical health variate (right). Cross-loadings greater or equal than $|0.1|$ (dashed line) were considered of interest.

Table E.25 displays the cross-loadings for the first mode of covariation for the CCA involving all physical health markers (i.e. including BMI, blood pressure and resting heart rate).

Table E.25: Cross-loadings mode 1 (complete-case)

Physical health	r	**Brain**	r
Diastolic BP	0.49	MD_left	0.12
BMI	0.36	MD_right	0.02
Systolic BP	0.2	fintra_left	-0.04
Self-reported PA	0.15	R_1_right	-0.08
Total weekly VPA	0.11	R_2^*_left	-0.09
Total weekly MPA	0.08	fintra_right	-0.12
$\dot{\mathbf{V}}O_2$peak	-0.13	VBM_right	-0.15
resting HR	-0.16	VBM_left	-0.16
Max workload	-0.29	R_2^*_right	-0.18
Total weekly sedentary	-0.34	perfusion_left	-0.25
		perfusion_right	-0.26
		R_1_left	-0.49

Abbreviations: bmi = body mass index, BP = blood pressure, fintra = intra-axonal compartment, HR = heart rate, MD = mean diffusivity, MPA = moderate physical activity, PA = physical activity, r = Pearson correlation, VBM = voxel-based morphometry, VPA = vigorous physical activity

The leave-one-school-out cross-validated correlation (ρ_{CV} = -0.21) is presented in Figure E.12.

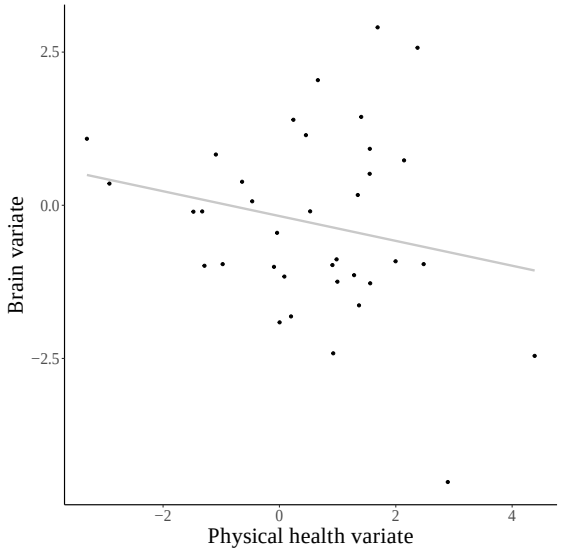


Figure E.12: Cross-validated canonical correlation between the first physical health variate and brain variate (complete-case).

While all results are non-significant and should not readily be interpreted, it is surprising that the cross-validated canonical correlations (ρ_{CV}) are negative compared to small positive values for the imputed dataset presented in chapter 5. This does not seem to be a result of the leave-one-school-out cross-validation procedure, given that a sensitivity analysis with ten-fold cross-validation (ignoring the school structure) resulted in correlation values with a similar sign (i.e. positive or negative). It is more likely, however, that the difference between the imputed and complete-case CCA findings is a result of a difference in sample (size and characteristics) considered in each dataset ($n_{imputed}$ = 61 versus $n_{complete\text{-}case}$ = 36).

Back to main text

This page is intentionally left blank.

F

Supplementary materials to chapter 6: *The effect of the Fit to Study intervention on cardiorespiratory fitness, cognitive performance and mental health*

Contents

F.1 Supplementary methods . . . 403
F.1.1 Actigraphy processing . . . 403
F.1.2 Minimum detectable effect size . . . 403
F.1.3 Data cleaning . . . 404
F.1.4 Complier-average causal effects analysis . . . 407
F.1.5 Software . . . 407
F.2 Missing data . . . 408
F.2.1 Missing data patterns . . . 410
F.2.2 Missing data approach . . . 410
F.3 Drop-out of schools at baseline assessments . . . 414
F.3.1 School drop-out pre-baseline . . . 414
F.4 Data cleaning . . . 417
F.5 Baseline PA and PE characteristics . . . 420
F.5.1 Baseline PA and PE characteristics . . . 420
F.5.2 PE lessons across schools . . . 421
F.6 Primary outcomes . . . 422
F.6.1 Fully adjusted models: multiply imputed data . . . 422
F.6.2 Sensitivity analysis: complete-cases . . . 423
F.7 Subgroup analysis . . . 431
F.8 Change - change analysis . . . 437
F.8.1 Primary analysis: multiply imputed data . . . 437

F.8.2 Sensitivity analysis: complete-cases 437
F.9 Intervention fidelity and CACE analysis 442
F.9.1 Intervention fidelity 442
F.9.2 CACE analysis 444
F.10 Additional analysis 445
F.10.1 Multiply imputed data 445
F.10.2 Sensitivity analysis 446

F.1 Supplementary methods

F.1.1 Actigraphy processing

Objective measures of PA during PE were collected using wrist-worn AX3 tri-axial accelerometers (Open Lab, University of Newcastle, UK). Participants wore the accelerometer on their non-dominant wrist. Accelerometers were set to record at 100 Hz and were attached after participants changed for PE and collected before they changed back into school clothes. Following conversion of raw accelerometer data into 1 s epochs, class average minutes of sedentary PA (SPA), moderate PA (MPA) and vigorous PA (VPA) during lesson time were calculated using established cut-points (Phillips *et al.*, 2013) and standardised to minutes per hour. We computed the full lesson PA estimates as well as the active, or "effective", PA estimates. The former represents PA estimates standardized to total time allocated for PE and the latter represent the amount of PA adjusted for change time.

Back to main text

F.1.2 Minimum detectable effect size

Sample size calculations were performed by NatCen for the primary intention-to-treat analysis, which compared maths test performance between intervention and control schools (Husain, 2017; Husain *et al.*, 2019). The initial sample size calculation demonstrated a range of minimum detectable effect sizes (MDES) varying between 0.18–0.26 standard deviations (SD) for various achieved sample sizes (80–100 schools), assuming 80% power, α = 0.05 (two-tailed), 90 pupils per school and baseline covariates explaining 50% of variance at pupil and school level. These estimates were updated (in February–March 2018) to reflect school drop-out (n = 18 schools) and a revised set of assumptions (Bartasevicius & Jabin, 2018). In particular, the new MDES were computed based on: 88 participating schools (47% intervention), on average 78 pupils per school, baseline covariates explaining 40% of pupil level variance and 10% of school level variance and an ICC of 0.16. The newly computed MDES was equal to 0.24 SD.

The MDES would have been 0.21 SD if all 106 recruited schools had participated in the trial. At the final analysis stage, the MDES was calculated to be 0.22 SD (Husain *et al.,* 2019). Power calculations were performed in PowerUp! (Dong & Peng, 2013). No a priori sample size calculations were performed for the secondary outcomes, but an indication of the MDES is provided below.

The MDES for the cognitive tests was calculated for the randomisation sample using the following set of assumptions: (1) 104 schools, 50% of which in the intervention group, (2) an average of 176 pupils per school (N = 18,261), (3) baseline covariates explaining 40% of pupil level variance and 10% of school level variance (based on NatCen's adjusted estimates), (4) an ICC of 0.02, following a previous study measuring EFs in a comparable sample (Aadland *et al.,* 2019) and (5) 80% power with α = 0.05 (two-tailed). Under these assumptions MDES was 0.07 SD (95% CI: 0.03–0.16; it would be 0.15 SD, 95% CI: 0.05–0.12, if the ICC was 0.06, Aadland *et al.,* 2017b). With a similar set of assumptions and an ICC = 0.25 (NB. teacher-rated; Berry *et al.,* 2016), the MDES for the SDQ was 0.23 SD. For CRF, the MDES was 0.08 SD with a similar set of assumptions and an ICC = 0.03 (Kriemler *et al.,* 2010). Notably, the trial was not designed to measure or affect cognitive, mental health, or CRF scores, hence this MDES solely serves as an *a-posteriori* indication of a possible detectable effect size under this particular set of assumptions. The MDES for the secondary outcomes was not updated *a-posteriori*, given the differential amounts of missingness across outcome measures and the lack of intervention fidelity of the trial.

Back to main text

F.1.3 Data cleaning

F.1.3.1 Cognitive assessments

Cognitive data were cleaned per assessment. For each assessment, we identified participants with fewer than 30% valid trials (i.e. out of the total number of trials), under the assumption that a proportion (>30%) of trials is required to achieve a reliable

estimate of each cognitive construct. Furthermore, for the relational memory and visual two-back task, we identified participants who consistently pressed a single response button throughout the task, as indicated by zero counts for false alarms and hits, or by zero counts for misses and correct rejections. We additionally inspected the primary outcome measures of each task and removed participants with values that were outside the expected range (i.e. 0–100 for accuracy outcomes, and mean RT outside the pre-defined RT bounds, for details see chapter 2) or were missing due to data collection errors.

Back to main text

F.1.3.2 Questionnaire data

Questionnaire measures included in this study were part of a larger battery that consisted of three blocks Likert-scale based items. Long strings of identical answers (i.e. same position on the scale) within a block, accounting for reversed items, were identified. The mean string length (MSL) per block was computed, and data were removed if participants gave the same answer throughout a block and recorded a string of MSL+/- 3 SD in at least one more block. In addition, we removed participant scores that were outside the range of individual items (e.g. score >5 on 5-point Likert scale) due to data collection errors.

Back to main text

F.1.3.3 Cardiorespiratory fitness assessments

Schools that completed the 12 min Cooper Run test (n = 4, of which 1 intervention school; 561 pupils) were excluded from the analysis. Due to the difference in scale and units of the outcome measures (20MSR: laps versus Cooper: total distance), the scores of these tests cannot easily be combined. Classically, researchers have used estimating equations to convert 20MSR (Batista *et al.,* 2013; Léger *et al.,* 1988; Mahar *et al.,* 2018; Matsuzaka *et al.,* 2004; Melo *et al.,* 2011; Tomkinson *et al.,* 2017) and Cooper Test performance (Cooper, 1968; Grant *et al.,* 1995) to estimated $\dot{V}O_2$max values. To this

end, they related field test performance to the gold-standard $\dot{V}O_2$max obtained with a maximal graded exercise test. While the conversion to a common metric ($\dot{V}O_2$max) is attractive, we decided not to pursue this for several reasons:

First, a large number of estimating equations exist, particularly for 20MSR performance (>17 equations, see Tomkinson *et al.,* 2017), but selecting one that is appropriate for the Fit to Study sample is not straightforward. Depending on the fitness protocol used, these equations vary in their predictors, most commonly including one or more of sex, age, speed or body weight, and more generally in their accuracy of predicting $\dot{V}O_2$max (Batista *et al.,* 2013; Melo *et al.,* 2011). Moreover, these equations have been established in samples that may not generalise to the population.

Second, estimating equations may underestimate (Grant *et al.,* 1995; Penry *et al.,* 2011) or overestimate (Scott *et al.,* 2019) an individual's real $\dot{V}O_2$max. This is particularly problematic if the degree of under / over estimation differs between the Cooper test and 20MSR. A comparison of estimated $\dot{V}O_2$max values from 20MSR and the Cooper test in healthy adults demonstrated that, although 20MSR underestimated an individual's level of CRF, the converted Cooper test results were biased: underestimation of $\dot{V}O_2$max values in low fit individuals (i.e. low real $\dot{V}O_2$max) and overestimation of $\dot{V}O_2$max in high fit individuals (Penry *et al.,* 2011).

Third, the $\dot{V}O_2$max data that has been collected as part of the brain-imaging sub-study (details in chapter 2) in a sub-sample of 61 participants is not appropriate for the creation of conversion equations. Despite high correlations between the 20MSR and $\dot{V}O_2$peak (Spearman's ρ = 0.77, $p \leq 0.01$), no such relationships were found in the small sample (n = 9) of participants who completed the Cooper test (Spearman's ρ = 0.23, p = 0.56). Moreover, participants of the sub-study are generally more fit compared to the main trial sample. Consequently, an equation that is based on the sub-sample may not be valid in the main trial sample. Nonetheless, the strong relationship between 20MSR and $\dot{V}O_2$peak suggests that 20MSR is a valid measure of

CRF.

Back to main text

F.1.4 Complier-average causal effects analysis

The ITT analysis will provide a conservative (or underestimated) treatment effect when schools do not adhere to the intervention (Hernán & Hernández-Díaz, 2012). We therefore estimated a complier-average causal effect (CACE, or local average causal effect) to estimate the average treatment effect in the population of compliers. The CACE was estimated under one-sided non-compliance, by fitting an instrumental variable (IV) model using the two-stage least squares (TSLS) method. The IV model included the assigned treatment as the instrumental variable and adjusted for the stratification variable (school gender type) and baseline values of the respective outcome measure. In line with the analysis of the primary outcome (Husain *et al.*, 2019), we used the compliance estimates of the post-intervention teacher-reported school survey as primary compliance metric. A total of 22 intervention schools (of 29 at posttest) provided compliance information. Intervention schools in which the intervention was delivered in >50% of PE lessons were classified as compliant. Schools that did not provide compliance data were assumed to be non-compliant. The control group was included in the IV model with their treatment received set to 0.

Back to main text

F.1.5 Software

The majority of the analyses were conducted in Rstudio (RStudio Team, 2016) with R (R Core Team, 2020, version 4.0) using the *estimatr* package (Blair *et al.*, 2020). Plots were created using *ggplot2* (Wickham, 2016). Missing data was imputed using *MICE* (Van Buuren & Groothuis-oudshoorn, 2011) on the high-performance cluster of the Wellcome Centre for Integrative Neuroimaging, Oxford UK (R, version 3.6). The

structural equation model was constructed using Mplus 8.4 (Muthén & Muthén, n.d.).

Back to main text

F.2 Missing data

Table F.1 and Table F.2 contain the number (% of total) of missing cases of, respectively, school- and pupil-level variables that were used in the analyses (arranged by increasing proportion of missingness). There was no missing data within each cognitive assessment, hence a single variable is used to represent each cognitive assessment. The proportion of missing values is > 70% for all cognitive assessments, but the reaction time task, at posttest.

Table F.1: Missing data (proportion) of school-level variables

	No. missing	% missing[a]
Gender status	0	0
School type	0	0
Establishment type	0	0
School size	0	0
IMD decile	0	0
Proportion of eFSM pupils	0	0
Teacher training received	0	0
Ofsted rating	1	1.08
Number of Year 7 form groups	2	2.15
School level averaged actigraphy pretest	3	3.23
School level averaged actigraphy posttest	22	23.66
Teacher-reported fidelity (NatCen recorded)	71	76.34

Abbreviations: IMD = index of multiple deprivation, eFSM = eligible for free school meals
[a] Of 93 schools that were part of the trial at baseline

Table F.2: Missing data (proportion) of pupil-level variables

	No. missing	% missing[a]
age	0	0
sex	0	0
eFSM	0	0
20MSR term posttest	0	0
20MSR term pretest	0	0
20MSR pretest	5,320	33.21
PE enjoyment pretest	6,243	38.98
Internalising score pretest	6,245	38.99
Externalising score pretest	6,246	39
Questionnaire date pretest	6,272	39.16
Questionnaire location pretest	6,272	39.16
Physical self-esteem pretest	6,296	39.31
Global self-esteem pretest	6,297	39.31
Self-reported PA pretest	6,318	39.45
20MSR posttest	7,903	49.34
Reaction time task pretest	8,077	50.43
Pupil-reported fidelity: warm-up	9,603	59.96
Pupil-reported fidelity: infusions	9,603	59.96
Pupil-reported fidelity: participation	9,603	59.96
Internalising score posttest	9,605	59.97
Externalising score posttest	9,609	59.99
Self-reported PA posttest	9,619	60.05
Questionnaire date posttest	9,623	60.08
Questionnaire location posttest	9,623	60.08
Global self-esteem posttest	9,868	61.61
Physical self-esteem posttest	9,868	61.61
Reaction time task posttest	9,878	61.67
Flanker task pretest	10,569	65.99
Relational memory task pretest	10,699	66.8
Two-back pretest	11,005	68.71
Colour-shape task switch pretest	11,295	70.52
Flanker task posttest	11,500	71.8
Two-back posttest	11,516	71.9
Relational memory task posttest	11,525	71.95
Teacher-reported fidelity	11,963	74.69
Colour-shape task switch posttest	12,081	75.43

Abbreviations: IMD = index of multiple deprivation, eFSM = eligible for free school meals
[a] Of 16,017 pupils that were part of the trial at baseline

F.2.1 Missing data patterns

A total of 1,590 unique missing data patterns were present in the data. To provide an overview of some of the patterns, we created an intersection plot (using UpsetR in R Conway *et al.,* 2017) displaying the patterns of complete data for those subjects who completed the assessments at both timepoints (Figure F.1).

F.2.2 Missing data approach

F.2.2.1 Primary analysis

The dataset was multiply imputed (m = 180 imputations) using multilevel multiple imputation by chained equations (MICE) in R. The imputation model included all variables that were part of the analysis model, as well as any variables that were associated (r > 0.4) with variables in those models or related with missingness in variables in the analysis model (Van Buuren *et al.,* 1999; Van Buuren, 2018). Models that contained all assessment period variables (i.e. period when assessments were conducted) at post-test resulted in errors. Given that (1) only 37 participants (of 6,174) completed a cognitive assessment during summer or autumn at posttest, and (2) the reaction time task was always completed first, we therefore decided to only include the variable indicating when the reaction time task was completed. An overview of all variables included in the imputation model is provided in Table F.3. Moreover, due to computational issues, we removed some variables from the final set of imputation models, reported in Table F.4.

Continuous variables were imputed using a two-level normal model with homogeneous within group variances (2l.pan) and categorical variables using two-level predictive mean matching (2l.pmm). The data was imputed separately by treatment group to allow group-specific covariance structures, thereby preserving all two-way interactions between the treatment group and other variables in the imputation model (Enders & Gottschall, 2011). The imputations were visually checked for convergence and estimates were pooled using Rubin's rules (Rubin, 1987).

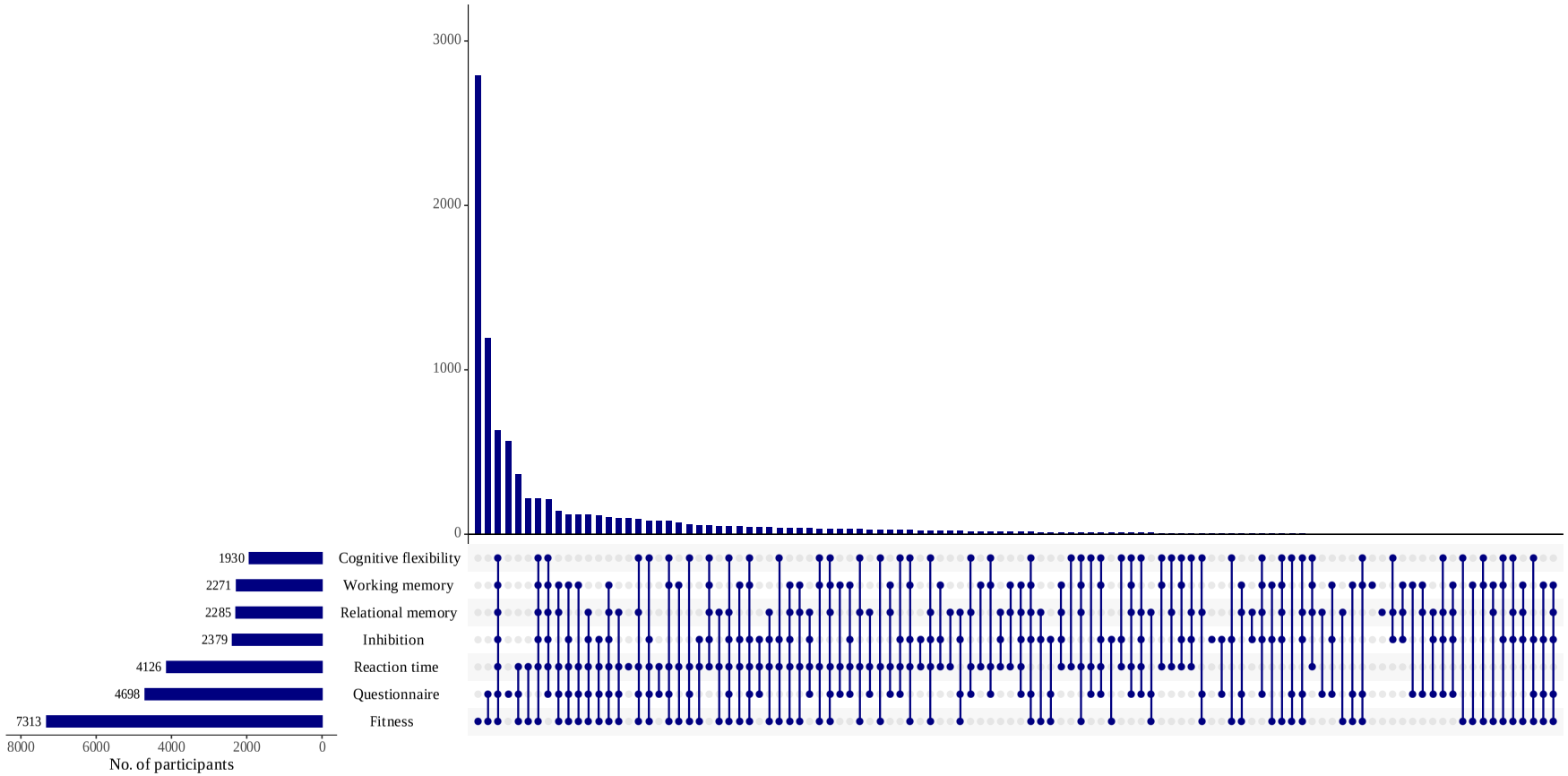


Figure F.1: Intersection plot demonstrating patterns of completion of assessments (pre and posttest). *Left bar plot*: This bar plot contains the total number of cases who completed each assessment, the name of which is displayed at the right side of the barplot. *Middle circle plot*: This plot shows the intersections, i.e. whether a participant completed one or more assessments. A closed dark blue circle denotes that a participant completed the task and vertical lines connecting circles refer to the set of assessments that were completed. *Top bar plot*: This bar plot shows the frequency of each intersection, i.e. the total number of participants who completed one or more assessments, as indicated by the circle plot below. For instance, the left most bar in the top barplot indicates that over 2,500 cases had data for the fitness assessment (pre and post) only.

411

F.2.2.2 Complier-average causal effect analysis

We separately imputed the data for the complier-average causal effect (CACE) analysis. The imputation models were identical to the ones included in the primary multiple imputation analysis, with the addition of a compliance indicator. Intervention schools in which the intervention was delivered in >50% of PE lessons were classified as compliant and all other schools were classified as non-compliant. We opted to multiply impute the data for the CACE analysis separately for several reasons: (1) the compliance measure that was used for the CACE analysis was missing for 24 out of 46 (i.e. 52%) intervention schools, that were assumed to be non-compliant with the intervention, and (2) the measure was deemed a proxy of compliance, given that it was reported by teachers following the intervention period, making it subject to various forms of bias (e.g. recall bias).

Table F.3: Variables included in multiple imputations

#	Variables
1	School
2	School gender type
3	School percentage eFSM upils
4	school size
5	age
6	sex
7	eFSM
8	Flanker RT congruent (tp1)
9	Flanker RT congruent (tp2)
10	Flanker RT incongruent (tp1)
11	Flanker RT incongruent (tp2)
12	Flanker date / term (tp1)
13	Flanker location (tp1)
14	Flanker location (tp2)
15	Flanker accuracy incongruent (tp1)
16	Flanker accuracy incongruent (tp2)
17	Flanker accuracy congruent (tp1)
18	Flanker accuracy congruent (tp2)
19	Two-back accuracy (tp1)
20	Two-back accuracy (tp2)
21	Two-back RT (tp1)
22	Two-back RT (tp2)
23	Two-back date / term (tp1)
24	Two-back location (tp1)
25	Two-back location (tp2)
26	Reaction time task RT (tp1)
27	Reaction time task date / term (tp1)
28	Reaction time task (tp1)

Continued on next page

Table F.3 – continued from previous page

#	Variables
29	Reaction time task RT (tp2)
30	Reaction time task date / term (tp2)
31	Reaction time task (tp2)
32	Task switch RT switch (tp1)
33	Task switch RT switch (tp2)
34	Task switch RT non-switch (tp1)
35	Task switch RT non-switch (tp2)
36	Task switch date / term (tp1)
37	Task switch location (tp1)
38	Task switch location (tp2)
39	Task switch accuracy switch (tp1)
40	Task switch accuracy non-switch (tp1)
41	Task switch accuracy switch (tp2)
42	Task switch accuracy non-switch (tp2)
43	Relational memory accuracy (tp1)
44	Relational memory accuracy (tp2)
45	Relational memory RT (tp1)
46	Relational memory RT (tp2)
47	Relational memory date / term (tp1)
48	Relational memory location (tp1)
49	Relational memory location (tp2)
50	Q internalising (tp1)
51	Q internalising (tp2)
52	Q externalising (tp1)
53	Q externalising (tp2)
54	Q global self esteem (tp1)
55	Q physical self esteem (tp1)
56	Q global self esteem (tp2)
57	Q physical self esteem (tp2)
58	Q location (tp1)
59	Q location (tp2)
60	20MSR term (tp1)
61	20MSR term (tp2)
62	20MSR (tp1)
63	20MSR (tp2)
64	Q date / term (tp1)
65	Q date / term (tp2)
66	Q PA past week (tp1)
67	Q PA past week (tp2)
68	Q attitude (tp1)
69	Q attitude (tp2)
70	Q enjoy PE (tp1)
71	Q enjoy PE (tp2)
72	Q habitual PA (tp1)
73	Q habitual PA (tp2)
74	Q willingness (tp1)
75	Q willingness (tp2)
76	Q self health (tp1)
77	Q self health (tp2)
78	Q similar to active (tp1)
79	Q similar to active (tp2)
80	Q intent to be active (tp1)
81	Q intent to be active (tp2)

Continued on next page

Table F.3 – continued from previous page

#	Variables
82	School level fidelity

Abbreviations: 20MSR = 20 meter shuttle run, eFSM = eligible for free school meals, Q = questionnaire, RT = reaction time, tp = timepoint

Table F.4: Excluded variables from imputation models

Variable name	Excluded from
Reaction time task date / term (tp2)	Imputation models for baseline (tp1) Flanker or task-switching performance
Q date / term (tp1)	Imputation models of task-switching (tp2) scores
20MSR term (tp1)	Only used to impute 20MSR scores
20MSR term (tp2)	

Abbreviations:: 20MSR = 20-meter shuttle run, tp = timepoint

Back to main text

F.3 Drop-out of schools at baseline assessments

F.3.1 School drop-out pre-baseline

Eleven schools (6 intervention, 5 control) dropped-out post-randomisation but prior to baseline assessments, because of time constraints (n = 3), staff changes (n = 2), inability to commit (n = 1) or unknown reasons (n = 4). These schools were unaware of their treatment allocation at the time of drop-out. Table F.5 presents characteristics of schools and pupils that remained in the trial and those that dropped out of the trial.

A comparison between schools that dropped-out and remained part of the trial showed no significant differences in any of the school-level or pupil level demographics (all p >0.05). In particular, a bias-reduced generalised linear model (Kosmidis, 2014; Kosmidis *et al.*, 2020) was fit to examine whether any of the school level variables predicted drop-out. The analysis of deviance table (type II Wald χ^2 tests) is provided in Table F.6. We then fit a multilevel logistic regression (with logit link) to examine whether any pupil level characteristics predicted drop-out Table F.7. The remaining

93 schools were available for baseline assessments and are considered the main trial baseline sample.

Table F.5: School level and pupil level characteristics of drop-outs and those that remained part of the trial.

	Control		Intervention		Overall	
	Included	**Drop-out**	**Included**	**Drop-out**	**Included**	**Drop-out**
School level						
No. schools	47	5	46	6	93	11
Gender status						
Co-ed	38 (80.9%)	4 (80.0%)	37 (80.4%)	5 (83.3%)	75 (80.6%)	9 (81.8%)
Female	8 (17.0%)	0 (0%)	9 (19.6%)	1 (16.7%)	17 (18.3%)	1 (9.1%)
Male	1 (2.1%)	1 (20.0%)	0 (0%)	0 (0%)	1 (1.1%)	1 (9.1%)
Establishment type						
Academy - Sponsor Led	9 (19.1%)	0 (0%)	7 (15.2%)	2 (33.3%)	16 (17.2%)	2 (18.2%)
Academy Converter	1 (2.1%)	0 (0%)	2 (4.3%)	0 (0%)	3 (3.2%)	0 (0%)
Academy Converter - Main	15 (31.9%)	3 (60.0%)	16 (34.8%)	1 (16.7%)	31 (33.3%)	4 (36.4%)
Community School	7 (14.9%)	1 (20.0%)	8 (17.4%)	2 (33.3%)	15 (16.1%)	3 (27.3%)
Foundation School	5 (10.6%)	1 (20.0%)	5 (10.9%)	0 (0%)	10 (10.8%)	1 (9.1%)
Free School - Mainstream	2 (4.3%)	0 (0%)	3 (6.5%)	0 (0%)	5 (5.4%)	0 (0%)
Voluntary aided school	7 (14.9%)	0 (0%)	4 (8.7%)	1 (16.7%)	11 (11.8%)	1 (9.1%)
Voluntary controlled school	1 (2.1%)	0 (0%)	1 (2.2%)	0 (0%)	2 (2.2%)	0 (0%)
School type						
Comprehensive to 16	9 (19.1%)	2 (40.0%)	7 (15.2%)	3 (50.0%)	16 (17.2%)	5 (45.5%)
Comprehensive to 18	33 (70.2%)	3 (60.0%)	34 (73.9%)	3 (50.0%)	67 (72.0%)	6 (54.5%)
Secondary Modern	5 (10.6%)	0 (0%)	5 (10.9%)	0 (0%)	10 (10.8%)	0 (0%)
School size						
Large (>200)	13 (27.7%)	1 (20.0%)	14 (30.4%)	4 (66.7%)	27 (29.0%)	5 (45.5%)
Medium (100-200)	27 (57.4%)	4 (80.0%)	27 (58.7%)	2 (33.3%)	54 (58.1%)	6 (54.5%)
Small (<100)	7 (14.9%)	0 (0%)	5 (10.9%)	0 (0%)	12 (12.9%)	0 (0%)
No. of form groups						
Median (range)	7 (3–14)	6 (5–8)	6 (3–12)	8.5 (5–12)	6 (3–14)	7 (5–12)
Missing, no. (%)	1 (2.1%)	0 (0%)	1 (2.2%)	0 (0%)	2 (2.2%)	0 (0%)
Ofsted rating[b]						
Outstanding	14 (29.8%)	2 (40.0%)	8 (17.4%)	2 (33.3%)	22 (23.7%)	4 (36.4%)
Good	24 (51.1%)	1 (20.0%)	29 (63.0%)	2 (33.3%)	53 (57.0%)	3 (27.3%)
Requires Improvement	6 (12.8%)	1 (20.0%)	4 (8.7%)	1 (16.7%)	10 (10.8%)	2 (18.2%)
Serious Weaknesses	0 (0%)	0 (0%)	1 (2.2%)	0 (0%)	1 (1.1%)	0 (0%)
Special Measures	3 (6.4%)	1 (20.0%)	3 (6.5%)	1 (16.7%)	6 (6.5%)	2 (18.2%)
Missing, no. (%)	0 (0%)	0 (0%)	1 (2.2%)	0 (0%)	1 (1.1%)	0 (0%)
Socioeconomic status[a]						
eFSM, mean (SD), %	17.7 (12.9)	17.4 (13.9)	17.0 (9.00)	18.8 (5.04)	17.3 (11.1)	18.2 (9.53)
IMD, median (range), decile	6 (1–10)	8 (1–10)	5 (1–10)	4.5 (3–7)	5 (1–10)	5 (1–10)
School location[c]						
Rural town and fringe	4 (8.5%)	0 (0%)	5 (10.9%)	0 (0%)	9 (9.7%)	0 (0%)
Rural village	0 (0%)	1 (20.0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.1%)
Urban city and town	20 (42.6%)	3 (60.0%)	19 (41.3%)	3 (50.0%)	39 (41.9%)	6 (54.5%)
Urban major conurbation	23 (48.9%)	1 (20.0%)	22 (47.8%)	3 (50.0%)	45 (48.4%)	4 (36.4%)
Pupil level						
No. participants	8157	878	7860	1366	16017	2244
Age[d], mean (SD), year	12.5 (0.29)	12.5 (0.29)	12.5 (0.3)	12.5 (0.23)	12.5 (0.3)	12.5 (0.3)
Female, no. (%)	4495 (55.1%)	340 (38.7%)	4466 (56.8%)	713 (52.2%)	8961 (55.9%)	1053 (46.9%)
eFSM, no. (%), yes	1422 (17.4%)	151 (17.2%)	1243 (15.8%)	252 (18.4%)	2665 (16.6%)	403 (18.0%)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, IMD = index of multiple deprivation, no. = number, SD = standard deviation

[a] Socioeconomic background was expressed for descriptive purposes using two complementary measures (1) average percentage of eFSM pupils within a school, and (2) using the postcode-derived index of multiple deprivation (deciles). A higher percentage of eFSM pupils reflects a lower socioeconomic background, whereas higher IMD scores reflect a higher socioeconomic background.

[b] Ofsted rating from schools was derived from the UK government website https://reports.ofsted.gov.uk

[c] School location was derived using: https://www.gov.uk/government/collections/rural-urban-classification

[d] Age at baseline (1 September 2017)

Table F.6: School level drop-out model: Analysis of Deviance (Type II Wald χ^2 test.

Predictor	df	χ^2	p
Gender type	2	0.73	0.69
School type	2	2.45	0.29
Establishment type	7	2.32	0.94
School size	2	0.97	0.62
Number of form groups	1	0.00	0.98
Ofsted rating	3	3.42	0.33
Percent eFSM	1	3.42	0.33
IMD decile	1	0.00	0.95
School location	3	3.72	0.29

Abbreviations: eFSM = eligible for free school meals, IMD = index of multiple deprivation

Table F.7: Pupil level predictors of school drop-out pre-baseline assessment (multilevel logistic regression with logit link function)

Predictor	OR	95% CI	p
Sex (reference = female)	1.18	0.001, 1043.13	0.96
Age	0.99	0.002, 452.8	1.00
eFSM (reference = no)	1.06	0.0002, 5870	0.99

Abbreviations: CI = confidence interval, eFSM = eligible for free school meals, OR = odds ratio
Note: intercept of model not displayed (all 0)

Back to main text

F.4 Data cleaning

An overview of participants who completed each assessment, per group and timepoint, is provided in Table F.8. While cognitive assessments and online questionnaires were cleaned using a conservative cleaning approach, no cleaning has been applied to the 20MSR data (for reasons, see F.1.3.3).

Given the large numbers of subjects that were removed from the dataset due to cleaning, we assessed whether participant characteristics could predict in- or exclusion during cleaning using linear mixed effects models (with logit link function, available in the *lme4* package in R). Results of linear mixed effects models are presented in Table F.9. Being male was predictive of exclusion due to cleaning for the majority

Table F.8: Total number of participants who completed each assessment, pre and post cleaning, per group and timepoint.

	Intervention							Control						
	Baseline			**Posttest**			**Pre-post (cleaned)**	**Baseline**			**Posttest**			**Pre-post (cleaned)**
	Raw	Clean	Excluded	Raw	Clean	Excluded		Raw	Clean	Excluded	Raw	Clean	Excluded	
Fitness	4927	—	—	3336	—	—	3049	5770	—	—	4778	—	—	4264
Reaction time task	3610	3562	48	2504	2490	14	1500	4449	4378	71	3676	3649	27	2626
Relational mem. task	2564	2471	93	1936	1788	148	840	2954	2847	107	2936	2704	232	1445
Two-back task	2395	2251	144	1938	1784	154	803	2928	2761	167	2937	2717	220	1468
Flanker task	2730	2608	122	1930	1806	124	890	2983	2840	143	2916	2711	205	1489
Switching task	2374	2169	205	1831	1561	270	709	2817	2553	264	2819	2375	444	1221
Questionnaire[a]	4753	4735	18	2578	2558	20	1899	5052	5040	12	3890	3856	34	2799

Abbreviations: Relational mem. = relational memory

[a] The values for the questionnaire may differ depending on the measure due sporadic missingness. Values presented here reflect the number of cases who completed the one or more sections of the questionnaire.

418

of cognitive assessments (i.e. relational memory task, Flanker task, two-back task, colour-shape switch task, and questionnaire at post-intervention), as well as being eligible for FSM (relational memory task at baseline, the Flanker task at baseline, the two-back task and colour-shape switch task) and older age (questionnaire at baseline).

Table F.9: Models testing whether baseline characteristics predict removal during cleaning

	OR	95%	*p*
Relational memory task —baseline			
sex (reference = female)	3.13	2.28, 4.3	<0.01
(reference = non eligible)	2.42	1.71, 3.42	<0.01
age	0.79	0.48, 1.29	0.35
Relational memory task—post-intervention			
sex (reference = female)	3.19	2.47, 4.13	<0.01
eFSM (reference = non eligible)	1.22	0.91, 1.64	0.19
age	1.01	0.69, 1.46	0.97
Reaction time task—baseline			
sex (reference = female)	1.27	0.87, 1.86	0.22
eFSM (reference = non eligible)	1.13	0.69, 1.87	0.62
age	0.91	0.49, 1.68	0.76
Reaction time task—post-intervention			
sex (reference = female)	1.34	0.71, 2.51	0.37
eFSM (reference = non eligible)	1.94	0.94, 4	0.07
age	1.31	0.45, 3.79	0.62
Flanker task—baseline			
sex (reference = female)	3.14	2.37, 4.17	<0.01
eFSM (reference = non eligible)	1.79	1.3, 2.46	<0.01
age	0.91	0.59, 1.4	0.68
Flanker task—post-intervention			
sex (reference = female)	2	1.54, 2.6	<0.01
eFSM (reference = non eligible)	1.25	0.91, 1.71	0.16
age	0.86	0.58, 1.28	0.45
Two-back task—baseline			
sex (reference = female)	1.64	1.28, 2.1	<0.01
eFSM (reference = non eligible)	1.9	1.4, 2.56	<0.01
age	1.16	0.78, 1.72	0.47
Two-back task—post-intervention			
sex (reference = female)	1.85	1.45, 2.35	<0.01
eFSM (reference = non eligible)	1.64	1.24, 2.16	<0.01
age	0.76	0.52, 1.11	0.15
Colour-shape task switching—baseline			
sex (reference = female)	2.24	1.81, 2.77	<0.01
eFSM (reference = non eligible)	1.84	1.43, 2.38	<0.01
age	1.01	0.72, 1.4	0.97
Colour-shape task switching—post-intervention			
sex (reference = female)	2.48	2.05, 3	<0.01
eFSM (reference = non eligible)	1.28	1.01, 1.62	0.04
age	0.9	0.68, 1.2	0.48

Continued on next page

Table F.9 – continued from previous page

	OR	95% CI	P
Questionnaire—baseline			
sex (reference = female)	3.03	1.39, 6.59	0.01
eFSM (reference = non eligible)	1.54	0.62, 3.83	0.36
age	7.55	2.06, 27.61	<0.01
Questionnaire—post-intervention			
sex (reference = female)	6.15	2.8, 13.51	<0.01
eFSM (reference = non eligible)	1.6	0.83, 3.09	0.16
age	0.66	0.25, 1.74	0.4

OR = odds ratio. Values are rounded to two decimal places.
eFSM = eligible for free school meals

Back to main text

F.5 Baseline PA and PE characteristics

F.5.1 Baseline PA and PE characteristics

PA characteristics of schools and pupils at baseline are presented in Table F.10. Estimates of vigorous PA (VPA), moderate-to-vigorous PA (MVPA), moderate PA (MPA) and sedentary PA (SPA) were equivalent across the groups.

Table F.10: Physical activity characteristics of schools and pupils at baseline (available cases only)

	Intervention	Control	Overall
School level			
No. schools	46	47	93
Full PE lesson (1 hr)			
VPA, mean (SD), min	3.08 (0.939)	3.31 (1.13)	3.20 (1.04)
MVPA, mean (SD), min	11.1 (2.16)	11.2 (2.27)	11.1 (2.20)
MPA, mean (SD), min	7.99 (1.54)	7.89 (1.52)	7.94 (1.52)
SPA, mean (SD), min	31.9 (4.26)	31.9 (3.20)	31.9 (3.75)
Active PE lesson (1 hr)			
VPA, mean (SD), min	3.93 (1.16)	4.39 (1.60)	4.16 (1.41)
MVPA, mean (SD), min	13.7 (2.66)	14.3 (3.00)	14.0 (2.83)
MPA, mean (SD), min	9.80 (1.93)	9.90 (1.85)	9.85 (1.88)
SPA, mean (SD), min	26.9 (4.65)	26.2 (3.96)	26.6 (4.31)
Missing, no. (%)	1 (2.2%)	2 (4.3%)	3 (3.2%)
Pupil level			
No. pupils with PA data (total)	4734	4965	9699
PA (past week), no (%), days 60 min MVPA/day			
0	156 (3.3%)	129 (2.6%)	285 (2.9%)
1	239 (5.0%)	239 (4.8%)	478 (4.9%)
2	468 (9.9%)	412 (8.3%)	880 (9.1%)
3	728 (15.4%)	730 (14.7%)	1458 (15.0%)
4	813 (17.2%)	915 (18.4%)	1728 (17.8%)
5	839 (17.7%)	877 (17.7%)	1716 (17.7%)
6	641 (13.5%)	666 (13.4%)	1307 (13.5%)
7	850 (18.0%)	997 (20.1%)	1847 (19.0%)
Habitual PA (6 months), no. (%), 60 min MVPA/day			
Never	80 (1.7%)	78 (1.6%)	158 (1.6%)
Rarely	139 (2.9%)	138 (2.8%)	277 (2.9%)
Occasionally	259 (5.5%)	255 (5.1%)	514 (5.3%)
Sometimes	1152 (24.3%)	1171 (23.6%)	2323 (24.0%)
Frequently	1059 (22.4%)	1073 (21.6%)	2132 (22.0%)
Usually	1158 (24.5%)	1265 (25.5%)	2423 (25.0%)
Always	887 (18.7%)	985 (19.8%)	1872 (19.3%)

Abbreviations: MPA = moderate intensity physical activity, MVPA = moderate-to-vigorous intensity physical activity, PA = physical activity, PE = physical education, SD = standard deviation, SPA = sedentary physical activity, VPA= vigorous intensity physical activity

F.5.2 PE lessons across schools

Table F.11 displays the average number of PE lessons, lesson duration and total minutes of PE per week in intervention and control schools.

Table F.11: PE lesson characteristics across intervention and control schools

	Intervention (n = 46)	Control (n = 47)	Overall (n = 93)
Number of PE lessons per week			
Mean (SD)	1.61 (0.493)	1.59 (0.498)	1.60 (0.493)
Median (range)	2 (1–2)	2 (1–2)	2 (1–2)
Missing, no. (%)	2 (4.3%)	1 (2.1%)	3 (3.2%)
Lesson duration, min			
Mean (SD)	72.3 (24.4)	72.2 (20.1)	72.3 (22.2)
Median (range)	60 (50–150)	60.0 (47.5–120)	60.0 (47.5–150)
Missing, no. (%)	1 (2.2%)	1 (2.1%)	2 (2.2%)
Total min of PE per week			
Mean (SD)	110 (31.0)	107 (20.7)	109 (26.2)
Median (range)	120 (50–240)	120 (47.5–150)	120 (7.5–240)
Missing, no. (%)	2 (4.3%)	1 (2.1%)	3 (3.2%)

Abbreviations: PE = physical education, SD = standard deviation

Back to main text

F.6 Primary outcomes

F.6.1 Fully adjusted models: multiply imputed data

Table F.12 displays the (un)standardized mean differences between the groups at post intervention (adjusted for baseline values of the outcome measure, school gender type and when and where assessments were completed at baseline and posttest).

Table F.12: The effect of the intervention on secondary outcomes (fully adjusted models)

	Adjusted mean difference (95% CI)[a]	
	Unstandardized	**Standardized[b]**
20MSR		
CRF, laps	0.43 (-2.61, 3.47)	0.02 (-0.12, 0.16)
Reaction time task		
RT[c], ms	3.85 (-6.8, 14.5)	0.04 (-0.07, 0.15)
Relational memory task		
Accuracy, %	-1.37 (-3, 0.27)	-0.1 (-0.23, 0.02)
Two-back task		
Accuracy, %	-1.04 (-3.47, 1.38)	-0.05 (-0.17, 0.07)
RT[c], ms	-5.84 (-24.79, 13.1)	-0.03 (-0.14, 0.07)
Flanker task		
Accuracy congruent, %	-0.8 (-2.28, 0.68)	-0.05 (-0.14, 0.04)
Accuracy incongruent, %	-1.14 (-2.97, 0.69)	-0.06 (-0.15, 0.03)
RT[c] congruent, ms	-2.75 (-10.63, 5.13)	-0.03 (-0.13, 0.06)
RT[c] incongruent, ms	-6.04 (-15.76, 3.68)	-0.06 (-0.14, 0.03)
Colour-shape switching task		
Accuracy non-switch, %	-0.83 (-2.7, 1.05)	-0.05 (-0.15, 0.06)
Accuracy switch, %	-0.42 (-2.26, 1.43)	-0.02 (-0.13, 0.08)
RT[c] non-switch, ms	-5.75 (-34.54, 23.05)	-0.02 (-0.11, 0.07)
RT[c] switch, ms	-14.63 (-68.45, 39.19)	-0.03 (-0.12, 0.07)
Psychosocial problems		
Internalising score3	0.19 (-0.19, 0.56)	0.05 (-0.05, 0.15)
Externalising score3	0.24 (-0.13, 0.6)	0.06 (-0.03, 0.16)
Self-esteem		
Global	-0.05 (-0.22, 0.12)	-0.05 (-0.21, 0.11)
Physical	-0.02 (-0.22, 0.19)	-0.01 (-0.16, 0.13)

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, RT = reaction time
[a] Adjusted mean difference, adjusted for baseline values of the outcome variable, school gender type, location and date variables, and including a random intercept term across schools.
[b] The outcome was standardized (mean = 0, SD = 1), prior to fitting the fully adjusted model.
[c] Lower scores represent better performance

Back to main text

F.6.2 Sensitivity analysis: complete-cases

F.6.2.1 Primary analysis models

Table F.13 displays the mean scores on all outcome measures by treatment group, as well as the (un)standardized mean differences between the groups at post intervention (adjusted for baseline values of the outcome measure and school gender type).

Back to main text

In Figure F.2, the distributions in change of outcome measures (post - pre test

Table F.13: Mean scores for outcome measures by group, with (un)standardized mean differences

	Intervention group			Control group			Mean difference (95% CI)[a]	
	N	Baseline M (SD)	Post M (SD)	N	Baseline M (SD)	Post M (SD)	Unstandardized	Standardized[b]
20MSR								
CRF, laps	3049	38.6 (21.3)	43.9 (21.9)	4264	38.3 (19.9)	42.5 (22.1)	1.12 (-2.03, 4.26)	0.05 (-0.09, 0.19)
Reaction time task								
RT[c], ms	1500	375 (89.2)	378 (95.1)	2626	372 (80.3)	371 (91.6)	6.63 (-5.26, 18.52)	0.034 (-0.05, 0.12)
Relational memory task								
Accuracy, %	840	60.9 (12.0)	62.6 (12.2)	1445	61.8 (12.0)	64.0 (12.9)	-1.33 (-3.22, 0.56)	-0.11 (-0.25, 0.04)
two-back task								
Accuracy, %	803	57.2 (17.2)	61.1 (19.5)	1468	57.9 (17.9)	62.3 (19.8)	-1.1 (-4.1, 1.9)	-0.05 (-0.20, 0.09)
RT[c], ms	803	881 (160)	824 (162)	1468	871 (160)	816 (168)	1.77 (-20.1, 23.66)	0.01 (-0.12, 0.14)
Flanker task								
Accuracy congruent, %	890	85.1 (16.3)	88.2 (14.5)	1489	85.4 (15.2)	88.6 (14.6)	-0.48 (-2.09, 1.14)	-0.03 (-0.13, 0.07)
Accuracy incongruent, %	890	63.9 (20.2)	67.9 (19.0)	1489	63.7 (19.6)	68.5 (18.7)	-0.94 (-2.87, 0.99)	-0.05 (-0.15, 0.05)
RT congruent[c], ms	890	492 (90.3)	477 (76.6)	1489	494 (94.1)	478 (78.1)	-1.29 (-8.24, 5.67)	-0.02 (-0.10, 0.07)
RT incongruent[c], ms	890	559 (124)	543 (95.0)	1489	561 (123)	547 (102)	-4.73 (-12.69, 3.22)	-0.04 (-0.12, 0.03)
Colour-shape switching task								
Accuracy non-switch, %	709	77.8 (16.9)	80.9 (16.5)	1221	78.1 (16.9)	82.3 (16.2)	-1.34 (-3.52, 0.84)	-0.08 (-0.21, 0.05)
Accuracy switch, %	709	73.7 (16.5)	77.5 (16.3)	1221	74.4 (16.8)	78.2 (16.7)	-0.56 (-2.64, 1.51)	-0.03 (-0.15, 0.09)
RT non-switch[c], ms	709	1090 (345)	967 (296)	1221	1080 (327)	953 (294)	9.27 (-25.22,43.75)	0.03 (-0.08, 0.14)
RT switch[c], ms	709	1540 (586)	1400 (497)	1221	1510 (549)	1390 (501)	-13.9 (-72.47, 44.67)	-0.03 (-0.14, 0.08)
Psychosocial problems								
Internalising score	1899	4.97 (3.47)	5.31 (3.59)	2795	5.16 (3.48)	5.38 (3.57)	0.02 (-0.27, 0.31)	0.004 (-0.08, 0.09)
Externalising score	1899	5.75 (3.59)	6.16 (3.58)	2790	5.88 (3.64)	6.19 (3.75)	0.03 (-0.27, 0.33)	0.009 (-0.07, 0.09)
Self-esteem								
Global	1886	4.45 (0.896)	4.32 (0.967)	2683	4.50 (0.865)	4.35 (0.951)	0.01 (-0.07, 0.08)	0.01 (-0.07, 0.09)
Physical	1886	4.39 (1.27)	4.07 (1.33)	2684	4.44 (1.25)	4.13 (1.32)	-0.02 (-0.09, 0.05)	-0.01 (-0.06, 0.04)

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, RT = reaction time

[a] Adjusted mean difference, adjusted for clustering, baseline values of the outcome variable and school gender type.

[b] The outcome was standardized (mean = 0, SD = 1), prior to fitting the baseline and stratification-variable adjusted model.

[c] Lower scores represent better performance.

424

scores) are visualised (left column), alongside the predicted outcomes (i.e. marginal means, at post-test; right column), using the primary ITT model. The predicted (marginal) effects are displayed with 95% confidence intervals, based on the standard error, and created with the *ggeffects* package in R (Lüdecke, 2018). All plots are based on available (complete-cases) only.

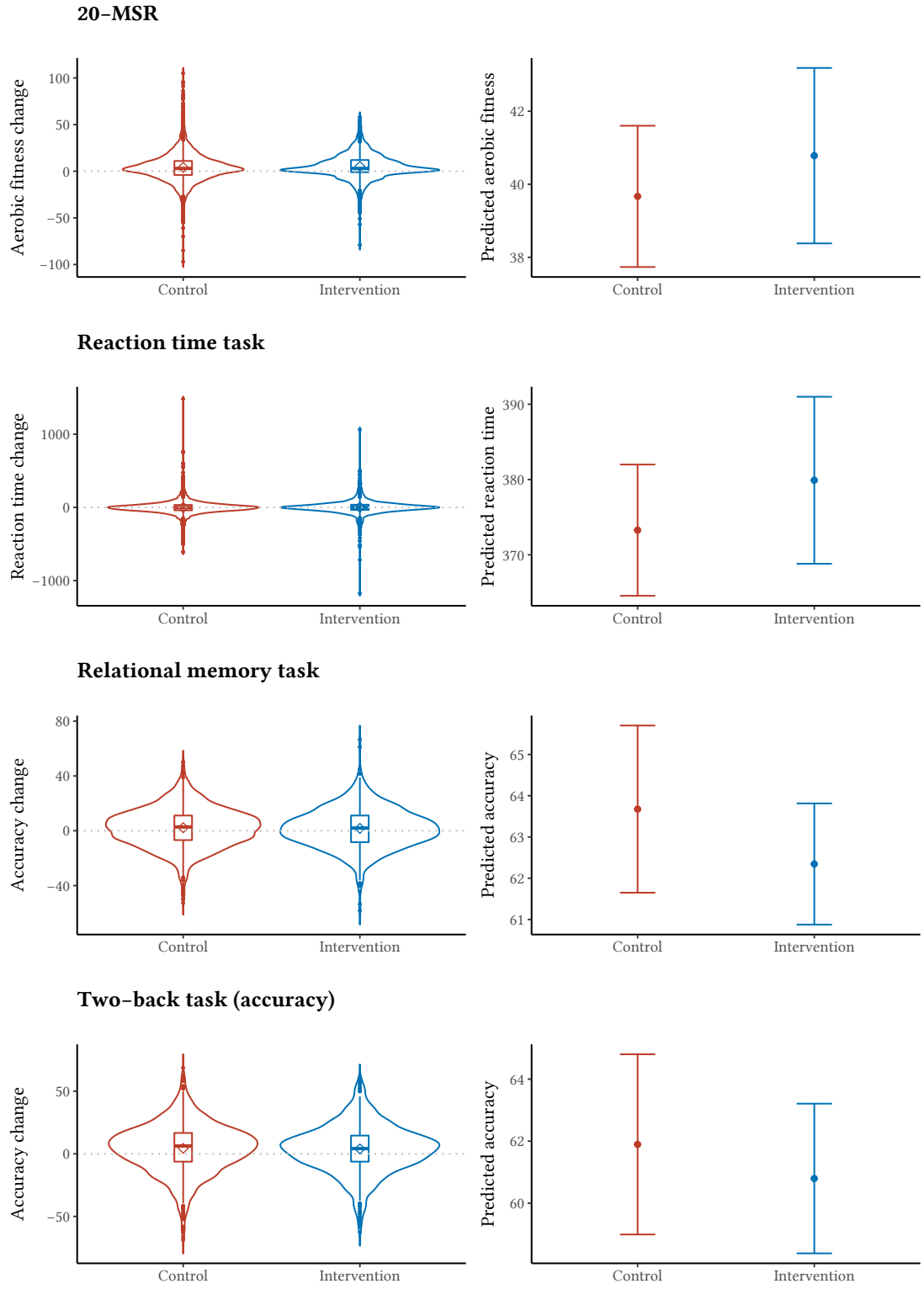


Figure F.2: Distributions of outcome variables, complete-cases (*cont.*)

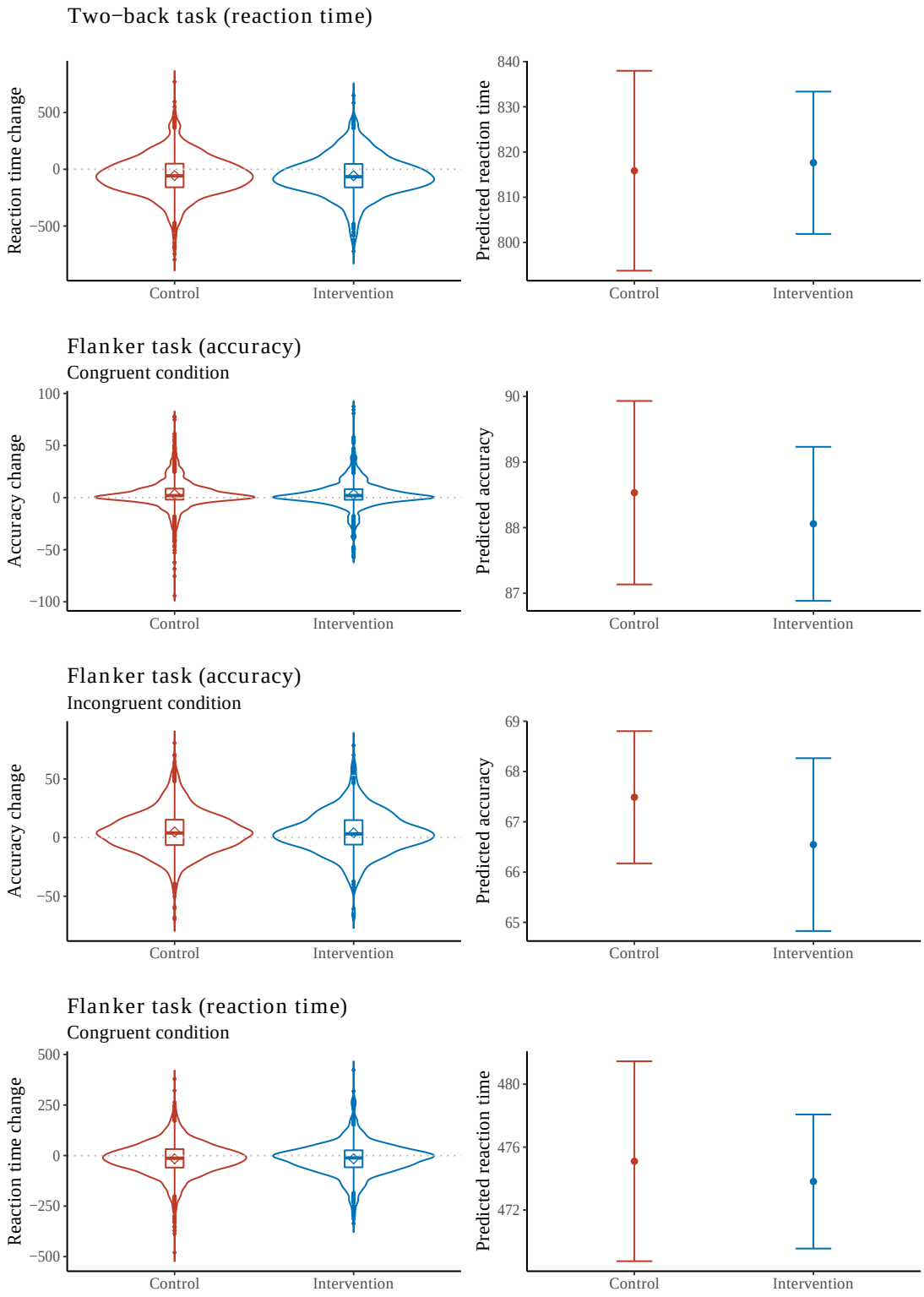


Figure F.2: (*cont.*) Distributions of outcome variables (complete-cases)

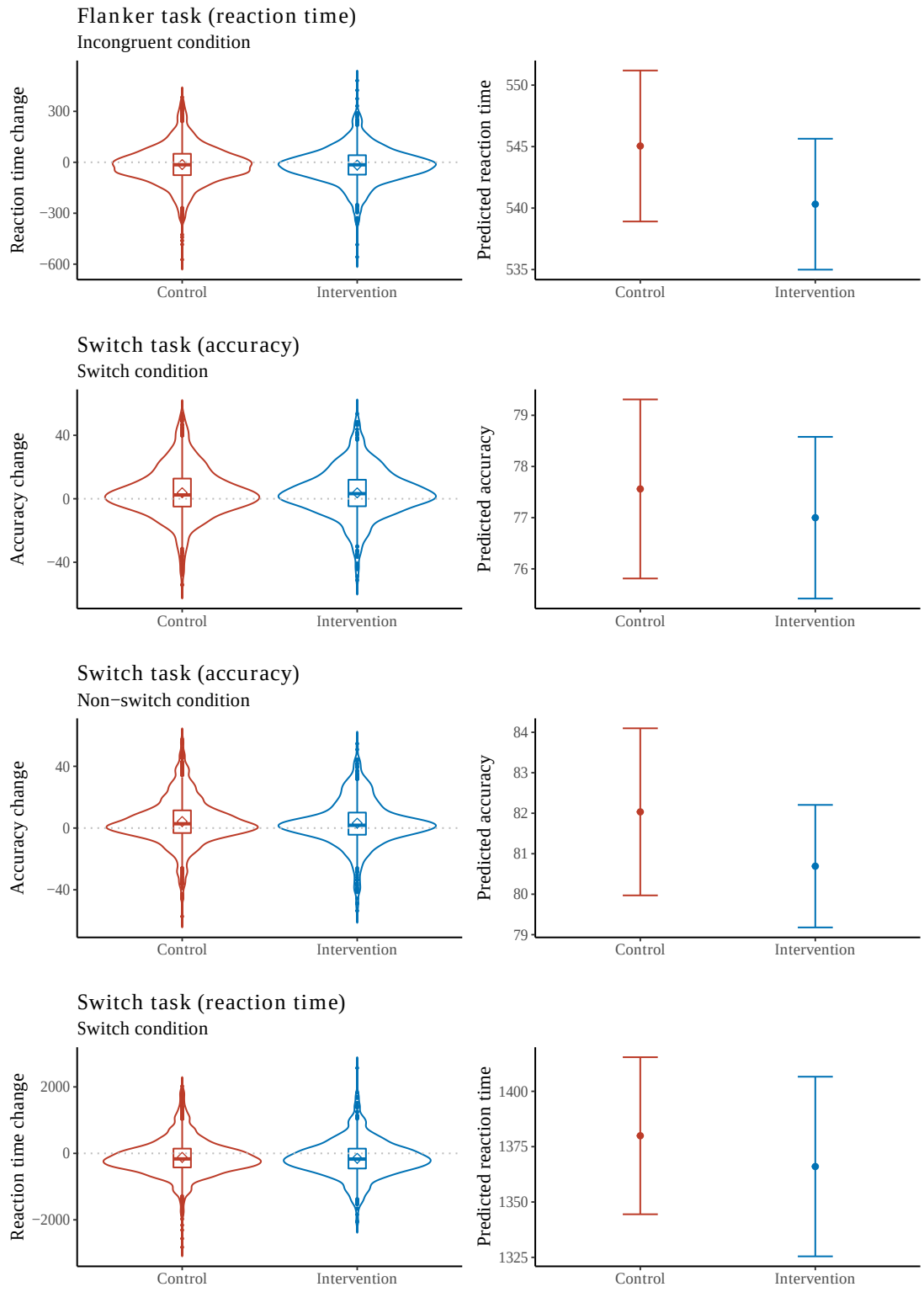


Figure F.2: (*cont.*) Distributions of outcome variables (complete-cases)

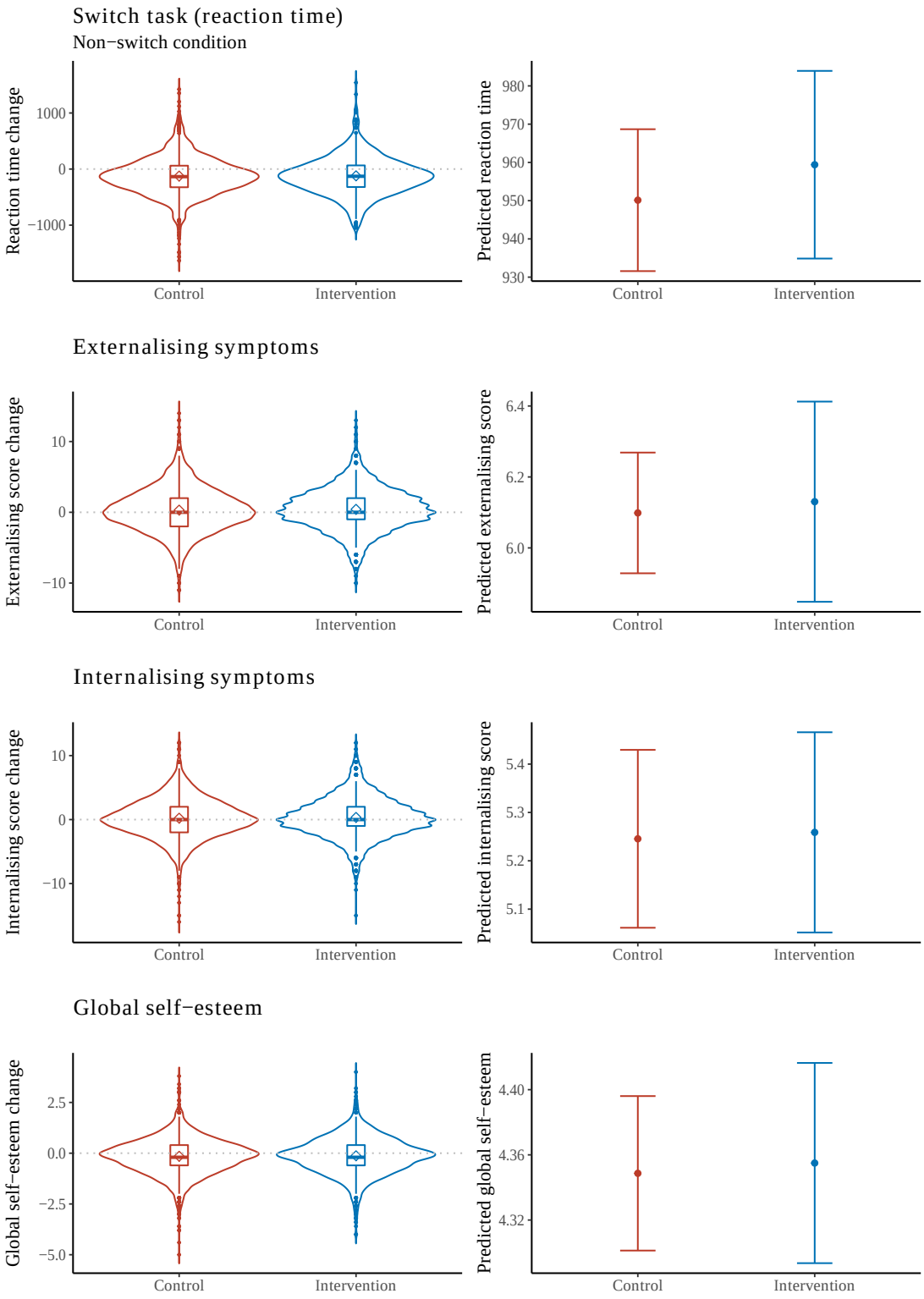


Figure F.2: (*cont.*) Distributions of outcome variables (complete-cases)

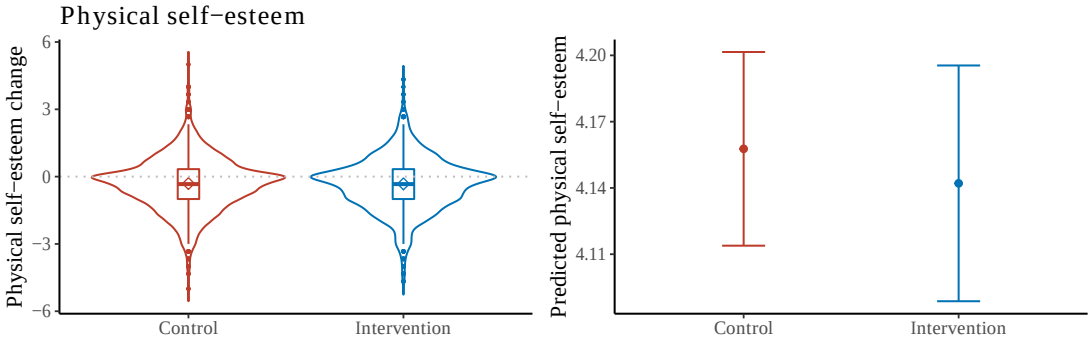


Figure F.2: (*cont.*) Distributions of outcome variables (complete-cases)

F.6.2.2 Fully adjusted models

Table F.14 presents the intervention standardized and unstandardized differences for the fully adjusted models in which we additionally controlled for location (school or home) and date on which the tasks were completed (summer, holidays and autumn; for CRF: summer and autumn only).

Table F.14: The effect of the intervention on secondary outcomes (fully adjusted models)

	N		Adjusted mean difference (95% CI)[a]	
	Participants	Clusters	Unstandardized	Standardized[b]
20MSR				
CRF, laps	7313	55	1.02 (-2.27, 4.3)	0.05 (-0.1, 0.2)
Reaction time task				
RT[c], ms	4126	55	8.17 (-2.63, 18.96)	0.09 (-0.03, 0.2)
Relational memory task				
Accuracy, %	2285	55	-1.41 (-3.35, 0.54)	-0.11 (-0.27, 0.04)
Two-back task				
Accuracy, %	2271	55	-1.71 (-4.86, 1.44)	-0.08 (-0.24, 0.07)
RT[c], ms	2271	55	-3.44 (-26.45, 19.56)	-0.02 (-0.16, 0.12)
Flanker task				
Accuracy congruent, %	2379	55	-0.5 (-2.26, 1.26)	-0.03 (-0.14, 0.08)
Accuracy incongruent, %	2379	55	-1.16 (-3.13, 0.8)	-0.06 (-0.16, 0.04)
RT congruent[c], ms	2379	55	-2.55 (-10.08, 4.98)	-0.03 (-0.13, 0.06)
RT incongruent[c], ms	2379	55	-6.26 (-15.09, 2.56)	-0.06 (-0.14, 0.02)
Colour-shape switching task				
Accuracy non-switch, %	1930	55	-1.82 (-3.88, 0.23)	-0.11 (-0.23, 0.01)
Accuracy switch, %	1930	55	-1.1 (-2.95, 0.75)	-0.06 (-0.17, 0.04)
RT non-switch[c], ms	1930	55	13.96 (-18.26, 46.17)	0.05 (-0.06, 0.15)
RT switch[c], ms	1930	55	-21.36 (-77.12, 34.4)	-0.04 (-0.14, 0.06)
Psychosocial problems				
Internalising score	4660	52	0.04 (-0.27, 0.35)	0.01 (-0.08, 0.1)
Externalising score	4655	52	0.08 (-0.2, 0.36)	0.02 (-0.05, 0.1)
Self-esteem				
Global	4536	52	0 (-0.08, 0.08)	0 (-0.08, 0.08)
Physical	4537	52	-0.02 (-0.09, 0.06)	-0.01 (-0.07, 0.04)

Abbreviations: 20MSR = 20 meter shuttle runCRF = cardiorespiratory fitness, RT = reaction time
[a] Adjusted mean difference, adjusted for baseline values of the outcome variable, school gender type, location and date variables, and including a random intercept term across schools.
[b] The outcome was standardized (mean = 0, SD = 1), prior to fitting the fully adjusted model.
[c] Lower scores represent better performance

Back to main text

F.7 Subgroup analysis

F.7.0.1 Primary analysis: multiply imputed data

We examined whether the effect of the intervention on CRF, cognitive outcomes and mental health was moderated by sex, socioeconomic status (eFSM), or baseline CRF levels (dichotomous: high or low). Table F.15 presents the standardized and unstandardized estimates of the interaction terms of each subgroup analysis, per outcome measure, on multiply imputed data. No significant moderation of sex,

socioeconomic status or baseline CRF levels was observed following correction for multiple comparisons (α = 0.002, with 24 tests). In uncorrected analyses, baseline CRF levels significantly moderated the effect of the intervention on CRF (uncorrected p = 0.03), with better performance (i.e. higher number of laps) for intervention participants with high baseline CRF levels. However, no significant effect of the intervention was observed when the model was run in each subgroup (p >0.05). Moreover, eFSM significantly moderated the effect of the intervention on processing speed (uncorrected p = 0.04), with worse performance (i.e. higher reaction times) for intervention participants that were eligible for free school meals (eFSM). A small significant effect of the intervention was found when the model was run in the eFSM subgroup only, with worse performance (i.e. higher reaction times) for pupils in the intervention group (p = 0.04).

A multigroup model, based on Figure 6.2 (main text chapter 6), was used to assess whether the effect of the intervention on the EF latent variable was equivalent for those classified as fit or unfit, high or low SES and males or females. We first checked for metric invariance as a function of each moderator (baseline CRF, SES and sex). We obtained full metric invariance between the groups for the CRF grouping (equal unstandardized factor loadings; test of the difference between configural and metric invariance $\chi^2(8)$ = 10.32, p = 0.24) and socioeconomic status (equal unstandardized factor loadings; test of the difference between configural and metric invariance $\chi^2(8)$ = 5.88, p= 0.66), but not for sex. The subgroup analysis by sex was therefore not pursued any further.

We next ran a model in which the regression coefficients from intervention to EF at posttest were freely estimated for high and low fitness groups, in this model the unstandardized regression coefficients for the intervention did not differ significantly between groups (Wald test $\chi^2(1)$ = 0.28, p = 0.60). The intervention produced no statistically significant change in EF in either group (unfit: SMD = -0.02, 95% CI [-0.19, 0.14]; fit: SMD = 0.01; 95% CI [-0.16, 0.18]) and the size of these effects did not differ

significantly between fitness groups.

The multigroup ANCOVA model in which the regression coefficients from intervention to EF at posttest were freely estimated for the high and low SES (i.e. non eFSM and eFSM respectively) groups showed that the unstandardized regression coefficients differed significantly between groups (Wald test $\chi^2(1) = 5.60$, $p = 0.02$). However, the intervention produced no statistically significant change in EF in either group (non eFSM: SMD = 0.03, 95% CI [-0.13, 0.19]; eFSM: SMD =-0.21; 95% CI [-0.43, 0.01]) although the intervention in the eFSM group tended to impair EF skills at posttest.

Table F.15: Subgroup analysis, examining whether sex, eFSM or baseline CRF moderate the effect of the intervention on fitness (high / low), cognitive and mental health outcomes.

Outcome	Moderator	Unstandardized		Standardized		p_{unc} [b]
		Estimate[a]	95% CI	Estimate[a]	95% CI	
20MSR	sex	-2.13	-5.12, 0.86	-0.1	-0.23, 0.04	0.16
20MSR	eFSM	-1.73	-4.22, 0.76	-0.08	-0.19, 0.03	0.17
20MSR	CRF (t_0)	2.97	0.34, 5.6	0.13	0.02, 0.25	0.03
Relational memory task, accuracy	sex	-0.54	-2.38, 1.29	-0.04	-0.18, 0.1	0.56
Relational memory task, accuracy	eFSM	0.43	-2.1, 2.95	0.03	-0.16, 0.23	0.74
Relational memory task, accuracy	CRF (t_0)	-0.88	-2.61, 0.85	-0.07	-0.2, 0.06	0.31
Reaction time task, RT	sex	-2.82	-14.8, 9.17	-0.03	-0.15, 0.09	0.64
Reaction time task, RT	eFSM	18.09	1.18, 35	0.19	0.01, 0.36	0.04
Reaction time task, RT	CRF (t_0)	5.59	-5.73, 16.91	0.06	-0.06, 0.17	0.33
Internalising score	sex	-0.09	-0.53, 0.34	-0.03	-0.14, 0.09	0.68
Internalising score	eFSM	0.06	-0.48, 0.61	0.02	-0.13, 0.17	0.82
Internalising score	CRF (t_0)	0.02	-0.34, 0.38	0.01	-0.09, 0.11	0.91
Externalising score	sex	-0.02	-0.43, 0.4	0	-0.11, 0.11	0.94
Externalising score	eFSM	0.17	-0.34, 0.68	0.04	-0.09, 0.18	0.52
Externalising score	CRF (t_0)	-0.06	-0.42, 0.3	-0.02	-0.11, 0.08	0.75
Global self-esteem	sex	0.11	-0.02, 0.24	0.11	-0.02, 0.24	0.11
Global self-esteem	eFSM	-0.06	-0.24, 0.13	-0.06	-0.24, 0.13	0.56
Global self-esteem	CRF (t_0)	-0.05	-0.19, 0.1	-0.05	-0.19, 0.09	0.52
Physical self-esteem	sex	0.12	-0.05, 0.29	0.09	-0.03, 0.21	0.16
Physical self-esteem	eFSM	-0.05	-0.27, 0.18	-0.03	-0.2, 0.13	0.69
Physical self-esteem	CRF (t_0)	-0.07	-0.25, 0.1	-0.05	-0.18, 0.07	0.41

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, RT = reaction time

[a] Displays the estimate of the interaction term between randomised group (reference = control) and baseline CRF status (reference = low fit), sex (reference = female) or eFSM (reference = not eFSM).

[b] The alpha level was adjusted for multiple comparisons (24 tests) to be 0.002. Uncorrected p-values are displayed.

434

Back to main text

F.7.0.2 Sensitivity analysis: complete-cases

Table F.16 presents the standardized and unstandardized estimates of the interaction terms of each subgroup analysis, per outcome measure. No significant moderation of sex, socioeconomic status or baseline CRF levels was observed following correction for multiple comparisons. Baseline CRF levels significantly moderated the effect of the intervention on processing speed (uncorrected $p = 0.02$), with better performance (i.e. lower reaction time) for control participants with high baseline CRF levels.

Table F.16: Subgroup analysis, examining whether sex, eFSM or baseline CRF moderate the effect of the intervention on CRF (high / low), cognitive and mental health outcomes.

Outcome	Moderator	N	Unstandardized		Standardized		P_{UNC} [b]
			Estimate[a]	95% CI	Estimate[a]	95% CI	
20MSR	sex	7313	-2.51	-7.17, 2.16	-0.11	-0.33, 0.1	0.28
20MSR	eFSM	7313	-2.07	-4.6, 0.46	-0.09	-0.21, 0.02	0.1
20MSR	CRF (t0)	7313	1.27	-2.87, 5.41	0.06	-0.13, 0.25	0.54
Relational memory task, accuracy	sex	2285	-1.47	-4.37, 1.42	-0.12	-0.35, 0.11	0.3
Relational memory task, accuracy	eFSM	2285	2.77	-1.81, 7.35	0.22	-0.14, 0.58	0.22
Relational memory task, accuracy	CRF (t0)	1981	-2.05	-4.9, 0.79	-0.16	-0.39, 0.06	0.15
Reaction time task, RT	sex	4126	4.26	-18.27, 26.8	0.04	-0.19, 0.28	0.7
Reaction time task, RT	eFSM	4126	-2.16	-27.54, 23.21	-0.02	-0.29, 0.24	0.86
Reaction time task, RT	CRF (t0)	3594	15.44	2.62, 28.26	0.16	0.03, 0.3	0.02
Internalising score	sex	4694	0.01	-0.46, 0.49	0	-0.13, 0.14	0.96
Internalising score	eFSM	4694	0.16	-0.38, 0.71	0.05	-0.11, 0.2	0.54
Internalising score	CRF (t0)	4024	0.14	-0.23, 0.51	0.04	-0.07, 0.14	0.46
Externalising score	sex	4689	0.15	-0.3, 0.61	0.04	-0.08, 0.17	0.49
Externalising score	eFSM	4689	0.18	-0.32, 0.68	0.05	-0.09, 0.18	0.46
Externalising score	CRF (t0)	4019	-0.15	-0.56, 0.26	-0.04	-0.15, 0.07	0.46
Global self-esteem	sex	4569	0.06	-0.06, 0.19	0.07	-0.06, 0.2	0.3
Global self-esteem	eFSM	4569	-0.09	-0.31, 0.12	-0.1	-0.32, 0.12	0.37
Global self-esteem	CRF (t0)	3916	-0.05	-0.17, 0.07	-0.05	-0.18, 0.08	0.42
Physical self-esteem	sex	4570	0.09	-0.05, 0.24	0.07	-0.04, 0.18	0.21
Physical self-esteem	eFSM	4570	-0.13	-0.36, 0.09	-0.1	-0.27, 0.07	0.23
Physical self-esteem	CRF (t0)	3917	-0.09	-0.26, 0.09	-0.06	-0.2, 0.07	0.33

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, RT = reaction time

[a] Displays the estimate of the interaction term between randomised group (reference = control) and baseline CRF status (reference = low fit). sex (reference = female) or eFSM (reference = not eFSM)

[b] The alpha level was adjusted for multiple comparisons (24 tests) to be 0.002. Uncorrected p-values are displayed.

436

Back to main text

F.8 Change - change analysis

F.8.1 Primary analysis: multiply imputed data

We explored whether a change in CRF was related to a change in cognitive or mental health outcomes, and whether this relationship was moderated by treatment status, adjusted for age (mean-centered), sex, eFSM, school gender type (boys-only, girls-only or co-eductional), and location (home or school) and period of completion (summer, holidays or autumn for cognitive and mental health measures; summer and autumn for CRF) of assessments. Table F.17 and Table F.18 show the model terms and estimates for respectively, cognitive and mental health outcomes on multiply imputed cases. The alpha level was adjusted for multiple comparisons (α = 0.008 with 6 comparisons). Model 1 contained an interaction term between CRF-change and the treatment indicator. If the interaction was not significant, the interaction term was removed and the model was reassessed (model 2). Treatment status did not significantly moderate relationships between changes in CRF and changes in cognitive or mental health outcomes, and changes in CRF were not significantly related with any of the outcomes. In uncorrected analyses, the relationship between changes in CRF and changes in processing speed was moderated by treatment (p = 0.04). However, no significant relationships were observed when the model was run in each subgroup (p >0.05).

Back to main text

F.8.2 Sensitivity analysis: complete-cases

Table F.19 and Table F.20 show the model terms and estimates for, respectively cognitive and mental health outcomes on complete-cases only. CRF assessments at post-test were completed during summer term only (for the available-case dataset), hence this confound was excluded from the models. The interaction of change in CRF by treatment group, nor the main effect of change in CRF were significantly related to changes in outcome measures.

Table F.17: Models examining whether a change in CRF is related to a change in cognitive outcomes.

	Relational memory			Processing speed		
	β	95% CI	p	β	95% CI	p
Model 1: CRF change by group (reference = control)	0.06	-0.02, 0.14	0.13	-0.57	-1.12, -0.03	0.04
Model 2: CRF change	-1.39	-3.02, 0.24	0.10	5.07	-6.28, 16.43	0.38

Abbreviations: CRF = cardiorespiratory fitness

Table F.18: Models examining whether a change in CRF is related to a change in mental health outcomes.

	Internalising symptoms			Externalising symptoms			Global self-esteem			Physical self-esteem		
	β	95% CI	p	β	95% CI	p	β	95% CI	p	β	95% CI	p
Model 1: CRF change by group (reference = control)	-0.002	-0.02, 0.01	0.79	0.001	-0.02, 0.02	0.91	-0.001	-0.01, 0.01	0.83	0.001	-0.01, 0.01	0.75
Model 2: CRF change	0.18	-0.21, 0.57	0.36	0.22	-0.16, 0.59	0.26	-0.03	-0.2, 0.14	0.71	0.02	-0.19, 0.22	0.88

Abbreviations: CRF = cardiorespiratory fitness

439

Table F.19: Models examining whether a change in CRF is related to a change in cognitive outcomes.

	Relational memory			Processing speed		
	β	95% CI	p	β	95% CI	p
CRF change by group (reference = control)	0.015	-0.081, 0.111	0.75	-0.418	-1.101, 0.266	0.22
CRF change	0.01	-0.041, 0.062	0.68	-0.043	-0.4, 0.314	0.81

Abbreviations: CRF = cardiorespiratory fitness

Table F.20: Models examining whether a change in CRF is related to a change in mental health outcomes.

	Internalising symptoms			Externalising symptoms			Global self-esteem			Physical self-esteem		
	β	95% CI	p	β	95% CI	p	β	95% CI	p	β	95% CI	p
CRF change by group (reference = control)	0.011	-0.014, 0.035	0.37	0.004	-0.01, 0.018	0.55	0	-0.005, 0.005	0.90	-0.002	-0.009, 0.005	0.5
CRF change	-0.002	-0.013, 0.008	0.65	-0.003	-0.009, 0.004	0.38	0.002	0, 0.005	0.06	0.002	-0.001, 0.005	0.2

Abbreviations: CRF = cardiorespiratory fitness

Back to main text

F.9 Intervention fidelity and CACE analysis

F.9.1 Intervention fidelity

Table F.21 presents school level fidelity measures collected at different times throughout the trial. In particular, teacher-reported compliance was collected post-intervention by NatCen Social Research using a survey asking intervention schools about the percentage of PE lessons in which the intervention was delivered as intended. Only a subset of 22 (of 29 still enrolled at post-intervention) intervention schools completed this survey. Pupil reported compliance with the intervention was collected using a brief three-question survey administered post-intervention, as part of the online questionnaire, asking whether (i) PE lessons started with a warm-up [1 = "never" to 5 = "always"], (ii) PE lessons included bursts of VPA that raised their heart rate and made them feel out of breath (i.e. infusions) [1 = "never" to 5 = "always"], and (iii) they took part in warm-ups or VPA bursts if asked by PE teachers (i.e. whether they "participated") [1 = "never" to 5 = "always"]. Given that the intervention was delivered on a school level and the questionnaire was completed by subsets of pupils within schools, we computed median values of questionnaire items within schools and present averages and median values across schools. In addition to self-report measures of fidelity, we collected an objective measure of VPA per hour of PE using the Axivity AX3 accelerometer, in a convenience sample of participants per school (at least 50% of the Year group). Data was collected during a single PE lesson between March–May 2018. The total minutes spent in VPA per hour of PE, with or without adjustment for change time (respectively, *full* lesson vs. *active* lesson) were computed per class and averaged within a school. Finally, Table F.21 contains information on whether PE teachers of intervention schools attended teacher training sessions face-to-face, online or neither (i.e. did not attend any session). Schools of which PE teachers did not attend any teacher training sessions (n = 5) were lost to follow-up.

Of 46 intervention schools, 17 (37%) implemented the intervention in over 50% of PE lessons during the year. Median and average pupil reported warm ups, infusions and participation on a school level are comparable across intervention and control schools. Average minutes spent in VPA per active hour of PE was slightly higher in intervention schools (M [SD] = 4.91 [1.77]) compared to control schools (M [SD] = 4.38 [1.93]).

Table F.21: School level fidelity measures

	Intervention (n = 46)	Control (n = 47)	p[b]
Teacher reported compliance (NatCen), no (%) of PE lessons			
0-10%	1 (2.2%)		
10-25%	1 (2.2%)		
25-50%	3 (6.5%)		
50-75%	8 (17.4%)		
75-90%	7 (15.2%)		
90-100%	2 (4.3%)		
Missing, no. (%)	24 (52.2%)	47 (100%)	
Pupil reported warm-up (school medians)[a]			
Mean (SD)	4.39 (0.533)	4.17 (0.747)	0.21
Median (range)	4 (3.5-4)	4 (3-5)	
Missing, no. (%)	24 (52.2%)	12 (25.5%)	
Pupil reported infusions (school medians)[a]			
Mean (SD)	3.91 (0.294)	3.83 (0.453)	0.42
Median (range)	4 (3-4)	4 (3-5)	
Missing, no. (%)	24 (52.2%)	12 (25.5%)	
Pupil reported participation (school medians)[a]			
Mean (SD)	4.52 (0.499)	4.49 (0.562)	0.8
Median (range)	4.75 (4-5)	5 (3-5)	
Missing, no. (%)	24 (52.2%)	12 (25.5%)	
VPA (PE lesson, school average)			
Full lesson, mean (SD), min	3.65 (1.09)	3.09 (1.43)	0.23
Active part, mean (SD), min	4.91 (1.77)	4.38 (1.93)	0.07
Missing, no. (%)	16 (34.8%)	6 (12.8%)	
Teacher training			
Face-to-face, no. (%)	15 (32.6%)		
Online, no. (%)	26 (56.5%)		
None, no. (%)	5 (10.9%)	47 (100%)	

Abbreviations: PE = physical education, SD = standard deviation VPA = vigorous intensity physical activity
[a] Median across pupils within a school. Measures were based on 5-point Likert-scales, ranging from never = 1 to always = 5
[b] Assessed using unpaired t-tests, not corrected for multiple testing

F.9.2 CACE analysis

The complier-average causal effect was determined using a two-stage least squares (TSLS) method. Compliers were defined as schools that delivered the intervention >50% of PE lessons. The CACE was first estimated on the multiply imputed sample (Table F.22) and then on the complete-case sample (Table F.23). CACE estimates were comparable in direction, though slightly larger in size, to ITT estimates provided in chapter 6. None of the analyses reached statistical significance.

F.9.2.1 Multiply imputed analysis

Table F.22: CACE estimates and confidence intervals (multiply imputed data)

	CACE (unstandardized)	95% CI
20MSR		
CRF, laps	0.03	-8.05, 8.11
Reaction time task		
RT, ms	4.44	-22.54, 31.43
Relational memory task		
Accuracy, %	-3.46	-7.95, 1.02
Two-back task		
Accuracy, %	-2.1	-8.18, 3.99
RT, ms	-9.11	-56.74, 38.52
Flanker task		
Accuracy congruent, %	-2.21	-6.31, 1.9
Accuracy incongruent, %	-3.3	-8.46, 1.86
RT congruent, ms	-9.41	-31.34, 12.52
RT incongruent, ms	-16.87	-42.79, 9.04
Colour-shape switching task		
Accuracy non-switch, %	-1.75	-6.59, 3.09
Accuracy switch, %	-0.76	-5.59, 4.07
RT non-switch, ms	-16.86	-96.72, 63
RT switch, ms	-50.33	-198.68, 98.02
Psychosocial problems		
Internalising score	0.44	-0.63, 1.51
Externalising score	0.67	-0.42, 1.75
Self-esteem		
Global	-0.16	-0.73, 0.41
Physical	-0.09	-0.71, 0.52

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, RT = reaction time

Back to main text

F.9.2.2 Sensitivity analysis: complete-cases

Table F.23: CACE estimates and confidence intervals (complete-case data)

	CACE (unstandardized)	95% CI
20MSR		
CRF, laps	1.84	-2.85, 6.54
Reaction time task		
RT, ms	11.7	-11.05, 34.45
Relational memory task		
Accuracy, %	-2.15	-5.5, 1.21
Two-back task		
Accuracy, %	-1.83	-7.06, 3.4
RT, ms	2.94	-33.31, 39.19
Flanker task		
Accuracy congruent, %	-0.79	-3.32, 1.75
Accuracy incongruent, %	-1.56	-4.84, 1.72
RT congruent, ms	-2.14	-12.93, 8.66
RT incongruent, ms	-7.84	-20.05, 4.36
Colour-shape switching task		
Accuracy non-switch, %	-2.27	-6.01, 1.48
Accuracy switch, %	-0.95	-4.35, 2.46
RT non-switch, ms	15.67	-38.57, 69.91
RT switch, ms	-23.53	-119.94, 72.87
Psychosocial problems		
Internalising score	0.02	-0.36, 0.39
Externalising score	0.04	-0.35, 0.44
Self-esteem		
Global	0.01	-0.09, 0.11
Physical	-0.02	-0.11, 0.07

Abbreviations: 20MSR = 20 meter shuttle run, CRF = cardiorespiratory fitness, RT = reaction time

Back to main text

F.10 Additional analysis

F.10.1 Multiply imputed data

We examined whether there was a difference in PE enjoyment or attitudes towards PA between the intervention and control group at baseline (Table F.24) and at posttest (Table F.25) on multiply imputed data, and using complete-case data (Table F.26 and Table F.27) as a sensitivity analysis.

Table F.24: Models examining baseline differences in PE enjoyment and attitudes towards PA

	Intervention	Control	Adjusted mean difference (95% CI)[a]		*p*
	M (SD)	M (SD)	**Unstandardized**	**Standardized** [b]	
PE enjoyment	5.81 (1.61)	5.81 (1.59)	0.01 (-0.14, 0.16)	0.01 (-0.09, 0.1)	0.91
Attitudes towards PA	4.97 (1.28)	5.01 (1.28)	-0.03 (-0.15, 0.08)	-0.03 (-0.12, 0.07)	0.57

[a] Models contained the intervention group and were adjusted for clustering and school-gender type
[b] Outcomes were standardized (mean = 0, SD = 1) prior to running the model

Table F.25: Models examining the effect of the intervention on PE enjoyment and attitudes towards PA

	Intervention group		Control group		Adjusted mean difference (95% CI)[a]		*p*
	Baseline M (SD)	**Post** M (SD)	**Baseline** M (SD)	**Post** M (SD)	**Unstandardized**	**Standardized**[b]	
PE enjoyment	5.81 (1.61)	5.49 (1.73)	5.81 (1.59)	5.56 (1.69)	-0.06 (-0.28, 0.16)	-0.04 (-0.16, 0.09)	0.58
Attitudes towards PA	4.97 (1.28)	4.52 (1.34)	5.01 (1.28)	4.7 (1.34)	-0.15 (-0.34, 0.03)	-0.11 (-0.25, 0.02)	0.1

[a] Models contained the intervention group and were adjusted for clustering and school-gender type
[b] Outcomes were standardized (mean = 0, SD = 1) prior to running the model

Back to main text

F.10.2 Sensitivity analysis

We examined whether there was a difference in PE enjoyment or attitudes towards PA between the intervention and control group at baseline (Table F.26) and at posttest (Table F.27). No significant difference between the groups were observed.

Table F.26: Models examining baseline differences in PE enjoyment and attitudes towards PA

	Treatment		Control		Adjusted mean difference (95% CI)[a]		*p*
	N	M (SD)	N	M (SD)	**Unstandardized**	**Standardized**[b]	
PE enjoyment	4734	5.78 (1.59)	5040	5.77 (1.57)	0.01 (-0.12, 0.14)	0.01 (-0.08, 0.09)	0.81
Attitudes towards PA	4734	4.97 (1.26)	5040	4.98 (1.27)	-0.01 (-0.11, 0.09)	-0.01 (-0.09, 0.07)	0.85

[a] Models contained the intervention group and were adjusted for clustering and school-gender type
[b] Outcomes were standardized (mean = 0, SD = 1) prior to running the model

Table F.27: Models examining the effect of the intervention on PE enjoyment and attitudes towards PA

	Intervention			Control group			Adjusted mean difference (95% CI)[a]		*p*
	N	**Baseline** M (SD)	**Post** M (SD)	N	**Baseline** M (SD)	**Post** M (SD)	**Unstandardized**	**Standardized**[b]	
PE enjoyment	1899	5.76 (1.59)	5.49 (1.64)	2797	5.76 (1.54)	5.54 (1.65)	-0.02 (-0.14, 0.09)	-0.01 (-0.08, 0.06)	0.69
Attitudes towards PA	1899	4.97 (1.25)	4.54 (1.26)	2797	5.02 (1.23)	4.65 (1.31)	-0.07 (-0.16, 0.01)	-0.06 (-0.12, 0.01)	0.08

[a] Models contained the intervention group and were adjusted for clustering and school-gender type
[b] Outcomes were standardized (mean = 0, SD = 1) prior to running the model

448

Back to main text

This page is intentionally left blank.

G

Supplementary materials to chapter 7: *The effect of the Fit to Study intervention on the anterior hippocampus, cardiorespiratory fitness and relational memory performance*

Contents

G.1 Methods 452
G.1.1 Missing data 452
G.1.2 Supplementary analysis 454
G.1.3 Software 455
G.2 Results 455
G.2.1 Demographics 455
G.2.2 Fidelity 456
G.2.3 Primary analysis 457
G.2.4 Sensitivity analyses 460
G.2.5 Supplementary analysis 472

G.1 Methods

G.1.1 Missing data

G.1.1.1 Primary analysis

For the primary analysis, missing data was multiply imputed separately for each outcome variable, using a joint modelling approach for multilevel multiple imputation available in the PAN package (Schafer & Graham, 2002; Zhao & Schafer, 2018).

The variables included in the imputation models for the univariate analyses are shown in Table G.1 for each outcome measure.

Table G.1: Variables used in imputation models per outcome measure (column name). Imputations were run separately in each treatment group.

CRF	Anterior hippocampus	Relational memory	Object-location task
School	School	School	School
group	group	group	group
sex	L_Hipp_ant.0	hours_weekTV.1	swap8s.1
paPastSixMonths.0	R_Hipp_ant.0	percentCorrectValid.0	id8s.1
timesPA.0	L_Hipp_ant.1	percentCorrectValid.1	error1s.1
vo2KgPeak.0	R_Hipp_ant.1	error1s.1	error8s.1
vo2KgPeak.1	dScoleSimilar.1	ni1s.1	swap8s.0
bmi.0	FA_left.1	percentCorrectSwitchValid.1	totWeekSed1s.1
dailyAvgVPA1s.0	VBMs_left.1	reactionTimeCorrectMean.1	sex
bmi.1	FA_right.1	reactionTimeHitMean.1	id8s.0
pubertalDevelopmentScale.1	VBMs_right.1	totWeekSed1s.1	
workloadMax.1	FA_left.0	sex	
dailyAvgVPA1s.1	MD_left.0	swap1s.0	
paPastSixMonths.1	VBMs_left.0	error8s.0	
paDaysPastWeek.1	FA_right.0	hours_weekTV.0	
selfRatedHealth.1	prosocialScale.1	hours_weekendTV.0	
timesPA.1	VBMs_right.0	timesVegetables.0	
totWeekSed1s.1	Vscaling.0	error1s.0	
	Vscaling.1		
	sex		
	id8s.0		
	paPastSixMonths.0		

Variable names ending with '.0' refer to variables collected at baseline, whereas those ending with '.1' refer to variables collected at post-intervention

The variables included in the imputation models for the multivariate MRI analy-

sis are shown in Table G.2. Note, this imputation model was not run separately in each treatment group due to the number of variables in the imputation model.

Table G.2: Variables used in imputation model for multivariate analysis.

Multivariate MRI markers	
School	perfusion_right.1
group	R1_right.1
paPastSixMonths.0	R2star_right.1
Vscaling.0	FA_right.1
Vscaling.1	MD_right.1
totWeekSed1s.0	fintra_right.1
percentCorrectValid.0	VBMs_right.1
peerScale.1	perfusion_right.0
perfusion_left.1	R1_right.0
R1_left.1	R2star_right.0
R2star_left.1	FA_right.0
FA_left.1	MD_right.0
MD_left.1	fintra_right.0
fintra_left.1	VBMs_right.0
VBMs_left.1	rerMax.1
perfusion_left.0	L_Hipp_ant.0
R1_left.0	L_Hipp_ant.1
R2star_left.0	R_Hipp_ant.1
FA_left.0	R_Hipp_ant.0
MD_left.0	sex
fintra_left.0	id8s.0
VBMs_left.0	

Variable names ending with '.0' refer to variables collected at baseline, whereas those ending with '.1' refer to variables collected at post-intervention

Back to main text

G.1.1.2 Sensitivity analysis

To assess the robustness of the multiple imputed results for the multivariate analysis, we repeated the analysis in a single imputed dataset. To this end, the SVD imputed dataset of chapter 5 was used (details of the method can be found in that chapter and appendix, Appendix E.1.6).

Back to main text

G.1.2 Supplementary analysis

In addition to examining the effect of the intervention on the anterior hippocampus, we performed a supplementary analysis to explore intervention effects on its subfields. Hippocampal subfields were segmented using Multiple Automatically Generated Templates (MAGeT) brain segmentation algorithm (Chakravarty *et al.,* 2013; Pipitone *et al.,* 2014, freely available via https://github.com/CobraLab/MAGeTbrain). MAGeTbrain has been used extensively to segment the entire hippocampus in various populations (e.g. Duerden *et al.,* 2020; Fish *et al.,* 2020), and increasingly to segment its subfields in adults (e.g. Patel *et al.,* 2017; Voineskos *et al.,* 2015) and pediatric samples (Decker *et al.,* 2017; Sussman *et al.,* 2015; Treadway *et al.,* 2015). The hippocampal subfield segmentations from MAGeT have been validated on 3T T_1-weighted images of an adolescent sample (Pipitone *et al.,* 2014).

MAGeT is a multi-atlas segmentation algorithm that, for the purpose of this study, uses five high resolution (0.3 mm isotropic) manually labelled hippocampal subfield atlases, created using 3T T_1w and T_2w images from five healthy volunteers (2 male, age range: 29–59, Winterburn *et al.,* 2013). Each atlas contains labels for the following subfields: Cornus Amonis (CA)1, CA2-CA3, CA4-dentate gyrus (DG), subiculum and stratum radiatum/stratum lacunosum/stratum moleculare. Whereas traditional multi-atlas segmentation algorithms propagate (using non-linear registration) the multiple atlas labels to individual subjects where they are combined into a single label using a label fusion procedure, MAGeT adds an additional step to this process that allows for fewer atlases to be used by leveraging the anatomical variability within the sample. In particular, MAGeT first propagates the atlas labels to a representative set of *template* images to create a *template library*. The template images are selected from within the sample and ideally reflect important characteristics of the sample as a whole (e.g. neuroanatomical variability, demographics, treatment groups, time). For purposes of this study, we performed random sampling stratified by sex, treatment group and timepoint using the *splitstackshape* package in R (Mahto, 2019), to select a total of 21 template images out of all available (n = 108) cohort 1 datasets (n

= 21 was demonstrated to be optimal by Pipitone *et al.*, 2014). Each template image is labelled by each of the atlases, yielding five candidate labels for each subfield per template (i.e. a total of 105 labels per subfield). The next steps of the algorithm are similar to traditional multi-atlas segmentation: the labels of each template image are propagated to each image in the dataset and combined into a single label using a majority vote procedure (Chakravarty *et al.*, 2013; Collins & Pruessner, 2010).

MAGeT was run on the supercomputer at the SciNet consortium (Loken *et al.*, 2010), using tools available in the MINC toolkit (https://bic-mni.github.io). Prior to running MAGeT, the T_1-weighted data was bias-field corrected and brain-extracted using FSL's FSL_ANAT pipeline, and converted to Medical Imaging NetCDF (MINC) format.

Back to main text

G.1.3 Software

Statistical analyses were performed in R (version 3.6) and Matlab. LMMs were fit using the *lme4* package (Bates2015). Missing data was imputed using *pan* available in *mitml* (Grund *et al.*, 2016).

Back to main text

G.2 Results

G.2.1 Demographics

Baseline school level demographics are displayed in Table G.3.

Table G.3: School-level baseline characteristics brain imaging sample by treatment group.

	Intervention	Control
No. schools	5	5
Gender type		
Co-ed	5 (100%)	4 (80.0%)
Female	0 (0%)	1 (20.0%)
Male	0 (0%)	0 (0%)
Type, no. (%)		
Comprehensive to 16	0 (0%)	0 (0%)
Comprehensive to 18	4 (80.0%)	3 (60.0%)
Secondary Modern	1 (20.0%)	2 (40.0%)
School establishment		
Academy Converter	1 (20.0%)	0 (0%)
Academy Convertor - Mainstream	2 (40.0%)	3 (60.0%)
Community School	2 (40.0%)	0 (0%)
Foundation School	0 (0%)	1 (20.0%)
Free School - Mainstream	0 (0%)	1 (20.0%)
Size, no. (%)		
Large (>200)	1 (20.0%)	0 (0%)
Medium (100-200)	4 (80.0%)	4 (80.0%)
Small (<100)	0 (0%)	1 (20.0%)
No. of form groups, median (range)	6 (6-11)	5 (4-8)
Socioeconomic status		
Percent eFSM[a], mean (SD)	8.62 (4.31)	8.07 (9.16)
IMD[b], median (range), decile	9 (7-10)	8 (3-10)
Location[c], no. (%)		
Rural town and fringe	3 (60.0%)	2 (40.0%)
Urban city and town	2 (40.0%)	3 (60.0%)
Urban major conurbation	0 (0%)	0 (0%)

Abbreviations: Co-ed = co-education, eFSM = eligible for free school meals, no. = number, SD = standard deviation

[a] Percent eFSM indicates the percentage of pupils within a school that were enrolled in Fit to Study who are eligible for free school meals

[b] Index of multiple deprivation was collected using the school's postcode, through Government records: http://imd-by-postcode.opendatacommunities.org/imd/2019

[c] Location was assessed using https://www.gov.uk/government/collections/rural-urban-classification

Back to main text

G.2.2 Fidelity

Table G.4 presents school level and pupil level fidelity data. No significant differences were observed in any of the fidelity metrics between intervention and control schools.

Table G.4: School level and pupil level fidelity data

	Intervention	Control	*p*
School level			
No. schools	5	5	
Average VPA during PE[a], min			
Mean (SD)	6.92 (2.33)	4.70 (1.80)	0.17
Missing, no. (%)	1 (20.0%)	0 (0%)	
Teacher-reported adherence			
Mean (SD)	4.33 (0.577)	—	
Median (range)	4.00 (4, 5)	—	
Missing, no. (%)	2 (40.0%)	5 (100%)	
Pupil level			
No. pupils	32	29	
Warm-up[b]			
Mean (SD)	4.29 (0.920)	4.38 (0.768)	0.78
Median (range)	5.00 (2, 5)	5.00 (3, 5)	
Missing, no. (%)	15 (46.9%)	16 (55.2%)	
Infusions during PE[b]			
Mean (SD)	3.76 (1.09)	3.69 (1.11)	0.86
Median (range)	4.00 (2, 5)	4.00 (1, 5)	
Missing, no. (%)	15 (46.9%)	16 (55.2%)	
Participation[b]			
Mean (SD)	4.59 (0.870)	4.38 (0.768)	0.51
Median (range)	5.00 (2, 5)	5.00 (3, 5)	
Missing, no. (%)	15 (46.9%)	16 (55.2%)	
Average VPA school time, min			
Mean (SD)	2.67 (2.06)	2.46 (1.77)	0.63
Missing, no. (%)	6 (18.8%)	4 (13.8%)	

Abbreviations: PA= physical activity, PE = physical education, SD = standard deviation, VPA = vigorous physical activity

[a] Active lesson PA (excluding change time), standardized to minutes per hour; data collected during a single PE lesson towards the end of the intervention period.

[b] Questionnaire-based items, see section 7.2.4 of the main text for details

Back to main text

G.2.3 Primary analysis

G.2.3.1 Multivariate analysis

Average values of MRI markers (multiply imputed) by treatment group are presented in Table G.5.

Table G.5: Means of multiple imputed MRI markers by treatment groups

	Intervention		Control	
	Baseline M (SD)	**Post** M (SD)	**Baseline** M (SD)	**Post** M (SD)
MD				
left	0.0007 (1.3e-05)	0.000696 (1.3e-05)	0.000705 (1.6e-05)	0.000698 (1.3e-05)
right	0.000706 (1.3e-05)	0.000705 (1.2e-05)	0.000708 (1.5e-05)	0.000702 (1.8e-05)
fintra				
left	0.382 (0.013)	0.384 (0.014)	0.376 (0.016)	0.381 (0.015)
right	0.371 (0.013)	0.372 (0.012)	0.369 (0.014)	0.375 (0.017)
R_1				
left	630.87 (21.10)	640.94 (22.9)	632.14 (15.25)	628.79 (24.27)
right	645.47 (20.12)	654.01 (22.12)	641.50 (15.05)	637.85 (23.04)
R_2^*				
left	0.015 (0.0016)	0.014 (0.0022)	0.015 (0.0011)	0.014 (0.0023)
right	0.015 (0.0017)	0.014 (0.0022)	0.015 (0.0015)	0.014 (0.0023)
Perfusion				
left	24.89 (6.72)	25.78 (5.65)	25.39 (5.31)	25.01 (4.32)
right	25.36 (6.15)	25.69 (5.57)	25.03 (4.90)	24.57 (4.72)
VBM				
left	0.83 (0.07)	0.84 (0.07)	0.85 (0.09)	0.84 (0.08)
right	0.82 (0.06)	0.81 (0.07)	0.82 (0.08)	0.81 (0.08)

Abbreviations: M = mean, MD = mean diffusivity, SD = standard deviation, VBM = voxel-based morphometry

The number and percent of significant partial tests by MRI marker is displayed in Table G.6. The distribution of p values by MRI markers are shown in Figure G.1. Both indicate that the intervention had a significant effect on left and right R_1 across a large percentage (respectively, 80% and 90%) of partial tests.

Table G.6: Number and percent of significant partial tests by MRI marker

Modality	**N**	**% of imputations**
VBM left	1	2
VBM right	0	0
R_1 left	40	80
R_1 right	45	90
R_2^* left	1	2
R_2^* right	0	0
fintra left	0	0
fintra right	21	42
MD left	0	0
MD right	10	20
perfusion left	0	0
perfusion right	0	0

Abbreviations: M = mean, MD = mean diffusivity, SD = standard deviation, VBM = voxel-based morphometry

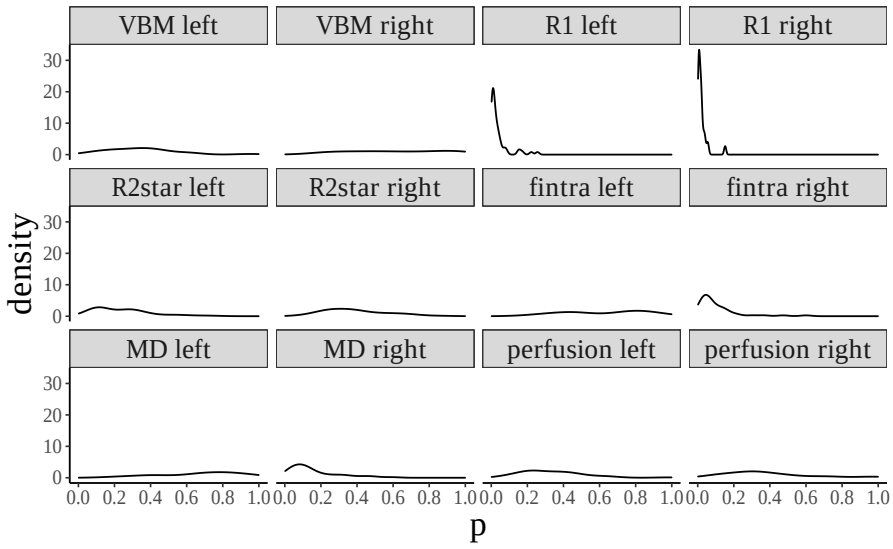


Figure G.1: Distribution of uncorrected p-values (frequency) of partial tests by MRI marker

Exploration of significant effect

To examine the relevance of the effect of the intervention on R_1, we explored whether changes in this metric were associated with changes in CRF or relational memory performance from pre to post intervention in each treatment group and whether these associations were significantly different between treatment groups (adjusted for, respectively, baseline $\dot{\mathbf{V}}O_2$peak or R_1 scores, baseline age, and pre-to-post changes in age). Table G.7 displays the associations between changes in $\dot{\mathbf{V}}O_2$peak and changes in R_1 (left and right) for each treatment group, as well as whether the correlation values were significantly different between treatment groups.

Table G.7: Associations between changes in $\dot{\mathbf{V}}O_2$peak and changes in R_1 by treatment group

	Control	Intervention	Difference in r values
	r	r	p
Change in $\dot{\mathbf{V}}O_2$peak with left R_1	-0.21	-0.05	0.54
Change in $\dot{\mathbf{V}}O_2$peak with right R_1	-0.18	-0.06	0.67

Abbreviations: M = mean, SD = standard deviation

Table G.8: Associations between changes in R_1 and changes in relational memory performance by treatment group

	Object location task			Relational memory paradigm		
	Control	**Intervention**	**Difference in r values**	**Control**	**Intervention**	**Difference in r values**
	r	r	p	r	r	p
Change in left R1 with memory performance	-0.08	-0.1	0.92	0.003	-0.32	0.72
Change in right R1 with memory performance	-0.08	-0.02	0.81	-0.28	-0.18	0.7

Abbreviations: M = mean, SD = standard deviation

Back to main text

G.2.4 Sensitivity analyses

G.2.4.1 Primary univariate analysis

We repeated the primary univariate analyses using complete (available) cases only. The results of these analyses are in line with those of the primary analyses on multiply imputed data and are presented in Table G.9.

Table G.9: The effect of the intervention on CRF, anterior hippocampal volume and relational memory

	Intervention			Control			Adjusted mean difference[a]		P
	N	**Baseline** M (SD)	**Posttest** M (SD)	**N**	**Baseline** M (SD)	**Posttest** M (SD)	**Unstandardized**	**Standardized[b]**	
Cardiorespiratory fitness									
$\dot{V}O_2$peak[c], ml/kg/min	25	36.0 (7.88)	37.0 (6.96)	24	39.5 (8.98)	40.8 (11.8)	-2.16 (-7.02, 2.71)	-0.22 (-0.72, 0.28)	0.44
Anterior hippocampal volume									
Left, mm^3	23	1430 (268)	1450 (266)	25	1400 (334)	1440 (342)	-27.31 (-64.7, 10.09)	-0.09 (-0.21, 0.03)	0.15
Right, mm^3	23	1550 (246)	1570 (252)	25	1510 (327)	1530 (320)	-1.76 (-38.1, 34.57)	-0.01 (-0.31, 0.12)	0.92
Cognitive performance									
OLT[c], % of items (errors)	25	13.4 (8.22)	12.8 (8.61)	25	13.1 (7.4)	10.8 (7.06)	2.9 (-4.00, 9.61)	0.36 (-0.51, 1.22)	0.48
Relational memory task, accuracy	23	71.9 (12.0)	71.7 (12.2)	21	71.1 (11.8)	70.8 (13.6)	0.63 (-6.81, 8.08)	0.05 (-0.55, 0.65)	0.86

Abbreviations: M = mean, OLT = object-location task, SD = standard deviation
[a] Mean difference adjusted for baseline values of the outcome
[b] Standardized by the variance of the outcome
[c] Fit using a multilevel model

In Figure G.2, Figure G.3 and Figure G.4 the distributions in, respectively, change (post - pre) in CRF, left and right anterior hippocampal volumes and memory performance are visualised (left column), alongside the predicted outcomes (i.e. marginal means, at post-test; right column), using the primary intention-to-treat model. The predicted (marginal) effects are displayed with 95% confidence intervals, based on the standard error, and created with the *ggeffects* package in R (Lüdecke, 2018). All plots are based on available (complete-cases) only.

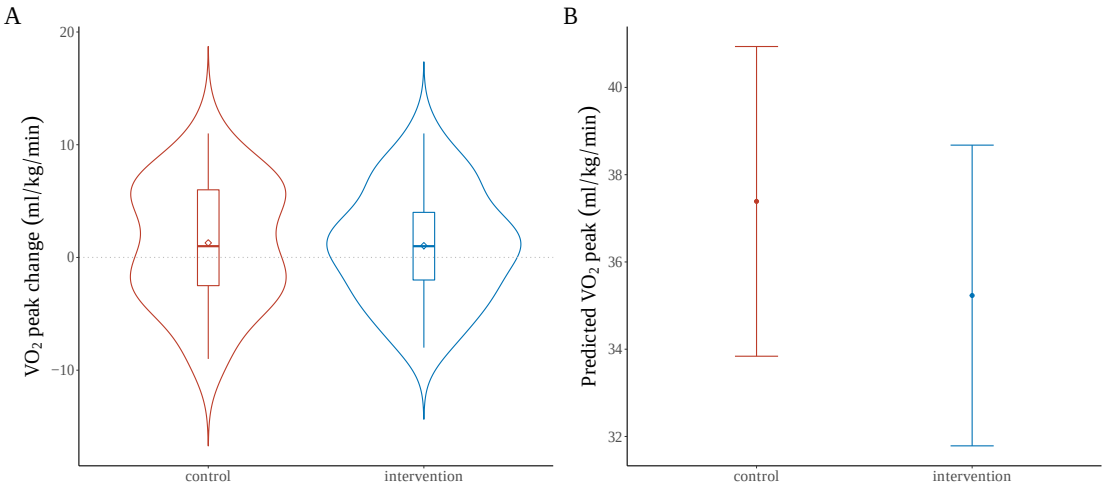


Figure G.2: Effect of the intervention on cardiorespiratory fitness (complete-cases)

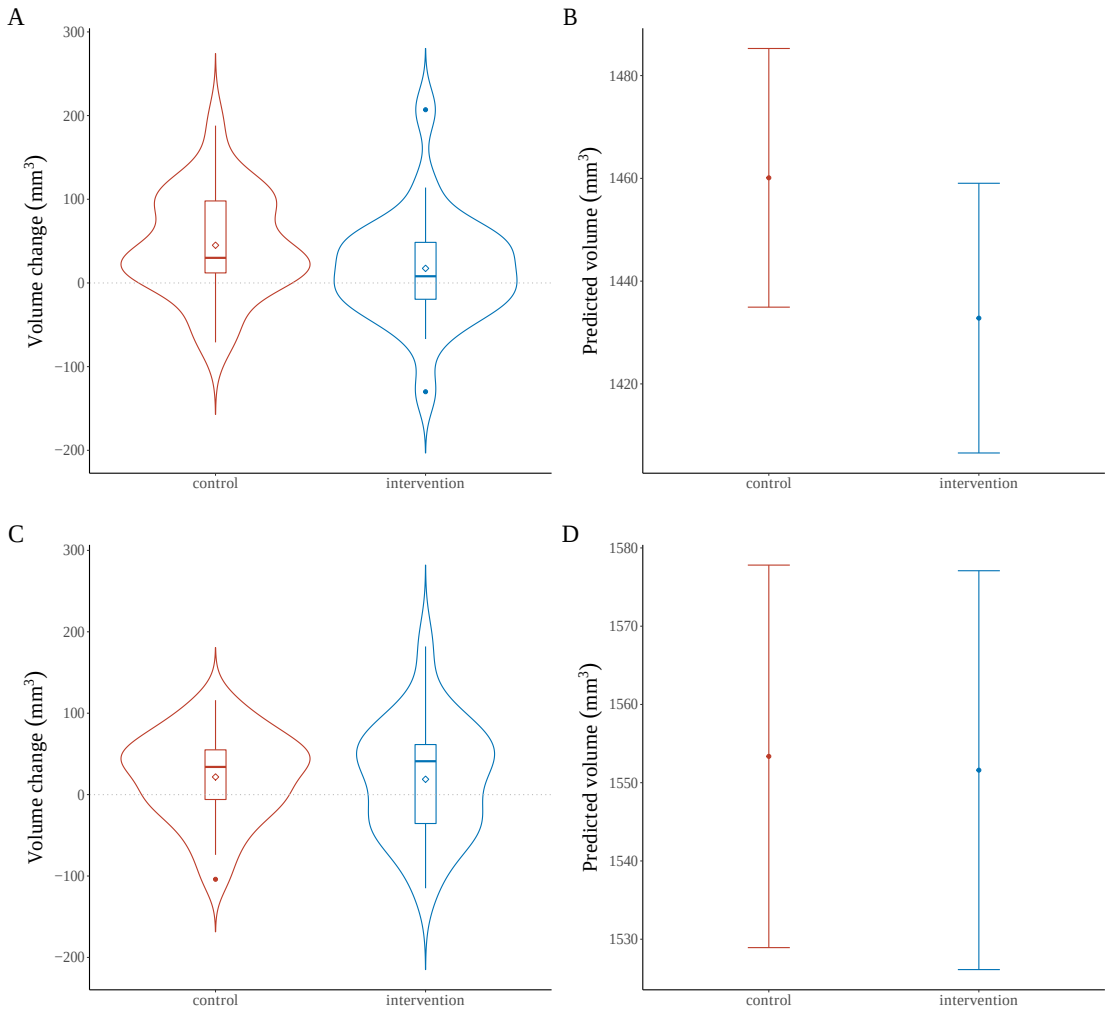


Figure G.3: Effect of the intervention on left and right anterior hippocampal volume (complete-cases). (*a*) and (*b*) display the distributions and predicted outcomes for left anterior hippocampal volume, whereas (*c*) and (*d*) display distributions and predicted outcomes for right anterior hippocampal volume.

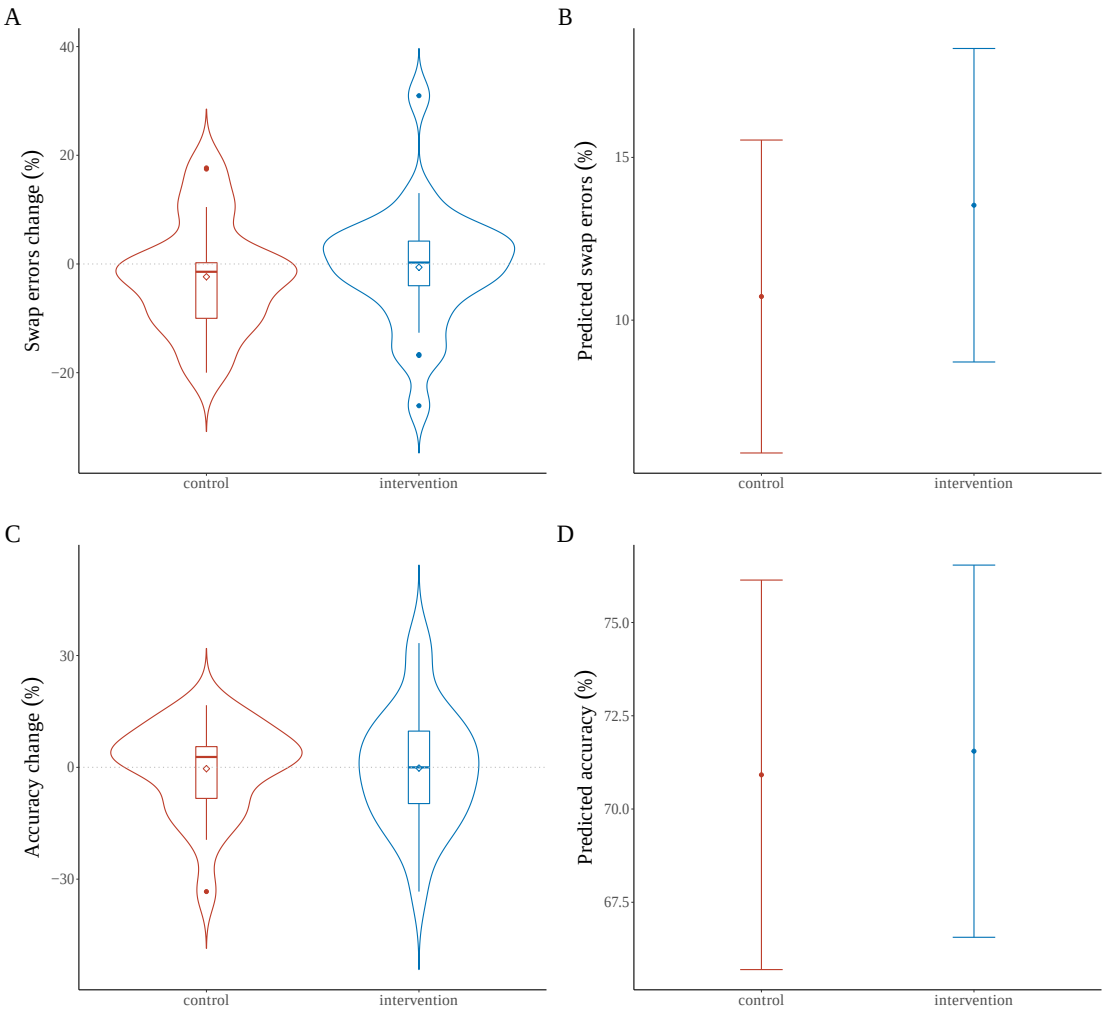


Figure G.4: Effect of the intervention on relational memory performance (complete-cases). (*a*) and (*b*) display the distributions and predicted outcomes for the object location task, whereas (*c*) and (*d*) display distributions and predicted outcomes for the relational memory paradigm.

Back to main text

G.2.4.2 Multivariate analysis

We repeated the primary multivariate analysis, testing for a joint effect of the intervention on MRI markers in the anterior hippocampus, using SVD imputed data and complete-cases.

SVD imputed data

In Figure G.5 the distributions of difference scores of MRI markers in the anterior hippocampus are shown by treatment group and hemisphere. The average values per timepoint are displayed in Table G.10. There was no significant joint effect of the intervention on anterior hippocampal MRI markers (NPC FWE-adjusted $p = 0.11$).

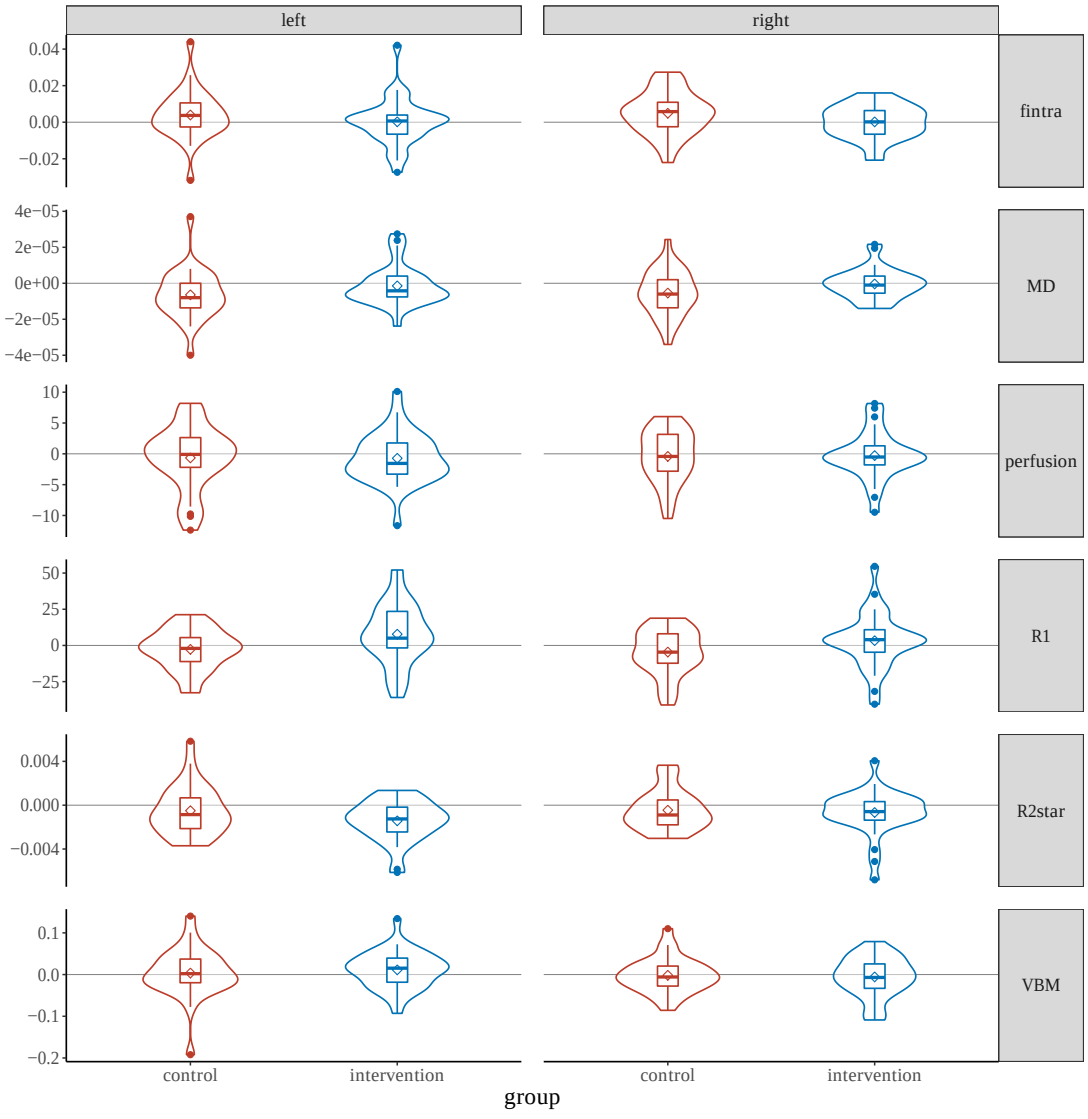


Figure G.5: Distributions of difference scores (post-pre) for SVD imputed MRI markers in the anterior hippocampus, by treatment group and hemisphere.

Table G.10: Means of single imputed MRI markers by treatment groups

	Intervention (N = 32)		Control (N = 29)	
	Baseline M (SD)	**Post** M (SD)	**Baseline** M (SD)	**Post** M (SD)
MD				
left	0.000698 (0.0000113)	0.000697 (0.0000110)	0.000704 (0.0000156)	0.000698 (0.0000113)
right	0.000705 (0.0000109)	0.000704 (0.00000982)	0.000708 (0.0000143)	0.000703 (0.0000150)
fintra				
left	0.383 (0.0120)	0.383 (0.0115)	0.376 (0.0154)	0.380 (0.0138)
right	0.372 (0.0115)	0.372 (0.00945)	0.369 (0.0137)	0.374 (0.0142)
R_1				
left	633 (18.9)	641 (17.4)	633 (14.1)	630 (15.3)
right	645 (18.3)	648 (14.4)	642 (14.4)	637 (15.5)
R_2^*				
left	0.0150 (0.00148)	0.0136 (0.00128)	0.0146 (0.00104)	0.0141 (0.00171)
right	0.0145 (0.00144)	0.0138 (0.00150)	0.0145 (0.00138)	0.0141 (0.00138)
Perfusion				
left	25.8 (4.95)	25.0 (3.80)	25.5 (5.13)	24.9 (3.69)
right	25.1 (4.24)	24.8 (3.42)	24.9 (4.62)	24.5 (3.96)
VBM				
left	0.831 (0.0646)	0.843 (0.0573)	0.846 (0.0888)	0.850 (0.0778)
right	0.819 (0.0561)	0.813 (0.0606)	0.820 (0.0774)	0.818 (0.0737)

Abbreviations: M = mean, SD = standard deviation

Complete-cases

In Figure G.6 the distributions of difference scores of MRI markers in the anterior hippocampus are shown by treatment group and hemisphere (N = 38, of which 17 intervention). The average values per timepoint are displayed in Table G.11. There was no significant joint effect of the intervention on anterior hippocampal MRI markers (NPC FWE-adjusted p = 0.32).

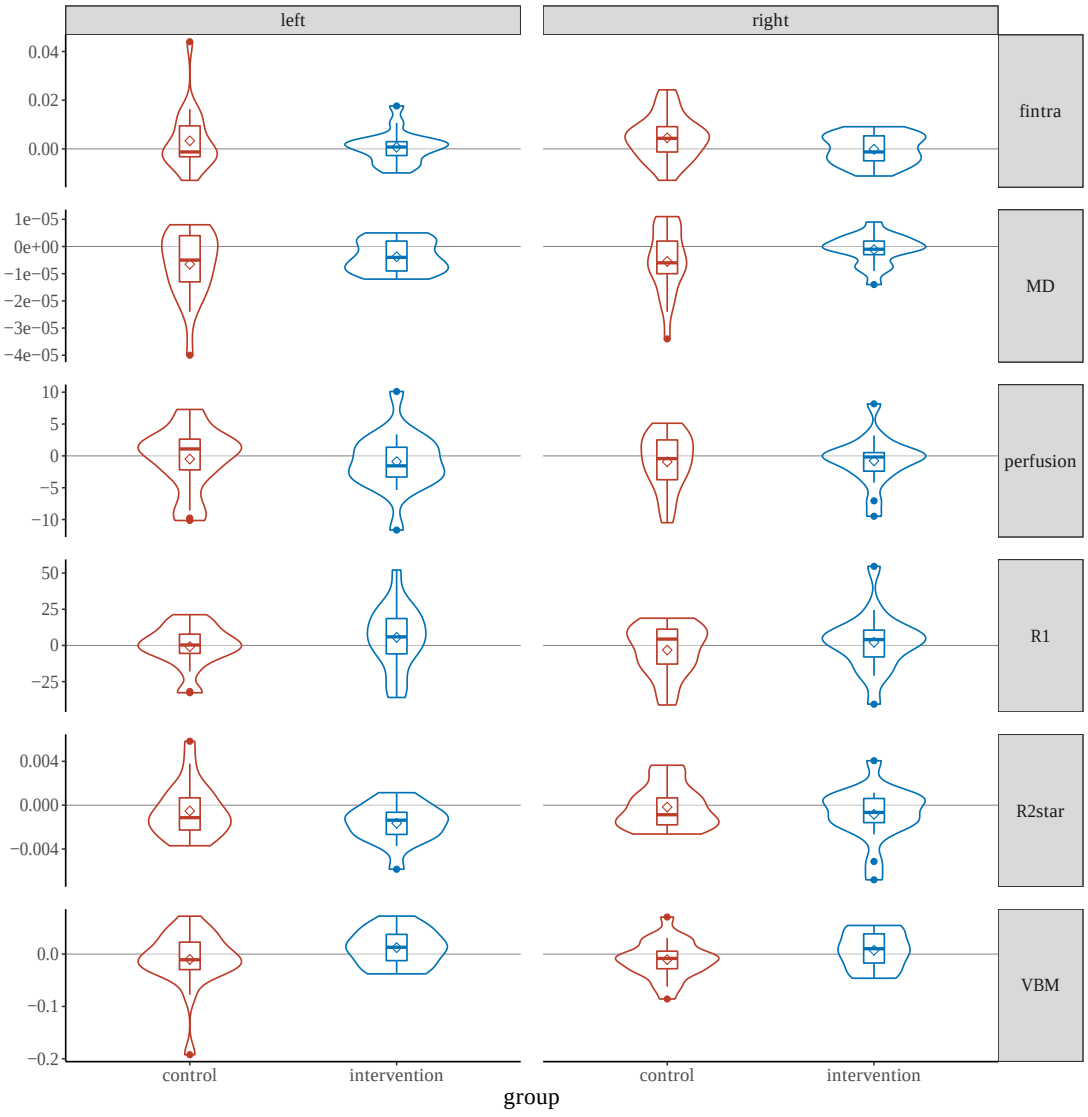


Figure G.6: Distributions of difference scores (post-pre) for MRI markers in the anterior hippocampus, by treatment group and hemisphere (complete-cases).

Table G.11: Means of MRI markers by treatment groups (complete-cases)

	Intervention (N = 17)		Control (N = 21)	
	Baseline M (SD)	**Post** M (SD)	**Baseline** M (SD)	**Post** M (SD)
MD				
left	0.000698 (0.0000113)	0.000695 (0.0000114)	0.000707 (0.0000142)	0.000701 (0.0000114)
right	0.000704 (0.0000115)	0.000703 (0.0000118)	0.000710 (0.0000134)	0.000705 (0.0000151)
fintra				
left	0.383 (0.0125)	0.384 (0.0110)	0.374 (0.0146)	0.377 (0.0134)
right	0.373 (0.0113)	0.373 (0.0108)	0.367 (0.0125)	0.372 (0.0137)
R_1				
left	635 (19.0)	641 (23.0)	630 (13.7)	629 (17.7)
right	649 (19.4)	651 (16.5)	640 (15.0)	637 (18.1)
R_2^*				
left	0.0151 (0.00118)	0.0135 (0.00169)	0.0147 (0.000997)	0.0141 (0.00198)
right	0.0147 (0.00154)	0.0139 (0.00205)	0.0143 (0.00140)	0.0141 (0.00160)
Perfusion				
left	25.1 (6.27)	24.2 (4.39)	25.3 (5.16)	24.9 (4.26)
right	24.9 (4.90)	24.1 (4.08)	25.0 (4.95)	24.0 (4.33)
VBM				
left	0.843 (0.0647)	0.855 (0.0629)	0.871 (0.0898)	0.861 (0.0836)
right	0.819 (0.0586)	0.826 (0.0692)	0.839 (0.0810)	0.829 (0.0814)

Abbreviations: M = mean, SD = standard deviation

Back to main text

G.2.5 Supplementary analysis

G.2.5.1 Hippocampal volume analysis

We assessed the effect of the intervention on the volume of the entire hippocampus in participants with complete datasets. Table G.12 presents average left and right hippocampal volumes per treatment group, as well as the (un)standardized mean difference scores. No significant effect on the left or right hippocampal volume was observed (p >0.05).

Table G.12: The effect of the intervention on hippocampal volumes

	Intervention (n = 23)		Control (n = 25)		Adjusted mean difference[a]		p
	Baseline M (SD)	**Posttest** M (SD)	**Baseline** M (SD)	**Posttest** M (SD)	**Unstandardized**	**Standardized[b]**	
Hippocampal volume							
Left[c], mm^3	2680 (347)	2730 (315)	2750 (380)	2780 (396)	-15.61 (-69.44, 38.23)	-0.04 (-0.2, 0.11)	0.56
Right, mm^3	2730 (339)	2790 (336)	2790 (378)	2800 (368)	-13.4 (-77.72, 50.96)	-0.04 (-0.22, 0.15)	0.71

Abbreviations: M = mean, SD = standard deviation
[a] Mean difference adjusted for baseline values of the outcome
[b] Scaled by the variance of the outcome
[c] Fit using a single level model

Back to main text

G.2.5.2 Hippocampal subfield volume analysis

The MAGeT segmentations were inspected for gross segmentation errors. A total of three subjects were excluded based on poor segmentation quality. The remaining subjects (n = 45) were included in the analysis. Initially, multilevel models were fit for each subfield. However, singular fits suggested that the random effects structure was too complex and single level linear models were used instead. Table G.13 displays the results of these models. No significant effects of the intervention on hippocampal subfields were observed following Bonferroni correction (10 tests, adjusted α = 0.005). Unadjusted p-values indicated that the right CA1 was larger and the right CA2-CA3 was smaller in the intervention group compared to control at post-test.

Table G.13: The effect of the intervention on volumes of hippocampal subfields

	Intervention (n = 23)		Control (n = 22)		Adjusted mean difference[a]		p^b
	Baseline M (SD)	**Posttest** M (SD)	**Baseline** M (SD)	**Posttest** M (SD)	**Unstandardized**	**Standardized[c]**	
CA1							
Right	664 (87.0)	673 (85.3)	661 (62.7)	659 (71.2)	10.78 (0.08 , 21.49)	0.14 (0, 0.27)	**0.05**
Left	640 (87.2)	650 (87.9)	650 (76.3)	654 (75.2)	5.26 (-3.6 , 14.11)	0.07 (-0.04, 0.18)	0.24
Subiculum							
Right	297 (42.3)	295 (43.7)	283 (30.2)	279 (30.9)	2.05 (-7.29 , 11.39)	0.05 (-0.19, 0.3)	0.66
Left	326 (47.6)	326 (47.9)	315 (46.0)	311 (42.8)	4.47 (-3.25 , 12.19)	0.1 (-0.07, 0.27)	0.25
CA4-DG							
Right	487 (58.4)	490 (59.6)	487 (57.7)	493 (62.4)	-1.49 (-12.97 , 9.99)	-0.02 (-0.22, 0.17)	0.79
Left	494 (61.3)	497 (58.3)	496 (51.8)	499 (59.9)	0.33 (-8.15 , 8.8)	0.01 (-0.14, 0.15)	0.94
CA2-CA3							
Right	134 (16.3)	136 (16.9)	133 (18.6)	142 (21.1)	-6.75 (-12.54 , -0.96)	-0.36 (-0.66, -0.05)	**0.02**
Left	129 (20.2)	131 (20.9)	128 (16.7)	132 (16.4)	-0.86 (-5.44 , 3.72)	-0.05 (-0.29, 0.2)	0.71
Stratum							
Right	439 (57.7)	447 (61.4)	441 (49.2)	445 (47.7)	3.34 (-6.37 , 13.04)	0.06 (-0.12, 0.24)	0.49
Left	469 (66.7)	472 (64.7)	475 (51.1)	480 (48.1)	-3.22 (-13.53 , 7.1)	-0.06 (-0.24, 0.12)	0.53

Abbreviations: CA = cornu Ammonis, DG = dentate gyrus, M = mean, N = number of participants, SD = standard deviation
[a] Mean difference adjusted for baseline values of the outcome
[b] Uncorrected p-values. P-values in bold are significant at the alpha = 0.05 level (uncorrected)
[c] Standardized mean difference, scaled by the variance of the outcome

Back to main text

This page is intentionally left blank.

References

2018 Physical Activity Guidelines Advisory Committee. *2018 Physical Activity Guidelines Advisory Committee Scientific Report* Washington, DC, 2018. https://health.gov/paguidelines/second-edition/report/pdf/PAG%7B%5C_%7DAdvisory%7B%5C_%7DCommittee%7B%5C_%7DReport.pdf.

Aadland, E., Kvalheim, O. M., Anderssen, S. A., Resaland, G. K. & Andersen, L. B. The multivariate physical activity signature associated with metabolic health in children. *International Journal of Behavioral Nutrition & Physical Activity* **15,** 77 (2018).

Aadland, K. N., Ommundsen, Y., Aadland, E., Brønnick, K. S., Lervåg, A., Resaland, G. K. & Moe, V. F. Executive functions do not mediate prospective relations between indices of physical activity and academic performance: The Active Smarter Kids (ASK) study. *Frontiers in Psychology* **8,** 1088 (2017).

Aadland, K. N., Moe, V. F., Aadland, E., Anderssen, S. A., Resaland, G. K. & Ommundsen, Y. Relationships between physical activity, sedentary time, aerobic fitness, motor skills and executive function and academic performance in children. *Mental Health and Physical Activity* **12,** 10–18 (2017).

Aadland, K. N., Ommundsen, Y., Anderssen, S. A., Brønnick, K. S., Moe, V. F., Resaland, G. K., Skrede, T., Stavnsbo, M. & Aadland, E. Effects of the Active Smarter Kids (ASK) Physical Activity School-based Intervention on Executive Functions: A Cluster-Randomized Controlled Trial. *Scandinavian Journal of Educational Research* **63,** 214–228 (2019).

Åberg, M. A. I., Pedersen, N. L., Toren, K., Svartengren, M., Backstrand, B., Johnsson, T., Cooper-Kuhn, C. M., Åberg, N. D., Nilsson, M., Kuhn, H. G., Åberg, N. D., Nilsson, M. & Kuhn, H. G. Cardiovascular fitness is associated with cognition in young adulthood. *Proceedings of the National Academy of Sciences* **106,** 20906–20911 (2009).

Aggio, D., Smith, L., Fisher, A. & Hamer, M. Context-specific associations of physical activity and sedentary behavior with cognition in children. *American Journal of Epidemiology* **183,** 1075–1082 (2016).

Ajzen, I. *From intentions to actions: A theory of planned behavior* (Springer, 1985).

Ajzen, I. The theory of planned behavior. *Organizational behavior and human decision processes* **50** (1991).

Akshoomoff, N., Newman, E., Thompson, W. K., McCabe, C., Bloss, C. S., Chang, L., Amaral, D. G., David, G., Casey, B. J., Ernst, T. M., Frazier, J. A., Gruen, J. R., Kaufmann, W. E., Kenet, T., Kennedy, D. N., Libiger, O., Mostofsky, S., Murray, S. S., Sowell, E. R., Schork, N., Dale, A. M. & Jernigan, T. L. The NIH toolbox cognition battery: Results from a large normative developmental sample (PING). *Neuropsychology* **28,** 1–10 (2014).

Alderson, P. Absence of Evidence Is Not Evidence of Absence. *BMJ* **328,** 477–9 (2004).

Alexander-Bloch, A., Clasen, L., Stockman, M., Ronan, L., Lalonde, F., Giedd, J. & Raznahan, A. Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI. *Human Brain Mapping* **37,** 2385–2397 (2016).

Alexander, A. L., Hurley, S. A., Samsonov, A. A., Adluru, N., Hosseinbor, A. P., Mossahebi, P., Tromp, d. P., Zakszewski, E. & Field, A. S. Characterization of Cerebral White Matter Properties Using Quantitative Magnetic Resonance Imaging Stains. *Brain Connectivity* **1,** 423–446 (2011).

Alexander, D. C., Dyrby, T. B., Nilsson, M. & Zhang, H. Imaging brain microstructure with diffusion MRI: Practicality and applications. *NMR in Biomedicine,* e3841 (2017).

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., Miller, K. L. & Smith, S. M. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage* **166,** 400–424 (2018).

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M., Miller, K. L., Nichols, T. E. & Smith, S. M. Confound modelling in UK Biobank brain imaging. *NeuroImage,* 117002 (2020).

Ali, J., Andrews, J. E., Somkin, C. P. & Rabinovich, C. E. Harms, benefits, and the nature of interventions in pragmatic clinical trials. *Clinical Trials* **12,** 467–475 (2015).

Allen, C. On (not) defining cognition. *Synthese* **194,** 4233–4249 (2017).

Almeida, J. R., Greenberg, T., Lu, H., Chase, H. W., Fournier, J., Cooper, C. M., Deckersbach, T., Adams, P., Carmody, T., Fava, M., Kurian, B., McGrath, P. J., McInnis, M. G., Oquendo, M. A., Parsey, R., Weissman, M., Trivedi, M. & Phillips, M. L. Test-retest reliability of cerebral blood flow in healthy individuals using arterial spin labeling: Findings from the EMBARC study. *Magnetic Resonance Imaging* **45,** 26–33 (2018).

Alnæs, D., Kaufmann, T., Marquand, A. F., Smith, S. M. & Westlye, L. T. Patterns of sociocognitive stratification and perinatal risk in the child brain. *Proceedings of the National Academy of Sciences* **117,** 12419–12427 (2020).

Althoff, T., Sosič, R., Hicks, J. L., King, A. C., Delp, S. L. & Leskovec, J. Large-scale physical activity data reveal worldwide activity inequality. *Nature* **547,** 336–339 (2017).

Álvarez-Bueno, C., Hillman, C. H., Cavero-Redondo, I., Sánchez-López, M., Pozuelo-Carrascosa, D. P. & Martínez-Vizcaíno, V. Aerobic fitness and academic achievement: A systematic review and meta-analysis. *Journal of Sports Sciences* **38,** 582–589 (2020).

Álvarez-Bueno, C., Pesce, C., Cavero-Redondo, I. I. I., Sánchez-López, M., Garrido-Miguel, M. & Martínez-Vizcaíno, V. Academic Achievement and Physical Activity: A Meta-analysis. *Pediatrics* **140,** e20171498 (2017).

Álvarez-Bueno, C., Pesce, C., Cavero-Redondo, I., Sánchez-López, M., Martínez-Hortelano, J. A. & Martínez-Vizcaíno, V. The Effect of Physical Activity Interventions on Children's Cognition and Metacognition: A Systematic Review and Meta-Analysis. *Journal of the American Academy of Child & Adolescent Psychiatry* **56,** 729–738 (2017).

Anderson, B. J. Plasticity of gray matter volume: The cellular and synaptic plasticity that underlies volumetric change. *Developmental Psychobiology* **53,** 456–465 (2011).

Anderson, P. Assessment and development of executive function (EF) during childhood. *Child Neuropsychology* **8,** 71–82 (2002).

Anderson, P., Anderson, V. & Lajoie, G. The tower of London test: Validation and standardization for pediatric populatons. *The Clinical Neuropsychologist* **10,** 54–65 (1996).

Andersson, J. L. R. & Sotiropoulos, S. N. Non-parametric representation and prediction of single-and multi-shell diffusion-weighted MRI data using Gaussian processes. *NeuroImage* **122,** 166–176 (2015).

Andersson, J. L. & Sotiropoulos, S. N. An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging. *NeuroImage* **125,** 1063–1078 (2016).

Andridge, R. R. Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biometrical Journal* **53,** 57–74 (2011).

Anwyl-Irvine, A. L., Dalmaijer, E., Hodges, N. & Evershed, J. Online Timing Accuracy and Precision: A comparison of platforms, browsers, and participant's devices. *PsyArXiv,* doi:10.31234/osf.io/jfeca (2020).

Armstrong, N., Welsman, J. & Winsley, R. Is peak VO2 a maximal index of children's aerobic fitness? *International Journal of Sports Medicine* **17,** 356–359 (1996).

Armstrong, N., Tomkinson, G. R. & Ekelund, U. Aerobic fitness and its relationship to sport, exercise training and habitual physical activity during Youth. *British Journal of Sports Medicine* **45,** 849–858 (2011).

Armstrong, N. & Welsman, J. Clarity and confusion in the development of youth aerobic fitness. *Frontiers in Physiology* **10** (2019).

Armstrong, N. & Welsman, J. Development of peak oxygen uptake from 11–16 years determined using both treadmill and cycle ergometry. *European Journal of Applied Physiology* **119,** 801–812 (2019).

Armstrong, N. & Welsman, J. Twenty-metre shuttle run: (mis)representation, (mis)interpretation and (mis)use. *British Journal of Sports Medicine* **53,** 1199 (2019).

Armstrong, N. & Welsman, J. Youth cardiorespiratory fitness: Evidence, myths and misconceptions. *Bulletin of the World Health Organization* **97,** 777–782 (2019).

Armstrong, N. & Welsman, J. R. Assessment and interpretation of aerobic fitness in children and adolescents. *Exercise and Sport Sciences Reviews* **22,** 435–476 (1994).

Armstrong, N. & Welsman, J. R. The physical activity patterns of European youth with reference to methods of assessment. *Sports Medicine* **36,** 1067–1086 (2006).

Artero, E. G., Espaa-Romero, V., Castro-Piero, J., Ortega, F. B., Suni, J., Castillo-Garzon, M. J. & Ruiz, J. R. Reliability of field-based fitness tests in youth. *International Journal of Sports Medicine* **32,** 159–169 (2011).

Ashburner, J. & Friston, K. J. Voxel-based morphometry - The methods. *NeuroImage* **11,** 805–821 (2000).

Assaf, Y., Berg, H. J. & Schotten, M. T. D. The role of diffusion MRI in neuroscience. *NMR in Biomedicine* **32,** e3762 (2019).

Audigier, V., Husson, F. & Josse, J. MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing* **27,** 501–518 (2017).

Audigier, V., White, I. R., Jolani, S., Debray, T. P. A., Quartagno, M., Carpenter, J., Van Buuren, S. & Resche-rigon, M. Multiple Imputation for Multilevel Data with Continuous and Binary Variables. *Statistical Science* **33,** 160–183 (2018).

Austin, P. C. & Leckie, G. The effect of number of clusters and cluster size on statistical power and Type I error rates when testing random effects variance components in multilevel linear and logistic regression models. *Journal of Statistical Computation and Simulation* **88,** 3151–3163 (2018).

Baayen, R. H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59,** 390–412 (2008).

Baddeley, A. D. & Hitch, G. *Psychology of learning and motivation* 47–89 (Elsevier, 1974).

Baggetta, P. & Alexander, P. A. Conceptualization and Operationalization of Executive Function. *Mind, Brain, and Education* **10,** 10–33 (2016).

Bagnato, F., Hametner, S., Boyd, E., Endmayr, V., Shi, Y., Ikonomidou, V., Chen, G., Pawate, S., Lassmann, H., Smith, S. & Brian Welch, E. Untangling the R2* contrast in multiple sclerosis: A combined MRI-histology study at 7.0 Tesla. *PLoS ONE* **13,** e0193839 (2018).

Bailey, R. C., Olson, J., Pepper, S. L., Porszasz, J., Barstow, T. J. & Cooper, D. M. The level and tempo of children's physical activities: an observational study. *Medicine and Science in Sports and Exercise* **27,** 1033–1041 (1995).

Banda, J. A., Haydel, K. F., Davila, T., Desai, M., Bryson, S., Haskell, W. L., Matheson, D. & Robinson, T. N. Effects of varying epoch lengths, wear time algorithms, and activity cut-points on estimates of child sedentary behavior and physical activity from accelerometer data. *PLoS ONE* **11,** e0150534 (2016).

Barcikowski, R. S. & Stevens, J. P. A Monte Carlo Study of the Stability of Canonical Correlations, Canonical Weights and Canonical Variate-Variable Correlations. *Multivariate Behavioral Research* **10,** 353–364 (1975).

Barha, C. K., Davis, J. C., Falck, R. S., Nagamatsu, L. S. & Liu-Ambrose, T. Sex differences in exercise efficacy to improve cognition: A systematic review and meta-analysis of randomized controlled trials in older humans. *Frontiers in Neuroendocrinology* **46,** 71–85 (2017).

Barha, C. K., Galea, L. A., Nagamatsu, L. S., Erickson, K. I. & Liu-Ambrose, T. Personalising exercise recommendations for brain health: Considerations and future directions. *British Journal of Sports Medicine* **51,** 636–639 (2017).

Barker, A. R., Williams, C. A., Jones, A. M. & Armstrong, N. Establishing maximal oxygen uptake in young people during a ramp cycle test to exhaustion. *British Journal of Sports Medicine* **45,** 498–503 (2011).

Barkovich, M. J., Li, Y., Desikan, R. S., Barkovich, A. J. & Xu, D. Challenges in pediatric neuroimaging. *NeuroImage* **185,** 793–801 (2019).

Barnes, J., Bartlett, J. W., Van De Pol, L. A., Loy, C. T., Scahill, R. I., Frost, C., Thompson, P. & Fox, N. C. A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiology of Aging* **30,** 1711–1723 (2009).

Barnes, J., Ridgway, G. R., Bartlett, J., Henley, S. M., Lehmann, M., Hobbs, N., Clarkson, M. J., MacManus, D. G., Ourselin, S. & Fox, N. C. Head size, age and gender adjustment in MRI studies: A necessary nuisance? *NeuroImage* **53,** 1244–1255 (2010).

Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68,** 255–278 (2013).

Barreira, T. V., Schuna, J. M., Mire, E. F., Katzmarzyk, P. T., Chaput, J. P., Leduc, G. & Tudor-Locke, C. Identifying children's nocturnal sleep using 24-h waist accelerometry. *Medicine and Science in Sports and Exercise* **47,** 937–943 (2015).

Bartasevicius, V. & Jabin, N. *Statistical Analysis Plan for the efficacy trial of the Fit to Study intervention* 2018. https://educationendowmentfoundation.org.uk/public/files/Projects/Fit%7B%5C_%7Dto%7B%5C_%7DStudy%7B%5C_%7DSAP%7B%5C_%7D2018.03.15%7B%5C_%7DFINAL.pdf.

Basser, P. J., Mattiello, J. & Lebihan, D. Estimation of the Effective Self-Diffusion Tensor from the NMR Spin Echo. *Journal of Magnetic Resonance, Series B* **103,** 247–254 (1994).

Basser, P. J., Mattiello, J. & Le Bihan, D. MR diffusion tensor spectroscopy and imaging. *Biophysical Journal* **66,** 259–267 (1994).

Basso, J. C. & Suzuki, W. A. The Effects of Acute Exercise on Mood, Cognition, Neurophysiology, and Neurochemical Pathways: A Review. *Brain Plasticity* **2,** 127–152 (2017).

Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67** (2015).

Batista, M. B., Cyrino, E. S., Arruda, M., Dourado, A. C., Coelho-E-Silva, M. J., Ohara, D., Romanzini, M. & Ronque, E. R. Validity of equations for estimating VO2peak from the 20-m shuttle run test in adolescents aged 11-13 years. *Journal of Strength and Conditioning Research* **27,** 2774–2781 (2013).

Batouli, S. A. H. & Saba, V. At least eighty percent of brain grey matter is modifiable by physical activity: A review study. *Behavioural Brain Research* **332,** 204–217 (2017).

Bauer, U. E., Briss, P. A., Goodman, R. A. & Bowman, B. A. Prevention of chronic disease in the 21st century: Elimination of the leading preventable causes of premature death and disability in the USA. *The Lancet* **384,** 45–52 (2014).

Bauman, A. E., Reis, R. S., Sallis, J. F., Wells, J. C., Loos, R. J. & Martin, B. W. Correlates of physical activity: Why are some people physically active and others not? *The Lancet* **380,** 258–271 (2012).

Bauman, A. E., Sallis, J. F., Dzewaltowski, D. A. & Owen, N. Toward a better understanding of the influences on physical activity: The role of determinants, correlates, causal variables, mediators, moderators, and confounders. *American Journal of Preventive Medicine* **23,** 5–14 (2002).

Bayne, T., Brainard, D., Byrne, R. W., Chittka, L., Clayton, N., Heyes, C., Mather, J., Ölveczky, B., Shadlen, M., Suddendorf, T. & Webb, B. What is cognition? *Current Biology* **29,** R608–R615 (2019).

Beaulieu, C. *Diffusion MRI: from quantitative measurement to in-vivo neuroanatomy* (eds Johansen-Berg, H. & Behrens, T. E.) (Elsevier, 2009).

Becht, A. I. & Mills, K. L. Modeling Individual Differences in Brain Development. *Biological Psychiatry* **88,** 63–69 (2020).

Bellg, A. J., Resnick, B., Minicucci, D. S., Ogedegbe, G., Ernst, D., Borrelli, B., Hecht, J., Ory, M., Orwig, D. & Czajkowski, S. Enhancing treatment fidelity in health behavior change studies: Best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology* **23,** 443–451 (2004).

Beltran-Valls, M. R., Janssen, X., Farooq, A., Adamson, A. J., Pearce, M. S., Reilly, J. J. K., Basterfield, L. & Reilly, J. J. K. Longitudinal changes in vigorous intensity physical activity from childhood to adolescence: Gateshead Millennium Study. *Journal of Science and Medicine in Sport* **22,** 450–455 (2019).

Beltz, N. M., Gibson, A. L., Janot, J. M., Kravitz, L., Mermier, C. M. & Dalleck, L. C. Graded Exercise Testing Protocols for the Determination of VO2max: Historical Perspectives, Progress, and Future Considerations. *Journal of Sports Medicine* **2016** (2016).

Benda, N. M., Seeger, J. P., Stevens, G. G., Hijmans-Kersten, B. T., Van Dijk, A. P., Bellersen, L., Lamfers, E. J., Hopman, M. T. & Thijssen, D. H. Effects of high-intensity interval training versus continuous training on physical fitness, cardiovascular function and quality of life in heart failure patients. *PLoS ONE* **10,** e0141256 (2015).

Benoy, A., Dasgupta, A. & Sajikumar, S. Hippocampal area CA2: an emerging modulatory gateway in the hippocampal circuit. *Experimental Brain Research* **236,** 919–931 (2018).

Berry, E. S., Jezzard, P. & Okell, T. W. The advantages of radial trajectories for vessel-selective dynamic angiography with arterial spin labeling. *Magnetic Resonance Materials in Physics, Biology and Medicine* **32,** 643–653 (2019).

Berry, V., Axford, N., Blower, S., Taylor, R. S., Edwards, R. T., Tobin, K., Jones, C. & Bywater, T. The Effectiveness and Micro-costing Analysis of a Universal, School-Based, Social–Emotional Learning Programme in the UK: A Cluster-Randomised Controlled Trial. *School Mental Health* **8,** 238–256 (2016).

Best, J. R. & Miller, P. H. A Developmental Perspective on Executive Function. *Child Development* **81,** 1641–1660 (2010).

Best, J. R., Miller, P. H. & Jones, L. L. Executive Functions after Age 5: Changes and Correlates. *Developmental Review* **29,** 180–200 (2009).

Bezerra, T. A., Clark, C. C. T., De Souza Filho, A. N., De Souza Fortes, L., Silva Mota, J. A. P., Duncan, M. J. & De Lucena Martins, C. M. 24-hour movement behaviour and executive function in preschoolers: a compositional and isotemporal reallocation analysis. *European Journal of Sport Science,* 1–9 (2020).

Biddle, S. J., Atkin, A. J., Cavill, N. & Foster, C. Correlates of physical activity in youth: A review of quantitative systematic reviews. *International Review of Sport and Exercise Psychology* **4,** 25–49 (2011).

Biddle, S. J. & Batterham, A. M. High-intensity interval exercise training for public health: A big HIT or shall we HIT it on the head? *International Journal of Behavioral Nutrition and Physical Activity* **12,** 1–8 (2015).

Biddle, S. J., Ciaccioni, S., Thomas, G. & Vergeer, I. Physical activity and mental health in children and adolescents: An updated review of reviews and an analysis of causality. *Psychology of Sport and Exercise* **42,** 146–155 (2019).

Black, J. E., Isaacs, K. R., Anderson, B. J., Alcantara, A. A. & Greenough, W. T. Learning causes synaptogenesis, whereas motor activity causes angiogenesis, in cerebellar cortex of adult rats. *Proceedings of the National Academy of Sciences* **87,** 5568–5572 (1990).

Blair, G., Cooper, J., Coppock, A., Humphreys, M. & Sonnet, L. *estimatr: Fast Estimators for Design-Based Inference* (2020). https://cran.r-project.org/package=estimatr.

Blair, S. N. & Powell, K. E. The Evolution of the Physical Activity Field. *Journal of Physical Education, Recreation & Dance* **85,** 9–12 (2014).

Blakemore, S.-J. Imaging brain development: The adolescent brain. *NeuroImage* **61,** 397–406 (2012).

Blakemore, S.-J. Adolescence and mental health. *The Lancet* **393,** 2030–2031 (2019).

Blokland, G. A., De Zubicaray, G. I., McMahon, K. L. & Wright, M. J. Genetic and environmental influences on neuroimaging phenotypes: A meta-analytical perspective on twin imaging studies. *Twin Research and Human Genetics* **15,** 351–371 (2012).

Boccardi, M., Bocchetta, M., Morency, F. C., Collins, D. L., Nishikawa, M., Ganzola, R., Grothe, M. J., Wolf, D., Redolfi, A., Pievani, M., Antelmi, L., Fellgiebel, A., Matsuda, H., Teipel, S., Duchesne, S., Jack, C. R. & Frisoni, G. B. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimer's and Dementia* **11,** 175–183 (2015).

Boccardi, M., Ganzola, R., Bocchetta, M., Pievani, M., Redolfi, A., Bartzokis, G., Camicioli, R., Csernansky, J. G., De Leon, M. J., Detoledo-Morrell, L., Killiany, R. J., Lehéricy, S., Pantel, J., Pruessner, J. C., Soininen, H., Watson, C., Duchesne, S., Jack, C. R. & Frisoni, G. B. Survey of protocols for the manual segmentation of the hippocampus: Preparatory steps towards a joint EADC-ADNI harmonized protocol. *Advances in Alzheimer's Disease* **2,** 111–125 (2011).

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J. S. S. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24,** 127–135 (2009).

Bookstein, F. L. "Voxel-Based Morphometry" Should Not Be Used With Imperfectly Registered Images. *NeuroImage* **14,** 1454–1462 (2001).

Booth, J. N., Tomporowski, P. D., Boyle, J. M., Ness, A. R., Joinson, C., Leary, S. D. & Reilly, J. J. Associations between executive attention and objectively measured physical activity in adolescence: Findings from ALSPAC, a UK cohort. *Mental Health and Physical Activity* **6,** 212–219 (2013).

Booth, M. L., Okely, A. D., Chey, T. & Bauman, A. The reliability and validity of the physical activity questions in the WHO health behaviour in schoolchildren (HBSC) survey: A population study. *British Journal of Sports Medicine* **35,** 263–267 (2001).

Booth, M. & Okely, A. Promoting physical activity among children and adolescents: the strengths and limitations of school-based approaches. *Health Promotion Journal of Australia* **16,** 52–54 (2005).

Bort-Roig, J., Gilson, N. D., Puig-Ribera, A., Contreras, R. S. & Trost, S. G. Measuring and influencing physical activity with smartphone technology: A systematic review. *Sports Medicine* **44,** 671–686 (2014).

Bradley, S. M., Michos, E. D. & Miedema, M. D. Physical Activity, Fitness, and Cardiovascular Health. *JAMA Network Open* **2,** e198343 (2019).

Bramen, J. E., Hranilovich, J. A., Dahl, R. E., Forbes, E. E., Chen, J., Toga, A. W., Dinov, I. D., Worthman, C. M. & Sowell, E. R. Puberty influences medial temporal lobe and cortical gray matter maturation differently in boys than girls matched for sexual maturity. *Cerebral Cortex* **21,** 636–646 (2011).

Brandes, M., Zijlstra, W., Heikens, S., Van Lummel, R. & Rosenbaum, D. Accelerometry based assessment of gait parameters in children. *Gait and Posture* **24,** 482–486 (2006).

Breitenstein, S. M., Gross, D., Garvey, C. A., Hill, C., Fogg, L. & Resnick, B. Implementation fidelity in community-based interventions. *Research in Nursing and Health* **33,** 164–173 (2010).

Brenhouse, H. C. & Andersen, S. L. Developmental trajectories during adolescence in males and females: A cross-species understanding of underlying brain changes. *Neuroscience and Biobehavioral Reviews* **35,** 1687–1703 (2011).

Brouwer, R. M., Panizzon, M. S., Glahn, D. C., Hibar, D. P., Hua, X., Jahanshad, N., Abramovic, L., De Zubicaray, G. I., Franz, C. E., Hansell, N. K., Hickie, I. B., Koenis, M. M., Martin, N. G., Mather, K. A., McMahon, K. L., Schnack, H. G., Strike, L. T., Swagerman, S. C., Thalamuthu, A., Wen, W., Gilmore, J. H., Gogtay, N., Kahn, R. S., Sachdev, P. S., Wright, M. J., Boomsma, D. I., Kremen, W. S., Thompson, P. M. & Hulshoff Pol, H. E. Genetic influences on individual differences in longitudinal changes in global and subcortical brain volumes: Results of the ENIGMA plasticity working group. *Human Brain Mapping* **38,** 4444–4458 (2017).

Brown, A. W., Li, P., Brown, M. M., Kaiser, K. A., Keith, S. W., Oakes, J. M. & Allison, D. B. Best (but oft-forgotten) practices: Designing, analyzing, and reporting cluster randomized controlled trials. *American Journal of Clinical Nutrition* **102,** 241–248 (2015).

Bruyer, R. & Brysbaert, M. Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica* **51,** 5–13 (2011).

Brydges, C. R., Fox, A. M., Reid, C. L. & Anderson, M. The differentiation of executive functions in middle and late childhood: A longitudinal latent-variable analysis. *Intelligence* **47,** 34–43 (2014).

Buchheit, M. & Laursen, P. B. High-intensity interval training, solutions to the programming puzzle: Part I: Cardiopulmonary emphasis. *Sports Medicine* **43,** 313–338 (2013).

Buchheit, M. & Laursen, P. B. High-intensity interval training, solutions to the programming puzzle: Part II: Anaerobic energy, neuromuscular load and practical applications. *Sports Medicine* **43,** 927–954 (2013).

Budde, H., Schwarz, R., Velasques, B., Ribeiro, P., Holzweg, M., Machado, S., Brazaitis, M., Staack, F. & Wegner, M. The need for differentiating between exercise, physical activity, and training. *Autoimmunity Reviews* **15,** 110–111 (2016).

Bull, F. C., Al-Ansari, S. S., Biddle, S., Borodulin, K., Buman, M. P., Cardon, G., Carty, C., Chaput, J. P., Chastin, S., Chou, R., Dempsey, P. C., Dipietro, L., Ekelund, U., Firth, J., Friedenreich, C. M., Garcia, L., Gichu, M., Jago, R., Katzmarzyk, P. T., Lambert, E., Leitzmann, M., Milton, K., Ortega, F. B., Ranasinghe, C., Stamatakis, E., Tiedemann, A., Troiano, R. P., Van Der Ploeg, H. P., Wari, V. & Willumsen, J. F. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *British Journal of Sports Medicine* **54,** 1451–1462 (2020).

Bull, R., Espy, K. A. & Wiebe, S. A. Short-Term Memory, Working Memory, and Executive Functioning in Preschoolers: Longitudinal Predictors of Mathematical Achievement at Age 7 Years. *Developmental Neuropsychology* **33,** 205–228 (2008).

Bull, R. & Scerif, G. Executive Functioning as a Predictor of Children's Mathematics Ability: Inhibition, Switching, and Working Memory. *Developmental Neuropsychology* **19,** 273–293 (2001).

Bulus, M., Dong, N., Kelcey, B. & Spybrook, J. *PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments* 2019. https://cran.r-project.org/package=PowerUpR.

Bunketorp Käll, L., Malmgren, H., Olsson, E., Lindén, T. & Nilsson, M. Effects of a Curricular Physical Activity Intervention on Children's School Performance, Wellness, and Brain Development. *Journal of School Health* **85,** 704–713 (2015).

Burgess, N. *Methodology of frontal and executive function* (ed Rabbitt, P.) 81–116 (Psychology Press, Hove, UK, 1997).

Burgette, L. F. & Reiter, J. P. Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology* **172,** 1070–1076 (2010).

Bustamante, E. E., Williams, C. F. & Davis, C. L. Physical Activity Interventions for Neurocognitive and Academic Performance in Overweight and Obese Youth. A Systematic Review. *Pediatric Clinics of North America* **63,** 459–480 (2016).

Butte, N. F., Watson, K. B., Ridley, K., Zakeri, I. F., McMurray, R. G., Pfeiffer, K. A., Crouter, S. E., Herrmann, S. D., Bassett, D. R., Long, A., Berhane, Z., Trost, S. G., Ainsworth, B. E., Berrigan, D. & Fulton, J. E. A youth compendium of physical activities: Activity codes and metabolic intensities. *Medicine and Science in Sports and Exercise* **50,** 246–256 (2018).

Buxton, R. B. Interpreting oxygenation-based neuroimaging signals: the importance and the challenge of understanding brain oxygen metabolism. *Frontiers in Neuroenergetics* **2,** 8 (2010).

Cabral, D. F., Rice, J., Morris, T. P., Rundek, T., Pascual-Leone, A. & Gomes-Osman, J. Exercise for Brain Health: An Investigation into the Underlying Mechanisms Guided by Dose. *Neurotherapeutics* **16,** 580–599 (2019).

Cain, K. L., Sallis, J. F., Conway, T. L., Van Dyck, D. & Calhoon, L. Using accelerometers in youth physical activity studies: a review of methods. *Journal of Physical Activity & Health* **10,** 437–450 (2013).

Callaghan, M. F., Helms, G., Lutti, A., Mohammadi, S. & Weiskopf, N. A general linear relaxometry model of R1 using imaging data. *Magnetic Resonance in Medicine* **73,** 1309–1314 (2015).

Callaghan, M. F., Freund, P., Draganski, B., Anderson, E., Cappelletti, M., Chowdhury, R., Diedrichsen, J., FitzGerald, T. H., Smittenaar, P., Helms, G., Lutti, A. & Weiskopf, N. Widespread age-related differences in the human brain microstructure revealed by quantitative magnetic resonance imaging. *Neurobiology of Aging* **35,** 1862–1872 (2014).

Calverley, T. A., Ogoh, S., Marley, C. J., Steggall, M., Marchi, N., Brassard, P., Lucas, S. J., Cotter, J. D., Roig, M., Ainslie, P. N., Wisløff, U. & Bailey, D. M. HIITing the brain with exercise; mechanisms, consequences and practical recommendations. *The Journal of Physiology* **598,** 2513–2530 (2020).

Campbell, J. P. & Turner, J. E. Debunking the myth of exercise-induced immune suppression: Redefining the impact of exercise on immunological health across the lifespan. *Frontiers in Immunology* **9,** 648 (2018).

Campbell, M. K., Mollison, J., Steen, N., Grimshaw, J. M. & Eccles, M. Analysis of cluster randomized trials in primary care: a practical approach. *Family Practice* **17,** 192–196 (2000).

Campbell, M. K., Grimshaw, J. M. & Elbourne, D. R. Intracluster correlation coefficients in cluster randomized trials: Empirical insights into how should they be reported. *BMC Medical Research Methodology* **4,** 9 (2004).

Campbell, M. K., Piaggio, G., Elbourne, D. R. & Altman, D. G. Consort 2010 statement: Extension to cluster randomised trials. *BMJ* **345,** e5661 (2012).

Canadian Society for Exercise Physiology. *Physical Activity Readiness Questionnaire (PAR-Q)* 2002.

Cancela, J., Burgo, H. & Sande, E. Physical fitness and executive functions in adolescents: cross-sectional associations with academic achievement. *Journal of Physical Therapy Science* **31,** 556–562 (2019).

Carmichael, O. T., Aizenstein, H. A., Davis, S. W., Becker, J. T., Thompson, P. M., Meltzer, C. C. & Liu, Y. Atlas-based hippocampus segmentation in Alzheimer's disease and mild cognitive impairment. *NeuroImage* **27,** 979–990 (2005).

Carpenter, J. & Kenward, M. *Multiple imputation and its application* (John Wiley & Sons, 2012).

Carpenter, T. P. T., Pogacar, R., Pullig, C., Isenberg, N., Kouril, M., LaBouff, J., Aguilar, S., Isenberg, N. & Chakroff, A. Building and Analyzing Implicit Association Tests for Online Surveys: A Tutorial and Open-source Tool. *PsyArXiv,* doi:10.31234/osf.io/hgy3z (2017).

Carson, V., Rinaldi, R. L., Torrance, B., Maximova, K., Ball, G. D., Majumdar, S. R., Plotnikoff, R. C., Veugelers, P., Boulé, N. G., Wozny, P., McCargar, L., Downs, S., Daymont, C., Lewanczuk, R. & McGavock, J. Vigorous physical activity and longitudinal associations with cardiometabolic risk factors in youth. *International Journal of Obesity* **38,** 16–21 (2014).

Carson, V., Tremblay, M. S. & Chastin, S. F. Associations between sleep duration, sedentary time, physical activity, and adiposity indicators among Canadian preschool-aged children using compositional analyses. *BMC Public Health* **41,** 294–302 (2016).

Carter, B. R. & Hood, K. Balance algorithm for cluster randomized trials. *BMC Medical Research Methodology* **8,** 65 (2008).

Casey, B. J., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., Soules, M. E., Teslovich, T., Dellarco, D. V., Garavan, H., Orr, C. A., Wager, T. D., Banich, M. T., Speer, N. K., Sutherland, M. T., Riedel, M. C., Dick, A. S., Bjork, J. M., Thomas, K. M., Chaarani, B., Mejia, M. H., Hagler, D. J., Daniela Cornejo, M., Sicat, C. S., Harms, M. P., Dosenbach, N. U., Rosenberg, M., Earl, E., Bartsch, H., Watts, R., Polimeni, J. R., Kuperman, J. M., Fair, D. A. & Dale, A. M. The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience* **32,** 43–54 (2018).

Caspersen, C. J., Powell, K. E. & Christenson, G. M. Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public health reports* **100,** 126–31 (1985).

Cassilhas, R. C., Tufik, S. & De Mello, M. T. Physical exercise, neuroplasticity, spatial learning and memory. *Cellular and Molecular Life Sciences* **73,** 975–983 (2016).

Castro-Piñero, J., Artero, E. G., España-Romero, V., Ortega, F. B., Sjöström, M., Suni, J. & Ruiz, J. R. Criterion-related validity of field-based fitness tests in youth: A systematic review. *British Journal of Sports Medicine* **44,** 934–943 (2010).

Cercignani, M. & Bouyagoub, S. Brain microstructure by multi-modal MRI: Is the whole greater than the sum of its parts? *NeuroImage* **182,** 117–127 (2018).

Chaddock-Heyman, L., Erickson, K. I., Chappell, M. A., Johnson, C. L., Kienzler, C., Knecht, A., Drollette, E. S., Raine, L. B., Scudder, M. R., Kao, S. C., Hillman, C. H. & Kramer, A. F. Aerobic fitness is associated with greater hippocampal cerebral blood flow in children. *Developmental Cognitive Neuroscience* **20,** 52–58 (2016).

Chaddock-Heyman, L., Erickson, K. I., Holtrop, J. L., Voss, M. W., Pontifex, M. B., Raine, L. B., Hillman, C. H. & Kramer, A. F. Aerobic fitness is associated with greater white matter integrity in children. *Frontiers in Human Neuroscience* **8,** 584 (2014).

Chaddock-Heyman, L., Erickson, K. I., Kienzler, C., King, M., Pontifex, M. B., Raine, L. B., Hillman, C. H. & Kramer, A. F. The role of aerobic fitness in cortical thickness and mathematics achievement in preadolescent children. *PLoS ONE* **10,** e0134115 (2015).

Chaddock, L., Erickson, K. I., Prakash, R. S., Kim, J. S., Voss, M. W., Vanpatter, M., Pontifex, M. B., Raine, L. B., Konkel, A., Hillman, C. H., Cohen, N. J. & Kramer, A. F. A neuroimaging investigation of the association between aerobic fitness, hippocampal volume, and memory performance in preadolescent children. *Brain Research* **1358,** 172–183 (2010).

Chaddock, L., Erickson, K. I., Prakash, R. S., Vanpatter, M., Voss, M. W., Pontifex, M. B., Raine, L. B., Hillman, C. H. & Kramer, A. F. Basal ganglia volume is associated with aerobic fitness in preadolescent children. *Developmental Neuroscience* **32,** 249–256 (2010).

Chaddock, L., Erickson, K. I., Prakash, R. S., Voss, M. W., VanPatter, M., Pontifex, M. B., Hillman, C. H. & Kramer, A. F. A functional MRI investigation of the association between childhood aerobic fitness and neurocognitive control. *Biological Psychology* **89,** 260–268 (2012).

Chakravarty, M. M., Steadman, P., Van Eede, M. C., Calcott, R. D., Gu, V., Shaw, P., Raznahan, A., Collins, D. L. & Lerch, J. P. Performing label-fusion-based segmentation using multiple automatically generated templates. *Human Brain Mapping* **34,** 2635–2654 (2013).

Chan, A.-W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J. A., Doré, C. J., Parulekar, W. R., Summerskill, W. S., Groves, T., Schulz, K. F., Sox, H. C., Rockhold, F. W., Rennie, D. & Moher, D. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. *Annals of Internal Medicine* **158,** 200 (2013).

Chang, Y.-K., Erickson, K. I., Stamatakis, E. & Hung, T.-M. How the 2018 US Physical Activity Guidelines are a Call to Promote and Better Understand Acute Physical Activity for Cognitive Function Gains. *Sports Medicine,* 1625–1627 (2019).

Chang, Y. K., Labban, J. D., Gapin, J. I. & Etnier, J. L. The effects of acute exercise on cognitive performance: A meta-analysis. *Brain Research* **1453,** 87–101 (2012).

Chang, Y. K., Chu, C. H., Chen, F. T., Hung, T. M. & Etnier, J. L. Combined Effects of Physical Activity and Obesity on Cognitive Function: Independent, Overlapping, Moderator, and Mediator Models. *Sports Medicine* **47,** 449–468 (2017).

Chappell, M. A., Groves, A. R., Whitcher, B. & Woolrich, M. W. Variational Bayesian Inference for a Nonlinear Forward Model. *IEEE Transactions on Signal Processing* **57,** 223–236 (2009).

Chaput, J. P., Carson, V., Gray, C. E. & Tremblay, M. S. Importance of all movement behaviors in a 24 hour period for overall health. *International Journal of Environmental Research and Public Health* **11,** 12575–12581 (2014).

Chen, A.-g. G., Zhu, L.-n. N., Yan, J., Yin, H.-c. C. & Chen, A.-g. G. Neural Basis of Working Memory Enhancement after Acute Aerobic Exercise: fMRI Study of Preadolescent Children. *Frontiers in Psychology* **7,** 1804 (2016).

Chen, A.-G., Yan, J., Yin, H.-c., Pan, C.-Y. & Chang, Y.-K. Effects of acute aerobic exercise on multiple aspects of executive function in preadolescent children. *Psychology of Sport & Exercise* **15,** 627–636 (2014).

Chen, J., Liu, J., Calhoun, V. D., Arias-Vasquez, A., Zwiers, M. P., Gupta, C. N., Franke, B. & Turner, J. A. Exploration of scanning effects in multi-site structural MRI studies. *Journal of Neuroscience Methods* **230,** 37–50 (2014).

Cherbuin, N., Anstey, K. J., Réglade-Meslin, C. & Sachdev, P. S. In vivo hippocampal measurement and memory: A comparison of manual tracing and automated segmentation in a large community-based sample. *PLoS ONE* **4,** e5265 (2009).

Chinapaw, M. J., Mokkink, L. B., Van Poppel, M. N., Van Mechelen, W. & Terwee, C. B. Physical activity questionnaires for youth: A systematic review of measurement properties. *Sports Medicine* **40,** 539–563 (2010).

Chinapaw, M., Klakk, H., Møller, N. C., Andersen, L. B., Altenburg, T. & Wedderkopp, N. Total volume versus bouts: prospective relationship of physical activity and sedentary time with cardiometabolic risk in children. *International Journal of Obesity* **42,** 1733–1742 (2018).

Chojnacki, M. R., Raine, L. B., Drollette, E. S., Scudder, M. R., Kramer, A. F., Hillman, C. H. & Khan, N. A. The Negative Influence of Adiposity Extends to Intraindividual Variability in Cognitive Control Among Preadolescent Children. *Obesity* **26,** 405–411 (2018).

Clark, C. C., Barnes, C. M., Stratton, G., McNarry, M. A., Mackintosh, K. A. & Summers, H. D. A Review of Emerging Analytical Techniques for Objective Physical Activity Measurement in Humans. *Sports Medicine* **47,** 439–447 (2017).

Clark, I. A., Monk, A. M., Hotchin, V., Pizzamiglio, G., Liefgreen, A., Callaghan, M. F. & Maguire, E. A. Does hippocampal volume explain performance differences on hippocampal-dependant tasks? *NeuroImage* **221,** 117211 (2020).

Clark, P. J., Brzezinska, W. J., Puchalski, E. K., Krone, D. A. & Rhodes, J. S. Functional analysis of neurovascular adaptations to exercise in the dentate gyrus of young adult mice associated with cognitive gain. *Hippocampus* **19,** 937–950 (2009).

Clifton, L. & Clifton, D. A. The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials* **20,** 4–9 (2019).

Cohen, N. J. & Eichenbaum, H. *Memory, amnesia, and the hippocampal system.* xii, 330–xii, 330 (The MIT Press, Cambridge, MA, US, 1993).

Colcombe, S. & Kramer, A. F. Fitness Effects on the Cognitive Function of Older Adults: A Meta-Analytic Study. *Psychological Science* **14,** 125–130 (2003).

Cole, J. H., Ritchie, S. J., Bastin, M. E., Valdés Hernández, M. C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S. E., Zhang, Q., Wray, N. R., Redmond, P., Marioni, R. E., Starr, J. M., Cox, S. R., Wardlaw, J. M., Sharp, D. J. & Deary, I. J. Brain age predicts mortality. *Molecular Psychiatry* **23,** 1385–1392 (2018).

Collins, D. L. & Pruessner, J. C. Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. *NeuroImage* **52,** 1355–1366 (2010).

Cong, S., Risacher, S. L., West, J. D., Wu, Y. C., Apostolova, L. G., Tallman, E., Rizkalla, M., Salama, P., Saykin, A. J. & Shen, L. Volumetric comparison of hippocampal subfields extracted from 4-minute accelerated vs. 8-minute high-resolution T2-weighted 3T MRI scans. *Brain Imaging and Behavior* **12,** 1583–1595 (2018).

Constantinidis, C. & Luna, B. Neural Substrates of Inhibitory Control Maturation in Adolescence. *Trends in Neurosciences* **42,** 604–616 (2019).

Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).

Cook, D. A. & Beckman, T. J. Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine* **119,** 166.e7 (2006).

Cooper, C., Moon, H. Y. & Van Praag, H. On the run for hippocampal plasticity. *Cold Spring Harbor Perspectives in Medicine* **8,** a029736 (2018).

Cooper, C. B. & Storer, T. W. *Exercise Testing and Interpretation: A Practical Approach* (Cambridge University Press, Cambridge, 2001).

Cooper, K. H. A Means of Assessing Maximal Oxygen Intake. *JAMA* **203,** 201–204 (1968).

Corbin, C. B. A multidimensional hierarchical model of physical fitness: A basis for integration and collaboration. *Quest* **43,** 296–306 (1991).

Corbin, C. B., Welk, G. J., Richardson, C., Vowell, C., Lambdin, D. & Wikgren, S. Youth Physical Fitness: Ten Key Concepts. *Journal of Physical Education, Recreation & Dance* **85,** 24–31 (2014).

Corder, K., Sharp, S. J., Atkin, A. J., Andersen, L. B., Cardon, G., Page, A., Davey, R., Grøntved, A., Hallal, P. C., Janz, K. F., Kordas, K., Kriemler, S., Puder, J. J., Sardinha, L. B., Ekelund, U., Van Sluijs, E. M., Cardon, G., Cooper, A., Puder, J. J., Sardinha, L. B., Andersen, L. B., Anderssen, S., Esliger, D. W., Grøntved, A., Hallal, P., Janz, K. F., Kordas, K., Kriemler, S., Page, A., Pate, R., Reilly, J., Salmon, J., Sherar, L. B. & Timperio, A. Age-related patterns of vigorous-intensity physical activity in youth: The International Children's Accelerometry Database. *Preventive Medicine Reports* **4,** 17–22 (2016).

Corder, K., Sharp, S. J., Atkin, A. J., Griffin, S. J., Jones, A. P., Ekelund, U. & Van Sluijs, E. M. Change in objectively measured physical activity during the transition to adolescence. *British Journal of Sports Medicine* **49,** 730–736 (2015).

Cornfield, J. Randomization by group: a formal analysis. *American Journal of Epidemiology* **108,** 100–102 (1978).

Costigan, S. A., Eather, N., Plotnikoff, R. C., Taaffe, D. R. & Lubans, D. R. High-intensity interval training for improving health-related fitness in adolescents: A systematic review and meta-analysis. *British Journal of Sports Medicine* **49,** 1253–1261 (2015).

Costigan, S. A., Eather, N., Plotnikoff, R. C., Hillman, C. H. & Lubans, D. R. High-Intensity Interval Training for Cognitive and Mental Health in Adolescents. *Medicine and Science in Sports and Exercise* **48,** 1985–1993 (2016).

Cotman, C. W. & Berchtold, N. C. Exercise: A behavioral intervention to enhance brain health and plasticity. *Trends in Neurosciences* **25,** 295–301 (2002).

Cotman, C. W., Berchtold, N. C. & Christie, L.-A. Exercise builds brain health: key roles of growth factor cascades and inflammation. *Trends in neurosciences* **30,** 464–472 (2007).

Creer, D. J., Romberg, C., Saksida, L. M., Van Praag, H. & Bussey, T. J. Running enhances spatial pattern separation in mice. *Proceedings of the National Academy of Sciences* **107,** 2367–2372 (2010).

Crombe, A., Planche, V., Raffard, G., Bourel, J., Dubourdieu, N., Panatier, A., Fukutomi, H., Dousset, V., Oliet, S., Hiba, B. & Tourdias, T. Deciphering the microstructure of hippocampal subfields with in vivo DTI and NODDI: Applications to experimental multiple sclerosis. *NeuroImage* **172,** 357–368 (2018).

Crone, E. A. & Steinbeis, N. Neural Perspectives on Cognitive Control Development during Childhood and Adolescence. *Trends in Cognitive Sciences* **21,** 205–215 (2017).

Cui, Z., Li, H., Xia, C. H., Larsen, B., Adebimpe, A., Baum, G. L., Cieslak, M., Gur, R. E., Gur, R. C., Moore, T. M., Oathes, D. J., Alexander-Bloch, A. F., Raznahan, A., Roalf, D. R., Shinohara, R. T., Wolf, D. H., Davatzikos, C., Bassett, D. S., Fair, D. A., Fan, Y. & Satterthwaite, T. D. Individual Variation in Functional Topography of Association Networks in Youth. *Neuron* **106,** 340–353 (2020).

Currie, C., Griebler, R., Inchley, J., Theunissen, A., Molcho, M., Samdal, O., Dür, W. & (eds.) Health Behaviour in School-aged Children (HBSC) Study Protocol: Background, Methodology and Mandatory Items for the 2009/10 Survey. *Edinburgh: CAHRU & Vienna: LBIHPR.* http://www.hbsc.org (2010).

D'Haese, S., Cardon, G., De Bourdeaudhuij, I., Deforche, B., De Meester, F. & Van Dyck, D. Changes in Individual and Social Environmental Characteristics in Relation to Changes in Physical Activity: a Longitudinal Study from Primary to Secondary School. *International Journal of Behavioral Medicine* **23,** 539–552 (2016).

Dahl, R. E., Allen, N. B., Wilbrecht, L. & Suleiman, A. B. Importance of investing in adolescence from a developmental science perspective. *Nature* **554,** 441–450 (2018).

Dal-Ré, R., Janiaud, P. & Ioannidis, J. P. Real-world evidence: How pragmatic are randomized controlled trials labeled as pragmatic? *BMC Medicine* **16,** 49 (2018).

Dale, A. M., Fischl, B., Sereno, M. I., Segmentation, I., Reconstruction, S. & Dale. Cortical Surface-Based Analysis. I. Segmentation and Surface Reconstruction. *NeuroImage* **9,** 179–194 (1999).

Dale, L. P., Vanderloo, L., Moore, S. & Faulkner, G. Physical activity and depression, anxiety, and self-esteem in children and youth: An umbrella systematic review. *Mental Health and Physical Activity* **16,** 66–79 (2019).

Daly-smith, A. J., Zwolinsky, S., Mckenna, J., Tomporowski, P. D., Defeyter, M. A. & Manley, A. Systematic review of acute physically active learning and classroom movement breaks on children's physical activity, cognition, academic performance and classroom behaviour: understanding critical design features. *BMJ Open Sport & Exercise Medicine* **4,** e000341 (2018).

Damadian, R. Tumor detection by nuclear magnetic resonance. *Science* **171,** 1151–1153 (1971).

Dane, A. V. & Schneider, B. H. Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review* **18,** 23–45 (1998).

Daniels, M. J. & Hogan, J. W. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis* (Chapman and Hall/CRC, 2008).

Darki, F. & Klingberg, T. The role of fronto-parietal and fronto-striatal networks in the development of working memory: A longitudinal study. *Cerebral Cortex* **25,** 1587–1595 (2015).

Darrow, C. L. The Effectiveness and Precision of Intervention Fidelity Measures in Preschool Intervention Research. *Early Education and Development* **24,** 1137–1160 (2013).

Daugherty, A. M., Flinn, R. & Ofen, N. Hippocampal CA3-dentate gyrus volume uniquely linked to improvement in associative memory from childhood to adulthood. *NeuroImage* **153,** 75–85 (2017).

Davidson, M. C., Amso, D., Cruess, L. & Diamond, A. Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* **44,** 2037–2078 (2006).

Davies, E. L., Martin, J. & Foxcroft, D. R. Age differences in alcohol prototype perceptions and willingness to drink in UK adolescents. *Psychology, Health & Medicine* **21,** 317–329 (2016).

Davis, C. L., Tomporowski, P. D., McDowell, J. E., Austin, B. P., Miller, P. H., Allison, J. D., Yanasak, N. E., Allison, J. D. & Naglieri, J. A. Exercises improves executive function and achievement and alters brain activation in overweight children A randomized controlled trial. *Health Psychology* **30,** 91–98 (2011).

De Almeida Mendes, M., Da Silva, I. C., Ramires, V. V., Reichert, F. F., Martins, R. C. & Tomasi, E. Calibration of raw accelerometer data to measure physical activity: A systematic review. *Gait and Posture* **61,** 98–110 (2018).

De Bruin, E. J., Van Run, C., Staaks, J. & Meijer, A. M. Effects of sleep manipulation on cognitive functioning of adolescents: A systematic review. *Sleep Medicine Reviews* **32,** 45–57 (2017).

De Flores, R., La Joie, R. & Chételat, G. Structural imaging of hippocampal subfields in healthy aging and Alzheimer's disease. *Neuroscience* **309,** 29–50 (2015).

De Greeff, J. W., Bosker, R. J., Oosterlaan, J., Visscher, C. & Hartman, E. Effects of physical activity on executive functions, attention and academic performance in preadolescent children: a meta-analysis. *Journal of Science and Medicine in Sport* **21,** 501–507 (2018).

De Jong, L. W., Vidal, J. S., Forsberg, L. E., Zijdenbos, A. P., Haight, T., Sigurdsson, S., Gudnason, V., Van Buchem, M. A. & Launer, L. J. Allometric scaling of brain regions to intra-cranial volume: An epidemiological MRI study. *Human Brain Mapping* **38,** 151–164 (2017).

De Leeuw, J. R. jsPsych: a JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods* **47,** 1–12 (2015).

De Leeuw, J. R. & Motz, B. A. Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods* **48,** 1–12 (2016).

De Master, D., Pathman, T., Lee, J. K. & Ghetti, S. Structural development of the hippocampus and episodic memory: Developmental differences along the anterior/posterior axis. *Cerebral Cortex* **24,** 3036–3045 (2014).

Decker, A. L., Szulc, K. U., Bouffet, E., Laughlin, S., Chakravarty, M. M., Skocic, J., De Medeiros, C. B. & Mabbott, D. J. Smaller hippocampal subfield volumes predict verbal associative memory in pediatric brain tumor survivors. *Hippocampus* **27,** 1140–1154 (2017).

Deighton, J., Humphrey, N., Belsky, J., Boehnke, J., Vostanis, P. & Patalay, P. Longitudinal pathways between mental health difficulties and academic performance during middle childhood and early adolescence. *British Journal of Developmental Psychology* **36,** 110–126 (2018).

DeKraker, J., Lau, J., Ferko, K., Khan, A. & Köhler, S. Hippocampal subfields revealed through unfolding and unsupervised clustering of laminar and morphological features in 3D BigBrain. *NeuroImage* **206,** 116328 (2019).

Demirakca, T., Brusniak, W., Tunc-Skarka, N., Wolf, I., Meier, S., Matthäus, F., Ende, G., Schulze, T. G. & Diener, C. Does body shaping influence brain shape? Habitual physical activity is linked to brain morphology independent of age. *World Journal of Biological Psychiatry* **15,** 387–396 (2014).

Den Heijer, A. E., Groen, Y., Tucha, L., Fuermaier, A. B. M., Koerts, J., Lange, K. W., Thome, J. & Tucha, O. Sweat it out? The effects of physical exercise on cognition and behavior in children and adults with ADHD: a systematic literature review. *Journal of Neural Transmission* **124,** 3–26 (2017).

Den Ouden, L., Kandola, A., Suo, C., Hendrikse, J., Costa, R. J., Watt, M. J., Lorenzetti, V., Chye, Y., Parkes, L., Sabaroedin, K. & Yücel, M. The Influence of Aerobic Exercise on Hippocampal Integrity and Function: Preliminary Findings of a Multi-Modal Imaging Analysis. *Brain Plasticity* **4,** 211–216 (2018).

Deoni, S. C. Quantitative Relaxometry of the Brain. *Topics in magnetic resonance imaging: TMRI* **21,** 101–113 (2010).

Deoni, S. C., Williams, S. C., Jezzard, P., Suckling, J., Murphy, D. G. & Jones, D. K. Standardized structural magnetic resonance imaging in multicentre studies using quantitative T1 and T2 imaging at 1.5 T. *NeuroImage* **40,** 662–671 (2008).

Destrieux, C., Bourry, D. & Velut, S. Surgical anatomy of the hippocampus. *Neurochirurgie* **59,** 149–158 (2013).

Detre, J. A., Leigh, J. S., Williams, D. S. & Koretsky, A. P. Perfusion imaging. *Magnetic Resonance in Medicine* **23,** 37–45 (1992).

Diamond, A. Developmental Time Course in Human Infants and Infant Monkeys, and the Neural Bases of, Inhibitory Control in Reaching. *Annals of the New York Academy of Sciences* **608,** 637–676 (1990).

Diamond, A. Executive Functions. *Annual Review of Psychology* **64,** 135–168 (2013).

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

Díaz-Ordaz, K., Kenward, M. G., Cohen, A., Coleman, C. L. & Eldridge, S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials* **11,** 590–600 (2014).

Dice, L. R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **26,** 297–302 (1945).

Dill, V., Franco, A. R. & Pinho, M. S. Automated Methods for Hippocampus Segmentation: the Evolution and a Review of the State of the Art. *Neuroinformatics* **13,** 133–150 (2015).

Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., Van Mechelen, W. & Pratt, M. The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet* **388,** 1311–1324 (2016).

Ding, D., Ramirez Varela, A., Bauman, A. E., Ekelund, U., Lee, I.-M., Heath, G., Katzmarzyk, P. T., Reis, R. & Pratt, M. Towards better evidence-informed global action: lessons learnt from the Lancet series and recent developments in physical activity and public health. *British Journal of Sports Medicine* **54,** 462–468 (2020).

Ding, Y.-H., Li, J., Zhou, Y., Rafols, J. a., Clark, J. C. & Ding, Y. Cerebral angiogenesis and expression of angiogenic factors in aging rats after exercise. *Current Neurovascular Research* **3,** 15–23 (2006).

Dinga, R., Schmaal, L., Penninx, B. W., Veltman, D. J. & Marquand, A. F. Controlling for effects of confounding variables on machine learning predictions. *bioRxiv,* doi: 10.1101/2020.08.17.255034 (2020).

Dishman, R. K., Washburn, R. A. & Schoeller, D. A. Measurement of physical activity. *Quest* **53,** 295–309 (2001).

Dobbins, M., De Corby, K., Robeson, P., Husson, H. & Tirilis, D. School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6-18. *Cochrane Database of Systematic Reviews* **18,** CD007651 (2009).

Dobbins, M., Husson, H., Decorby, K. & Rl, L. School-based physical activity programs for promoting physical activity and fitness in children and adolescents aged 6 to 18 (Review). *Cochrane Database of Systematic Reviews* **18,** CD007651 (2013).

Does, M. D. Inferring brain tissue composition and microstructure via MR relaxometry. *NeuroImage* **182,** 136–148 (2018).

Dong, Y. & Peng, C.-Y. J. Principled missing data methods for researchers. *SpringerPlus* **2,** 222 (2013).

Donnelly, J. E., Hillman, C. H., Castelli, D., Etnier, J. L., Lee, S., Tomporowski, P., Lambourne, K. & Szabo-Reed, A. N. Physical activity, fitness, cognitive function, and academic achievement in children: A systematic review. *Medicine and Science in Sports and Exercise* **48,** 1197–1222 (2016).

Donner, A., Birkett, N. & Buck, C. Randomization by cluster. *American Journal of Epidemiology* **114,** 906–914 (1981).

Donner, A. & Klar, N. Pitfalls of and Controversies in Cluster Randomization Trials. *American Journal of Public Health* **94,** 416–422 (2004).

Doove, L. L., Van Buuren, S. & Dusseldorp, E. Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics and Data Analysis* **72,** 92–104 (2014).

Dosenbach, N. U., Koller, J. M., Earl, E. A., Miranda-Dominguez, O., Klein, R. L., Van, A. N., Snyder, A. Z., Nagel, B. J., Nigg, J. T., Nguyen, A. L., Wesevich, V., Greene, D. J. & Fair, D. A. Real-time motion analytics during brain MRI improve data quality and reduce costs. *NeuroImage* **161,** 80–93 (2017).

Dow-Edwards, D., MacMaster, F. P., Peterson, B. S., Niesink, R., Andersen, S. & Braams, B. R. Experience during adolescence shapes brain development: From synapses and networks to normal and pathological behavior. *Neurotoxicology and Teratology* **76,** 106834 (2019).

Downing, S. M. Reliability: on the reproducibility of assessment data. *Medical Education* **38,** 1006–1012 (2004).

Draheim, C., Hicks, K. L. & Engle, R. W. Combining Reaction Time and Accuracy: The Relationship Between Working Memory Capacity and Task Switching as a Case Example. *Perspectives on Psychological Science* **11,** 133–155 (2016).

Draheim, C., Mashburn, C. A., Martin, J. D. & Engle, R. W. Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin* **145,** 508–535 (2019).

Drake, C., Nickel, C., Burduvali, E., Roth, T., Jefferson, C., Badia, P., Pietro, B. & Badia, P. The pediatric daytime sleepiness scale (PDSS): sleep habits and school outcomes in middle-school children. *Sleep* **26,** 455–458 (2003).

Drechsler, J. Multiple Imputation of Multilevel Missing Data — Rigor Versus Simplicity. *Journal of Educational and Behavioral Statistics* **40,** 69–95 (2015).

Dreher, G. F. & Bretz, R. D. Cognitive ability and career attainment: The moderating effects of early career success (CAHRS Working Paper #90-10). *Ithaca, NY: Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies.* http://digitalcommons.ilr.cornell.edu/cahrswp/375 (1990).

Drenowatz, C., Eisenmann, J., Pfeiffer, K., Heelan, K., Gentile, D. & Walsh, D. Influence of socio-economic status on habitual physical activity and sedentary behavior in 8- to 11-year old children. *BMC Public Health* **10,** 1–11 (2010).

Drollette, E. S., Scudder, M. R., Raine, L. B., Davis Moore, R., Pontifex, M. B., Erickson, K. I. & Hillman, C. H. The sexual dimorphic association of cardiorespiratory fitness to working memory in children. *Developmental Science* **19,** 90–108 (2016).

Ducharme, S., Albaugh, M. D., Nguyen, T. V., Hudziak, J. J., Mateos-Pérez, J. M., Labbe, A., Evans, A. C., Karama, S. & Group, B. D. C. Trajectories of cortical thickness maturation in normal brain development – The importance of quality control procedures. *Neuroimage* **125,** 267–279 (2016).

Duerden, E. G., Chakravarty, M. M., Lerch, J. P. & Taylor, M. J. Sex-Based Differences in Cortical and Subcortical Development in 436 Individuals Aged 4–54 Years. *Cerebral Cortex* **30,** 2854–2866 (2020).

Dumith, S. C., Gigante, D. P., Domingues, M. R. & Kohl, H. W. Physical activity change during adolescence: A systematic review and a pooled analysis. *International Journal of Epidemiology* **40,** 685–698 (2011).

Dumuid, D., Pedišić, Ž., Stanford, T. E., Martín-Fernández, J. A., Hron, K., Maher, C. A., Lewis, L. K. & Olds, T. The compositional isotemporal substitution model: A method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Statistical Methods in Medical Research* **28,** 846–857 (2019).

Dumville, J. C., Torgerson, D. J. & Hewitt, C. E. Reporting attrition in randomised controlled trials. *British Medical Journal* **332,** 969–971 (2006).

Duncan, M. J., Roscoe, C. M., Faghy, M., Tallis, J. & Eyre, E. L. Estimating physical activity in children aged 8-11 years using accelerometry: Contributions from fundamental movement skills and different accelerometer placements. *Frontiers in Physiology* **10,** 242 (2019).

DuPaul, G., Power, T., Anastopoulos, A. & Reid, R. *ADHD Rating Scale-IV: Checklists, Norms, and Clinical Interpretation* (Guilford, New York, 1998).

DuPre, E. & Spreng, R. N. Structural covariance networks across the life span, from 6 to 94 years of age. *Network Neuroscience* **1,** 302–323 (2017).

Durston, S., Davidson, M. C., Tottenham, N., Galvan, A., Spicer, J., Fossella, J. A. & Casey, B. J. A shift from diffuse to focal cortical activity with development: Commentary. *Developmental Science* **9,** 1–8 (2006).

Duvernoy, H. M. *The human hippocampus: functional anatomy, vascularization and serial sections with MRI.* (Springer Science & Business Media, 2005).

Dziura, J. D., Post, L. A., Zhao, Q., Fu, Z. & Peduzzi, P. Strategies for dealing with missing data in clinical trials: From design to analysis. *Yale Journal of Biology and Medicine* **86,** 343–358 (2013).

Eadie, B. D., Redila, V. A. & Christie, B. R. Voluntary exercise alters the cytoarchitecture of the adult dentate gyrus by increasing cellular proliferation, dendritic complexity, and spine density. *Journal of Comparative Neurology* **486,** 39–47 (2005).

Eddolls, W. T. B., McNarry, M. A., Stratton, G., Winn, C. O. N. & Mackintosh, K. A. High-Intensity Interval Training Interventions in Children and Adolescents: A Systematic Review. *Sports Medicine* **47,** 2363–2374 (2017).

Edelman, R. R. The history of MR imaging as seen through the pages of radiology. *Radiology* **273,** S181–S200 (2014).

Edwards, L. J., Kirilina, E., Mohammadi, S. & Weiskopf, N. Microstructural imaging of human neocortex in vivo. *NeuroImage* **182,** 184–206 (2018).

Eekhout, I., De Boer, M. R., Twisk, J. W., De Vet, Henrica, C. & Heymans, M. W. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology* **23,** 729–732 (2012).

Eichenbaum, H., Amaral, D. G., Buffalo, E. A., Buzsáki, G., Cohen, N., Davachi, L., Frank, L., Heckers, S., Morris, R. G., Moser, E. I., Nadel, L., O'Keefe, J., Preston, A., Ranganath, C., Silva, A. & Witter, M. Hippocampus at 25. *Hippocampus* **26,** 1238–1249 (2016).

Eime, R. M., Charity, M. J., Harvey, J. T. & Payne, W. R. Participation in sport and physical activity: Associations with socio-economic status and geographical remoteness. *BMC Public Health* **15,** 434 (2015).

Ekeland, E., Heian, F. & Hagen, K. B. Can exercise improve self esteem in children and young people? A systematic review of randomised controlled trials. *British Journal of Sports Medicine* **39,** 792–798 (2005).

Ekelund, U., Anderssen, S. A., Froberg, K., Sardinha, L. B., Andersen, L. B. & Brage, S. Independent associations of physical activity and cardiorespiratory fitness with metabolic risk factors in children: The European youth heart study. *Diabetologia* **50,** 1832–1840 (2007).

Ekelund, U., Steene-Johannessen, J., Brown, W. J., Fagerland, M. W., Owen, N., Powell, K. E., Bauman, A. & Lee, I. M. Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women. *The Lancet* **388,** 1302–1310 (2016).

Ekstrom, A. How and when the fMRI BOLD signal relates to underlying neural activity: The danger in dissociation. *Brain Research Reviews* **62,** 233–244 (2010).

Eldridge, S. M., Ashby, D. & Kerry, S. Sample size for cluster randomized trials: Effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology* **35,** 1292–1300 (2006).

Eldridge, S. & Kerry, S. *A Practical Guide to Cluster Randomised Trials in Health Services Research* (John Wiley & Sons, 2012).

Eldridge, S., Kerry, S. & Torgerson, D. J. Bias in identifying and recruiting participants in cluster randomised trials: What can be done? *BMJ (Online)* **339,** 36–39 (2009).

Elliott, M. L., Belsky, D. W., Knodt, A. R., Ireland, D., Melzer, T. R., Poulton, R., Ramrakha, S., Caspi, A., Moffitt, T. E. & Hariri, A. R. Brain-age in midlife is associated with accelerated biological aging and cognitive decline in a longitudinal birth cohort. *Molecular Psychiatry,* 1–10 (2019).

EMA. Guideline on adjustment for baseline covariates in clinical trials. **44,** 11 (2015).

Enders, C. K. *Applied missing data analysis* (Guilford press, 2010).

Enders, C. K., Du, H. & Keller, B. T. A Model-Based Imputation Procedure for Multilevel Regression Models With Random Coefficients, Interaction Effects, and Nonlinear Terms. *Psychological Methods* **25,** 88–112 (2020).

Enders, C. K. & Gottschall, A. C. Multiple imputation strategies for multiple group structural equation models. *Structural Equation Modeling* **18,** 35–54 (2011).

Enders, C. K., Keller, B. T. & Levy, R. A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods* **23,** 298–317 (2018).

Enders, C. K., Keller, B. T. & Levy, R. A Fully Conditional Specification Approach to Multilevel Imputation of Categorical and Continuous Variables. *Psychological Methods* **23,** 298–317 (2018).

Erickson, K. I., Voss, M. W., Prakash, R. S., Basak, C., Szabo, A., Chaddock, L., Kim, J. S., Heo, S., Alves, H., White, S. M., Wojcicki, T. R., Mailey, E., Vieira, V. J., Martin, S. A., Pence, B. D., Woods, J. A., McAuley, E. & Kramer, A. F. Exercise training increases size of hippocampus and improves memory. *Proceedings of the National Academy of Sciences* **108,** 3017–3022 (2011).

Erickson, K. I., Hillman, C. H. & Kramer, A. F. Physical activity, brain, and cognition. *Current Opinion in Behavioral Sciences* **4,** 27–32 (2015).

Erickson, K. I., Hillman, C., Stillman, C. M., Ballard, R. M., Bloodgood, B., Conroy, D. E., Macko, R., Marquez, D. X., Petruzzello, S. J. & Powell, K. E. Physical Activity, Cognition, and Brain Outcomes: A Review of the 2018 Physical Activity Guidelines. *Medicine and Science in Sports and Exercise* **51,** 1242–1251 (2019).

Erickson, K. I., Leckie, R. L. & Weinstein, A. M. Physical activity, fitness, and gray matter volume. *Neurobiology of Aging* **35,** S20–S28 (2014).

Erickson, K. I., Prakash, R. S., Voss, M. W., Chaddock, L., Hu, L., Morris, K. S., White, S. M., Wo, T. R., Mcauley, E. & Kramer, A. F. Aerobic Fitness is Associated With Hippocampal Volume in Elderly Humans. *Hippocampus* **19,** 1030–1039 (2009).

Erler, N. S., Rizopoulos, D. & Lesaffre, E. M. JointAI: Joint Analysis and Imputation of Incomplete Data in R. *arXiv e-prints,* arXiv:1907.10867 (2019).

Erler, N. S., Rizopoulos, D., Van Rosmalen, J., Jaddoe, V. W., Franco, O. H. & Lesaffre, E. M. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine* **35,** 2955–2974 (2016).

Eshaghi, A., Kievit, R. A., Prados, F., Sudre, C. H., Nicholas, J., Cardoso, M. J., Chan, D., Nicholas, R., Ourselin, S., Greenwood, J., Thompson, A. J., Alexander, D. C., Barkhof, F., Chataway, J. & Ciccarelli, O. Applying causal models to explore the mechanism of action of simvastatin in progressive multiple sclerosis. *Proceedings of the National Academy of Sciences* **166,** 11020–11027 (2019).

Espie, C. A., Kyle, S. D., Hames, P., Gardani, M., Fleming, L. & Cape, J. The Sleep Condition Indicator: A clinical screening tool to evaluate insomnia disorder. *BMJ Open* **4** (2014).

Esser, P., Dawes, H., Collett, J., Feltham, M. G. & Howells, K. Assessment of spatio-temporal gait parameters using inertial measurement units in neurological populations. *Gait and Posture* **34,** 558–560 (2011).

Esser, P., Dawes, H., Collett, J. & Howells, K. IMU: Inertial sensing of vertical CoM movement. *Journal of Biomechanics* **42,** 1578–1581 (2009).

Esteban-Cornejo, I., Cadenas-Sanchez, C., Contreras-Rodriguez, O., Verdejo-Roman, J., Mora-Gonzalez, J., Migueles, J. H., Henriksson, P., Davis, C. L., Verdejo-Garcia, A., Catena, A. & Ortega, F. B. A whole brain volumetric approach in overweight/obese children: Examining the association with different physical fitness components and academic performance. The ActiveBrains project. *NeuroImage* **159,** 346–354 (2017).

Esteban-Cornejo, I., Catena, A., Hillman, C. H., Kramer, A. F., Erickson, K. I. & Ortega, F. B. Commentary: At least eighty percent of brain grey matter is modifiable by physical activity: A review study. *Frontiers in Human Neuroscience* **12,** 195 (2018).

Esteban-Cornejo, I., Mora-Gonzalez, J., Cadenas-Sanchez, C., Contreras-Rodriguez, O., Verdejo-Román, J., Henriksson, P., Migueles, J. H., Rodriguez-Ayllon, M., Molina-García, P., Suo, C., Hillman, C. H., Kramer, A. F., Erickson, K. I., Catena, A., Verdejo-García, A. & Ortega, F. B. Fitness, cortical thickness and surface area in overweight/obese children: The mediating role of body composition and relationship with intelligence. *NeuroImage* **186,** 771–781 (2019).

Esteban-Cornejo, I., Rodriguez-Ayllon, M., Verdejo-Roman, J., Cadenas-Sanchez, C., Mora-Gonzalez, J., Chaddock-Heyman, L., Raine, L. B., Stillman, C. M., Kramer, A. F., Erickson, K. I., Catena, A., Ortega, F. B. & Hillman, C. H. Physical fitness, white matter volume and academic performance in children: Findings from the activebrains and FITKids2 projects. *Frontiers in Psychology* **10,** 208 (2019).

Esteban-Cornejo, I., Stillman, C. M., Rodriguez-Ayllon, M., Kramer, A. F., Hillman, C. H., Catena, A., Erickson, K. I. & Ortega, F. B. Physical fitness, hippocampal functional connectivity and academic performance in children with overweight/obesity: The ActiveBrains project. *Brain, Behavior, and Immunity* **91,** 284–295 (2021).

Esteban-Cornejo, I., Tejero-Gonzalez, C. M., Sallis, J. F. & Veiga, O. L. Physical activity and cognition in adolescents: A systematic review. *Journal of Science and Medicine in Sport* **18,** 534–539 (2015).

Fairclough, S. J., Rowlands, A. V., Taylor, S. & Boddy, L. M. Cut-point-free accelerometer metrics to assess children's physical activity: An example using the school day. *Scandinavian Journal of Medicine and Science in Sports* **30,** 117–125 (2020).

Fairclough, S. J., Taylor, S., Rowlands, A. V., Boddy, L. M. & Noonan, R. J. Average acceleration and intensity gradient of primary school children and associations with indicators of health and well-being. *Journal of Sports Sciences* **37,** 2159–2167 (2019).

Fairclough, S. J., Noonan, R. J., Rowlands, A. V., Van Hees, V., Knowles, Z. & Boddy, L. M. Wear Compliance and Activity in Children Wearing Wrist and Hip-Mounted Accelerometers. *Medicine & Science in Sports & Exercise* **48,** 245–253 (2016).

Fairclough, S. & Stratton, G. Physical activity levels in middle and high school physical education: a review. *Pediatric exercise science* **17,** 217–236 (2005).

Falck, R. S., Davis, J. C., Best, J. R., Crockett, R. A. & Liu-Ambrose, T. Impact of exercise training on physical and cognitive function among older adults: a systematic review and meta-analysis. *Neurobiology of Aging* **79,** 119–130 (2019).

Fanning, J., Porter, G., Awick, E. A., Ehlers, D. K., Roberts, S. A., Cooke, G., Burzynska, A. Z., Voss, M. W., Kramer, A. F. & McAuley, E. Replacing sedentary time with sleep, light, or moderate-to-vigorous physical activity: effects on self-regulation and executive functioning. *Journal of Behavioral Medicine* **40,** 332–342 (2017).

Farah, M. J. Socioeconomic status and the brain: prospects for neuroscience-informed policy. *Nature Reviews Neuroscience* **19,** 428–438 (2018).

Farooq, A., Martin, A., Janssen, X., Wilson, M. G., Gibson, A. M., Hughes, A. & Reilly, J. J. Longitudinal changes in moderate-to-vigorous-intensity physical activity in children and adolescents: A systematic review and meta-analysis. *Obesity Reviews* **21,** e12953 (2020).

Farooq, M. A., Parkinson, K. N., Adamson, A. J., Pearce, M. S., Reilly, J. K., Hughes, A. R., Janssen, X., Basterfield, L. & Reilly, J. J. Timing of the decline in physical activity in childhood and adolescence: Gateshead Millennium Cohort Study. *British Journal of Sports Medicine* **52,** 1002–1006 (2018).

Farrahi, V., Niemelä, M., Kangas, M., Korpelainen, R. & Jämsä, T. Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches. *Gait and Posture* **68,** 285–299 (2019).

Fawns-Ritchie, C. & Deary, I. J. Reliability and validity of the UK Biobank cognitive tests. *PLoS ONE* **15,** e0231627 (2020).

Fedewa, A. L. & Ahn, S. The effects of physical activity and physical fitness on children's achievement and cognitive outcomes: A meta-analysis. *Research quarterly for exercise and sport* **82,** 521–535 (2011).

Fernandes, D. J., Ellegood, J., Askalan, R., Blakely, R. D., Dicicco-Bloom, E., Egan, S. E., Osborne, L. R., Powell, C. M., Raznahan, A., Robins, D. M., Salter, M. W., Sengar, A. S., Veenstra-VanderWeele, J. & Lerch, J. P. Spatial gene expression analysis of neuroanatomical differences in mouse models. *NeuroImage* **163,** 220–230 (2017).

Ferreira, I., Van Der Horst, K., Wendel-Vos, W., Kremers, S., Van Lenthe, F. J. & Brug, J. Environmental correlates of physical activity in youth - A review and update. *Obesity Reviews* **8,** 129–154 (2007).

Fielding, S., Maclennan, G., Cook, J. A. & Ramsay, C. R. A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* **9,** 51 (2008).

Fiero, M. H., Huang, S., Oren, E. & Bell, M. L. Statistical analysis and handling of missing data in cluster randomized trials: A systematic review. *Trials* **17,** 72 (2016).

Filley, C. M. & Fields, R. D. White matter and cognition: Making the connection. *Journal of Neurophysiology* **116,** 2093–2104 (2016).

Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80,** 27–38 (1993).

Firth, J., Stubbs, B., Vancampfort, D., Schuch, F., Lagopoulos, J., Rosenbaum, S., Ward, P. B., Feter, N., Penny, J., Freitas, M. & Rombaldi, A. Effect of aerobic exercise on hippocampal volume in humans: A systematic review and meta-analysis. *NeuroImage* **166,** 230–238 (2018).

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B. & Dale, A. M. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33,** 341–355 (2002).

Fish, A. M., Nadig, A., Seidlitz, J., Reardon, P. K., Mankiw, C., McDermott, C. L., Blumenthal, J. D., Clasen, L. S., Lalonde, F., Lerch, J. P., Chakravarty, M. M., Shinohara, R. T. & Raznahan, A. Sex-biased trajectories of amygdalo-hippocampal morphology change over human development. *NeuroImage* **204,** 116122 (2020).

Fiuza-Luces, C., Garatachea, N., Berger, N. A. & Lucia, A. Exercise is the real polypill. *Physiology* **28,** 330–358 (2013).

Fjell, A. M., Sneve, M. H., Sederevicius, D., Sørensen, Ø., Krogsrud, S. K., Mowinckel, A. M. & Walhovd, K. B. Volumetric and microstructural regional changes of the hippocampus underlying development of recall performance after extended retention intervals. *Developmental Cognitive Neuroscience* **40,** 100723 (2019).

Focke, N. K., Helms, G., Kaspar, S., Diederich, C., Tóth, V., Dechent, P., Mohr, A. & Paulus, W. Multi-site voxel-based morphometry - Not quite there yet. *NeuroImage* **56,** 1164–1170 (2011).

Ford, I. & Norrie, J. Pragmatic Trials. *New England Journal of Medicine* **375,** 454–463 (2016).

Fortin, J. P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R. & Shinohara, R. T. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* **161,** 149–170 (2017).

Fotuhi, M., Do, D. & Jack, C. Modifiable factors that alter the size of the hippocampus with ageing. *Nature Reviews Neurology* **8,** 189–202 (2012).

Foulkes, L. & Blakemore, S. J. Studying individual differences in human adolescent brain development. *Nature Neuroscience* **21,** 315–323 (2018).

Franklin, B. A., Whaley, M. H., Howley, E. T. & Balady, G. J. *ACSM's Guidelines for Exercise Testing and Prescription (7th ed.)* (Lippincott Williams & Wilkins, Philadelphia, 2000).

Freedson, P., Pober, D. & Janz, K. F. Calibration of accelerometer output for children. *Medicine and Science in Sports and Exercise* **37,** 523–530 (2005).

Friedman, N. P., Miyake, A., Altamirano, L. J., Corley, R. P., Young, S. E., Rhea, S. A. & Hewitt, J. K. Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. *Developmental Psychology* **52,** 326–340 (2016).

Friedman, N. P. & Miyake, A. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex* **86,** 186–204 (2017).

Friedman, N. P., Miyake, A., Young, S. E., Defries, J. C., Corley, R. P. & Hewitt, J. K. Individual Differences in Executive Functions Are Almost Entirely Genetic in Origin. *Journal of Experimental psychology genetics* **137,** 201–225 (2008).

Friendly, M. & Fox, J. *candisc: Visualizing Generalized Canonical Discriminant and Canonical Correlation Analysis. R package version 0.8-3* 2020. https://cran.r-project.org/package=candisc.

Frodl, T., Strehl, K., Carballedo, A., Tozzi, L., Doyle, M., Amico, F., Gormley, J., Lavelle, G. & O'Keane, V. Aerobic exercise increases hippocampal subfield volumes in younger adults and prevents volume decline in the elderly. *Brain Imaging and Behavior,* 1–11 (2019).

Fuhrmann, D., Knoll, L. J. & Blakemore, S. J. Adolescence as a Sensitive Period of Brain Development. *Trends in Cognitive Sciences* **19,** 558–566 (2015).

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B. & Pee, D. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine* **15,** 1069–1092 (1996).

Galván, A., Van Leijenhorst, L. & McGlennen, K. M. Considerations for imaging the adolescent brain. *Developmental Cognitive Neuroscience* **2,** 293–302 (2012).

Gearing, R. E., El-Bassel, N., Ghesquiere, A., Baldwin, S., Gillies, J. & Ngeow, E. Major ingredients of fidelity: A review and scientific guide to improving quality of intervention research implementation. *Clinical Psychology Review* **31,** 79–88 (2011).

Gelman, A. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* **27,** 2865–2873 (2008).

Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press, Cambridge, 2006).

Germine, L., Nakayama, K., Duchaine, B. C. & Wilmer, J. B. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive / perceptual experiments. *Psychonomic Bulletin and Review* **19,** 847–857 (2012).

Gerrard, M., Gibbons, F. X., Houlihan, A. E., Stock, M. L. & Pomery, E. A. A dual-process approach to health risk decision making: The prototype willingness model. *Developmental Review* **28,** 29–61 (2008).

Geuze, E., Vermetten, E. & Bremner, J. D. MR-based in vivo hippocampal volumetrics: 1. Review of methodologies currently employed. *Molecular Psychiatry* **10,** 147–159 (2005).

Gevers, S., Van Osch, M. J., Bokkers, R. P., Kies, D. A., Teeuwisse, W. M., Majoie, C. B., Hendrikse, J. & Nederveen, A. J. Intra-and multicenter reproducibility of pulsed, continuous and pseudo-continuous arterial spin labeling methods for measuring cerebral perfusion. *Journal of Cerebral Blood Flow and Metabolism* **31,** 1706–1715 (2011).

Gibala, M. J., Little, J. P., Van Essen, M., Wilkin, G. P., Burgomaster, K. A., Safdar, A., Raha, S. & Tarnopolsky, M. A. Short-term sprint interval versus traditional endurance training: Similar initial adaptations in human skeletal muscle and exercise performance. *Journal of Physiology* **575,** 901–911 (2006).

Gibbons, F. X., Gerrard, M., Blanton, H. & Russell, D. W. Reasoned action and social reaction: willingness and intention as independent predictors of health risk. *Journal of personality and social psychology* **74,** 1164 (1998).

Gibbons, F. X., Gerrard, M. & McCoy, S. B. Prototype perception predicts (lack of) pregnancy prevention. *Personality and Social Psychology Bulletin* **21,** 85–93 (1995).

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., Paus, T., Evans, A. C. & Rapoport, J. L. Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience* **2,** 861–863 (1999).

Giedd, J. N. & Rapoport, J. L. Structural MRI of Pediatric Brain Development: What Have We Learned and Where Are We Going? *Neuron* **67,** 728–734 (2010).

Gilbert, P. E., Kesner, R. P. & Lee, I. Dissociating hippocampal subregions: A double dissociation between dentate gyrus and CA1. *Hippocampus* **11,** 626–636 (2001).

Gilmore, J. H., Knickmeyer, R. C. & Gao, W. Imaging structural and functional brain development in early childhood. *Nature Reviews Neuroscience* **19,** 123–137 (2018).

Giovanello, K. S., Schnyer, D. M. & Verfaellie, M. A critical role of the anterior hippocampus in relational memory: Evidence from an fMRI study comparing associative and item recognition. *Hippocampus* **14,** 5–8 (2004).

Giuliano, A., Donatelli, G., Cosottini, M., Tosetti, M., Retico, A. & Fantacci, M. E. Hippocampal subfields at ultra high field MRI: An overview of segmentation and measurement methods. *Hippocampus* **27,** 481–494 (2017).

GL Assessment. *Progress Test in Maths, Level 13* https://www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/ (2018).

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C. & Jenkinson, M. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80,** 105–124 (2013).

Global Council on Brain Health. *Brain health and mental well-being: GCBH recommendations on feeling good and functioning well* 2018. www . GlobalCouncilOnBrainHealth.org.

Goddings, A. L., Beltz, A., Peper, J. S., Crone, E. A. & Braams, B. R. Understanding the Role of Puberty in Structural and Functional Development of the Adolescent Brain. *Journal of Research on Adolescence* **29,** 32–53 (2019).

Goddings, A. L., Mills, K. L., Clasen, L. S., Giedd, J. N., Viner, R. M. & Blakemore, S. J. The influence of puberty on subcortical brain development. *NeuroImage* **88,** 242–251 (2014).

Goddings, A.-l., Roalf, D., Lebel, C. & Tamnes, C. K. Development of white matter microstructure and executive functions during childhood and adolescence: a review of diffusion MRI studies. *PsyArXiv* **October 28,** doi:10.31234/osf.io/kyxbq. (2020).

Godfrey, S. & Mearns, M. Pulmonary function and response to exercise in cystic fibrosis. *Archives of Disease in Childhood* **46,** 144–151 (1971).

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., Herman, D. H., Clasen, L. S., Toga, A. W., Rapoport, J. L. & Thompson, P. M. Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences* **101,** 8174–8179 (2004).

Goldstein, H., Carpenter, J., Kenward, M. G. & Levin, K. A. Multilevel models with multivariate mixed response types. *Statistical Modelling* **9,** 173–197 (2009).

Gomez-Pinilla, F. & Hillman, C. H. The influence of exercise on cognitive abilities. *Comprehensive Physiology* **3,** 403–428 (2013).

González-Villà, S., Oliver, A., Valverde, S., Wang, L., Zwiggelaar, R. & Lladó, X. A review on brain structures segmentation in magnetic resonance imaging. *Artificial Intelligence in Medicine* **73,** 45–69 (2016).

Good, C. D., Johnsrude, I. S., Ashburner, J., Henson, R. N., Friston, K. J. & Frackowiak, R. S. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage* **14,** 21–36 (2001).

Goodman, A., Lamping, D. L. & Ploubidis, G. B. When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the strengths and difficulties questionnaire (SDQ): Data from british parents, teachers and children. *Journal of Abnormal Child Psychology* **38,** 1179–1191 (2010).

Goodman, R. The Strengths and Difficulties Questionnaire: a research note. *Journal of Child Psychology & Psychiatry & Allied Disciplines* **38,** 581–586 (1997).

Gorham, L. S., Jernigan, T., Hudziak, J. & Barch, D. M. Involvement in Sports, Hippocampal Volume, and Depressive Symptoms in Children. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* **4,** 484–492 (2019).

Gosling, S. D., Vazire, S., Srivastava, S. & John, O. P. Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist* **59,** 93–104 (2004).

Graham, J. W. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology* **60,** 549–576 (2009).

Graham, J. W., Olchowski, A. E. & Gilreath, T. D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* **8,** 206–213 (2007).

Grant, S., Corbett, K., Amjad, A. M., Wilson, J. & Aitchison, T. A comparison of methods of predicting maximum oxygen uptake. *British Journal of Sports Medicine* **29,** 147–152 (1995).

Gray, H. M., LaPlante, D. A., Bannon, B. L., Ambady, N. & Shaffer, H. J. Development and validation of the alcohol identity implicit associations test (AI-IAT). *Addictive behaviors* **36,** 919–926 (2011).

Green, P. & Macleod, C. J. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution* **7,** 493–498 (2016).

Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., Shimony, J. S., Petersen, S. E., Schlaggar, B. L., Fair, D. A. & Dosenbach, N. U. Behavioral interventions for reducing head motion during MRI scans in children. *NeuroImage* **171,** 234–245 (2018).

Greenwald, A. G. & Farnham, S. D. Using the implicit association test to measure self-esteem and self-concept. *Journal of personality and social psychology* **79,** 1022 (2000).

Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. K. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* **74,** 1464 (1998).

Greenwald, A. G., Nosek, B. A. & Banaji, M. R. Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of personality and social psychology* **85,** 197 (2003).

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L. & Banaji, M. R. Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology* **97,** 17 (2009).

Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48,** 63–72 (2009).

Grgic, J., Dumuid, D., Bengoechea, E. G., Shrestha, N., Bauman, A., Olds, T. & Pedisic, Z. Health outcomes associated with reallocations of time between sleep, sedentary behaviour, and physical activity: A systematic scoping review of isotemporal substitution studies. *International Journal of Behavioral Nutrition and Physical Activity* **15,** 69 (2018).

Grissom, J. B. Physical fitness and academic achievement. *Journal of Exercise Physiology* **8,** 11–25 (2005).

Gronwald, T., Törpel, A., Herold, F. & Budde, H. Perspective of Dose and Response for Individualized Physical Exercise and Training Prescription. *Journal of Functional Morphology and Kinesiology* **5,** 48 (2020).

Grove, J. R. & Prapavessis, H. Preliminary evidence for the reliability and validity of an abbreviated Profile of Mood States. *International Journal of Sport Psychology* **23,** 93–109 (1992).

Grover, V. P., Tognarelli, J. M., Crossey, M. M., Cox, I. J., Taylor-Robinson, S. D. & McPhail, M. J. Magnetic Resonance Spectroscopy: Principles and Techniques: Lessons for Clinicians. *Journal of Clinical and Experimental Hepatology* **5,** 320–328 (2015).

Grund, S., Lüdtke, O. & Robitzsch, A. Multiple Imputation of Multilevel Missing Data: An Introduction to the R Package pan. *SAGE Open* **6,** 2158244016668220 (2016).

Grund, S., Lüdtke, O. & Robitzsch, A. Multiple Imputation of Missing Data at Level 2: A Comparison of Fully Conditional and Joint Modeling in Multilevel Designs. *Journal of Educational and Behavioral Statistics* **43,** 316–353 (2018).

Grund, S., Ludtke, O., Robitzsch, A., Lüdtke, O. & Robitzsch, A. Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. *Organizational Research Methods* **21,** 111–149 (2018).

Grussu, F., Schneider, T., Tur, C., Yates, R. L., Tachrount, M., Ianus, A., Yiannakas, M. C., Newcombe, J., Zhang, H., Alexander, D. C., DeLuca, G. C. & Gandini Wheeler-Kingshott, C. A. Neurite dispersion: a new marker of multiple sclerosis spinal cord pathology? *Annals of Clinical and Translational Neurology* **4,** 663–679 (2017).

Guiney, H. & Machado, L. Benefits of regular aerobic exercise for executive functioning in healthy populations. *Psychonomic Bulletin and Review* **20,** 73–86 (2013).

Gunnell, K. E., Poitras, V. J., LeBlanc, A., Schibli, K., Barbeau, K., Hedayati, N., Ponitfex, M. B., Goldfield, G. S., Dunlap, C., Lehan, E. & Tremblay, M. S. Physical activity and brain structure, brain function, and cognition in children and youth: A systematic review of randomized controlled trials. *Mental Health and Physical Activity* **16,** 105–127 (2018).

Gupta, N., Mathiassen, S. E., Mateu-Figueras, G., Heiden, M., Hallman, D. M., Jørgensen, M. B. & Holtermann, A. A comparison of standard and compositional data analysis in studies addressing group differences in sedentary behavior and physical activity. *International Journal of Behavioral Nutrition and Physical Activity* **15,** 1–12 (2018).

Guthold, R., Stevens, G. A., Riley, L. M. & Bull, F. C. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *The Lancet Global Health* **6,** e1077–e1086 (2018).

Guthold, R., Stevens, G. A., Riley, L. M. & Bull, F. C. Global trends in insufficient physical activity among adolescents: a pooled analysis of 298 population-based surveys with 1.6 million participants. *The Lancet Child & Adolescent Health* **9,** 335–338 (2020).

Haacke, E. M., Cheng, N. Y., House, M. J., Liu, Q., Neelavalli, J., Ogg, R. J., Khan, A., Ayaz, M., Kirsch, W. & Obenaus, A. Imaging iron stores in the brain using magnetic resonance imaging. *Magnetic Resonance Imaging* **23,** 1–25 (2005).

Hackert, V. H., Den Heijer, T., Oudkerk, M., Koudstaal, P. J., Hofman, A. & Breteler, M. M. Hippocampal head size associated with verbal memory performance in nondemented elderly. *NeuroImage* **17,** 1365–1372 (2002).

Hackman, D. A., Farah, M. J. & Meaney, M. J. Socioeconomic status and the brain: Mechanistic insights from human and animal research. *Nature Reviews Neuroscience* **11,** 651–659 (2010).

Hagger, M. S., Chatzisarantis, N., Biddle, S. J. & Orbell, S. Antecedents of children's physical activity intentions and behaviour: Predictive validity and longitudinal effects. *Psychology and Health* **16,** 391–407 (2001).

Hahn, S., Puffer, S., Torgerson, D. J. & Watson, J. Methodological bias in cluster randomised trials. *BMC Medical Research Methodology* **5,** 1–8 (2005).

Hallal, P. C., Andersen, L. B., Bull, F. C., Guthold, R., Haskell, W., Ekelund, U., Alkandari, J. R., Bauman, A. E., Blair, S. N., Brownson, R. C., Craig, C. L., Goenka, S., Heath, G. W., Inoue, S., Kahlmeier, S., Katzmarzyk, P. T., Kohl, H. W., Lambert, E. V., Lee, I. M., Leetongin, G., Lobelo, F., Loos, R. J., Marcus, B., Martin, B. W., Owen, N., Parra, D. C., Pratt, M., Puska, P., Ogilvie, D., Reis, R. S., Sallis, J. F., Sarmiento, O. L. & Wells, J. C. Global physical activity levels: Surveillance progress, pitfalls, and prospects. *The Lancet* **380,** 247–257 (2012).

Hamer, M., Sharma, N. & Batty, G. D. Association of objectively measured physical activity with brain structure: UK Biobank study. *Journal of Internal Medicine* **284,** 439–443 (2018).

Han, X., Jovicich, J., Salat, D., Van Der Kouwe, A., Quinn, B., Czanner, S., Busa, E., Pacheco, J., Albert, M., Killiany, R., Maguire, P., Rosas, D., Makris, N., Dale, A., Dickerson, B. & Fischl, B. Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage* **32,** 180–194 (2006).

Harding, S. K., Page, A. S., Falconer, C. & Cooper, A. R. Longitudinal changes in sedentary time and physical activity during adolescence. *International Journal of Behavioral Nutrition and Physical Activity* **12,** 44 (2015).

Harkins, K. D., Xu, J., Dula, A. N., Li, K., Valentine, W. M., Gochberg, D. F., Gore, J. C. & Does, M. D. The microstructural correlates of T1 in white matter. *Magnetic Resonance in Medicine* **75,** 1341–1345 (2016).

Harrell, J. S., Mcmurray, R. G., Baggett, C. D., Pennell, M. L., Pearce, P. F. & Bangdiwala, S. I. Energy Costs of Physical Activities in Children and Adolescents. *Medicine & Science in Sports & Exercise* **37,** 329–336 (2005).

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., Robinson, B. S., Hodgson, D. J. & Inger, R. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ* **6,** e4794 (2018).

Hart, P. D. A Canonical Correlation Analysis of Physical Activity Parameters and Body Composition Measures in College Students. *American Journal of Sports Science and Medicine* **5,** 64–68 (2017).

Harter, S. *Causes and consequences of low self-esteem in children and adolescents.* 87–116 (Springer, 1993).

Hartung, J., Engelhardt, L. E., Thibodeaux, M. L., Harden, K. P. & Tucker-Drob, E. M. Developmental transformations in the structure of executive functions. *Journal of Experimental Child Psychology* **189,** 104681 (2020).

Hassevoort, K. M., Khan, N. A., Hillman, C. H., Kramer, A. F. & Cohen, N. J. Relational Memory is Associated with Academic Achievement in Preadolescent Children. *Trends in Neuroscience and Education* **13,** 8–16 (2018).

Hassevoort, K. M., Khazoum, S. E., Walker, J. A., Barnett, S. M., Raine, L. B., Hammond, B. R., Renzi-Hammond, L. M., Kramer, A. F., Khan, N. A., Hillman, C. H. & Cohen, N. J. Macular Carotenoids, Aerobic Fitness, and Central Adiposity Are Associated Differentially with Hippocampal-Dependent Relational Memory in Preadolescent Children. *The Journal of Pediatrics* **183,** 108–114 (2017).

Hatch, S. L., Jones, P. B., Kuh, D., Hardy, R., Wadsworth, M. E. & Richards, M. Childhood cognitive ability and adult mental health in the British 1946 birth cohort. *Social Science and Medicine* **64,** 2285–2296 (2007).

Haverkamp, B. F., Wiersma, R., Vertessen, K., Van Ewijk, H., Oosterlaan, J. & Hartman, E. Effects of physical activity interventions on cognitive outcomes and academic performance in adolescents and young adults: A meta-analysis. *Journal of Sports Sciences* **38,** 2637–2660 (2020).

Hayat, S. A., Luben, R., Dalzell, N., Moore, S., Hogervorst, E., Matthews, F. E., Wareham, N., Brayne, C. & Khaw, K. T. Understanding the relationship between cognition and death: a within cohort examination of cognitive measures and mortality. *European Journal of Epidemiology* **33,** 1049–1062 (2018).

Hayes, A. F. A primer on multilevel modeling. *Human Communication Research* **32,** 385–410 (2006).

Hayes, R. & Moulton, L. H. *Cluster Randomised Trials* (Chapman and Hall/CRC, 2017).

Heale, R. & Twycross, A. Validity and reliability in quantitative studies. *Evidence-Based Nursing* **18,** 66–67 (2015).

Heath, F., Hurley, S. A., Johansen-Berg, H. & Sampaio-Baptista, C. Advances in noninvasive myelin imaging. *Developmental Neurobiology* **78,** 136–151 (2018).

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S. & Sumner, P. Low and variable correlation between reaction time costs and accuracy costs explained by accumulation models: Meta-analysis and simulations. *Psychological Bulletin* **144,** 1200–1227 (2018).

Hedge, C., Powell, G. & Sumner, P. The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods* **50,** 1166–1186 (2018).

Hedges, L. V. & Hedberg, E. C. Intraclass Correlation Values for Planning Group-Randomized Trials in Education. *Educational Evaluation and Policy Analysis* **29,** 60–87 (2007).

Hedman, A. M., Van Haren, N. E., Schnack, H. G., Kahn, R. S. & Hulshoff Pol, H. E. Human brain changes across the life span: A review of 56 longitudinal magnetic resonance imaging studies. *Human Brain Mapping* **33,** 1987–2002 (2012).

Heinze, K., Cumming, J., Dosanjh, A., Palin, S., Poulton, S., Bagshaw, A. P. & Broome, M. R. Neurobiological evidence of longer-term physical activity interventions on mental health outcomes and cognition in young people: A systematic review of randomised controlled trials. *Neuroscience & Biobehavioral Reviews* (2020).

Heitz, R. P. The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience* **8,** 150 (2014).

Helmer, M., Warrington, S., Lisa, J., Howell, A. & Rosand, B. Reliable estimation of Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS) with application to brain-behavior associations. *bioRxiv,* doi: https://doi.org/10.1101/2020.08.25.265546 (2020).

Helms, G., Dathe, H. & Dechent, P. Quantitative FLASH MRI at 3T using a rational approximation of the Ernst equation. *Magnetic Resonance in Medicine* **59,** 667–672 (2008).

Helms, G., Dathe, H., Kallenberg, K. & Dechent, P. High-resolution maps of magnetization transfer with inherent correction for RF inhomogeneity and T1 relaxation obtained from 3D FLASH MRI. *Magnetic Resonance in Medicine* **60,** 1396–1407 (2008).

Hemming, K., Eldridge, S., Forbes, G., Weijer, C. & Taljaard, M. How to design efficient cluster randomised trials. *BMJ* **358,** j3064 (2017).

Henley, S. M., Ridgway, G. R., Scahill, R. I., Klöppel, S., Tabrizi, S. J., Fox, N. C. & Kassubek, J. Pitfalls in the use of voxel-based morphometry as a biomarker: Examples from Huntington disease. *American Journal of Neuroradiology* **31,** 711–719 (2010).

Hernán, M. A. & Hernández-Díaz, S. Beyond the intention-to-treat in comparative effectiveness research. *Clinical Trials* **9,** 48–55 (2012).

Hernandez-Fernandez, M. *cuDIMOT* https://users.fmrib.ox.ac.uk/%7B~%7Dmoisesf/cudimot/DesignModel.html (2018).

Herold, F., Gronwald, T., Scholkmann, F., Zohdi, H., Wyser, D., Müller, N. G. & Hamacher, D. New directions in exercise prescription: Is there a role for brain-derived parameters obtained by functional near-infrared spectroscopy? *Brain Sciences* **10,** 14–23 (2020).

Herold, F., Müller, P., Gronwald, T. & Müller, N. G. Dose–Response Matters! – A Perspective on the Exercise Prescription in Exercise–Cognition Research. *Frontiers in Psychology* **10** (2019).

Herold, F., Törpel, A., Schega, L. & Müller, N. G. Functional and/or structural brain changes in response to resistance exercises and resistance training lead to cognitive improvements - A systematic review. *European Review of Aging and Physical Activity* **16,** 10 (2019).

Herting, M. M., Johnson, C., Mills, K. L., Vijayakumar, N., Dennison, M., Liu, C., Goddings, A. L., Dahl, R. E., Sowell, E. R., Whittle, S., Allen, N. B. & Tamnes, C. K. Development of subcortical volumes across adolescence in males and females: A multisample study of longitudinal changes. *NeuroImage* **172,** 194–205 (2018).

Herting, M. M., Keenan, M. F. & Nagel, B. J. Aerobic Fitness Linked to Cortical Brain Development in Adolescent Males: Preliminary Findings Suggest a Possible Role of BDNF Genotype. *Frontiers in Human Neuroscience* **10,** 327 (2016).

Herting, M. M., Kim, R., Uban, K. A., Kan, E., Binley, A. & Sowell, E. R. Longitudinal changes in pubertal maturation and white matter microstructure. *Psychoneuroendocrinology* **81,** 70–79 (2017).

Herting, M. M. & Nagel, B. J. Aerobic fitness relates to learning on a virtual Morris Water Task and hippocampal volume in adolescents. *Behavioural Brain Research* **233,** 517–525 (2012).

Herting, M. M., Colby, J. B., Sowell, E. R. & Nagel, B. J. White matter connectivity and aerobic fitness in male adolescents. *Developmental Cognitive Neuroscience* **7,** 65–75 (2014).

Hidding, L. M., Chinapaw, M. J., Van Poppel, M. N., Mokkink, L. B. & Altenburg, T. M. An Updated Systematic Review of Childhood Physical Activity Questionnaires. *Sports Medicine* **48,** 2797–2842 (2018).

Hill, A. V. & Lupton, H. Muscular Exercise, Lactic Acid, and the Supply and Utilization of Oxygen. *QJM: An International Journal of Medicine* **16,** 135–171 (1923).

Hill, A. V., Long, C. N. H. & Lupton, H. Muscular Exercise, Lactic Acid, and the Supply and Utilisation of Oxygen. *Proceedings of the Royal Society of London* **96,** 438–475 (1924).

Hillman, C. H., Buck, S. M., Themanson, J. R., Pontifex, M. B. & Castelli, D. M. Aerobic Fitness and Cognitive Development: Event-Related Brain Potential and Task Performance Indices of Executive Control in Preadolescent Children. *Developmental Psychology* **45,** 114–129 (2009).

Hillman, C. H., Erickson, K. I. & Hatfield, B. D. Run for Your Life! Childhood Physical Activity Effects on Brain and Cognition. *Kinesiology Review* **6,** 12–21 (2017).

Hillman, C. H., Pontifex, M. B., Castelli, D. M., Khan, N. A., Raine, L. B., Scudder, M. R., Drollette, E. S., Moore, R. D., Wu, C.-T. C. T. & Kamijo, K. Effects of the FITKids Randomized Controlled Trial on Executive Control and Brain Function. *Pediatrics* **134,** e1063–e1071 (2014).

Hillman, C. H., Erickson, K. I. & Kramer, A. F. Be smart, exercise your heart: exercise effects on brain and cognition. *Nature Review Neuroscience* **9,** 58–65 (2008).

Hillman, C., Kamijo, K. & Scudder, M. A Review of Chronic and Acute Physical Activity Participation on Neuroelectric Measures of Brain Health and Cognition during Childhood. *Preventive Medicine* **52,** S21–S28 (2011).

Hills, A. P., Dengel, D. R. & Lubans, D. R. Supporting Public Health Priorities: Recommendations for Physical Education and Physical Activity Promotion in Schools. *Progress in Cardiovascular Diseases* **57,** 368–374 (2015).

Hills, A. P., Mokhtar, N. & Byrne, N. M. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. *Frontiers in Nutrition* **1,** 1–16 (2014).

Hindmarch, P., Hawkins, A., McColl, E., Hayes, M., Majsak-Newman, G., Ablewhite, J., Deave, T. & Kendrick, D. Recruitment and retention strategies and the examination of attrition bias in a randomised controlled trial in children's centres serving families in disadvantaged areas of England. *Trials* **16,** 79 (2015).

Hobbiss, M. H., Massonnié, J., Tokuhama-Espinosa, T., Gittner, A., De Sousa Lemos, M. A., Tovazzi, A., Hindley, C., Baker, S., Sumeracki, M. A., Wassenaar, T. & Gous, I. "UNIFIED": Bridging the Researcher- Practitioner Divide in Mind, Brain, and Education. *Mind, Brain, and Education* **13,** 298–312 (2019).

Hobbs, G. & Vignoles, A. *Is Free School Meal Status a Valid Proxy for Socioeconomic Status (in Schools Research)?* 46 (Centre for the Economics of Education, London School of Economics and Political Science, 2007).

Hodkinson, D. J., O'Daly, O., Zunszain, P. A., Pariante, C. M., Lazurenko, V., Zelaya, F. O., Howard, M. A. & Williams, S. C. Circadian and homeostatic modulation of functional connectivity and regional cerebral blood flow in humans under normal entrained conditions. *Journal of Cerebral Blood Flow and Metabolism* **34,** 1493–1499 (2014).

Hogan, M. J., O'Hora, D., Kiefer, M., Kubesch, S., Kilmartin, L., Collins, P. & Dimitrova, J. The effects of cardiorespiratory fitness and acute aerobic exercise on executive functioning and EEG entropy in adolescents. *Frontiers in Human Neuroscience* **9,** 538 (2015).

Högström, G., Nordström, A. & Nordström, P. High aerobic fitness in late adolescence is associated with a reduced risk of myocardial infarction later in life: a nationwide cohort study in men. *European Heart Journal* **35,** 3133–3140 (2014).

Hollis, J. L., Sutherland, R., Williams, A. J., Campbell, E., Nathan, N., Wolfenden, L., Morgan, P. J., Lubans, D. R., Gillham, K. & Wiggers, J. A systematic review and meta-analysis of moderate-to-vigorous physical activity levels in secondary school physical education lessons. *International Journal of Behavioral Nutrition and Physical Activity* **14,** 52 (2017).

Horien, C., Shen, X., Scheinost, D. & Constable, R. T. The individual functional connectome is unique and stable over months to years. *NeuroImage* **189,** 676–687 (2019).

Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28,** 321 (1936).

Hötting, K. & Röder, B. Beneficial effects of physical exercise on neuroplasticity and cognition. *Neuroscience and Biobehavioral Reviews* **37,** 2243–2257 (2013).

Howley, E. T. Type of activity: Resistance, aerobic and leisure versus occupational physical activity. *Medicine and Science in Sports and Exercise* **33,** S364–S369 (2001).

Huang, T., Larsen, K. T., Jepsen, J. R. M., Møller, N. C., Thorsen, A. K., Mortensen, E. L. & Andersen, L. B. Effects of an obesity intervention program on cognitive function in children: A randomized controlled trial. *Obesity* **23,** 2101–2108 (2015).

Hudziak, J. J., Albaugh, M. D., Ducharme, S., Karama, S., Spottswood, M., Crehan, E., Evans, A. C. & Botteron, K. N. Cortical thickness maturation and duration of music training: Health-promoting activities shape brain development. *Journal of the American Academy of Child and Adolescent Psychiatry* **53,** 1153–1161 (2014).

Hughes, M. M., Linck, J. A., Bowles, A. R., Koeth, J. T. & Bunting, M. F. Alternatives to switch-cost scoring in the task-switching paradigm: Their reliability and increased validity. *Behavior Research Methods* **46,** 702–721 (2014).

Hughes, R. A., Heron, J., Sterne, J. A. & Tilling, K. Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *International Journal of Epidemiology* **48,** 1294–1304 (2019).

Huh, M. H. & Jhun, M. Random permutation testing in multiple linear regression. *Communications in Statistics - Theory and Methods* **30,** 2023–2032 (2001).

Huizinga, M., Dolan, C. V. & Van Der Molen, M. W. Age-related change in executive function: Developmental trends and a latent variable analysis. *Neuropsychologia* **44,** 2017–2036 (2006).

Huque, H., Carlin, J. B., Simpson, J. A. & Lee, K. J. A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology* **18,** 168 (2018).

Husain, F. *Fit to Study. Evaluation protocol* tech. rep. (NatCen, 2017). https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/fit-to-study/.

Husain, F., Bartasevicius, V., Marshall, L., Chidley, S. & Forsyth, E. *Fit to study. Evaluation report.* tech. rep. September (NatCen Social Research, 2019). https://educationendowmentfoundation.org.uk/public/files/FitToStudy.pdf.

Husson, F., Josse, J., Narasimhan, B. & Robin, G. Imputation of Mixed Data With Multilevel Singular Value Decomposition. *Journal of Computational and Graphical Statistics* **28,** 552–566 (2019).

Hutchinson, E. B., Schwerin, S. C., Avram, A. V., Juliano, S. L. & Pierpaoli, C. Diffusion MRI and the detection of alterations following traumatic brain injury. *Journal of Neuroscience Research* **96,** 612–625 (2018).

Hutton, C., Draganski, B., Ashburner, J. & Weiskopf, N. A comparison between voxel-based cortical thickness and voxel-based morphometry in normal aging. *NeuroImage* **48,** 371–380 (2009).

Ilie, S., Sutherland, A. & Vignoles, A. Revisiting free school meal eligibility as a proxy for pupil socio-economic deprivation. *British Educational Research Journal* **43,** 253–274 (2017).

Inchley, J., Currie, D., Young, T., Samdal, O., Torsheim, T., Augustson, L., Mathison, F., Aleman-Diaz, A., Molcho, M., Weber, M. & Barnekow, V. Growing up unequal: Health Behaviour in School-aged Children (HBSC) study: International report from the 2013/2014 survey. *Health Policy for Children and Adolescents,* 1–277 (2016).

Ishihara, T., Sugasawa, S., Matsuda, Y. & Mizuno, M. Improved Executive Functions in 6–12-Year-Old Children Following Cognitively Engaging Tennis Lessons. *Journal of sports sciences* **35,** 2014–2020 (2017).

Ishihara, T., Sugasawa, S., Matsuda, Y. & Mizuno, M. Relationship between sports experience and executive function in 6–12-year-old children: independence from physical fitness and moderation by gender. *Developmental Science* **21,** e12555 (2018).

Ivers, N. M., Halperin, I. J., Barnsley, J., Grimshaw, J. M., Shah, B. R., Tu, K., Upshur, R. & Zwarenstein, M. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials* **13,** 1–9 (2012).

Jack, C. R., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Lowe, V., Kantarci, K., Bernstein, M. A., Senjem, M. L., Gunter, J. L., Boeve, B. F., Trojanowski, J. Q., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C. & Knopman, D. S. Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of Neurology* **69,** 856–867 (2012).

Jackson, W. M., Davis, N., Sands, S. A., Whittington, R. A. & Sun, L. S. Physical Activity and Cognitive Development: A Meta-Analysis. *Journal of Neurosurgical Anesthesiology* **28,** 373–380 (2016).

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

Jager, K. J., Zoccali, C., MacLeod, A. & Dekker, F. W. Confounding: What it is and how to deal with it. *Kidney International* **73,** 256–260 (2008).

Jago, R., Fairclough, S. J., Mackintosh, K. A., Mccrorie, P., Sebire, S. J., Sherar, L. B., Van Sluijs, E., Williams, C. A. & Banfield, K. *Expert Working Group Working Paper Children and Young People UK physical activity guidelines: draft review and recommendations for Children and Young People* tech. rep. (2018). http://www.bristol.ac.uk/media-library/sites/sps/documents/cmo/children-young-people-technical-report.pdf.

Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkel, P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. *BMC Medical Research Methodology* **17,** 162 (2017).

Janssen, I. & LeBlanc, a. G. Systematic review of the health benefits of physical activity and fitness in school-aged children and youth. *International Journal of Behavioral Nutrition and Physical Activity* **7,** 40 (2010).

Janssen, M., Toussaint, H. M., Van Mechelen, W. & Verhagen, E. A. Effects of acute bouts of physical activity on children's attention: a systematic review of the literature. *SpringerPlus* **3,** 410 (2014).

Janszky, I., Ahlbom, A. & Svensson, A. C. The Janus face of statistical adjustment: Confounders versus colliders. *European Journal of Epidemiology* **25,** 361–363 (2010).

Jbabdi, S. & Johansen-Berg, H. Tractography: Where Do We Go from Here? *Brain Connectivity* **1,** 169–183 (2011).

Jelescu, I. O. & Budde, M. D. Design and Validation of Diffusion MRI Models of White Matter. *Frontiers in Physics* **5,** 61 (2017).

Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* **17,** 825–841 (2002).

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W. & Smith, S. M. FSL. *NeuroImage* **62,** 782–90 (2012).

Jeon, Y. K. & Ha, C. H. The effect of exercise intensity on brain derived neurotrophic factor and memory in adolescents. *Environmental Health and Preventive Medicine* **22,** 1–6 (2017).

Jezzard, P., Chappell, M. A. & Okell, T. W. Arterial spin labeling for the measurement of cerebral perfusion and angiography. *Journal of Cerebral Blood Flow & Metabolism* **38,** 603–626 (2017).

Jiang, C., Zhang, L., Zou, C., Long, X., Liu, X., Zheng, H., Liao, W. & Diao, Y. Diurnal microstructural variations in healthy adult brain revealed by diffusion tensor imaging. *PLoS ONE* **9,** e84822 (2014).

Johansen-Berg, H., Baptista, C. S. & Thomas, A. G. Human Structural Plasticity at Record Speed. *Neuron* **73,** 1058–1060 (2012).

Johnson, M. H. Functional brain development in humans. *Nature Reviews Neuroscience* **2,** 475–483 (2001).

Jones, D. K., Knösche, T. R. & Turner, R. White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *NeuroImage* **73,** 239–254 (2013).

Jones, M., Defever, E., Letsinger, A., Steele, J. & Mackintosh, K. A. A mixed-studies systematic review and meta-analysis of school-based interventions to promote physical activity and/or reduce sedentary time in children. *Journal of Sport and Health Science* **9,** 3–17 (2019).

Josse, J., Pagès, J. & Husson, F. Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification* **5,** 231–246 (2011).

Kaczkurkin, A. N., Raznahan, A. & Satterthwaite, T. D. Sex differences in the developing brain: insights from multimodal neuroimaging. *Neuropsychopharmacology* **44,** 71–85 (2019).

Kahan, B. C., Forbes, G., Ali, Y., Jairath, V., Bremner, S., Harhay, M. O., Hooper, R., Wright, N., Eldridge, S. M. & Leyrat, C. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: A review, reanalysis, and simulation study. *Trials* **17,** 438 (2016).

Kahan, B. C. & Morris, T. P. Improper analysis of trials randomised using stratified blocks or minimisation. *Statistics in Medicine* **31,** 328–340 (2012).

Kahlert, J., Gribsholt, S. B., Gammelager, H., Dekkers, O. M. & Luta, G. Control of confounding in the analysis phase – an overview for clinicians. *Clinical Epidemiology* **9,** 195–204 (2017).

Kail, R. Development of processing speed in childhood and adolescence. *Advances in child development and behavior* **23,** 151–185 (1991).

Kamijo, K. & Masaki, H. Fitness and ERP Indices of Cognitive Control Mode during Task Preparation in Preadolescent Children. *Frontiers in Human Neuroscience* **10,** 441 (2016).

Kandola, A., Hendrikse, J., Lucassen, P. J. & Yücel, M. Aerobic Exercise as a Tool to Improve Hippocampal Plasticity and Function in Humans: Practical Implications for Mental Health Treatment. *Frontiers in Human Neuroscience* **10,** 373 (2016).

Kao, S. C., Cadenas-Sanchez, C., Shigeta, T. T., Walk, A. M., Chang, Y. K., Pontifex, M. B. & Hillman, C. H. A systematic review of physical activity and cardiorespiratory fitness on P3b. *Psychophysiology* **57,** e13425 (2020).

Kao, S.-C., Drollette, E. S., Scudder, M. R., Raine, L. B., Westfall, D. R., Pontifex, M. B. & Hillman, C. H. Aerobic Fitness Is Associated With Cognitive Control Strategy in Preadolescent Children. *Journal of Motor Behavior* **49,** 150–162 (2017).

Karch, J. D., Filevich, E., Wenger, E., Lisofsky, N., Becker, M., Butler, O., Mårtensson, J., Lindenberger, U., Brandmaier, A. M. & Kühn, S. Identifying predictors of within-person variance in MRI-based brain volume estimates. *NeuroImage* **200,** 575–589 (2019).

Karr, J. E., Areshenkoff, C. N., Rast, P., Hofer, S. M., Iverson, G. L. & Garcia-Barrera, M. A. The unity and diversity of executive functions: A systematic review and re-analysis of latent variable studies. *Psychological Bulletin* **144,** 1147–1185 (2018).

Keenan, K. E., Gimbutas, Z., Dienstfrey, A. & Stupic, K. F. Assessing effects of scanner upgrades for clinical studies. *Journal of Magnetic Resonance Imaging* **50,** 1948–1954 (2019).

Keith, N. R., Clark, D. O., Stump, T. E., Miller, D. K. & Callahan, C. M. Validity and Reliability of the Self-Reported Physical Fitness (SRFit) Survey. *Journal of Physical Activity & Health* **11,** 853–859 (2015).

Kempermann, G. The neurogenic reserve hypothesis: what is adult hippocampal neurogenesis good for? *Trends in Neurosciences* **31,** 163–169 (2008).

Kenward, M. G. & Carpenter, J. Multiple imputation: current perspectives. *Statistical Methods in Medical Research* **16,** 199–218 (2007).

Khan, A., Lee, E. Y. & Tremblay, M. S. Meeting 24-h movement guidelines and associations with health related quality of life of Australian adolescents. *Journal of Science and Medicine in Sport* (2020).

Khan, N. A. & Hillman, C. H. The relation of childhood physical activity and aerobic fitness to brain function and cognition: a review. *Pediatric exercise science* **26,** 138–46 (2014).

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M. & Jost, K. Control and Interference in Task Switching-A Review. *Psychological Bulletin* **136,** 849–874 (2011).

Killgore, W. D. S., Olson, E. A. & Weber, M. Physical exercise habits correlate with gray matter volume of the hippocampus in healthy adult humans. *Scientific Reports* **3,** 1–6 (2013).

Kinghorn, A., Shanaube, K., Toska, E., Cluver, L. & Bekker, L. G. Defining adolescence: priorities from a global health perspective. *The Lancet Child and Adolescent Health* **2,** e10 (2018).

Kleemeyer, M. M., Kühn, S., Prindle, J., Bodammer, N. C., Brechtel, L., Garthe, A., Kempermann, G., Schaefer, S. & Lindenberger, U. Changes in fitness are associated with changes in hippocampal microstructure and hippocampal volume among older adults. *NeuroImage* **131,** 155–161 (2016).

Knickmeyer, R. C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J. K., Hamer, R. M., Lin, W., Gerig, G. & Gilmore, J. H. A structural MRI study of human brain development from birth to 2 years. *Journal of Neuroscience* **28,** 12176–12182 (2008).

Kohl, H. W., Craig, C. L., Lambert, E. V., Inoue, S., Alkandari, J. R., Leetongin, G., Kahlmeier, S., Andersen, L. B., Bauman, A. E., Blair, S. N., Brownson, R. C., Bull, F. C., Ekelund, U., Goenka, S., Guthold, R., Hallal, P. C., Haskell, W. L., Heath, G. W., Katzmarzyk, P. T., Lee, I. M., Lobelo, F., Loos, R. J., Marcus, B., Martin, B. W., Owen, N., Parra, D. C., Pratt, M., Puska, P., Ogilvie, D., Reis, R. S., Sallis, J. F., Sarmiento, O. L. & Wells, J. C. The pandemic of physical inactivity: Global action for public health. *The Lancet* **380,** 294–305 (2012).

Konkel, A. & Cohen, N. J. Relational memory and the hippocampus: Representations and methods. *Frontiers in Neuroscience* **3,** 166–174 (2009).

Konrad, C., Ukas, T., Nebel, C., Arolt, V., Toga, A. W. & Narr, K. L. Defining the human hippocampus in cerebral magnetic resonance images-An overview of current segmentation protocols. *NeuroImage* **47,** 1185–1195 (2009).

Koo, T. K. & Li, M. Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine* **15,** 155–63 (2016).

Kosmidis, I. Bias in parametric estimation: Reduction and useful side-effects. *Wiley Interdisciplinary Reviews: Computational Statistics* **6,** 185–196 (2014).

Kosmidis, I. *brglm2: Bias Reduction in Generalized Linear Models, R package version 0.7.0* 2020. https://cran.r-project.org/package=brglm2.

Kosmidis, I., Kenne Pagui, E. C. & Sartori, N. Mean and median bias reduction in generalized linear models. *Statistics and Computing* **30,** 43–59 (2020).

Kovalchik, S. *RISmed: download content from NCBI databases. R package version 2.1.7* 2017. https://cran.r-project.org/package=RISmed.

Kozareva, D. A., Cryan, J. F. & Nolan, Y. M. Born this way: Hippocampal neurogenesis across the lifespan. *Aging Cell* **18,** 1–18 (2019).

Krafft, C. E., Schaeffer, D. J., Schwarz, N. F., Chi, L., Weinberger, A. L., Pierce, J. E., Rodrigue, A. L., Allison, J. D., Yanasak, N. E., Liu, T., Davis, C. L. & McDowell, J. E. Improved frontoparietal white matter integrity in overweight children is associated with attendance at an after-school exercise program. *Developmental Neuroscience* **36,** 1–9 (2014).

Krafft, C. E., Schwarz, N. F., Chi, L., Weinberger, A. L., Schaeffer, D. J., Pierce, J. E., Rodrigue, A. L., Yanasak, N. E., Miller, P. H., Tomporowski, P. D., Davis, C. L. & McDowell, J. E. An 8-Month Randomized Controlled Exercise Trial Alters Brain Activation During Cognitive Tasks in Overweight Children. *Obesity* **22,** 232–242 (2014).

Kramer, A. F., Colcombe, S. & Kramer, A. F. Fitness Effects on the Cognitive Function of Older Adults: A Meta-Analytic Study—Revisited. *Perspectives on Psychological Science* **14,** 213–217 (2018).

Kriemler, S., Meyer, U., Martin, E., Van Sluijs, E. M., Andersen, L. B. & Martin, B. W. Effect of school-based interventions on physical activity and fitness in children and adolescents: A review of reviews and systematic update. *British Journal of Sports Medicine* **45,** 923–930 (2011).

Kriemler, S., Zahner, L., Schindler, C., Meyer, U., Hartmann, T., Hebestreit, H., Brunner-La Rocca, H. P., Van Mechelen, W. & Puder, J. J. Effect of school based physical activity programme (KISS) on fitness and adiposity in primary schoolchildren: cluster randomised controlled trial. *BMJ* **340,** c785 (2010).

Kropko, J., Goodrich, B., Gelman, A. & Hill, J. Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Political Analysis* **22,** 497–519 (2014).

Krueger, G., Granziera, C., Jack, C. R., Gunter, J. L., Littmann, A., Mortamet, B., Kannengiesser, S., Sorensen, A. G., Ward, C. P., Reyes, D. A., Britson, P. J., Fischer, H. & Bernstein, M. A. Effects of MRI scan acceleration on brain volume measurement consistency. *Journal of Magnetic Resonance Imaging* **36,** 1234–1240 (2012).

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82,** 1–26 (2017).

Kwak, L., Kremers, S. P., Bergman, P., Ruiz, J. R., Rizzo, N. S. & Sjöström, M. Associations between physical activity, fitness, and academic achievement. *Journal of Pediatrics* **155,** 19–24 (2009).

Kyllonen, P. C. & Zu, J. Use of response time for measuring cognitive ability. *Journal of Intelligence* **4,** 14 (2016).

Laird, N. M. & Ware, J. H. Random-Effects Models for Longitudinal Data. *Biometrics* **38,** 963–974 (1982).

Lambert, J. D., Greaves, C. J., Farrand, P., Cross, R., Haase, A. M. & Taylor, A. H. Assessment of fidelity in individual level behaviour change interventions promoting physical activity among adults: A systematic review. *BMC Public Health* **17,** 765 (2017).

Lambourne, K. & Tomporowski, P. The effect of exercise-induced arousal on cognitive task performance : A meta-regression analysis. *Brain Research* **1341,** 12–24 (2010).

Lang, C., Kalak, N., Brand, S., Holsboer-Trachsler, E., Pühse, U. & Gerber, M. The relationship between physical activity and sleep from mid adolescence to early adulthood. A systematic review of methodological approaches and meta-analysis. *Sleep Medicine Reviews* **28,** 32–45 (2016).

Lang, J. J., Tomkinson, G. R., Janssen, I., Ruiz, J. R., Ortega, F. B., Léger, L. & Tremblay, M. S. Making a Case for Cardiorespiratory Fitness Surveillance among Children and Youth. *Exercise and Sport Sciences Reviews* **46,** 66–75 (2018).

Lang, J. J., Tremblay, M. S., Ortega, F. B., Ruiz, J. R. & Tomkinson, G. R. Review of criterion-referenced standards for cardiorespiratory fitness: what percentage of 1 142 026 international children and youth are apparently healthy? *British Journal of Sports Medicine* **53,** 953–958 (2019).

Lang, K. M. & Little, T. D. Principled missing data treatments. *Prevention Science* **19,** 284–294 (2018).

Langkammer, C., Krebs, N., Goessler, W., Scheurer, E., Ebner, F., Yen, K., Fazekas, F. & Ropele, S. Quantitative MR imaging of brain iron: A postmortem validation study. *Radiology* **257,** 455–462 (2010).

Langnes, E., Sneve, M. H., Sederevicius, D., Amlien, I. K., Walhovd, K. B. & Fjell, A. M. Lifespan trajectories and relationships to memory of the macro- and microstructure of the anterior and posterior hippocampus – a longitudinal multimodal imaging study. *bioRxiv,* doi: https://doi.org/10.1101/564732 (2019).

Larsen, B. & Luna, B. Adolescence as a neurobiological critical period for the development of higher-order cognition. *Neuroscience and Biobehavioral Reviews* **94,** 179–195 (2018).

Lauterbur, P. C. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature* **242,** 190–191 (1973).

Le Bihan, D. & Johansen-Berg, H. Diffusion MRI at 25: Exploring brain tissue structure and function. *NeuroImage* **61,** 324–341 (2012).

Le Bihan, D., Mangin, J. F., Poupon, C., Clark, C. a., Pappata, S., Molko, N. & Chabriat, H. Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* **13,** 534–46 (2001).

Leahy, A. A., Eather, N., Smith, J. J., Hillman, C. H., Morgan, P. J., Plotnikoff, R. C., Nilsson, M., Costigan, S. A., Noetel, M. & Lubans, D. R. Feasibility and Preliminary Efficacy of a Teacher-facilitated High-Intensity Interval Training Intervention for Older Adolescents. *Pediatric Exercise Science* **31,** 107–117 (2019).

Leahy, A. A., Michels, M. F., Eather, N., Hillman, C. H., Shigeta, T. T., Lubans, D. R. & Smith, J. J. Feasibility of test administration and preliminary findings for cognitive control in the Burn 2 learn pilot randomised controlled trial. *Journal of Sports Sciences* **38,** 1708–1716 (2020).

Lear, S. A., Hu, W., Rangarajan, S., Gasevic, D., Leong, D., Iqbal, R., Casanova, A., Swaminathan, S., Anjana, R. M., Kumar, R., Rosengren, A., Wei, L., Yang, W., Chuangshi, W., Huaxing, L., Nair, S., Diaz, R., Swidon, H., Gupta, R., Mohammadifard, N., Lopez-Jaramillo, P., Oguz, A., Zatonska, K., Seron, P., Avezum, A., Poirier, P., Teo, K. & Yusuf, S. The effect of physical activity on mortality and cardiovascular disease in 130 000 people from 17 high-income, middle-income, and low-income countries: the PURE study. *The Lancet* **390,** 2643–2654 (2017).

Lebel, C., Gee, M., Camicioli, R., Wieler, M., Martin, W. & Beaulieu, C. Diffusion tensor imaging of white matter tract evolution over the lifespan. *NeuroImage* **60,** 340–352 (2012).

Lebel, C. & Deoni, S. The development of brain white matter microstructure. *NeuroImage* **182,** 207–218 (2018).

Lebel, C., Treit, S. & Beaulieu, C. A review of diffusion MRI of typical white matter development from early childhood to young adulthood. *NMR in Biomedicine* **32,** e3778 (2019).

LeBlanc, A. G. & Janssen, I. Difference between self-reported and accelerometer measured moderate-to-vigorous physical activity in youth. *Pediatric Exercise Science* **22,** 523–534 (2010).

Leckie, R. L., Weinstein, A. M., Hodzic, J. C. & Erickson, K. I. Potential moderators of physical activity on brain health. *Journal of Aging Research* **948981** (2012).

Ledford, H. The Shifting Boundaries Of Adolescence. *Nature* **554,** 429–431 (2018).

Lee, I.-M. M. & Shiroma, E. J. Using Accelerometers to Measure Physical Activity in Large- Scale Epidemiologic Studies: Issues and Challenges. *British Journal of Sports Medicine* **48,** 197–201 (2014).

Lee, I. M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., Alkandari, J. R., Andersen, L. B., Bauman, A. E., Brownson, R. C., Bull, F. C., Craig, C. L., Ekelund, U., Goenka, S., Guthold, R., Hallal, P. C., Haskell, W. L., Heath, G. W., Inoue, S., Kahlmeier, S., Kohl, H. W., Lambert, E. V., Leetongin, G., Loos, R. J., Marcus, B., Martin, B. W., Owen, N., Parra, D. C., Pratt, M., Ogilvie, D., Reis, R. S., Sallis, J. F., Sarmiento, O. L. & Wells, J. C. Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *The Lancet* **380,** 219–229 (2012).

Lee, K. J. & Carlin, J. B. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* **171,** 624–632 (2010).

Lee, K. J. & Carlin, J. B. Multiple imputation in the presence of non-normal data. *Statistics in Medicine* **36,** 606–617 (2017).

Lee, K., Bull, R. & Ho, R. M. Developmental changes in executive functioning. *Child Development* **84,** 1933–1953 (2013).

Lees, C. & Hopkins, J. Effect of aerobic exercise on cognition, academic achievement, and psychosocial function in children: A systematic review of randomized control trials. *Preventing Chronic Disease* **10** (2013).

Léger, L. A., Mercier, D., Gadoury, C. & Lambert, J. The multistage 20 metre shuttle run test for aerobic fitness. *Journal of Sports Sciences* **6,** 93–101 (1988).

Lenroot, R. K. & Giedd, J. N. Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience and Biobehavioral Reviews* **30,** 718–729 (2006).

Lerch, J. P., Van Der Kouwe, A. J., Raznahan, A., Paus, T., Johansen-Berg, H., Miller, K. L., Smith, S. M., Fischl, B. & Sotiropoulos, S. N. Studying neuroanatomy using MRI. *Nature Neuroscience* **20,** 314–326 (2017).

Lerch, J. P., Yiu, A. P., Martinez-Canabal, A., Pekar, T., Bohbot, V. D., Frankland, P. W., Henkelman, R. M., Josselyn, S. A. & Sled, J. G. Maze training in mice induces MRI-detectable brain shape changes specific to the type of learning. *NeuroImage* **54,** 2086–2095 (2011).

Li, C.-L., Deng, Y.-J., He, Y.-H., Zhai, H.-C. & Jia, F.-C. The development of brain functional connectivity networks revealed by resting-state functional magnetic resonance imaging. *Neural Regeneration Research* **14,** 1419–1429 (2019).

Li, F., Turner, E. L., Heagerty, P. J., Murray, D. M., Vollmer, W. M. & DeLong, E. R. An evaluation of constrained randomization for the design and analysis of group-randomized trials with binary outcomes. *Statistics in Medicine* **36,** 3791–3806 (2017).

Li, J. W., O'Connor, H., O'Dwyer, N. & Orr, R. The effect of acute and chronic exercise on cognitive function and academic performance in adolescents: A systematic review. *Journal of Science and Medicine in Sport* **20,** 841–848 (2017).

Li, X., Li, Q., Wang, X., Li, D. & Li, S. Differential age-related changes in structural covariance networks of human anterior and posterior hippocampus. *Frontiers in Physiology* **9,** 518 (2018).

Liang, K. Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22 (1986).

Liang, Y., Pertzov, Y., Nicholas, J. M., Henley, S. M., Crutch, S., Woodward, F., Leung, K., Fox, N. C. & Husain, M. Visual short-term memory binding deficit in familial Alzheimer's disease. *Cortex* **78,** 150–164 (2016).

Libby, L. A., Ekstrom, A. D., Daniel Ragland, J. & Ranganath, C. Differential connectivity of perirhinal and parahippocampal cortices within human hippocampal subregions revealed by high-resolution functional imaging. *Journal of Neuroscience* **32,** 6550–6560 (2012).

Liesefeld, H. R. & Janczyk, M. Combining speed and accuracy to control for speed-accuracy trade-offs(?) *Behavior Research Methods* **51,** 40–60 (2019).

Lightfoot, J. T., De Geus, E. J., Booth, F. W., Bray, M. S., Den Hoed, M., Kaprio, J., Kelly, S. A., Pomp, D., Saul, M. C., Thomis, M. A., Garland, T. & Bouchard, C. Biological / Genetic Regulation of Physical Activity Level: Consensus from GenBioPAC. *Medicine & Science in Sports & Exercise* **50,** 863–873 (2018).

Lin, W. C. & Tsai, C. F. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review* **53,** 1487–1509 (2020).

Lindenberger, U. & Lövdén, M. Brain Plasticity in Human Lifespan Development: The Exploration – Selection – Refinement Model. *Annual Review ofDevelopmental Psychology,* 197–224 (2019).

Lindquist, C. H., Reynolds, K. D. & Goran, M. I. Sociocultural determinants of physical activity among children. *Preventive Medicine* **29,** 305–312 (1999).

Liu, M., Wu, L. & Ming, Q. How does physical activity intervention improve self-esteem and self-concept in children and adolescents? Evidence from a meta-analysis. *PLoS ONE* **10,** e0134804 (2015).

Liu, N. Y.-S. Y., Plowman, S. A. & Looney, M. A. The reliability and validity of the 20-meter shuttle test in american students 12 to 15 years old. *Research Quarterly for Exercise and Sport* **63,** 360–365 (1992).

Logan, G. R. M., Harris, N., Duncan, S. & Schofield, G. A review of adolescent high-intensity interval training. *Sports Medicine* **44,** 1071–1085 (2014).

Logan, G. R., Duncan, S., Harris, N. K., Hinckson, E. A. & Schofield, G. Adolescent physical activity levels: discrepancies with accelerometer data analysis. *Journal of Sports Sciences* **34,** 2047–2053 (2016).

Loken, C., Gruner, D., Groer, L., Peltier, R., Bunn, N., Craig, M., Henriques, T., Dempsey, J., Yu, C. H., Chen, J., Dursi, L. J., Chong, J., Northrup, S., Pinto, J., Knecht, N. & Zon, R. V. SciNet: Lessons learned from building a power-efficient top-20 system and data centre. *Journal of Physics: Conference Series* **256,** 012026 (2010).

Lonsdale, C., Rosenkranz, R. R., Peralta, L. R., Bennie, A., Fahey, P. & Lubans, D. R. A systematic review and meta-analysis of interventions designed to increase moderate-to-vigorous physical activity in school physical education lessons. *Preventive Medicine* **56,** 152–161 (2013).

López-Vicente, M., Tiemeier, H., Wildeboer, A., Muetzel, R. L., Verhulst, F. C., Jaddoe, V. W., Sunyer, J. & White, T. Cortical Structures Associated With Sports Participation in Children: A Population-Based Study. *Developmental Neuropsychology* **42,** 58–69 (2017).

Lorio, S., Kherif, F., Ruef, A., Melie-Garcia, L., Frackowiak, R., Ashburner, J., Helms, G., Lutti, A. & Draganski, B. Neurobiological origin of spurious brain morphological changes: A quantitative MRI study. *Human Brain Mapping* **37,** 1801–1815 (2016).

Lövdén, M., Wenger, E., Mårtensson, J., Lindenberger, U. & Bäckman, L. Structural brain plasticity in adult learning and development. *Neuroscience and Biobehavioral Reviews* **37,** 2296–2310 (2013).

Love, R., Adams, J., Atkin, A. & Van Sluijs, E. Socioeconomic and ethnic differences in children's vigorous intensity physical activity: A cross-sectional analysis of the UK Millennium Cohort Study. *BMJ Open* **9,** e027627 (2019).

Lubans, D., Richards, J., Hillman, C., Faulkner, G., Beauchamp, M., Nilsson, M., Kelly, P., Smith, J., Raine, L. & Biddle, S. Physical activity for cognitive and mental health in youth: A systematic review of mechanisms. *Pediatrics* **138,** e20161642 (2016).

Lüdecke, D. ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *Journal of Open Source Software* **3,** 772 (2018).

Lüdtke, O., Robitzsch, A. & Grund, S. Multiple imputation of missing data in multilevel designs: A comparison of different strategies. *Psychological Methods* **22,** 141–165 (2017).

Ludyga, S., Gerber, M., Pühse, U., Looser, V. N. & Kamijo, K. Systematic review and meta-analysis investigating moderators of long-term effects of exercise on cognition in healthy individuals. *Nature Human Behaviour* **4,** 603–612 (2020).

Luna, B., Garver, K. E., Urban, T. A., Lazar, N. A. & Sweeney, J. A. Maturation of cognitive processes from late childhood to adulthood. *Child Development* **75,** 1357–1372 (2004).

Luna, B., Marek, S., Larsen, B., Tervo-Clemmens, B. & Chahal, R. An Integrative Model of the Maturation of Cognitive Control. *Annual Review of Neuroscience* **38,** 151–170 (2015).

Maass, A., Düzel, S., Goerke, M., Becke, A., Sobieray, U., Neumann, K., Lövden, M., Lindenberger, U., Bäckman, L., Braun-Dullaeus, R., Ahrens, D., Heinze, H. J., Müller, N. G. & Düzel, E. Vascular hippocampal plasticity after aerobic exercise in older adults. *Molecular Psychiatry* **20,** 585–593 (2015).

Madhyastha, T., Peverill, M., Koh, N., McCabe, C., Flournoy, J., Mills, K., King, K., Pfeifer, J. & McLaughlin, K. A. Current methods and limitations for longitudinal fMRI analysis across development. *Developmental Cognitive Neuroscience* **33,** 118–128 (2018).

Magar, E. C., Phillips, L. H. & Hosie, J. A. Brief report: Cognitive-regulation across the adolescent years. *Journal of Adolescence* **33,** 779–781 (2010).

Magnus, J. R., Sinha, A. K. & Journal, E. On Theil' s errors. *The Econometrics Journal* **8,** 39–54 (2005).

Mah, A., Geeraert, B. & Lebel, C. Detailing neuroanatomical development in late childhood and early adolescence using NODDI. *PLoS ONE* **12,** e0182340 (2017).

Mahar, M. T., Guerieri, A. M., Hanna, M. S. & Kemble, C. D. Estimation of aerobic fitness from 20-m multistage shuttle run test performance. *American Journal of Preventive Medicine* **41,** S117–S123 (2011).

Mahar, M. T., Welk, G. J. & Rowe, D. A. Estimation of aerobic fitness from PACER performance with and without body mass index. *Measurement in Physical Education and Exercise Science* **22,** 239–249 (2018).

Mahto, A. *splitstackshape: Stack and Reshape Datasets After Splitting Concatenated Values, R package version 1.4.8* 2019. https://cran.r-project.org/package=splitstackshape.

Malykhin, N. V., Bouchard, T. P., Ogilvie, C. J., Coupland, N. J., Seres, P. & Camicioli, R. Three-dimensional volumetric analysis and reconstruction of amygdala and hippocampal head, body and tail. *Psychiatry Research - Neuroimaging* **155,** 155–165 (2007).

Mancini, M., Karakuzu, A., Cohen-Adad, J., Cercignani, M., Nichols, T. E. & Stikov, N. An interactive meta-analysis of MRI biomarkers of Myelin. *eLife* **9,** e61523 (2020).

Marascuilo, L. A. & Levin, J. R. *Multivariate Statistics in the Social Sciences: A Researcher's Guide.* (Brooks/ Cole Publishing Company, Monterey, CA, 1983).

Marques, A., Gómez, F., Martins, J., Catunda, R. & Sarmento, H. Association between physical education, school-based physical activity, and academic performance: a systematic review. *Retos* **31,** 316–320 (2017).

Marques, A., Santos, D. A., Hillman, C. H. & Sardinha, L. B. L. B. How does academic achievement relate to cardiorespiratory fitness, self-reported physical activity and objectively reported physical activity: a systematic review in children and adolescents aged 6-18 years. *British Journal of Sports Medicine* **52,** 1039 (2018).

Mars, T., Ellard, D., Carnes, D., Homer, K., Underwood, M. & Taylor, S. J. Fidelity in complex behaviour change interventions: A standardised approach to evaluate intervention integrity. *BMJ Open* **3** (2013).

Marsh, H. W., Martin, A. J. & Jackson, S. Introducing a short version of the physical self description questionnaire: new strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of sport and exercise psychology* **32,** 438–482 (2010).

Martin, A., Booth, J. N., Laird, Y., Sproule, J., Reilly, J. J. & Saunders, D. H. Physical activity, diet and other behavioural interventions for improving cognition and school achievement in children and adolescents with obesity or overweight. *Cochrane Database of Systematic Reviews* **3,** CD009728 (2018).

Martínez-Gómez, D., Ruiz, J. R., Gómez-Martínez, S., Chillón, P., Rey-López, J. P., Díaz, L. E., Castillo, R., Veiga, O. L. & Marcos, A. Active commuting to school and cognitive performance in adolescents: The AVENA study. *Archives of Pediatrics and Adolescent Medicine* **165,** 300–305 (2011).

Martland, R., Mondelli, V., Gaughran, F. & Stubbs, B. Can high-intensity interval training improve physical and mental health outcomes? A meta-review of 33 systematic reviews across the lifespan. *Journal of Sports Sciences* **38,** 430–469 (2020).

Matsumoto, Y., Tsunekawa, Y., Nomura, T., Suto, F., Matsumata, M., Tsuchiya, S. & Osumi, N. Differential Proliferation Rhythm of Neural Progenitor and Oligodendrocyte Precursor Cells in the Young Adult Hippocampus. *PLoS ONE* **6,** e27628 (2011).

Matsuzaka, A., Takahashi, Y., Yamazoe, M., Kumakura, N., Ikeda, A., Wilk, B. & Bar-Or, O. Validity of the multistage 20-m shuttle-run test for Japanese children, adolescents, and adults. *Pediatric Exercise Science* **16,** 113–125 (2004).

Matthews, C. E. Calibration of accelerometer output for adults. *Medicine and Science in Sports and Exercise* **37,** S512–22 (2005).

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H. & Bates, D. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* **94,** 305–315 (2017).

Mayorga-Vega, D., Aguilar-Soto, P. & Viciana, J. Criterion-Related Validity of the 20-M Shuttle Run Test for Estimating Cardiorespiratory Fitness: A Meta-Analysis. *Journal of Sports Science and Medicine* **14,** 536–547 (2015).

Mayorga-Vega, D., Bocanegra-Parrilla, R., Ornelas, M. & Viciana, J. Criterion-related validity of the distance- and time-based walk/run field tests for estimating cardiorespiratory fitness: A systematic review and meta-analysis. *PLoS ONE* **11,** e0151671 (2016).

McArdle, W. D., Katch, F. I. & Katch, V. L. *Exercise physiology: nutrition, energy, and human performance* 7th (Lippincott Williams & Wilkins, 2010).

McKercher, C., Sanderson, K., Schmidt, M. D., Otahal, P., Patton, G. C., Dwyer, T. & Venn, A. J. Physical activity patterns and risk of depression in young adulthood: a 20-year cohort study since childhood. *Social Psychiatry and Psychiatric Epidemiology* **49,** 1823–1834 (2014).

McMahon, E. M., Corcoran, P., O'Regan, G., Keeley, H., Cannon, M., Carli, V., Wasserman, C., Hadlaczky, G., Sarchiapone, M. & Apter, A. Physical activity in European adolescents and associations with anxiety, depression and well-being. *European child & adolescent psychiatry* **26,** 111–122 (2017).

McMorris, T. Developing the catecholamines hypothesis for the acute exercise-cognition interaction in humans: Lessons from animal studies. *Physiology & Behavior* **165,** 291–299 (2016).

McNeish, D. Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research* **52,** 661–670 (2017).

McNeish, D. M., Stapleton, L. & Silverman, R. On the Unnecessary Ubiquity of Hierarchical Linear Modeling. *Psychological Methods* **22,** 114–140 (2016).

McNeish, D. M. & Stapleton, L. M. The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educational Psychology Review* **28,** 295–314 (2016).

McPherson, A., Mackay, L., Kunkel, J. & Duncan, S. Physical activity, cognition and academic performance: An analysis of mediating and confounding relationships in primary school children. *BMC Public Health* **18,** 936 (2018).

Meijer, A., Königs, M., Vermeulen, G. T., Visscher, C., Bosker, R. J., Hartman, E. & Oosterlaan, J. The effects of physical activity on brain structure and neurophysiological functioning in children: A systematic review and meta-analysis. *Developmental Cognitive Neuroscience* **45,** 100828 (2020).

Melo, X., Santa-Clara, H., Almeida, J. P., Carnero, E. A., Sardinha, L. B., Bruno, P. M. & Fernhall, B. Comparing several equations that predict peak VO2 using the 20-m multistage-shuttle run-test in 8-10-year-old children. *European Journal of Applied Physiology* **111,** 839–849 (2011).

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M. & Culhane, A. C. Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* **17,** 628–641 (2016).

Meng, X. L. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9,** 538–558 (1994).

Meredith-Jones, K., Galland, B., Haszard, J., Gray, A., Sayers, R., Hanna, M., Taylor, B. & Taylor, R. Do young children consistently meet 24-h sleep and activity guidelines? A longitudinal analysis using actigraphy. *International Journal of Obesity* **43,** 2555–2564 (2019).

Michie, S., Johnston, M., Francis, J., Hardeman, W. & Eccles, M. From Theory to Intervention: Mapping Theoretically Derived Behavioural Determinants to Behaviour Change Techniques. *Applied Psychology* **57,** 660–680 (2008).

Middleton, L. E., Barnes, D. E., Lui, L. Y. & Yaffe, K. Physical activity over the life course and its association with cognitive performance and impairment in old age. *Journal of the American Geriatrics Society* **58,** 1322–1326 (2010).

Mietchen, D. & Gaser, C. Computational morphometry for detecting changes in brain structure due to development, aging, learning, disease and evolution. *Frontiers in Neuroinformatics* **3,** 25 (2009).

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Delisle Nyström, C., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J. R. & Ortega, F. B. Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Medicine* **47,** 1821–1845 (2017).

Migueles, J. H., Cadenas-Sanchez, C., Tudor-Locke, C., Löf, M., Esteban-Cornejo, I., Molina-Garcia, P., Mora-Gonzalez, J., Rodriguez-Ayllon, M., Garcia-Marmol, E., Ekelund, U. & Ortega, F. B. Comparability of published cut-points for the assessment of physical activity: Implications for data harmonization. *Scandinavian Journal of Medicine and Science in Sports* **29,** 566–574 (2019).

Migueles, J. H., Cadenas-sanchez, C., Esteban-cornejo, I., Torres-lopez, L. V., Aadland, E., Chastin, S. F., Erickson, K. I., Catena, A. & Ortega, F. B. Associations of Objectively-Assessed Physical Activity and Sedentary Time with Hippocampal Gray Matter Volume in Children with Overweight / Obesity. *Journal of Clinical Medicine* **9,** 1080 (2020).

Miller, J. The sampling distribution of d'. *Perception & Psychophysics* **58,** 65–72 (1996).

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P. M. & Smith, S. M. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience* **19,** 1523–1536 (2016).

Miller, M. A., Leckie, R. L., Donofry, S. D., Gianaros, P. J., Erickson, K. I., Manuck, S. B. & Roecklein, K. A. Photoperiod is associated with hippocampal volume in a large community sample. *Hippocampus* **25,** 534–543 (2015).

Mills, K. L., Goddings, A. L., Herting, M. M., Meuwese, R., Blakemore, S. J., Crone, E. A., Dahl, R. E., Güroğlu, B., Raznahan, A., Sowell, E. R. & Tamnes, C. K. Structural brain development between childhood and adulthood: Convergence across four longitudinal samples. *NeuroImage* **141,** 273–281 (2016).

Mills, K. L. & Tamnes, C. K. Longitudinal structural and functional brain development in childhood and adolescence. *PsyArXiv,* doi: 10.31234/osf.io/87kft (2018).

Mistler, S. A. & Enders, C. K. A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data. *Journal of Educational and Behavioral Statistics* **42,** 432–466 (2017).

Miyake, A. & Friedman, N. P. The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions. *Current directions in psychological science* **21,** 8–14 (2012).

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A. & Wager, T. D. The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology* **41,** 49–100 (2000).

Moerbeek, M. & Teerenstra, S. *Power Analysis of Trials with Multilevel Data* (CRC Press, 2015).

Molenberghs, G., Fitzmaurice, G. G., Kenward, M., Anastasios, T. & Verbeke, G. *Handbook of Missing Data Methodology* (Chapman and Hall/CRC, 2014).

Molenberghs, G. & Verbeke, G. *Linear mixed models for longitudinal data.* (Springer New York, New York, 2000).

Mollink, J., Kleinnijenhuis, M., Cappellen Van Walsum, A. M. V., Sotiropoulos, S. N., Cottaar, M., Mirfin, C., Heinrich, M. P., Jenkinson, M., Pallebage-Gamarallage, M., Ansorge, O., Jbabdi, S. & Miller, K. L. Evaluating fibre orientation dispersion in white matter: Comparison of diffusion MRI, histology and polarized light imaging. *NeuroImage* **157,** 561–574 (2017).

Moncher, F. J. & Prinz, R. J. Treatment Fidelity in Outcome Studies. *Clinical Psychology Review* **11,** 247–266 (1991).

Monsell, S. Task switching. *Trends in Cognitive Sciences* **7,** 134–140 (2003).

Montero, D. & Lundby, C. Refuting the myth of non-response to exercise training: 'non-responders' do respond to higher dose of training. *Journal of Physiology* **595,** 3377–3387 (2017).

Monti, J. M., Hillman, C. H. & Cohen, N. J. Aerobic fitness enhances relational memory in preadolescent children: the FITKids randomized control trial. *Hippocampus* **22,** 1876–1882 (2012).

Montoye, H. J., Kemper, H. C. G., Saris, W. H. & Washburn, R. A. *Measuring physical activity and energy expenditure* (ed Montoye, H. J.) (Champaign, IL : Human Kinetics, 1996).

Mora-Gonzalez, J., Esteban-Cornejo, I., Cadenas-Sanchez, C., Migueles, J. H., Molina-Garcia, P., Rodriguez-Ayllon, M., Henriksson, P., Pontifex, M. B., Catena, A. & Ortega, F. B. Physical Fitness, Physical Activity, and the Executive Function in Children with Overweight and Obesity. *Journal of Pediatrics* **208,** 50–56 (2019).

Mora-Gonzalez, J., Esteban-Cornejo, I., Solis-Urra, P., Migueles, J. H., Cadenas-Sanchez, C., Molina-Garcia, P., Rodriguez-Ayllon, M., Hillman, C. H., Catena, A., Pontifex, M. B. & Ortega, F. B. Fitness, physical activity, sedentary time, inhibitory control, and neuroelectric activity in children with overweight or obesity: The ActiveBrains project. *Psychophysiology* **57,** e13579 (2020).

Moreau, D. & Chou, E. The Acute Effect of High-Intensity Exercise on Executive Function: A Meta-Analysis. *Perspectives on Psychological Science* **14,** 734–764 (2019).

Moreau, D., Kirk, I. J. & Waldie, K. E. High-intensity training enhances executive function in children in a randomized , placebo-controlled trial. *eLife* **6,** e25062 (2017).

Morey, R. A., Petty, C. M., Xu, Y., Pannu Hayes, J., Wagner, H. R., Lewis, D. V., LaBar, K. S., Styner, M. & McCarthy, G. A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *NeuroImage* **45,** 855–866 (2009).

Mori, S. & Zhang, J. Principles of Diffusion Tensor Imaging and Its Applications to Basic Neuroscience Research. *Neuron* **51,** 527–539 (2006).

Morland, C., Andersson, K. A., Haugen, Ø. P., Hadzic, A., Kleppa, L., Gille, A., Rinholm, J. E., Palibrk, V., Diget, E. H., Kennedy, L. H., Stølen, T., Hennestad, E., Moldestad, O., Cai, Y., Puchades, M., Offermanns, S., Vervaeke, K., Bjørås, M., Wisløff, U., Storm-Mathisen, J. & Bergersen, L. H. Exercise induces cerebral VEGF and angiogenesis via the lactate receptor HCAR1. *Nature Communications* **8,** 1–9 (2017).

Mota, J. G., Clark, C. C. T., Bezerra, T. A., Lemos, L., Reuter, C. P., Mota, J. A. P. S., Duncan, M. J. & Martins, C. M. D. L. Twenty-four-hour movement behaviours and fundamental movement skills in preschool children: A compositional and isotemporal substitution analysis. *Journal of Sports Sciences* **38,** 2071–2079 (2020).

Mulder, E. R., De Jong, R. A., Knol, D. L., Van Schijndel, R. A., Cover, K. S., Visser, P. J., Barkhof, F. & Vrenken, H. Hippocampal volume change measurement: Quantitative assessment of the reproducibility of expert manual outlining and the automated methods FreeSurfer and FIRST. *NeuroImage* **92,** 169–181 (2014).

Mura, G., Vellante, M., Nardi, A. E., Machado, S. & Carta, M. G. Effects of school-based physical activity interventions on cognition and academic achievement: a systematic review. *CNS & Neurological Disorders - Drug Targets* **14,** 1194–1208 (2015).

Murray, D. M., Varnell, S. P. & Blitstein, J. L. Design and analysis of group-randomised trials: a review of recent methodological developments. *American Journal of Public Health* **94,** 423–432 (2004).

Murray, D. M. *Design and Analysis of Group-Randomized Trials* (Oxford University Press, 1998).

Muthén, L. & Muthén, B. *Mplus User's Guide. Eighth Edition.* Los Angeles, CA, USA.

Nader, P. R., Bradley, R. H., Houts, R. M., Mcritchie, S. L. & Brien, M. O. Moderate-to-Vigorous Physical Activity From Ages 9 to 15 Years. *JAMA* **300,** 295–305 (2008).

Nakagawa, S., Johnson, P. C. & Schielzeth, H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface* **14,** 20170213 (2017).

Nakamura, K., Brown, R. A., Narayanan, S., Collins, D. L. & Arnold, D. L. Diurnal fluctuations in brain volume: Statistical analyses of MRI from large populations. *NeuroImage* **118,** 126–132 (2015).

Narvacan, K., Treit, S., Camicioli, R., Martin, W. & Beaulieu, C. Evolution of deep gray matter volume across the human lifespan. *Human Brain Mapping* **38,** 3771–3790 (2017).

Naylor, P.-J. J., Nettlefold, L., Race, D., Hoy, C., Ashe, M. C., Wharf Higgins, J. & McKay, H. A. Implementation of school based physical activity interventions: A systematic review. *Preventive Medicine* **72,** 95–115 (2015).

Nelson, M. C., Cordray, D. S., Hulleman, C. S., Darrow, C. L. & Sommer, E. C. A procedure for assessing intervention fidelity in experiments testing educational and behavioral interventions. *Journal of Behavioral Health Services and Research* **39,** 374–396 (2012).

Nguyen, C. D., Carlin, J. B. & Lee, K. J. Model checking in multiple imputation: An overview and case study. *Emerging Themes in Epidemiology* **14,** 8 (2017).

Nieman, D. C. & Wentz, L. M. The compelling link between physical activity and the body's defense system. *Journal of Sport and Health Science* **8,** 201–217 (2019).

Nilsson, A., Ekelund, U., Yngve, A. & Sjöström, M. Assessing physical activity among children with accelerometers using different time sampling intervals and placements. *Pediatric Exercise Science* **14,** 87–96 (2002).

Noble, S., Scheinost, D., Finn, E. S., Shen, X., Papademetris, X., McEwen, S. C., Bearden, C. E., Addington, J., Goodyear, B., Cadenhead, K. S., Mirzakhanian, H., Cornblatt, B. A., Olvet, D. M., Mathalon, D. H., McGlashan, T. H., Perkins, D. O., Belger, A., Seidman, L. J., Thermenos, H., Tsuang, M. T., Van Erp, T. G., Walker, E. F., Hamann, S., Woods, S. W., Cannon, T. D. & Constable, R. T. Multisite reliability of MR-based functional connectivity. *NeuroImage* **146,** 959–970 (2017).

Nordenskjöld, R., Malmberg, F., Larsson, E. M., Simmons, A., Ahlström, H., Johansson, L. & Kullberg, J. Intracranial volume normalization methods: Considerations when investigating gender differences in regional brain volume. *Psychiatry Research - Neuroimaging* **231,** 227–235 (2015).

Norris, E., Steen, T. V., Direito, A. & Stamatakis, E. Physically active lessons in schools and their impact on physical activity, educational, health and cognition outcomes: A systematic review and meta-analysis. *British Journal of Sport Medicine* **15,** 826–838 (2020).

North, C. T., McCullagh, P. & Tran, Z. V. Effect of Exercise on Depression. *Exercise and sport sciences reviews* **18,** 379–416 (1990).

Northey, J. M., Cherbuin, N., Pumpa, K. L., Smee, D. J. & Rattray, B. Exercise interventions for cognitive function in adults older than 50: A systematic review with meta-Analysis. *British Journal of Sports Medicine* **52,** 154–160 (2018).

Norton, K., Norton, L. & Sadgrove, D. Position statement on physical activity and exercise intensity terminology. *Journal of Science and Medicine in Sport* **13,** 496–502 (2010).

Nosek, B. A., Greenwald, A. G. & Banaji, M. R. Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin* **31,** 166–180 (2005).

Novikov, D. S., Fieremans, E., Jespersen, S. N. & Kiselev, V. G. Quantifying brain microstructure with diffusion MRI: Theory and parameter estimation. *NMR in Biomedicine* **32,** e3998 (2018).

Novikov, D. S., Kiselev, V. G. & Jespersen, S. N. On modeling. *Magnetic Resonance in Medicine* **79,** 3172–3193 (2018).

Nugent, A. C., Luckenbaugh, D. A., Wood, S. E., Bogers, W., Zarate, C. A. & Drevets, W. C. Automated subcortical segmentation using FIRST: Test-retest reliability, interscanner reliability, and comparison to manual segmentation. *Human Brain Mapping* **34,** 2313–2329 (2013).

Nyaradi, A., Li, J., Hickling, S., Foster, J. & Oddy, W. H. The role of nutrition in children's neurocognitive development, from pregnancy through childhood. *Frontiers in Human Neuroscience* **7,** 97 (2013).

Nyongesa, M. K., Ssewanyana, D., Mutua, A. M., Chongwo, E., Scerif, G., Newton, C. R. & Abubakar, A. Assessing executive function in adolescence: A scoping review of existing measures and their psychometric robustness. *Frontiers in Psychology* **10,** 311 (2019).

O'Brien, L. M., Ziegler, D. A., Deutsch, C. K., Frazier, J. A., Herbert, M. R. & Locascio, J. J. Statistical adjustments for brain size in volumetric neuroimaging studies: Some practical implications in methods. *Psychiatry Research - Neuroimaging* **193,** 113–122 (2011).

O'Donnell, C. L. Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research* **78,** 33–84 (2008).

O'Kelly, M. & Ratitch, B. *Clinical trials with missing data: A guide for practitioners* (John Wiley & Sons, 2014).

O'Leary, J. D., Hoban, A. E., Cryan, J. F., O'Leary, O. F. & Nolan, Y. M. Differential effects of adolescent and adult-initiated voluntary exercise on context and cued fear conditioning. *Neuropharmacology* **145,** 49–58 (2019).

O'Leary, J. D., Hoban, A. E., Murphy, A., O'Leary, O. F., Cryan, J. F. & Nolan, Y. M. Differential effects of adolescent and adult-initiated exercise on cognition and hippocampal neurogenesis. *Hippocampus* **29,** 352–365 (2019).

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H. & Ugurbil, K. Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences* **89,** 5951–5955 (1992).

Okell, T. W., Chappell, M. A., Kelly, M. E. & Jezzard, P. Cerebral blood flow quantification using vessel-encoded arterial spin labeling. *Journal of Cerebral Blood Flow and Metabolism* **33,** 1716–1724 (2013).

Olsen, R. K., Carr, V. A., Daugherty, A. M., La Joie, R., Amaral, R. S., Amunts, K., Augustinack, J. C., Bakker, A., Bender, A. R., Berron, D., Boccardi, M., Bocchetta, M., Burggren, A. C., Chakravarty, M. M., Chételat, G., De Flores, R., DeKraker, J., Ding, S. L., Geerlings, M. I., Huang, Y., Insausti, R., Johnson, E. G., Kanel, P., Kedo, O., Kennedy, K. M., Keresztes, A., Lee, J. K., Lindenberger, U., Mueller, S. G., Mulligan, E. M., Ofen, N., Palombo, D. J., Pasquini, L., Pluta, J., Raz, N., Rodrigue, K. M., Schlichting, M. L., Lee Shing, Y., Stark, C. E., Steve, T. A., Suthana, N. A., Wang, L., Werkle-Bergner, M., Yushkevich, P. A., Yu, Q. & Wisse, L. E. Progress update from the hippocampal subfields group. *Alzheimer's and Dementia: Diagnosis, Assessment and Disease Monitoring* **11,** 439–449 (2019).

Opel, N., Martin, S., Meinert, S., Redlich, R., Enneking, V., Richter, M., Goltermann, J., Johnen, A., Dannlowski, U. & Repple, J. White matter microstructure mediates the association between physical fitness and cognition in healthy, young adults. *Scientific Reports* **9,** 1–9 (2019).

Orban, C., Kong, R., Li, J., Chee, M. W. & Yeo, B. T. Time of day is associated with paradoxical reductions in global signal fluctuation and functional connectivity. *PLoS biology* **18,** e3000602 (2020).

Orben, A. & Przybylski, A. K. The association between adolescent well-being and digital technology use. *Nature Human Behaviour* **3,** 173–182 (2019).

Ordaz, S. J., Foran, W., Velanova, K. & Luna, B. Longitudinal growth curves of brain function underlying inhibitory control through adolescence. *Journal of Neuroscience* **33,** 18109–18124 (2013).

Orme, M., Wijndaele, K., Sharp, S. J., Westgate, K., Ekelund, U. & Brage, S. Combined influence of epoch length, cut-point and bout duration on accelerometry-derived physical activity. *International Journal of Behavioral Nutrition and Physical Activity* **11,** 34 (2014).

Ortega, F. B., Artero, E. G., Ruiz, J. R., Vicente-Rodriguez, G., Bergman, P., Hagströmer, M., Ottevaere, C., Nagy, E., Konsta, O., Rey-López, J. P., Polito, A., Dietrich, S., Plada, M., Béghin, L., Manios, Y., Sjöström, M. & Castillo, M. J. Reliability of health-related physical fitness tests in European adolescents. The HELENA Study. *International Journal of Obesity* **32,** S49–S57 (2008).

Ortega, F. B., Ruiz, J. R., Castillo, M. J., Sjöström, M. & Sjostrom, M. Physical fitness in childhood and adolescence: a powerful marker of health. *International Journal of Obesity* **32,** 1–11 (2008).

Ortega, F. B., Campos, D., Cadenas-Sanchez, C., Altmäe, S., Martínez-Zaldívar, C., Martín-Matillas, M., Catena, A. & Campoy, C. Physical fitness and shapes of subcortical brain structures in children. *British Journal of Nutrition* **122,** S49–S58 (2019).

Ortega, F. B., Ruiz, J. R., España-Romero, V., Vicente-Rodriguez, G., Martínez-Gómez, D., Manios, Y., Béghin, L., Molnar, D., Widhalm, K., Moreno, L. A., Sjöström, M. & Castillo, M. J. The International Fitness Scale (IFIS): Usefulness of self-reported fitness in youth. *International Journal of Epidemiology* **40,** 701–711 (2011).

Owens, S., Galloway, R. & Gutin, B. The case for vigorous physical activity in youth. *American Journal of Lifestyle Medicine* **11,** 96–115 (2017).

Paap, K. R., Anders-Jefferson, R., Zimiga, B., Mason, L. & Mikulinsky, R. Interference scores have inadequate concurrent and convergent validity: Should we stop using the flanker, Simon, and spatial Stroop tasks? *Cognitive Research: Principles and Implications* **5,** 1–27 (2020).

Paap, K. R. & Oliver, S. The role of test-retest reliability in measuring individual and group differences in executive functioning. *Journal of Neuroscience Methods* **274,** 81–93 (2016).

Paap, K. R. & Sawi, O. Bilingual advantages in executive functioning: Problems in convergent validity, discriminant validity, and the identification of the theoretical constructs. *Frontiers in Psychology* **5,** 962 (2014).

Packer, M. & Pitt, B. Future large-scale clinical trials in cardiovascular medicine: challenges and uncertainties. *European Journal of Heart Failure* **20,** 1645–1648 (2018).

Packwood, S., Hodgetts, H. M. & Tremblay, S. A multiperspective approach to the conceptualization of executive functions. *Journal of Clinical and Experimental Neuropsychology* **33,** 456–470 (2011).

Pajonk, F. G., Wobrock, T., Gruber, O., Scherk, H., Berner, D., Kaizl, I., Kierer, A., Müller, S., Oest, M., Meyer, T., Backens, M., Schneider-Axmann, T., Thornton, A. E., Honer, W. G. & Falkai, P. Hippocampal plasticity in response to exercise in schizophrenia. *Archives of General Psychiatry* **67,** 133–143 (2010).

Paradisis, G. P., Zacharogiannis, E., Mandila, D., Smirtiotou, A., Argeitaki, P. & Cooke, C. B. Multi-stage 20-m shuttle run fitness test, maximal oxygen uptake and velocity at maximal oxygen uptake. *Journal of Human Kinetics* **41,** 81–87 (2014).

Pardoe, H. R., Kucharsky Hiess, R. & Kuzniecky, R. Motion and morphometry in clinical and nonclinical populations. *NeuroImage* **135,** 177–185 (2016).

Pardoe, H. R., Pell, G. S., Abbott, D. F. & Jackson, G. D. Hippocampal volume assessment in temporal lobe epilepsy: How good is automated segmentation? *Epilepsia* **50,** 2586–2592 (2009).

Park, C. A., Kang, C. K., Kim, Y. B. & Cho, Z. H. Advances in MR angiography with 7T MRI: From microvascular imaging to functional angiography. *NeuroImage* **168,** 269–278 (2018).

Pate, R. R. The evolving definition of physical fitness. *Quest* **40,** 174–179 (1988).

Patel, S., Park, M. T. M., Devenyi, G. A., Patel, R., Masellis, M., Knight, J. & Chakravarty, M. M. Heritability of hippocampal subfield volumes using a twin and non-twin siblings design. *Human Brain Mapping* **38,** 4337–4352 (2017).

Patenaude, B., Smith, S. M., Kennedy, D. N. & Jenkinson, M. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage* **56,** 907–922 (2011).

Patton, G. C., Olsson, C. A., Skirbekk, V., Saffery, R., Wlodek, M. E., Azzopardi, P. S., Stonawski, M., Rasmussen, B., Spry, E., Francis, K., Bhutta, Z. A., Kassebaum, N. J., Mokdad, A. H., Murray, C. J., Prentice, A. M., Reavley, N., Sheehan, P., Sweeny, K., Viner, R. M. & Sawyer, S. M. Adolescence and the next generation. *Nature* **554,** 458–466 (2018).

Patton, G. C. & Viner, R. Pubertal transitions in health. *The Lancet* **369,** 1130–1139 (2007).

Patton, G. C., Sawyer, S. M., Santelli, J. S., Ross, D. A., Afifi, R., Allen, N. B., Arora, M., Azzopardi, P., Baldwin, W., Bonell, C., Kakuma, R., Kennedy, E., Mahon, J., McGovern, T., Mokdad, A. H., Patel, V., Petroni, S., Reavley, N., Taiwo, K., Waldfogel, J., Wickremarathne, D., Barroso, C., Bhutta, Z., Fatusi, A. O., Mattoo, A., Diers, J., Fang, J., Ferguson, J., Ssewamala, F. & Viner, R. M. Our future: a Lancet commission on adolescent health and wellbeing. *The Lancet* **387,** 2423–2478 (2016).

Paus, T. How environment and genes shape the adolescent brain. *Hormones and Behavior* **64,** 195–202 (2013).

Paus, T., Keshavan, M. & Giedd, J. N. Why do many psychiatric disorders emerge during adolescence? *Nature Reviews Neuroscience* **9,** 947–957 (2008).

Paus, T., Wong, A. P. Y., Syme, C. & Pausova, Z. Sex differences in the adolescent brain and body: Findings from the saguenay youth study. *Journal of Neuroscience Research* **95,** 362–370 (2017).

Pedersen, B. K. & Saltin, B. Exercise as medicine - Evidence for prescribing exercise as therapy in 26 different chronic diseases. *Scandinavian Journal of Medicine and Science in Sports* **25,** 1–72 (2015).

Penry, J. T., Wilcox, A. R. & Yun, J. Validity and reliability analysis of Cooper's 12-minute run and the multistage shuttle run in healthy adults. *Journal of Strength and Conditioning Research* **25,** 597–605 (2011).

Peper, J. S., Brouwer, R. M., Boomsma, D. I., Kahn, R. S. & Hulshoff Pol, H. E. Genetic influences on human brain structure: A review of brain imaging studies in twins. *Human Brain Mapping* **28,** 464–473 (2007).

Pereira, A. C., Huddleston, D. E., Brickman, A. M., Sosunov, A. A., Hen, R., McKhann, G. M., Sloan, R., Gage, F. H., Brown, T. R. & Small, S. A. An in vivo correlate of exercise-induced neurogenesis in the adult dentate gyrus. *Proceedings of the National Academy of Sciences* **104,** 5638–5643 (2007).

Perepletchikova, F., Treat, T. A. & Kazdin, A. E. Treatment Integrity in Psychotherapy Research: Analysis of the Studies and Examination of the Associated Factors. *Journal of Consulting and Clinical Psychology* **75,** 829–841 (2007).

Perez, E. C., Bravo, D. R., Rodgers, S. P., Khan, A. R. & Leasure, J. L. Shaping the adult brain with exercise during development: Emerging evidence and knowledge gaps. *International Journal of Developmental Neuroscience* **78,** 147–155 (2019).

Perkins, N. J., Cole, S. R., Harel, O., Tchetgen Tchetgen, E. J., Sun, B., Mitchell, E. M. & Schisterman, E. F. Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology* **187,** 568–575 (2018).

Perri, M. G., Anton, S. D., Durning, P. E., Ketterson, T. U., Sydeman, S. J., Berlant, N. E., Kanasky, W. F., Newton, R. L., Limacher, M. C. & Daniel Martin, A. Adherence to exercise prescriptions: Effects of prescribing moderate versus higher levels of intensity and frequency. *Health Psychology* **21,** 452–458 (2002).

Persson, J., Spreng, R. N., Turner, G., Herlitz, A., Morell, A., Stening, E., Wahlund, L. O., Wikström, J. & Söderlund, H. Sex differences in volume and structural covariance of the anterior and posterior hippocampus. *NeuroImage* **99,** 215–225 (2014).

Pertzov, Y., Dong, M. Y., Peich, M. C. & Husain, M. Forgetting What Was Where: The Fragility of Object-Location Binding. *PLoS ONE* **7,** e48214 (2012).

Pertzov, Y., Miller, T. D., Gorgoraptis, N., Caine, D., Schott, J. M., Butler, C. & Husain, M. Binding deficits in memory following medial temporal lobe damage in patients with voltage-gated potassium channel complex antibody-associated limbic encephalitis. *Brain* **136,** 2474–2485 (2013).

Pesarin, F. On a nonparametric combination method for dependent permutation tests with applications. *Psychotherapy and Psychosomatics* **54,** 172–179 (1990).

Pesce, C. Shifting the Focus from Quantitative to Qualitative Exercise Characteristics in Exercise and Cognition Research. *Journal of Sport and Exercise Psychology* **34,** 766–786 (2012).

Peterson, A., Crockett, L., Richards, M. & Boxer, A. A self-report measure of pubertal status. *Journal of Youth and Adolescence* **17,** 117–133 (1988).

Peugh, J. L. A practical guide to multilevel modeling. *Journal of School Psychology* **48,** 85–112 (2010).

Phillips, L. R. S., Parfitt, G. & Rowlands, A. V. Calibration of the GENEA accelerometer for assessment of physical activity intensity in children. *Journal of Science & Medicine in Sport* **16,** 124–128 (2013).

Pickering, C. & Kiely, J. Do Non-Responders to Exercise Exist—and If So, What Should We Do About Them? *Sports Medicine* **49,** 1–7 (2019).

Pierpaoli, C., Jezzard, P., Basser, P. J., Barnett, A. & Di Chiro, G. Diffusion Tensor MR Imaging of the Human Brain. *Radiology* **201,** 637–648 (1996).

Pindus, D. M., Davis, R. D., Hillman, C. H., Bandelow, S., Hogervorst, E., Biddle, S. J. & Sherar, L. B. The relationship of moderate-to-vigorous physical activity to cognitive processing in adolescents: findings from the ALSPAC birth cohort. *Psychological Research* **79,** 715–728 (2015).

Pindus, D. M., Drollette, E. S., Raine, L. B., Kao, S. C., Khan, N., Westfall, D. R., Hamill, M., Shorin, R., Calobrisi, E., John, D., Kramer, A. F. & Hillman, C. H. Moving fast, thinking fast: The relations of physical activity levels and bouts to neuroelectric indices of inhibitory control in preadolescents. *Journal of Sport and Health Science* **8,** 301–314 (2019).

Pindus, D. M., Drollette, E. S., Scudder, M. R., Khan, N. A., Raine, L. B., Sherar, L. B., Esliger, D. W., Kramer, A. F. & Hillman, C. H. Moderate-to-vigorous physical activity, indices of cognitive control, and academic achievement in preadolescents. *Journal of Pediatrics* **173,** 136–142 (2016).

Pipitone, J., Park, M. T. M., Winterburn, J., Lett, T. A., Lerch, J. P., Pruessner, J. C., Lepage, M., Voineskos, A. N. & Chakravarty, M. M. Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *NeuroImage* **101,** 494–512 (2014).

Piredda, G. F., Hilbert, T., Thiran, J. P. & Kober, T. Probing myelin content of the human brain with MRI: A review. *Magnetic Resonance in Medicine* **85,** 627–652 (2021).

Polderman, T. J., Posthuma, D., De Sonneville, L. M., Stins, J. F., Verhulst, F. C. & Boomsma, D. I. Genetic analyses of the stability of executive functioning during childhood. *Biological Psychology* **76,** 11–20 (2007).

Poldrack, R. A., Huckins, G. & Varoquaux, G. Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry* **77,** 534–540 (2020).

Pontifex, M. B., McGowan, A. L., Chandler, M. C., Gwizdala, K. L., Parks, A. C., Fenn, K. & Kamijo, K. A primer on investigating the after effects of acute bouts of physical activity on cognition. *Psychology of Sport and Exercise* **40,** 1–22 (2019).

Pontifex, M. B., Raine, L. B., Johnson, C. R., Chaddock, L., Voss, M. W., Cohen, N. J., Kramer, A. F. & Hillman, C. H. Cardiorespiratory Fitness and the Flexible Modulation of Cognitive Control in Preadolescent Children. *Journal of Cognitive Neuroscience* **23,** 1332–1345 (2011).

Poon, K. Hot and cool executive functions in adolescence: Development and contributions to important developmental outcomes. *Frontiers in Psychology* **8,** 2311 (2018).

Poppenk, J., Evensmoen, H. R., Moscovitch, M. & Nadel, L. Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences* **17,** 230–240 (2013).

Poppenk, J. & Moscovitch, M. A hippocampal marker of recollection memory ability among healthy young adults: Contributions of posterior and anterior segments. *Neuron* **72,** 931–937 (2011).

Pössel, P., Smith, E. & Alexander, O. LARS&LISA: a universal school-based cognitive-behavioral program to prevent adolescent depression. *Psicologia: Reflexao e Critica* **31,** 23 (2018).

Powney, M., Williamson, P., Kirkham, J. & Kolamunnage-Dona, R. A review of the handling of missing longitudinal outcome data in clinical trials. *Trials* **15,** 237 (2014).

Pozuelo-Carrascosa, D. P., García-Hermoso, A., Álvarez-Bueno, C., Sánchez-López, M. & Martinez-Vizcaino, V. Effectiveness of school-based physical activity programmes on cardiorespiratory fitness in children: A meta-Analysis of randomised controlled trials. *British Journal of Sports Medicine* **52,** 1234–1240 (2018).

Prakash, R. S., Voss, M. W., Erickson, K. I. & Kramer, A. F. Physical activity and cognitive vitality. *Annual Review of Psychology* **66,** 769–797 (2015).

Prochaska, J. J., Sallis, J. F., Slymen, D. J. & McKenzie, T. L. A longitudinal study of children's enjoyment of physical education. *Pediatric exercise science* **15,** 170–178 (2003).

Pugh, R. C. & Hu, Y. Use and Interpretation of Canonical Correlation Analyses in Journal of Educational Research Articles: 1978-1989. *Journal of Educational Research* **84,** 147–152 (1991).

Pustejovsky, J. E. & Tipton, E. Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models. *Journal of Business and Economic Statistics* **36,** 672–683 (2018).

Quan, M., Zhang, H., Zhang, J., Zhou, T., Zhang, J., Zhao, G., Fang, H., Sun, S., Wang, R. & Chen, P. Preschoolers' Technology-Assessed Physical Activity and Cognitive Function: A Cross-Sectional Study. *Journal of Clinical Medicine* **7,** 108 (2018).

Quartagno, M. & Carpenter, J. R. Multiple imputation for IPD meta-analysis : allowing for heterogeneity and studies with missing covariates. *Statistics in medicine* **35,** 2938–2954 (2016).

Quartagno, M. & Carpenter, J. R. *Multilevel multiple imputation in presence of interactions, non-linearities and random slopes Studies in Theoretical and Applied Statistics-SIS2018-49th Meeting of the Italian Statistical Society, Palermo* (2018), 20–22. http://researchonline.lshtm.ac.uk/4648780/.

R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2020). https://www.r-project.org/.

Raghunathan, T., Lepkowski, J., Van Hoewyk, J. & Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology* **27,** 85–96 (2001).

Raichlen, D. A., Klimentidis, Y. C., Bharadwaj, P. K. & Alexander, G. E. Differential associations of engagement in physical activity and estimated cardiorespiratory fitness with brain volume in middle-aged to older adults. *Brain Imaging and Behavior,* 1–10 (2019).

Raichlen, D. A. & Alexander, G. E. Adaptive Capacity: An Evolutionary Neuroscience Model Linking Exercise, Cognition, and Brain Health. *Trends in Neurosciences* **40,** 408–421 (2017).

Raine, L. B., Kao, S.-C., Pindus, D., Westfall, D. R., Shigeta, T. T., Logan, N., Cadenas-Sanchez, C., Li, J., Drollette, E. S. & Pontifex, M. B. A Large-Scale Reanalysis of Childhood Fitness and Inhibitory Control. *Journal of Cognitive Enhancement* **2,** 170–192 (2018).

Rathbone, J., Hoffmann, T. & Glasziou, P. Faster title and abstract screening? Evaluating Abstrackr , a semi-automated online screening program for systematic reviewers. *Systematic Reviews* **4,** 1–7 (2015).

Rathore, A. & Lom, B. The effects of chronic and acute physical activity on working memory performance in healthy participants: A systematic review with meta-analysis of randomized controlled trials. *Systematic Reviews* **6,** 124 (2017).

Raudenbush, S. & Bryk, A. S. A Hierarchical Model for Studying School Effects. *Sociology of Education* **59,** 1–17 (1986).

Raykov, T. On testability of missing data mechanisms in incomplete data sets. *Structural Equation Modeling* **18,** 419–429 (2011).

Reich, N. G., Myers, J. A., Obeng, D., Milstone, A. M. & Perl, T. M. Empirical power and sample size calculations for cluster-randomized and cluster-randomized crossover studies. *PLoS ONE* **7,** e35564 (2012).

Reilly, J. J., Penpraze, V., Hislop, J., Davies, G., Grant, S. & Paton, J. Y. Objective measurement of physical activity and sedentary behaviour: Review with new data. *Archives of Disease in Childhood* **93,** 614–619 (2008).

Reis, R. S., Salvo, D., Ogilvie, D., Lambert, E. V., Goenka, S. & Brownson, R. C. Scaling up physical activity interventions worldwide: stepping up to larger and smarter approaches to get people moving. *The Lancet* **388,** 1337–1348 (2016).

Rendeiro, C. & Rhodes, J. S. A new perspective of the hippocampus in the origin of exercise–brain interactions. *Brain Structure and Function* **223,** 2527–2545 (2018).

Resaland, G. K., Aadland, E., Fusche, V., Aadland, K. N., Skrede, T., Stavnsbo, M., Suominen, L., Steene-johannessen, J., Glosvik, Ø., Andersen, J. R., Kvalheim, O. M., Engelsrud, G., Andersen, L. B., Holme, I. M., Ommundsen, Y., Kriemler, S., Mechelen, W. V., Mckay, H. A., Ekelund, U. & Anderssen, S. A. Effects of physical activity on schoolchildren's academic performance: The Active Smarter Kids (ASK) cluster-randomized controlled trial. *Preventive Medicine* **91,** 322–328 (2016).

Resche-Rigon, M. & White, I. R. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Statistical Methods in Medical Research* **27,** 1634–1649 (2018).

Resche-Rigon, M., White, I. R., Bartlett, J. W., Peters, S. A. & Thompson, S. G. Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Statistics in Medicine* **32,** 4890–4905 (2013).

Reuter, M., Tisdall, M. D., Qureshi, A., Buckner, R. L., Van Der Kouwe, A. J. & Fischl, B. Head motion during MRI acquisition reduces gray matter volume and thickness estimates. *NeuroImage* **107,** 107–115 (2015).

Revelle, W. *psych: Procedures for Psychological, Psychometric, and Personality Research, R package version 1.8.12* Evanston, Illinois, 2018. https://cran.r-project.org/package=psych.

Rey-Mermet, A., Gade, M. & Oberauer, K. Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning Memory and Cognition* **44,** 501–526 (2018).

Richardson, M., Garner, P. & Donegan, S. Cluster randomised trials in cochrane reviews: Evaluation of methodological and reporting practice. *PLoS ONE* **11,** e0151818 (2016).

Richter, M., Erhart, M., Vereecken, C. A., Zambon, A., Boyce, W. & Gabhainn, S. N. The role of behavioural factors in explaining socio-economic differences in adolescent health: A multilevel study in 33 countries. *Social Science and Medicine* **69,** 396–403 (2009).

Ridley, K., Ainsworth, B. E. & Olds, T. S. Development of a Compendium of Energy Expenditures for youth. *International Journal of Behavioral Nutrition and Physical Activity* **5,** 45 (2008).

Riebe, D., Franklin, B. A., Thompson, P. D., Garber, C. E., Whitfield, G. P., Magal, M. & Pescatello, L. S. Updating ACSM's recommendations for exercise preparticipation health screening. *Medicine and Science in Sports and Exercise* **47,** 2473–2479 (2015).

Rioux, C. & Little, T. D. Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development,* 0165025419880609 (2019).

Rivis, A. & Sheeran, P. Social influences and the theory of planned behaviour: Evidence for a direct relationship between prototypes and young people's exercise behaviour. *Psychology and Health* **18,** 567–583 (2003).

Rodriguez-Ayllon, M., Esteban-Cornejo, I., Verdejo-Román, J., Muetzel, R. L., Mora-Gonzalez, J., Cadenas-Sanchez, C., Plaza-Florido, A., Molina-Garcia, P., Kramer, A. F., Catena, A. & Ortega, F. B. Physical fitness and white matter microstructure in children with overweight or obesity: the ActiveBrains project. *Scientific Reports* **10,** 1–9 (2020).

Rodriguez-Ayllon, M., Derks, I. P., Van Den Dries, M. A., Esteban-Cornejo, I., Labrecque, J. A., Yang-Huang, J., Raat, H., Vernooij, M. W., White, T., Ortega, F. B., Tiemeier, H. & Muetzel, R. L. Associations of physical activity and screen time with white matter microstructure in children from the general population. *NeuroImage* **205,** 116258 (2020).

Rogers, R. D. & Monsell, S. Costs of a Predictable Switch Between Simple Cognitive Tasks. *Journal of Experimental Psychology* **124,** 207–231 (1995).

Rohde, T. E. & Thompson, L. A. Predicting academic achievement with cognitive ability. *Intelligence* **35,** 83–92 (2007).

Rollo, S., Antsygina, O. & Tremblay, M. S. The whole day matters: Understanding 24-hour movement guideline adherence and relationships with health indicators across the lifespan. *Journal of Sport and Health Science* (2020).

Rombach, I., Rivero-Arias, O., Gray, A. M., Jenkinson, C. & Burke, Ó. The current practice of handling and reporting missing outcome data in eight widely used PROMs in RCT publications: a review of the current literature. *Quality of Life Research* **25,** 1613–1623 (2016).

Romine, C. B. & Reynolds, C. R. A model of the development of frontal lobe functioning: Findings from a meta-analysis. *Applied Neuropsychology* **12,** 190–201 (2005).

Rosano, C., Guralnik, J., Pahor, M., Glynn, N. W., Newman, A. B., Ibrahim, T. S., Erickson, K., Cohen, R., Shaaban, C. E., MacCloud, R. L. & Aizenstein, H. J. Hippocampal Response to a 24-Month Physical Activity Intervention in Sedentary Older Adults. *American Journal of Geriatric Psychiatry* **25,** 209–217 (2017).

Ross, N., Yau, P. L. & Convit, A. Obesity, fitness, and brain integrity in adolescence. *Appetite* **93,** 44–50 (2015).

Roth, T. C., Brodin, A., Smulders, T. V., Ladage, L. D. & Pravosudov, V. V. Is bigger always better? A critical appraisal of the use of volumetric analysis in the study of the hippocampus. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365,** 915–931 (2010).

Rouder, J. N., Kumar, A. & Haaf, J. M. Why most studies of individual differences with inhibition tasks are bound to fail. *PsyArXiv,* doi: 10.31234/osf.io/3cjr5 (2019).

Routen, A. C., Upton, D., Edwards, M. G. & Peters, D. M. Discrepancies in accelerometer-measured physical activity in children due to cut-point non-equivalence and placement site. *Journal of Sports Sciences* **30,** 1303–1310 (2012).

Rowland, T. W. Does peak VO2 reflect VO2max in children?: Evidence from supra-maximal testing. *Medicine and Science in Sports and Exercise* **25,** 689–693 (1993).

Rowlands, A. V., Edwardson, C. L., Davies, M. J., Khunti, K., Harrington, D. M. & Yates, T. Beyond Cut Points: Accelerometer Metrics that Capture the Physical Activity Profile. *Medicine and Science in Sports and Exercise* **50,** 1323–1332 (2018).

Rowlands, A. V., Mirkes, E. M., Yates, T., Clemes, S., Davies, M., Khunti, K. & Edwardson, C. L. Accelerometer-assessed physical activity in epidemiology: Are monitors equivalent? *Medicine and Science in Sports and Exercise* **50,** 257–265 (2018).

RStudio Team. *RStudio: Integrated Development Environment for R* RStudio, Inc. (Boston, MA, 2016). http://www.rstudio.com/.

Rubin, D. & Little, R. *Statistical Analysis with Missing Data* (John Wiley & Sons, Hoboken, 2002).

Rubin, D. B. Inference and Missing Data. *Biometrika* **63,** 581–592 (1976).

Rubin, D. B. *Multiple Imputation for NonResponse in Surveys* (Wiley, New-York, 1987).

Ruiz-Ariza, A., Grao-Cruces, A., Marques De Loureiro, N. E. & Martínez-López, E. J. Influence of physical fitness on cognitive and academic performance in adolescents: A systematic review from 2005–2015. *International Review of Sport and Exercise Psychology* **10,** 108–133 (2017).

Ruiz, J. R., Castro-Piñero, J., Artero, E. G., Ortega, F. B., Sjöström, M., Suni, J., Castillo, M. J., Sjo, M., Suni, J. & Castillo, M. J. Predictive validity of health-related fitness in youth: A systematic review. *British Journal of Sports Medicine* **43,** 909–923 (2009).

Ruiz, J. R., Ortega, F. B., Castillo, R., Martín-Matillas, M., Kwak, L., Vicente-Rodríguez, G., Noriega, J., Tercedor, P., Sjöström, M. & Moreno, L. A. Physical activity, fitness, weight status, and cognitive performance in adolescents. *Journal of Pediatrics* **157,** 917–922 (2010).

Ruiz, J. R., Silva, G., Oliveira, N., Ribeiro, J. C., Oliveira, J. F. & Mota, J. Criterion-related validity of the 20-m shuttle run test in youths aged 13-19 years. *Journal of Sports Sciences* **27,** 899–906 (2009).

Ruotsalainen, I., Gorbach, T., Perkola, J., Renvall, V., Syväoja, H. J., Tammelin, T. H., Karvanen, J. & Parviainen, T. Physical activity, aerobic fitness, and brain white matter: Their role for executive functions in adolescence. *Developmental Cognitive Neuroscience* **42,** 100765 (2020).

Ruotsalainen, I., Renvall, V., Gorbach, T., Syväoja, H. J., Tammelin, T. H., Karvanen, J. & Parviainen, T. Aerobic fitness, but not physical activity, is associated with grey matter volume in adolescents. *Behavioural Brain Research* **362,** 122–130 (2019).

Rutterford, C., Copas, A. & Eldridge, S. Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology* **44,** 1051–1067 (2015).

Sabia, S., Van Hees, V. T., Shipley, M. J., Trenell, M. I., Hagger-Johnson, G., Elbaz, A., Kivimaki, M. & Singh-Manoux, A. Association between questionnaire- and accelerometer-assessed physical activity: The role of sociodemographic factors. *American Journal of Epidemiology* **179,** 781–790 (2014).

Sagi, Y., Tavor, I., Hofstetter, S., Tzur-Moryosef, S., Blumenfeld-Katzir, T. & Assaf, Y. Learning in the Fast Lane: New Insights into Neuroplasticity. *Neuron* **73,** 1195–1203 (2012).

Saint-Maurice, P. F., Kim, Y., Welk, G. J. & Gaesser, G. A. Kids are not little adults: what MET threshold captures sedentary behavior in children? *European Journal of Applied Physiology* **116,** 29–38 (2016).

Sallis, J. F. Self-Report Measures of Children's Physical Activity. *Journal of School Health* **61,** 215–220 (1991).

Sallis, J. F., Bull, F., Guthold, R., Heath, G. W., Inoue, S., Kelly, P., Oyeyemi, A. L., Perez, L. G., Richards, J. & Hallal, P. C. Progress in physical activity over the Olympic quadrennium. *The Lancet* **388,** 1325–1336 (2016).

Sallis, J. F. & Saelens, B. E. Assessment of physical activity by self-report: status, limitations, and future directions. *Research Quarterly for Exercise & Sport* **71,** S1–14 (2000).

Salvan, P., Wassenaar, T., Wheatley, C., Beale, N., Cottaar, M., Papp, D., Bastiani, M., Fitzgibbon, S., Duff, E., Andersson, J., Winkler, A. M., Douaud, G., Nichols, T. E., Smith, S., Dawes, H. & Johansen-Berg, H. Multimodal imaging brain markers in early adolescence are linked with a physically active lifestyle. *bioRxiv,* doi: https://doi.org/10.1101/2020.09.03.269571 (2020).

Sampaio-Baptista, C., Khrapitchev, A. A., Foxley, S., Schlagheck, T., Scholz, J., Jbabdi, S., DeLuca, G. C., Miller, K. L., Taylor, A., Thomas, N., Kleim, J., Sibson, N. R., Bannerman, D. & Johansen-Berg, H. Motor skill learning induces changes in white matter microstructure and myelination. *Journal of Neuroscience* **33,** 19499–19503 (2013).

Sampasa-Kanyinga, H., Standage, M., Tremblay, M. S., Katzmarzyk, P. T., Hu, G., Kuriyan, R., Maher, C., Maia, J., Olds, T., Sarmiento, O. L., Tudor-Locke, C. & Chaput, J. P. Associations between meeting combinations of 24-h movement guidelines and health-related quality of life in children from 12 countries. *Public Health* **153,** 16–24 (2017).

Sánchez-Benavides, G., Gómez-Ansón, B., Sainz, A., Vives, Y., Delfino, M. & Peña-Casanova, J. Manual validation of FreeSurfer's automated hippocampal segmentation in normal aging, mild cognitive impairment, and Alzheimer Disease subjects. *Psychiatry Research - Neuroimaging* **181,** 219–225 (2010).

Sánchez-López, M., Martínez-Vizcaíno, V., García-Hermoso, A., Jiménez-Pavón, D. & Ortega, F. B. Construct validity and test-retest reliability of the International Fitness Scale (IFIS) in Spanish children aged 9-12 years. *Scandinavian Journal of Medicine and Science in Sports* **25,** 543–551 (2015).

Sandercock, G. R. H. & Cohen, D. D. Temporal trends in muscular fitness of English 10-year-olds 1998 – 2014: An allometric approach. *Journal of Science and Medicine in Sport* **22,** 201–205 (2019).

Sato, K., Kerever, A., Kamagata, K., Tsuruta, K., Irie, R., Tagawa, K., Okazawa, H., Arikawa-Hirasawa, E., Nitta, N., Aoki, I. & Aoki, S. Understanding microstructure of the brain by comparison of neurite orientation dispersion and density imaging (NODDI) with transparent mouse brain. *Acta Radiologica Open* **6,** 205846011770381 (2017).

Satterthwaite, T. D., Shinohara, R. T., Wolf, D. H., Hopson, R. D., Elliott, M. A., Vandekar, S. N., Ruparel, K., Calkins, M. E., Roalf, D. R., Gennatas, E. D., Jackson, C., Erus, G., Prabhakaran, K., Davatzikos, C., Detre, J. A., Hakonarson, H., Gur, R. C. & Gur, R. E. Impact of puberty on the evolution of cerebral perfusion during adolescence. *Proceedings of the National Academy of Sciences* **111,** 8643–8648 (2014).

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., Gur, R. C. & Gur, R. E. Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage* **60,** 623–632 (2012).

Saunders, T. J., Gray, C. E., Poitras, V. J., Chaput, J. P., Janssen, I., Katzmarzyk, P. T., Olds, T., Connor Gorber, S., Kho, M. E., Sampson, M., Tremblay, M. S. & Carson, V. Combinations of physical activity, sedentary behaviour and sleep: Relationships with health indicators in school-aged children and youth. *Applied Physiology, Nutrition and Metabolism* **41,** S283–S293 (2016).

Savage, N. Made to measure. *Nature* **527,** S12 (2015).

Sawyer, S. M., Afifi, R. A., Bearinger, L. H., Blakemore, S. J., Dick, B., Ezeh, A. C. & Patton, G. C. Adolescence: A foundation for future health. *The Lancet* **379,** 1630–1640 (2012).

Sawyer, S. M., Azzopardi, P. S., Wickremarathne, D. & Patton, G. C. The age of adolescence. *The Lancet child and Adolescent Health* **2,** 223–228 (2018).

Schafer, J. L. *Analysis of incomplete multivariate data* (Chapman and Hall/CRC, 1997).

Schafer, J. L. J. L. & Graham, J. W. Missing data: Our view of the state of the art. *Psychological Methods* **7,** 147–177 (2002).

Schafer, J. L. & Yucel, R. M. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics* **11,** 437–457 (2002).

Scheuren, F. Multiple imputation: How it began and continues. *American Statistician* **59,** 315–319 (2005).

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z. & Araya-Ajoy, Y. G. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution* **11,** 1141–1152 (2020).

Schilling, K. G., Janve, V., Gao, Y., Stepniewska, I., Landman, B. A. & Anderson, A. W. Histological validation of diffusion MRI fiber orientation distributions and dispersion. *NeuroImage* **165,** 200–221 (2018).

Schmidt, M., Egger, F., Roebers, C. M. & Conzelmann, A. Cognitively Engaging Chronic Physical Activity, But Not Aerobic Exercise, Affects Executive Functions in Primary School Children: A Group-Randomized Controlled Trial. *Journal of Sport and Exercise Psychology* **37,** 575–591 (2015).

Schmitt, J. A., Benton, D. & Kallus, K. W. General methodological considerations for the assessment of nutritional influences on human cognitive functions. *European Journal of Nutrition* **44,** 459–464 (2005).

Schmitz, J., Fraenz, C., Schlüter, C., Friedrich, P., Jung, R. E., Güntürkün, O., Genç, E. & Ocklenburg, S. Hemispheric asymmetries in cortical gray matter microstructure identified by neurite orientation dispersion and density imaging. *NeuroImage* **189,** 667–675 (2019).

Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., Gauthier, S. & Pruessner, J. C. Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *NeuroImage* **129,** 1–14 (2016).

Schoenwald, S. K. & Garland, A. F. A Review of Treatment Adherence Measurement Methods. *Psychological Assessment* **25,** 146–156 (2013).

Scholz, J., Allemang-Grand, R., Dazai, J. & Lerch, J. P. Environmental enrichment is associated with rapid volumetric brain changes in adult mice. *NeuroImage* **109,** 190–198 (2015).

Schouten, J. F. & Bekker, J. A. M. Reactlon tine and accuracy. *Acta psychologica* **27,** 143–153 (1967).

Schultes, M. T., Jöstl, G., Finsterwald, M., Schober, B. & Spiel, C. Measuring intervention fidelity from different perspectives with multiple methods: The Reflect program as an example. *Studies in Educational Evaluation* **47,** 102–112 (2015).

Schulze-Neick, I., Wessel, H. & Paul, M. Heart rate and oxygen uptake response to exercise in children with low peak exercise heart rate. *European journal of pediatrics* **151,** 160–166 (1992).

Schutte, N. M., Nederend, I., Hudziak, J. J., Bartels, M. & De Geus, E. J. Twin-sibling study and meta-analysis on the heritability of maximal oxygen consumption. *Physiological Genomics* **48,** 210–219 (2016).

Schwarzfischer, P., Gruszfeld, D., Stolarczyk, A., Ferre, N., Escribano, J., Rousseaux, D., Moretti, M., Mariani, B., Verduci, E., Koletzko, B. & Grote, V. Physical activity and sedentary behavior from 6 to 11 years. *Pediatrics* **143** (2019).

Scott, J. J., Rowlands, A. V., Cliff, D. P., Morgan, P. J., Plotnikoff, R. C. & Lubans, D. R. Comparability and feasibility of wrist- and hip-worn accelerometers in free-living adolescents. *Journal of Science and Medicine in Sport* **20,** 1101–1106 (2017).

Scott, J. J., Morgan, P. J., Plotnikoff, R. C. & Lubans, D. R. Reliability and validity of a single-item physical activity measure for adolescents. *Journal of paediatrics and child health* **51,** 787–793 (2015).

Scott, S. N., Springer, C. M., Oody, J. F., McClanahan, M. S., Wiseman, B. D., Kybartas, T. J. & Coe, D. P. Development and validation of a PACER prediction equation for VO2peak in 10-to 15-year-old youth. *Pediatric exercise science* **31,** 223–228 (2019).

Seaman, S. R., Bartlett, J. W. & White, I. R. Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology* **12,** 46 (2012).

Sedaghat, S., Cremers, L. G., De Groot, M., Hofman, A., Van Der Lugt, A., Niessen, W. J., Franco, O. H., Dehghan, A., Ikram, M. A. & Vernooij, M. W. Lower microstructural integrity of brain white matter is related to higher mortality. *Neurology* **87,** 927–934 (2016).

Sellström, E. & Bremberg, S. Is there a "school effect" on pupil outcomes? A review of multilevel studies. *Journal of Epidemiology and Community Health* **60,** 149–155 (2006).

Sember, V., Jurak, G., Kovač, M., Morrison, S. A. & Starc, G. Children's Physical Activity, Academic Performance, and Cognitive Functioning: A Systematic Review and Meta-Analysis. *Frontiers in Public Health* **8** (2020).

Senjem, M. L., Gunter, J. L., Shiung, M. M., Petersen, R. C. & Jack, C. R. Comparison of different methodological implementations of voxel-based morphometry in neurodegenerative disease. *NeuroImage* **26,** 600–608 (2005).

Sexton, C. E., Betts, J. F., Demnitz, N., Dawes, H., Ebmeier, K. P. & Johansen-Berg, H. A systematic review of MRI studies examining the relationship between physical fitness and activity and the white matter of the ageing brain. *NeuroImage* **131,** 81–90 (2016).

Sexton, C. E., Betts, J. F., Dennis, A., Doherty, A., Leeson, P., Holloway, C., Dall'Armellina, E., Winkler, A. M., Demnitz, N., Wassenaar, T., Dawes, H. & Johansen-Berg, H. The effects of an aerobic training intervention on cognition, grey matter volumes and white matter microstructure. *Physiology and Behavior* **223,** 112923 (2020).

Shallice, T. Specific impairments of planning. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **298,** 199–209 (1982).

Shallice, T. The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* **351,** 1405–1412 (1996).

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., Evans, A., Rapoport, J. & Giedd, J. Intellectual ability and cortical development in children and adolescents. *Nature* **440,** 676–679 (2006).

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E. & Henry, D. A. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* **358,** j4008 (2017).

Shephard, R. J. Limits to the measurement of habitual physical activity by questionnaires: Commentary. *British Journal of Sports Medicine* **37,** 197–206 (2003).

Sherry, A. & Henson, R. K. Conducting and Interpreting Canonical Correlation Analysis in Personality Research: A User-Friendly Primer. *Journal of Personality Assessment* **84,** 37–48 (2005).

Shigeta, T. T., Leahy, A. A., Smith, J. J., Eather, N., Lubans, D. R. & Hillman, C. H. Cardiorespiratory and muscular fitness associations with older adolescent cognitive control. *Journal of Sport and Health Science* (2020).

Shrout, P. E. & Fleiss, J. L. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* **86,** 420–428 (1979).

Sibley, B. A. & Etnier, J. L. The relationship between physical activity and cognition in children: A meta-analysis. *Pediatric exercise science* **15,** 243–256 (2003).

Sierra, A., Gröhn, O. & Pitkänen, A. Imaging microstructural damage and plasticity in the hippocampus during epileptogenesis. *Neuroscience* **309,** 162–172 (2015).

Silk, T. J. & Wood, A. G. Lessons About Neurodevelopment From Anatomical Magnetic Resonance Imaging. *Journal of Deve* **32,** 158–168 (2011).

Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal,* https://doi.org/10.2139/ssrn.2694998 (2015). (2015).

Simonsohn, U., Simmons, J. P. & Nelson, L. D. Specification curve analysis. *Nature Human Behaviour* (2020).

Simpson, R. J., Campbell, J. P., Gleeson, M., Krüger, K., Nieman, D. C., Pyne, D. B., Turner, J. E. & Walsh, N. P. Can exercise affect immune function to increase susceptibility to infection? *Exercise Immunology Review* **26,** 8–22 (2020).

Singh, A. S., Saliasi, E., Van Den Berg, V., Uijtdewilligen, L., De Groot, R. H., Jolles, J., Andersen, L. B., Bailey, R., Chang, Y. K., Diamond, A., Ericsson, I., Etnier, J. L., Fedewa, A. L., Hillman, C. H., McMorris, T., Pesce, C., Pühse, U., Tomporowski, P. D. & Chinapaw, M. J. Effects of physical activity interventions on cognitive and academic performance in children and adolescents: A novel combination of a systematic review and recommendations from an expert panel. *British Journal of Sports Medicine* **53,** 640–647 (2019).

Singmann, H. & Kellen, D. *New Methods in Cognitive Psychology* (eds Spieler, D. H. & Schumacher, E.) 4–31 (Psychology Press, 2019).

Skog, H., Lintu, N., Haapala, H. L. & Haapala, E. A. Associations of cardiorespiratory fitness, adiposity, and arterial stiffness with cognition in youth. *Physiological Reports* **8,** e14586 (2020).

Smith, J. J., Morgan, P. J., Plotnikoff, R. C., Dally, K. A., Salmon, J., Okely, A. D., Finn, T. L. & Lubans, D. R. Smart-phone obesity prevention trial for adolescent boys in low-income communities: The ATLAS RCT. *Pediatrics* **134,** e723–e731 (2014).

Smith, J. J., Eather, N., Morgan, P. J. & Lubans, D. R. The Health Benefits of Muscular Fitness for Children and Adolescents: A Systematic Review and Meta-Analysis. *Sports Medicine* **44,** 1209–1223 (2014).

Smith, M. P., Standl, M., Heinrich, J. & Schulz, H. Accelerometric estimates of physical activity vary unstably with data handling. *PLoS ONE* **12,** e0187706 (2017).

Smith, S. M. Fast robust automated brain extraction. *Human Brain Mapping* **17,** 143–155 (2002).

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M. & Behrens, T. E. J. Tract-based spatial statistics: Voxelwise analysis of multi-subject diffusion data. *NeuroImage* **31,** 1487–1505 (2006).

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M. & Matthews, P. M. Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23,** 208–219 (2004).

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., J Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C. & Miller, K. L. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience* **18,** 1565–1567 (2015).

Sneck, S., Viholainen, H., Syväoja, H., Kankaanpää, A., Hakonen, H., Poikkeus, A.-M. & Tammelin, T. Effects of school-based physical activity on mathematics performance in children: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity* **16,** 109 (2019).

Snijders, T. A. & Bosker, R. J. *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (Sage, 2012).

Snoek, L., Miletić, S. & Scholte, H. S. How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage* **184,** 741–760 (2019).

Soga, K., Masaki, H., Gerber, M. & Ludyga, S. Acute and Long-term Effects of Resistance Training on Executive Function. *Journal of Cognitive Enhancement* **2,** 200–207 (2018).

Solis-Urra, P., Olivares-Arancibia, J., Suarez-Cadenas, E., Sanchez-Martinez, J., Rodríguez-Rodríguez, F., Ortega, F. B., Esteban-Cornejo, I., Cadenas-Sanchez, C., Castro-Piñero, J., Veloz, A., Chabert, S., Saradangani, K. P., Zavala-Crichton, J. P., Migueles, J. H., Mora-Gonzalez, J., Quiroz-Escobar, M., Almonte-Espinoza, D., Urzúa, A., Dragicevic, C. D., Astudillo, A., Méndez-Gassibe, E., Riquelme-Uribe, D., Azagra, M. J. & Cristi-Montero, C. Study protocol and rationale of the "Cogni-action project" a cross-sectional and randomized controlled trial about physical activity, brain health, cognition, and educational achievement in schoolchildren. *BMC Pediatrics* **19,** 260 (2019).

Sørensen, Ø., Walhovd, K. B. & Fjell, A. A Recipe for Accurate Estimation of Lifespan Brain Trajectories, Distinguishing Longitudinal and Cohort Effects. *NeuroImage,* 117596 (2021).

Sowell, E. R., Peterson, B. S., Thompson, P. M., Welcome, S. E., Henkenius, A. L. & Toga, A. W. Mapping cortical change across the human life span. *Nature Neuroscience* **6,** 309–315 (2003).

Spallazzi, M., Dobisch, L., Becke, A., Berron, D., Stucht, D., Oeltze-Jafra, S., Caffarra, P., Speck, O. & Düzel, E. Hippocampal vascularization patterns: A high-resolution 7 Tesla time-of-flight magnetic resonance angiography study. *NeuroImage: Clinical* **21,** 101609 (2019).

Spartano, N. L., Davis-Plourde, K. L., Himali, J. J., Andersson, C., Pase, M. P., Maillard, P., DeCarli, C., Murabito, J. M., Beiser, A. S., Vasan, R. S. & Seshadri, S. Association of Accelerometer-Measured Light-Intensity Physical Activity With Brain Volume: The Framingham Heart Study. *JAMA network open* **2,** e192745 (2019).

Spearman, C. Correlation Calculated From Faulty Data. *British Journal of Psychology* **3,** 271–295 (1910).

Spilsbury, J. C., Drotar, D., Rosen, C. L. & Redline, S. The Cleveland Adolescent Sleepiness Questionnaire: A new measure to assess excessive daytime sleepiness in adolescents. *Journal of Clinical Sleep Medicine* **3,** 603–612 (2007).

Spirduso, W. W. *Physical fitness, aging, and psychomotor speed: A review* 1980.

Sport England. *Active lives children and young people survey attitudes towards sport and physical activity (academic year 2017/18)* tech. rep. December (2018), 1–34. https://www.sportengland.org/media/13698/active-lives-children-survey-academic-year-17-18.pdf.

Squeglia, L. M., Tapert, S. F., Sullivan, E. V., Jacobus, J., Meloy, M. J., Rohlfing, T. & Pfefferbaum, A. Brain development in heavy-drinking adolescents. *American Journal of Psychiatry* **172,** 531–542 (2015).

Stekhoven, D. J. & Bühlmann, P. Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28,** 112–118 (2012).

Stenling, A., Sörman, D. E., Lindwall, M., Hansson, P., Körning Ljungberg, J. & Machado, L. Physical activity and cognitive function: between-person and within-person associations and moderators. *Aging, Neuropsychology, and Cognition,* 1–26 (2020).

Sterdt, E., Liersch, S. & Walter, U. Correlates of Physical Activity of Children and Adolescents: A Systematic Review of Reviews. *Health Education Journal* **73,** 72–89 (2014).

Stern, Y. Cognitive reserve in ageing and Alzheimer's disease. *The Lancet Neurology* **11,** 1006–1012 (2012).

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. & Carpenter, J. R. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* **338** (2009).

Stewart, D. & Love, W. A General Canonical Correlation Index. *Psychological Bulletin* **70,** 160 (1968).

Stiles, J. & Jernigan, T. L. The basics of brain development. *Neuropsychology Review* **20,** 327–348 (2010).

Stillman, C. M., Cohen, J., Lehman, M. E. & Erickson, K. I. Mediators of Physical Activity on Neurocognitive Function: A Review at Multiple Levels of Analysis. *Frontiers in Human Neuroscience* **10,** 626 (2016).

Stillman, C. M., Donofry, S. D. & Erickson, K. I. Exercise, Fitness and the Aging Brain: A Review of Functional Connectivity in Aging. *Archives of Psychology* **3** (2019).

Stillman, C. M. & Erickson, K. I. Physical activity as a model for health neuroscience. *Annals of the New York Academy of Sciences* **1428,** 103–111 (2018).

Stillman, C. M., Esteban-Cornejo, I., Brown, B., Bender, C. M. & Erickson, K. I. Effects of Exercise on Brain and Cognition Across Age Groups and Health States. *Trends in Neurosciences* (2020).

Stillman, C. M., Uyar, F., Huang, H., Grove, G. A., Watt, J. C., Wollam, M. E. & Erickson, K. I. Cardiorespiratory fitness is associated with enhanced hippocampal functional connectivity in healthy young adults. *Hippocampus* **28,** 239–247 (2018).

Strack, F. & Deutsch, R. Reflective and impulsive determinants of social behavior. *Personality and social psychology review* **8,** 220–247 (2004).

Strange, B. A., Witter, M. P., Lein, E. S. & Moser, E. I. Functional organization of the hippocampal longitudinal axis. *Nature Reviews Neuroscience* **15,** 655–669 (2014).

Streitbürger, D. P., Möller, H. E., Tittgemeyer, M., Hund-Georgiadis, M., Schroeter, M. L. & Mueller, K. Investigating Structural Brain Changes of Dehydration Using Voxel-Based Morphometry. *PLoS ONE* **7,** e44195 (2012).

Stroth, S., Kubesch, S., Dieterle, K., Ruchsow, M., Heim, R. & Kiefer, M. Physical fitness, but not acute exercise modulates event-related potential indices for executive control in healthy adolescents. *Brain Research* **1269,** 114–124 (2009).

Stüber, C., Morawski, M., Schäfer, A., Labadie, C., Wähnert, M., Leuze, C., Streicher, M., Barapatre, N., Reimann, K., Geyer, S., Spemann, D. & Turner, R. Myelin and iron concentration in the human brain: A quantitative study of MRI contrast. *NeuroImage* **93,** 95–106 (2014).

Suarez-Manzano, S., Ruiz-Ariza, A., De La Torre-Cruz, M. & Martínez-López, E. J. Acute and chronic effect of physical activity on cognition and behaviour in young people with ADHD: A systematic review of intervention studies. *Research in Developmental Disabilities* **77,** 12–23 (2018).

Sui, J., Adali, T., Yu, Q., Chen, J. & Calhoun, V. D. A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of Neuroscience Methods* **204,** 68–81 (2012).

Sullivan, R. A., Kuzel, A. M. H. M. H., Vaandering, M. E. & Chen, W. The Association of Physical Activity and Academic Behavior: A Systematic Review. *Journal of School Health* **87,** 388–398 (2017).

Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P. & Lee, K. J. Should multiple imputation be the method of choice for handling missing data in randomized trials? *Statistical Methods in Medical Research* **27,** 2610–2626 (2018).

Sun, H., Walsh, A. J., Lebel, R. M., Blevins, G., Catz, I., Lu, J. Q., Johnson, E. S., Emery, D. J., Warren, K. G. & Wilman, A. H. Validation of quantitative susceptibility mapping with Perls' iron staining for subcortical gray matter. *NeuroImage* **105,** 486–492 (2015).

Sussman, D., Leung, R. C., Vogan, V. M., Lee, W., Trelle, S., Lin, S., Cassel, D. B., Chakravarty, M. M., Lerch, J. P., Anagnostou, E. & Taylor, M. J. The autism puzzle: Diffuse but not pervasive neuroanatomical abnormalities in children with ASD. *NeuroImage: Clinical* **8,** 170–179 (2015).

Suttorp, M. M., Siegerink, B., Jager, K. J., Zoccali, C. & Dekker, F. W. Graphical presentation of confounding in directed acyclic graphs. *Nephrology Dialysis Transplantation* **30,** 1418–1423 (2015).

Sylvia, L. G., Bernstein, E. E., Hubbard, J. L., Keating, L. & Anderson, E. J. A Practical Guide to Measuring Physical Activity. *Journal of the Academy of Nutrition and Dietetics* **114,** 199–208 (2014).

Syväoja, H. J., Kantomaa, M. T., Ahonen, T., Hakonen, H., Kankaanpää, A. & Tammelin, T. H. Response. *Medicine and Science in Sports and Exercise* **46,** 841 (2014).

Syväoja, H. J., Tammelin, T. H., Ahonen, T., Kankaanpää, A. & Kantomaa, M. T. The Associations of Objectively Measured Physical Activity and Sedentary Time with Cognitive Functions in School-Aged Children. *PLoS ONE* **9,** e103559 (2014).

Szuhany, K. L., Bugatti, M. & Otto, M. W. A meta-analytic review of the effects of exercise on brain-derived neurotrophic factor. *Journal of Psychiatric Research* **60,** 56–64 (2015).

Taljaard, M., Donner, A. & Klar, N. Imputation Strategies for Missing Continuous Outcomes in Cluster Randomized Trials. *Biometrical Journal* **50,** 329–345 (2008).

Taljaard, M., Weijer, C., Grimshaw, J. M., Eccles, M. P. & Ottawa Ethics of Cluster Randomised Trials Consensus Group. The Ottawa Statement on the ethical design and conduct of cluster randomised trials: precis for researchers and research ethics committees. *BMJ* **346,** f2838 (2013).

Tamnes, C. K., Bos, M. G., Van De Kamp, F. C., Peters, S. & Crone, E. A. Longitudinal development of hippocampal subregions from childhood to adulthood. *Developmental Cognitive Neuroscience* **30,** 212–222 (2018).

Tamnes, C. K., Fjell, A. M., Westlye, L. T., Østby, Y. & Walhovd, K. B. Becoming consistent: Developmental reductions in intraindividual variability in reaction time are related to white matter integrity. *Journal of Neuroscience* **32,** 972–982 (2012).

Tamnes, C. K., Herting, M. M., Goddings, A. L., Meuwese, R., Blakemore, S. J., Dahl, R. E., Güroğlu, B., Raznahan, A., Sowell, E. R., Crone, E. A. & Mills, K. L. Development of the cerebral cortex across adolescence: A multisample study of inter-related longitudinal changes in cortical volume, surface area, and thickness. *Journal of Neuroscience* **37,** 3402–3412 (2017).

Tamnes, C. K., Roalf, D. R., Goddings, A. L. & Lebel, C. Diffusion MRI of white matter microstructure development in childhood and adolescence: Methods, challenges and progress. *Developmental Cognitive Neuroscience* **33,** 161–175 (2018).

Tamnes, C. K., Walhovd, K. B., Dale, A. M., Østby, Y., Grydeland, H., Richardson, G., Westlye, L. T., Roddey, J. C., Hagler, D. J., Due-Tønnessen, P., Holland, D. & Fjell, A. M. Brain development and aging: Overlapping and unique patterns of change. *NeuroImage* **68,** 63–74 (2013).

Tamnes, C. K., Walhovd, K. B., Engvig, A., Grydeland, H., Krogsrud, S. K., Østby, Y., Holland, D., Dale, A. M. & Fjell, A. M. Regional hippocampal volumes and development predict learning and memory. *Developmental Neuroscience* **36,** 161–174 (2014).

Tamnes, C. K., Walhovd, K. B., Grydeland, H., Holland, D., Østby, Y., Dale, A. M. & Fjell, A. M. Longitudinal Working Memory Development Is Related to Structural Maturation of Frontal and Parietal Cortices. *Journal of Cognitive Neuroscience* **25,** 1611–1623 (2013).

Tanner, J. M. Fallacy of Per-Weight and Per-Surface Area Standards, and Their Relation to Spurious Correlation. *Journal of Applied Physiology* **2,** 1–15 (1949).

Tardif, C. L., Collins, D. L. & Pike, G. B. Sensitivity of voxel-based morphometry analysis to choice of imaging protocol at 3 T. *NeuroImage* **44,** 827–838 (2009).

Tardif, C. L., Collins, D. L. & Pike, G. B. Regional impact of field strength on voxel-based morphometry results. *Human Brain Mapping* **31,** 943–957 (2010).

Tardif, C. L., Gauthier, C. J., Steele, C. J., Bazin, P. L., Schäfer, A., Schaefer, A., Turner, R. & Villringer, A. Advanced MRI techniques to improve our understanding of experience-induced neuroplasticity. *NeuroImage* **131,** 55–72 (2016).

Tariq, M., Schneider, T., Alexander, D. C., Gandini Wheeler-Kingshott, C. A. & Zhang, H. Bingham-NODDI: Mapping anisotropic orientation dispersion of neurites using diffusion MRI. *NeuroImage* **133,** 207–223 (2016).

Tarp, J., Child, A., White, T., Westgate, K., Bugge, A., Grøntved, A., Wedderkopp, N., Andersen, L. B., Cardon, G., Davey, R., Janz, K. F., Kriemler, S., Northstone, K., Page, A. S., Puder, J. J., Reilly, J. J., Sardinha, L. B., Van Sluijs, E. M., Ekelund, U., Wijndaele, K. & Brage, S. Physical activity intensity, bout-duration, and cardiometabolic risk markers in children and adolescents. *International Journal of Obesity* **42,** 1639–1650 (2018).

Tarp, J., Domazet, S. L., Froberg, K., Hillman, C. H., Andersen, L. B. & Bugge, A. Effectiveness of a School-Based Physical Activity Intervention on Cognitive Performance in Danish Adolescents: LCoMotion — Learning , Cognition and Motion – A Cluster Randomized Controlled Trial. *PLoS ONE* **11,** e0158087 (2016).

Tau, G. Z. & Peterson, B. S. Normal development of brain circuits. *Neuropsychopharmacology* **35,** 147–168 (2010).

Taylor, C. The Reliability of Free School Meal Eligibility as a Measure of Socio-Economic Disadvantage: Evidence from the Millennium Cohort Study in Wales. *British Journal of Educational Studies* **66,** 29–51 (2018).

Taylor, K. L., Weston, M. & Batterham, A. M. Evaluating intervention fidelity: An example from a high-intensity interval training study. *PLoS ONE* **10,** e0125166 (2015).

Telzer, E. H., McCormick, E. M., Peters, S., Cosme, D., Pfeifer, J. H. & Van Duijvenvoorde, A. C. Methodological considerations for developmental longitudinal fMRI research. *Developmental Cognitive Neuroscience* **33,** 149–160 (2018).

Templer, K. J. & Lange, S. R. Internet testing: Equivalence between proctored lab and unproctored field conditions. *Computers in Human Behavior* **24,** 1216–1228 (2008).

Theil, H. The Analysis of Disturbances in Regression Analysis. *Journal of the American Statistical Association* **60,** 1067–1079 (1965).

Theil, H. A Simplification of the Blus Procedure for Analyzing Regression Disturbances. *Journal of the American Statistical Association* **63,** 242–251 (1968).

Theodoraki, T. E., McGeown, S. P., Rhodes, S. M. & MacPherson, S. E. Developmental changes in executive functions during adolescence: A study of inhibition, shifting, and working memory. *British Journal of Developmental Psychology* **38,** 74–89 (2020).

Thomas, A. G., Dennis, A., Bandettini, P. A. & Johansen-Berg, H. The effects of aerobic activity on brain structure. *Frontiers in Psychology* **3,** 86 (2012).

Thomas, A. G., Dennis, A., Rawlings, N. B., Stagg, C. J., Matthews, L., Morris, M., Kolind, S. H., Foxley, S., Jenkinson, M., Nichols, T. E., Dawes, H., Bandettini, P. A. & Johansen-Berg, H. Multi-modal characterization of rapid anterior hippocampal volume increase associated with aerobic exercise. *NeuroImage* **131,** 162–170 (2016).

Thomas, B. P., Welch, E. B., Niederhauser, B. D., Whetsell, W. O., Anderson, A. W., Gore, J. C., Avison, M. J. & Creasy, J. L. High-resolution 7T MRI of the human hippocampus in vivo. *Journal of Magnetic Resonance Imaging* **28,** 1266–1272 (2008).

Thomas, C., Sadeghi, N., Nayak, A., Trefler, A., Sarlls, J., Baker, C. I. & Pierpaoli, C. Impact of time-of-day on diffusivity measures of brain tissue derived from diffusion tensor imaging. *NeuroImage* **173,** 25–34 (2018).

Thompson, A. & Steinbeis, N. Sensitive periods in executive function development. *Current Opinion in Behavioral Sciences* **36,** 98–105 (2020).

Thompson, D., Peacock, O., Western, M. & Batterham, A. M. Multidimensional physical activity: An opportunity, not a problem. *Exercise and Sport Sciences Reviews* **43,** 67–74 (2015).

Timmer, K., Calabria, M., Branzi, F. M., Baus, C. & Costa, A. On the reliability of switching costs across time and domains. *Frontiers in Psychology* **9,** 1032 (2018).

Tipton, C. M. The history of "Exercise Is Medicine" in ancient civilizations. *Advances in Physiology Education* **38,** 109–117 (2015).

Tomarken, A. J. & Waller, N. G. Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology* **1,** 31–65 (2005).

Tomkinson, G. R., Carver, K. D., Atkinson, F., Daniell, N. D., Lewis, L. K., Fitzgerald, J. S., Lang, J. J. & Ortega, F. B. European normative values for physical fitness in children and adolescents aged 9-17 years: results from 2779165 Eurofit performances representing 30 countries. *British Journal of Sports Medicine* **52,** 1445–1456 (2018).

Tomkinson, G. R., Lang, J. J. & Tremblay, M. S. Temporal trends in the cardiorespiratory fitness of children and adolescents representing 19 high- income and upper middle-income countries between 1981 and 2014. *British Journal of Sport Medicine* **53,** 478–486 (2019).

Tomkinson, G. R., Lang, J. J., Tremblay, M. S., Dale, M., Leblanc, A. G., Belanger, K., Ortega, F. B. & Léger, L. International normative 20 m shuttle run values from 1142026 children and youth representing 50 countries. *British Journal of Sports Medicine* **51,** 1545–1554 (2017).

Tomkinson, G. R. & Olds, T. S. *Paediatric Exercise Science and Medicine* (eds and Armstrong, N. & Van Mechelen, W.) 2nd, 109–128 (Oxford University Press, 2008).

Tomlinson, L., Leiton, C. V. & Colognato, H. Behavioral experiences as drivers of oligodendrocyte lineage dynamics and myelin plasticity. *Neuropharmacology* **110,** 548–562 (2016).

Tomporowski, P. D., Lambourne, K. & Okumura, M. S. Physical activity interventions and children's mental function: An introduction and overview. *Preventive Medicine* **52,** S3–S9 (2011).

Tomporowski, P. D., McCullick, B., Pendleton, D. M. & Pesce, C. Exercise and children's cognition: The role of exercise characteristics and a place for metacognition. *Journal of Sport and Health Science* **4,** 47–55 (2015).

Tooley, U. A., Mackey, A. P., Ciric, R., Ruparel, K., Moore, T. M., Gur, R. C., Gur, R. E., Satterthwaite, T. D. & Bassett, D. S. Associations between Neighborhood SES and Functional Brain Network Development. *Cerebral Cortex* **30,** 1–19 (2020).

Townsend, J. T. & Ashby, F. G. *Methods of Modeling Capacity in Simple Processing Systems* (eds Castellan, J. N. J. & Restle, F.) 199–239 (Lawrence Erlbaum Associates, New York, 1978).

Treadway, M. T., Waskom, M. L., Dillon, D. G., Holmes, A. J., Park, M. T. M., Chakravarty, M. M., Dutra, S. J., Polli, F. E., Iosifescu, D. V., Fava, M., Gabrieli, J. D. & Pizzagalli, D. A. Illness progression, recent stress, and morphometry of hippocampal subfields and medial prefrontal cortex in major depression. *Biological Psychiatry* **77,** 285–294 (2015).

Trefler, A., Sadeghi, N., Thomas, A. G., Pierpaoli, C., Baker, C. I. & Thomas, C. Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging. *NeuroImage* **133,** 41–52 (2016).

Tremblay, M. S., Carson, V., Chaput, J. P., Connor Gorber, S., Dinh, T., Duggan, M., Faulkner, G., Gray, C. E., Grube, R., Janson, K., Janssen, I., Katzmarzyk, P. T., Kho, M. E., Latimer-Cheung, A. E., LeBlanc, C., Okely, A. D., Olds, T., Pate, R. R., Phillips, A., Poitras, V. J., Rodenburg, S., Sampson, M., Saunders, T. J., Stone, J. A., Stratton, G., Weiss, S. K. & Zehr, L. Canadian 24-hour movement guidelines for children and youth: An integration of physical activity, sedentary behaviour, and sleep. *Applied Physiology, Nutrition and Metabolism* **41,** S311–S327 (2016).

Tremblay, M. S., Carson, V. & Chaput, J. P. Introduction to the Canadian 24-Hour Movement Guidelines for Children and Youth: An Integration of Physical Activity, Sedentary Behaviour, and Sleep. *Applied Physiology, Nutrition and Metabolism* **41,** iii–iv (2016).

Troiano, R. P., McClain, J. J., Brychta, R. J. & Chen, K. Y. Evolution of accelerometer methods for physical activity research. *British Journal of Sports Medicine* **48,** 1019–1023 (2014).

Trost, S. G., Loprinzi, P. D., Moore, R. & Pfeiffer, K. A. Comparison of accelerometer cut points for predicting activity intensity in youth. *Medicine and Science in Sports and Exercise* **43,** 1360–1368 (2011).

Tucker-Drob, E. M. & Briley, D. A. Continuity of Genetic and Environmental Influences on Cognition across the Life Span: A Meta-Analysis of Longitudinal Twin and Adoption Studies. *Psychological Bulletin* **140,** 949–979 (2014).

Turner, E. L., Li, F., Gallis, J. A., Prague, M. & Murray, D. M. Review of recent methodological developments in group-randomized trials: Part 1 - Design. *American Journal of Public Health* **107,** 907–915 (2017).

Turner, E. L., Li, F., Gallis, J. A., Prague, M. & Murray, D. M. Review of recent methodological developments in group-randomized trials: Part 2 - Analysis. *American Journal of Public Health* **107,** 907–915 (2017).

Tuvey, S., Steele, J., Horton, E., Mayo, X., Liguori, G., Mann, S., Willinger, N. & Jimenez, A. Do changes in cardiorespiratory fitness resulting from physical activity interventions impact academic performance and executive function in children and adolescents? A systematic review, meta-analysis, and meta-regression. *OSF,* 10.31236/osf.io/4j2sa (2019).

Twisk, J. W. R., Kemper, H. & Van Mechelen, W. Tracking of activity and fitness and the relationship with cardiovascular disease risk factors. *Medicine & Science in Sports & Exercise* **32,** 1455–1461 (2000).

Twisk, J., Bosman, L., Hoekstra, T., Rijnhart, J., Welten, M. & Heymans, M. Different ways to estimate treatment effects in randomised controlled trials. *Contemporary Clinical Trials Communications* **10,** 80–85 (2018).

Uijtdewilligen, L., Nauta, J., Singh, A. S., Van Mechelen, W., Twisk, J. W., Van Der Horst, K. & Chinapaw, M. J. Determinants of physical activity and sedentary behaviour in young people: A review and quality synthesis of prospective studies. *British Journal of Sports Medicine* **45,** 896–905 (2011).

Ukoumunne, O. C., Gulliford, M. C., Chinn, S., Sterne, J. A., Burney, P. G. & Donner, A. Methods in health service research. Evaluation of health interventions at area and organisation level. *British Medical Journal* **319,** 376–379 (1999).

Utter, A. C., Robertson, R. J., Nieman, D. C. & Kang, J. I. E. Children's OMNI scale of perceived exertion : walking / running evaluation. *Medicine & Science in Sports & Exercise* **34,** 139–144 (2002).

Valkenborghs, S. R., Noetel, M., Hillman, C. H., Nilsson, M., Smith, J. J., Ortega, F. B. & Lubans, D. R. The Impact of Physical Activity on Brain Structure and Function in Youth: A Systematic Review. *Pediatrics* **144,** e20184032 (2019).

Van Breukelen, G. J. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology* **59,** 920–925 (2006).

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G. & Rubin, D. B. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76,** 1049–1064 (2006).

Van Buuren, S., Boshuizen, H. C. & Knook, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18,** 681–694 (1999).

Van Buuren, S. *Flexible imputation of missing data* (Chapman and Hall/CRC, 2018).

Van Buuren, S. & Groothuis-oudshoorn, K. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45,** 1–67 (2011).

Van Der Borght, K., Kóbor-Nyakas, D. É., Klauke, K., Eggen, B. J., Nyakas, C., Van Der Zee, E. A. & Meerlo, P. Physical exercise leads to rapid adaptations in hippocampal vasculature: Temporal dynamics and relationship to cell proliferation and neurogenesis. *Hippocampus* **19,** 928–936 (2009).

Van Der Kouwe, A. J., Benner, T. & Dale, A. M. Real-time rigid body motion correction and shimming using cloverleaf navigators. *Magnetic Resonance in Medicine* **56,** 1019–1032 (2006).

Van Der Kouwe, A. J., Benner, T., Salat, D. H. & Fischl, B. Brain morphometry with multiecho MPRAGE. *NeuroImage* **40,** 559–569 (2008).

Van Der Meer, D., Rokicki, J., Kaufmann, T., Córdova-Palomera, A., Moberget, T., Alnæs, D., Bettella, F., Frei, O., Doan, N. T., Sønderby, I. E., Smeland, O. B., Agartz, I., Bertolino, A., Bralten, J., Brandt, C. L., Buitelaar, J. K., Djurovic, S., Van Donkelaar, M., Dørum, E. S., Espeseth, T., Faraone, S. V., Fernández, G., Fisher, S. E., Franke, B., Haatveit, B., Hartman, C. A., Hoekstra, P. J., Håberg, A. K., Jönsson, E. G., Kolskår, K. K., Le Hellard, S., Lund, M. J., Lundervold, A. J., Lundervold, A., Melle, I., Monereo Sánchez, J., Norbom, L. C., Nordvik, J. E., Nyberg, L., Oosterlaan, J., Papalino, M., Papassotiropoulos, A., Pergola, G., De Quervain, D. J., Richard, G., Sanders, A. M., Selvaggi, P., Shumskaya, E., Steen, V. M., Tønnesen, S., Ulrichsen, K. M., Zwiers, M. P., Andreassen, O. A. & Westlye, L. T. Brain scans from 21,297 individuals reveal the genetic architecture of hippocampal subfield volumes. *Molecular Psychiatry* **25,** 3053–3065 (2020).

Van Der Niet, A. G., Smith, J., Scherder, E. J., Oosterlaan, J., Hartman, E. & Visscher, C. Associations between daily physical activity and executive functioning in primary school-aged children. *Journal of Science and Medicine in Sport* **18,** 673–677 (2015).

Van Ginkel, J. R., Linting, M., Rippe, R. C. & Van Der Voort, A. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *Journal of Personality Assessment* **102,** 297–308 (2020).

Van Hees, V. T., Van Lummel, R. C. & Westerterp, K. R. Estimating activity-related energy expenditure under sedentary conditions using a tri-axial seismic accelerometer. *Obesity* **17,** 1287–1292 (2009).

Van Hees, V. T., Sabia, S., Jones, S. E., Wood, A. R., Anderson, K. N., Kivimäki, M., Frayling, T. M., Pack, A. I., Bucan, M., Trenell, M. I., Mazzotti, D. R., Gehrman, P. R., Singh-Manoux, B. A. & Weedon, M. N. Estimating sleep parameters using an accelerometer without sleep diary. *Scientific Reports* **8,** 1–11 (2018).

Van Heuvelen, M. J., Kempen, G. I., Ormel, J. & De Greef, M. H. Self-reported physical fitness of older persons: A substitute for performance-based measures of physical fitness? *Journal of Aging and Physical Activity* **5,** 298–310 (1997).

Van Leeuwen, M., Van Den Berg, S. M., Hoekstra, R. A. & Boomsma, D. I. Endophenotypes for intelligence in children and adolescents. *Intelligence* **35,** 369–380 (2007).

Van Petten, C. Relationship between hippocampal volume and memory ability in healthy individuals across the lifespan: Review and meta-analysis. *Neuropsychologia* **42,** 1394–1413 (2004).

Van Praag, H. Exercise Enhances Learning and Hippocampal Neurogenesis in Aged Mice. *Journal of Neuroscience* **25,** 8680–8685 (2005).

Van Praag, H., Kempermann, G. & Gage, F. H. Running increases cell proliferation and neurogenesis in the adult mouse dentate gyrus. *Nature Neuroscience* **2,** 266–270 (1999).

Van Sluijs, E. M. & Kriemler, S. Reflections on physical activity intervention research in young people - dos, don'ts, and critical thoughts. *International Journal of Behavioral Nutrition and Physical Activity* **13,** 25 (2016).

Van Strien, N. M., Cappaert, N. L. & Witter, M. P. The anatomy of memory: An interactive overview of the parahippocampal- hippocampal network. *Nature Reviews Neuroscience* **10,** 272–282 (2009).

Van Waelvelde, H., Vanden Wyngaert, K., Mariën, T., Baeyens, D. & Calders, P. The relation between children's aerobic fitness and executive functions: A systematic review. *Infant and Child Development* **29,** e2163 (2020).

Vance, D. E., Wadley, V. G., Ball, K. K., Roenker, D. L. & Rizzo, M. The effects of physical activity and sedentary behavior on cognitive health in older adults. *Journal of Aging and Physical Activity* **13,** 294–313 (2005).

Vandierendonck, A. Further Tests of the Utility of Integrated Speed-Accuracy Measures in Task Switching. *Journal of Cognition* **1** (2018).

Vanhees, L., Lefevre, J., Philippaerts, R., Martens, M., Huygens, W., Troosters, T. & Beunen, G. How to assess physical activity? How to assess physical fitness? *European Journal of Preventive Cardiology* **12,** 102–114 (2005).

Vanhove, J. Analyzing randomized controlled interventions: Three notes for applied linguists. *Studies in Second Language Learning and Teaching* **5,** 135–152 (2015).

Varela, A. R., Pratt, M., Harris, J., Lecy, J., Salvo, D., Brownson, R. C. & Hallal, P. C. Mapping the historical development of physical activity and health research: A structured literature review and citation network analysis. *Preventive Medicine* **111,** 466–472 (2018).

Vaynman, S., Ying, Z. & Gomez-Pinilla, F. Hippocampal BDNF mediates the efficacy of exercise on synaptic plasticity and cognition. *European Journal of Neuroscience* **20,** 2580–2590 (2004).

Vazou, S., Pesce, C., Lakes, K. & Smiley-Oyen, A. More than one road leads to Rome: A narrative review and meta-analysis of physical activity intervention effects on cognition in youth. *International Journal of Sport and Exercise Psychology* **17,** 153–178 (2019).

Veldman, K., Bültmann, U., Stewart, R. E., Ormel, J., Verhulst, F. C. & Reijneveld, S. A. Mental health problems and educational attainment in adolescence: 9-year follow-up of the TRAILS study. *PLoS ONE* **9,** e101751 (2014).

Venckunas, T., Snieckus, A., Trinkunas, E., Baranauskiene, N., Solianik, R., Juodsnukis, A., Streckis, V. & Kamandulis, S. Interval Running Training Improves Cognitive Flexibility and Aerobic Power of Young Healthy Adults. *Journal of Strength and Conditioning Research* **30,** 2114–2121 (2016).

Verburgh, L., Königs, M., Scherder, E. J. A. & Oosterlaan, J. Physical exercise and executive functions in preadolescent children, adolescents and young adults: a meta-analysis. *British Journal of Sports Medicine* **48,** 973–979 (2014).

Verde, M. F. Measures of sensitivity based on a single hit rate and false alarm rate: The accuracy, precision, and robustness of d' , Az , and A'. *Perception & Psychophysics* **68,** 643–654 (2006).

Vickers, A. J. & Altman, D. G. Analysing controlled trials with baseline and follow up measurements. *British Medical Journal* **323,** 1123–1124 (2001).

Vijayakumar, N., Mills, K. L., Alexander-Bloch, A., Tamnes, C. K. & Whittle, S. Structural brain development: A review of methodological approaches and best practices. *Developmental Cognitive Neuroscience* **33,** 129–148 (2018).

Vijayakumar, N., Op De Macks, Z., Shirtcliff, E. A. & Pfeifer, J. H. Puberty and the human brain: Insights into adolescent development. *Neuroscience and Biobehavioral Reviews* **92,** 417–436 (2018).

Vijayakumar, N., Whittle, S., Yücel, M., Dennison, M., Simmons, J. & Allen, N. B. Prefrontal Structural Correlates of Cognitive Control during Adolescent Development: A 4-Year Longitudinal Study. *Journal of Cognitive Neuroscience* **26,** 1118–1130 (2014).

Visser, E., Keuken, M. C., Douaud, G., Gaura, V., Bachoud-Levi, A. C., Remy, P., Forstmann, B. U. & Jenkinson, M. Automatic segmentation of the striatum and globus pallidus using MIST: Multimodal Image Segmentation Tool. *NeuroImage* **125,** 479–497 (2016).

Vivar, C., Peterson, B. D. & Van Praag, H. Running rewires the neuronal network of adult-born dentate granule cells. *NeuroImage* **131,** 29–41 (2016).

Vivar, C. & Van Praag, H. Running Changes the Brain: the Long and the Short of It. *Physiology* **32,** 410–424 (2017).

Voelcker-Rehage, C. & Niemann, C. Structural and functional brain changes related to different types of physical activity across the life span. *Neuroscience and Biobehavioral Reviews* **37,** 2268–2295 (2013).

Voets, N. L., Zamboni, G., Stokes, M. G., Carpenter, K., Stacey, R. & Adcock, J. E. Aberrant Functional Connectivity in Dissociable Hippocampal Networks Is Associated with Deficits in Memory. *Journal of Neuroscience* **34,** 4920–4928 (2014).

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., Wahlund, L., Larsson, E., Westman, E. & Initiative, A. D. N. The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Frontiers in Aging Neuroscience* **6,** 264 (2014).

Voineskos, A. N., Winterburn, J. L., Felsky, D., Pipitone, J., Rajji, T. K., Mulsant, B. H. & Chakravarty, M. M. Hippocampal (subfield) volume and shape in relation to cognitive performance across the adult lifespan. *Human Brain Mapping* **36,** 3020–3037 (2015).

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C. & Vandenbroucke, J. P. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *The Lancet* **370,** 1453–1457 (2007).

Von Hippel, P. 8. How to impute interactions, squares, and other transformed variables. *Sociological methodology* **39,** 265–291 (2009).

Von Hippel, P. T. Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociological Methods and Research* **42,** 105–138 (2013).

Von Hippel, P. T. How Many Imputations Do You Need? A Two-stage Calculation Using a Quadratic Rule. *Sociological Methods & Research* **49,** 699–718 (2020).

Voss, M. W., Chaddock, L., Kim, J. S., Vanpatter, M., Pontifex, M. B., Raine, L. B., Cohen, N. J., Hillman, C. H. & Kramer, A. F. Aerobic fitness is associated with greater efficiency of the network underlying cognitive control in preadolescent children. *Neuroscience* **199,** 166–176 (2011).

Voss, M. W., Nagamatsu, L. S., Liu-Ambrose, T. & Kramer, A. F. Exercise, brain, and cognition across the life span. *Journal of Applied Physiology* **111,** 1505–1513 (2011).

Voss, M. W., Soto, C., Yoo, S., Sodoma, M., Vivar, C. & Van Praag, H. Exercise and Hippocampal Memory Systems. *Trends in Cognitive Sciences* **23,** 318–333 (2019).

Voss, M. W., Vivar, C., Kramer, A. F. & Van Praag, H. Bridging animal and human models of exercise-induced brain plasticity. *Trends in Cognitive Sciences* **17,** 525–544 (2013).

Voss, M. W., Weng, T. B., Burzynska, A. Z., Wong, C. N., Cooke, G. E., Clark, R., Fanning, J., Awick, E., Gothe, N. P., Olson, E. A., McAuley, E. & Kramer, A. F. Fitness, but not physical activity, is related to functional integrity of brain networks associated with aging. *NeuroImage* **131,** 113–125 (2016).

Vovk, U., Pernuš, F. & Likar, B. A review of methods for correction of intensity inhomogeneity in MRI. *IEEE Transactions on Medical Imaging* **26,** 405–421 (2007).

Wade, L., Leahy, A., Lubans, D. R., Smith, J. J. & Duncan, M. J. A systematic review of cognitive assessment in physical activity research involving children and adolescents. *Journal of Science and Medicine in Sport* **23,** 740–745 (2020).

Wagner, G., Herbsleb, M., De La Cruz, F., Schumann, A., Brünner, F., Schachtzabel, C., Gussew, A., Puta, C., Smesny, S., Gabriel, H. W., Reichenbach, J. R. & Bär, K. J. Hippocampal structure, metabolism, and inflammatory response after a 6-week intense aerobic exercise in healthy young adults: A controlled trial. *Journal of Cerebral Blood Flow and Metabolism* **35,** 1570–1578 (2015).

Walhovd, K. B., Johansen-Berg, H. & Káradóttir, R. T. Unraveling the secrets of white matter - Bridging the gap between cellular, animal and human imaging studies. *Neuroscience* **276,** 2–13 (2014).

Walhovd, K. B., Fjell, A. M., Giedd, J., Dale, A. M. & Brown, T. T. Through Thick and Thin: a Need to Reconcile Contradictory Results on Trajectories in Human Cortical Development. *Cerebral Cortex* **27,** 1472–1481 (2017).

Walhovd, K. B., Krogsrud, S. K., Amlien, I. K., Bartsch, H., Bjørnerud, A., Due-Tønnessen, P., Grydeland, H., Hagler, D. J., Håberg, A. K., Kremen, W. S., Ferschmann, L., Nyberg, L., Panizzon, M. S., Rohani, D. A., Skranes, J., Storsve, A. B., Sølsnes, A. E., Tamnes, C. K., Thompson, W. K., Reuter, C., Dale, A. M. & Fjell, A. M. Neurodevelopmental origins of lifespan changes in brain and cognition. *Proceedings of the National Academy of Sciences* **113,** 9357–9362 (2016).

Walhovd, K. B., Tamnes, C. K. & Fjell, A. M. Brain structural maturation and the foundations of cognitive behavioral development. *Current Opinion in Neurology* **27,** 176–184 (2014).

Walsh, E. I., Smith, L., Northey, J., Rattray, B. & Cherbuin, N. Towards an understanding of the physical activity-BDNF-cognition triumvirate: A review of associations and dosage. *Ageing Research Reviews* **60,** 101044 (2020).

Walsh, J. J., Barnes, J. D., Cameron, J. D., Goldfield, G. S., Chaput, J. P., Gunnell, K. E., Ledoux, A. A., Zemek, R. L. & Tremblay, M. S. Associations between 24-hour movement behaviours and global cognition in a cross-sectional sample of American children. *The Lancet Child & Adolescent Health* **2,** 783–791 (2018).

Wang, H.-T., Smallwood, J., Mourao-Miranda, J., Xia, C. H., Satterthwaite, T. D., Bassett, D. S. & Bzdok, D. Finding the needle in high-dimensional haystack: A tutorial on canonical correlation analysis. *arXiv.* arXiv: 1812.02598 (2018).

Wang, N., Zhang, J., Cofer, G., Qi, Y., Anderson, R. J., White, L. E. & Allan Johnson, G. Neurite orientation dispersion and density imaging of mouse brain microstructure. *Brain Structure and Function* **224,** 1797–1813 (2019).

Warburton, D. E., Nicol, C. W. & Bredin, S. S. Health benefits of physical activity: The evidence. *CMAJ* **174,** 801–809 (2006).

Warren, J. M., Ekelund, U., Besson, H., Mezzani, A., Geladas, N. & Vanhees, L. Assessment of physical activity - A review of methodologies with reference to epidemiological research: A report of the exercise physiology section of the European Association of Cardiovascular Prevention and Rehabilitation. *European Journal of Cardiovascular Prevention and Rehabilitation* **17,** 127–139 (2010).

Wassenaar, T. M., Wheatley, C. M., Beale, N., Salvan, P., Meaney, A., Possee, J. B., Atherton, K. E., Duda, J. L., Dawes, H. & Johansen-Berg, H. Effects of a programme of vigorous physical activity during secondary school physical education on academic performance, fitness, cognition, mental health and the brain of adolescents (Fit to Study): Study protocol for a cluster-randomised trial. *Trials* **20,** 189 (2019).

Wassenaar, T. M., Yaffe, K., Van Der Werf, Y. D. & Sexton, C. E. Associations between modifiable risk factors and white matter of the aging brain: insights from diffusion tensor imaging studies. *Neurobiology of Aging* **80,** 56–70 (2019).

Wassenaar, T. M., Williamson, W., Johansen-Berg, H., Dawes, H., Roberts, N., Foster, C. & Sexton, C. E. A critical evaluation of systematic reviews assessing the effect of chronic physical activity on academic achievement, cognition and the brain in children and adolescents: a systematic review. *International Journal of Behavioral Nutrition & Physical Activity* **17,** 1–18 (2020).

Watson, A., Timperio, A., Brown, H., Best, K. & Hesketh, K. D. Effect of classroom-based physical activity interventions on academic and physical activity outcomes: A systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity* **14,** 114 (2017).

Wei, J., Hou, R., Xie, L., Chandrasekar, E. K., Lu, H., Wan, T., Li, C. & Xu, H. Sleep, Sedentary activity, physical activity, and cognitive function among older adults: The National Health and Nutrition Examination Survey, 2011-2014. *Journal of Science and Medicine in Sport* (2020).

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., Havlik, R., Conway, K., Edwards, E., Freund, L., King, J. W., Moy, C., Witt, E. & Gershon, R. C. Cognition assessment using the NIH Toolbox Sandra. *Neurology* **80,** 54–64 (2013).

Weiskopf, N., Mohammadi, S., Lutti, A. & Callaghan, M. F. Advances in MRI-based computational neuroanatomy: From morphometry to in-vivo histology. *Current Opinion in Neurology* **28,** 313–322 (2015).

Weiskopf, N., Suckling, J., Williams, G., Correia M., M. M., Inkster, B., Tait, R., Ooi, C., Bullmore T., E. T. & Lutti, A. Quantitative multi-parameter mapping of R1, PD*, MT, and R2* at 3T: A multi-center validation. *Frontiers in Neuroscience* **7,** 95 (2013).

Weiss, A. P., Dewitt, I., Goff, D., Ditman, T. & Heckers, S. Anterior and posterior hippocampal volumes in schizophrenia. *Schizophrenia Research* **73,** 103–112 (2005).

Welk, G. J., Corbin, C. B. & Dale, D. Measurement issues in the assessment of physical activity in children. *Research Quarterly for Exercise and Sport* **71,** 59–73 (2000).

Welsman, J. & Armstrong, N. Interpreting aerobic fitness in youth: The fallacy of ratio scaling. *Pediatric Exercise Science* **31,** 184–190 (2019).

Welsman, J. & Armstrong, N. Interpreting Cardiorespiratory Fitness in Young Clinical Populations—Folklore and Fallacy. *JAMA Pediatrics* **173,** 713–714 (2019).

Wenger, E., Mårtensson, J., Noack, H., Bodammer, N. C., Kühn, S., Schaefer, S., Heinze, H. J., Düzel, E., Bäckman, L., Lindenberger, U. & Lövdén, M. Comparing manual and automatic segmentation of hippocampal volumes: Reliability and validity issues in younger and older brains. *Human Brain Mapping* **35,** 4236–4248 (2014).

Westfall, D. R., Gejl, A. K., Tarp, J., Wedderkopp, N., Kramer, A. F., Hillman, C. H. & Bugge, A. Associations between aerobic fitness and cognitive control in adolescents. *Frontiers in Psychology* **9,** 1298 (2018).

Weston, M., Taylor, K. L., Batterham, A. M. & Hopkins, W. G. Effects of low-volume high-intensity interval training (HIT) on fitness in adults: A meta-analysis of controlled and non-controlled trials. *Sports Medicine* **44,** 1005–1017 (2014).

Whalley, L. J., Dick, F. D. & McNeill, G. A life-course approach to the aetiology of late-onset dementias. *The Lancet Neurology* **5,** 87–96 (2006).

Wheatley, C. M., Davies, E. L. & Dawes, H. Unspoken Playground Rules Discourage Adolescent Physical Activity in School: A Focus Group Study of Constructs in the Prototype Willingness Model. *Qualitative health research* **28,** 624–632 (2017).

Wheatley, C. M., Wassenaar, T. & Johansen-Berg, H. *Educational Neuroscience. Development Across the Life Span* (eds Thomas, M. S. C., Mareschal, D. & Dumontheil, I.) 359–385 (Taylor & Francis, New York, 2020).

Wheatley, C., Beale, N., Wassenaar, T., Graham, M., Eldridge, E., Dawes, H. & Johansen-Berg, H. Fit to Study: Reflections on designing and implementing a large-scale randomized controlled trial in secondary schools. *Trends in Neuroscience and Education* **20,** 100134 (2020).

Wheatley, C., Wassenaar, T., Salvan, P., Beale, N., Nichols, T., Dawes, H. & Johansen-berg, H. Associations between fitness, physical activity and mental health in a community sample of young British adolescents: baseline data from the Fit to Study trial. *BMJ Open Sport & Exercise Medicine* **6,** e000819 (2020).

White, I. R. Uses and limitations of randomization-based efficacy estimators. *Statistical Methods in Medical Research* **14,** 327–347 (2005).

White, I. R. & Carlin, J. B. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine* **29,** 2920–2931 (2010).

White, I. R., Royston, P. & Wood, A. M. Tutorial in Biostatistics Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30** (2011).

Whittle, S., Vijayakumar, N., Simmons, J. G., Dennison, M., Schwartz, O., Pantelis, C., Sheeber, L., Byrne, M. L. & Allen, N. B. Role of positive parenting in the association between neighborhood social disadvantage and brain development across adolescence. *JAMA Psychiatry* **74,** 824–832 (2017).

Wickel, E. E. Sedentary time, physical activity, and executive function in a longitudinal study of youth. *Journal of Physical Activity and Health* **14,** 222–228 (2017).

Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* 2016. https://cran.r-project.org/package=ggplot2.

Wiebe, S. A., Espy, K. A. & Charak, D. Using confirmatory factor analysis to understand executive control in preschool children: I. Latent Structure. *Developmental psychology* **44,** 575 (2008).

Wierenga, L., Langen, M., Ambrosino, S., Van Dijk, S., Oranje, B. & Durston, S. Typical development of basal ganglia, hippocampus, amygdala and cerebellum from age 7 to 24. *NeuroImage* **96,** 67–72 (2014).

Wilckens, K. A., Erickson, K. I. & Wheeler, M. E. Physical Activity and Cognition: A Mediating Role of Efficient Sleep. *Behavioral Sleep Medicine* **16,** 569–586 (2018).

Williams, D. S., Detre, J. A., Leigh, J. S. & Koretsky, A. P. Magnetic resonance imaging of perfusion using spin inversion of arterial water. *Proceedings of the National Academy of Sciences* **89,** 4220–4220 (1992).

Williams, T. C., Bach, C. C., Matthiesen, N. B., Henriksen, T. B. & Gagliardi, L. Directed acyclic graphs: a tool for causal studies in paediatrics. *Pediatric Research* **84,** 487–493 (2018).

Winkler, A. M., Renaud, O., Smith, S. M. & Nichols, T. E. Permutation inference for Canonical Correlation Analysis. *arXiv.* arXiv: 2002.10046 (2020).

Winkler, A. M., Webster, M. A., Brooks, J. C., Tracey, I., Smith, S. M. & Nichols, T. E. Non-parametric combination and related permutation tests for neuroimaging. *Human Brain Mapping* **37,** 1486–1511 (2016).

Winkler, A. M., Webster, M. A., Vidaurre, D., Nichols, T. E. & Smith, S. M. Multi-level block permutation. *NeuroImage* **123,** 253–268 (2015).

Winston, G. P. The physical and biological basis of quantitative parameters derived from diffusion MRI. *Quantitative imaging in medicine and surgery* **2,** 254–265 (2012).

Winter, E., Jones, A., Davidson, R., Bromley, P. & Mercer, T. *Sport and Exercise Physiology Testing Guidelines: Volume II - Exercise and Clinical Testing. The British Association of Sport and Exercise Sciences Guide.* (eds Winter, E., Jones, A., Davidson, R., Bromley, P. & Mercer, T.) (Routledge, London, 2007).

Winterburn, J. L., Pruessner, J. C., Chavez, S., Schira, M. M., Lobaugh, N. J., Voineskos, A. N. & Chakravarty, M. M. A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. *NeuroImage* **74,** 254–265 (2013).

Wisse, L. E., Chételat, G., Daugherty, A. M., De Flores, R., La Joie, R., Mueller, S. G., Stark, C. E., Wang, L., Yushkevich, P. A., Berron, D., Raz, N., Bakker, A., Olsen, R. K. & Carr, V. A. Hippocampal subfield volumetry from structural isotropic 1 mm3 MRI scans: A note of caution. *Human Brain Mapping* (2020).

Woltz, D. J. & Was, C. A. Availability of related long-term memory during and after attention focus in working memory. *Memory and Cognition* **34,** 668–684 (2006).

World Health Organization. *Global action plan on physical activity 2018–2030: more active people for a healthier world.* 2018. https://www.who.int/ncds/prevention/physical-activity/gappa/action-plan (2019).

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H. J. & Ettinger, U. Reliability and plasticity of response inhibition and interference control. *Brain and Cognition* **81,** 82–94 (2013).

Wylie, G. & Allport, A. Task switching and the measurement of "switch costs". *Psychological Research* **63,** 212–233 (2000).

Xue, Y., Yang, Y. & Huang, T. Effects of chronic exercise interventions on executive function among children and adolescents: A systematic review with meta-analysis. *British Journal of Sports Medicine* **53,** 1297–1404 (2019).

Yang, C. C. & Hsu, Y. L. A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors* **10,** 7772–7788 (2010).

Yeaton, W. H. & Sechrest, L. Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology* **49,** 156–167 (1981).

Yildirim, M., Van Stralen, M. M., Chinapaw, M. J., Brug, J., Van Mechelen, W., Twisk, J. W. & Te Velde, S. J. For whom and under what circumstances do school-based energy balance behavior interventions work? Systematic review on moderators. *International Journal of Pediatric Obesity* **6,** 46–57 (2011).

Young, J., Angevaren, M., Rusted, J. & Tabet, N. Aerobic exercise to improve cognitive function in older people without known cognitive impairment (Review). *Cochrane Database of Systematic Reviews,* CD005381 (2015).

Youngjohn, J. R., Larrabee, G. J. & Crook, T. H. Test-retest reliability of computerized, everyday memory measures and traditional memory tests. *Clinical Neuropsychologist* **6,** 276–286 (1992).

Yu, H., Li, F., Gallis, J. A. & Turner, E. L. *cvcrand: Efficient Design and Analysis of Cluster Randomized Trials* 2020. https://cran.r-project.org/package=cvcrand.

Yu, Q., Daugherty, A. M., Anderson, D. M., Nishimura, M., Brush, D., Hardwick, A., Lacey, W., Raz, S. & Ofen, N. Socioeconomic status and hippocampal volume in children and young adults. *Developmental Science* **21,** e12561 (2018).

Yucel, R. M. Multiple imputation inference for multivariate multilevel continuous data with ignorable non-response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **366,** 2389–2403 (2008).

Yushkevich, P. A., Amaral, R. S., Augustinack, J. C., Bender, A. R., Bernstein, J. D., Boccardi, M., Bocchetta, M., Burggren, A. C., Carr, V. A., Chakravarty, M. M., Chételat, G., Daugherty, A. M., Davachi, L., Ding, S. L., Ekstrom, A., Geerlings, M. I., Hassan, A., Huang, Y., Iglesias, J. E., La Joie, R., Kerchner, G. A., LaRocque, K. F., Libby, L. A., Malykhin, N., Mueller, S. G., Olsen, R. K., Palombo, D. J., Parekh, M. B., Pluta, J. B., Preston, A. R., Pruessner, J. C., Ranganath, C., Raz, N., Schlichting, M. L., Schoemaker, D., Singh, S., Stark, C. E., Suthana, N., Tompary, A., Turowski, M. M., Van Leemput, K., Wagner, A. D., Wang, L., Winterburn, J. L., Wisse, L. E., Yassa, M. A. & Zeineh, M. M. Quantitative comparison of 21 protocols for labeling hippocampal subfields and parahippocampal subregions in in vivo MRI: Towards a harmonized segmentation protocol. *NeuroImage* **111,** 526–541 (2015).

Zaitsev, M., Akin, B., LeVan, P. & Knowles, B. R. Prospective motion correction in functional MRI. *NeuroImage* **154,** 33–42 (2017).

Zatorre, R. J., Fields, R. D. & Johansen-berg, H. Plasticity in gray and white : neuroimaging changes in brain structure during learning. *Nature neuroscience* **15,** 528–536 (2012).

Zeger, S. L., Liang, K.-Y. & Albert, P. S. Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics* **44,** 1049–1060 (1988).

Zeidman, P. & Maguire, E. A. Anterior hippocampus: The anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience* **17,** 173–182 (2016).

Zelazo, P. D. & Carlson, S. M. Hot and Cool Executive Function in Childhood and Adolescence: Development and Plasticity. *Child Development Perspectives* **6,** 354–360 (2012).

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A. & Alexander, D. C. NODDI: Practical in vivo neurite orientation dispersion and density imaging of the human brain. *NeuroImage* **61,** 1000–1016 (2012).

Zhang, Y., Brady, M. & Smith, S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging* **20,** 45–57 (Jan. 2001).

Zhao, J. H. & Schafer, J. L. *pan: Multiple imputation for multivariate panel or clustered data* 2018. https://cran.r-project.org/package=pan.

Zucker, D. M. An Analysis of Variance Pitfall: The Fixed Effects Analysis in a Nested Design. *Educational and Psychological Measurement* **50,** 731–738 (1990).