The role of spatial cognition in children's science learning

Alex Hodgkiss
UCL

Thesis submitted for the degree of Doctor of Philosophy

2019

I hereby declare that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature ……………………………………………………..

Word count (exclusive of abstract, impact statement, publication list, acknowledgements, appendices, table of contents, reference list and lists of figures and tables):

81,655

Abstract

Spatial cognition incorporates the human capacity to use both spatial thinking skills (e.g., mental rotation of objects) and symbolic spatial tools (e.g., gesture, diagrams). Prior research with adults indicates that spatial thinking skills predict science learning outcomes. However, research in the primary school years is sparse and includes a restricted range of spatial thinking and science measures. The main aim of this thesis was to comprehensively investigate the relationship between spatial thinking skills and science learning in middle childhood (7-11 years). Children completed a selection of spatial thinking tasks and science assessments, across varying contexts. Spatial thinking skills predicted performance on a curriculum-based science assessment (Chapter 3). Mental folding and spatial scaling were the strongest predictors, with mental folding a stronger predictor than spatial scaling. Spatial thinking skills also predicted the learning outcomes of whole-class instruction on sound (Chapter 4). Mental folding again emerged as the strongest spatial predictor. The relationship was evident for tasks involving the application of conceptual understanding, rather than factual recall. An additional aim was to investigate gesture as a spatial tool in supporting children's science learning. No strong evidence emerged that gesture was a more effective spatial learning tool than teaching via the use of concrete models, or by using verbal descriptions and diagrams (Chapter 5). Consistent with Chapter 3 and 4, mental folding also predicted learning. However, for problems closely related to the lesson content, this relationship was stronger in the conditions where children had no experience with concrete models. This may reflect the use of mental models to scaffold learning. Despite evidence for certain spatial skills being stronger predictors of science learning, psychometric analyses best supported a one-factor model of spatial thinking. Nevertheless, the findings of the thesis suggest that targeting mental folding skills within spatial training interventions may yield the greatest benefits.

Impact Statement

The findings and conclusions presented in this thesis are beneficial through contributions to the wider knowledge base, and also through potential wider societal and economic impact. Considering first academic impact, the contribution of spatial cognition to science learning in the primary school years is an under-researched area. This is also the case for cognitive predictors of science learning within this age range, more generally. The research presented in this thesis therefore has an academic impact by adding to the existing limited evidence base. In addition, the research findings also present several avenues for future related research. The thesis therefore provides a base from which future research can expand upon. Beyond academia, the findings have potential benefits to education and education policy. There are currently concerns regarding the recruitment of STEM professionals in the UK (Baker, 2018). The results of the thesis indicate that individual differences in spatial thinking skills are a predictor of science learning in the primary-school years. Therefore, in the short-term, supporting spatial thinking in middle childhood has the potential to support children's confidence and success in science. Furthermore, children who are more confident and successful in science at a younger age may be more likely to opt for further study in science at the post-16 level. Moreover, supporting spatial thinking skills in childhood may, in the long-term, reduce the rate at which students drop out at the initial stages of science undergraduate degrees, due to not having the supportive spatial skills. In both cases, this has the potential to increase current STEM graduate numbers, and a potential economic impact may be seen through the reduction in the current STEM shortfall. The thesis provides more specific insights into the possible methods through which spatial skills might be supported. At the school level, the impact could be focused on individual teaching practice. For example, informing teachers about the role of spatial thinking in science learning could result in adjustments being made to lesson planning and delivery. Furthermore, at a national level policy level, spatial thinking is currently not a component of the National Curriculum. The findings of the thesis suggest that incorporating spatial thinking activities into the curriculum may be beneficial. However, overall, further research is needed to bridge these findings to the larger-scale policy level; for example, through small and large scale training and intervention studies.

Publications

The spatial scaling task used in Chapters 2, 3 and 4, and the developmental data for this task in Chapter 2, is published as:

Gilligan, K. A*., Hodgkiss, A*., Thomas, M. S., & Farran, E. K. (2018). The use of discrimination scaling tasks: A novel perspective on the development of spatial scaling in children. *Cognitive Development*, *47*, 133-145. https://doi.org/10.1016/j.cogdev.2018.04.001

* denotes joint first authorship

The data in Chapter 3 are published as:

Hodgkiss, A., Gilligan, K. A., Tolmie, A. K., Thomas, M. S., & Farran, E. K. (2018). Spatial cognition and science achievement: The contribution of intrinsic and extrinsic spatial skills from 7 to 11 years. *British Journal of Educational Psychology*, *88(4),* 675-697. https://doi.org/10.1111/bjep.12211

The spatial thinking tasks reported in Chapter 2 and 3 are also reported in relation to mathematics measures in:

Gilligan, K. A., Hodgkiss, A., Thomas, M. S., & Farran, E. K. (2019). The developmental relations between spatial cognition and mathematics in primary school children. *Developmental Science, 22(4),* 1-19. https://doi.org/10.1111/desc.12786

Acknowledgements

I would first like to thank the Economic and Social Research Council for funding the studentship which made this research possible. I am also grateful to the children and schools who participated in the studies within this thesis. Particular thanks go to the teachers for their flexibility, and also for making me feel so welcome.

I would also like to express my sincere thanks my three supervisors: Professor Emily Farran, Professor Andrew Tolmie and Professor Michael Thomas for their invaluable experience, guidance and advice. I would particularly like to thank Andrew for sharing his insights on science learning. I would also particularly like to extend my deepest gratitude to Emily, for her unrelenting support and encouragement throughout the past three years.

Many thanks also to members of the Centre for Educational Neuroscience, particularly the PhD group, and also to other members of the CoGDevLab, for their support and feedback on this research. I am extremely grateful to Katie Gilligan, for her valuable advice and suggestions, and also for being such a supportive friend.

I would also like to extend my deepest thanks to my family and friends for their support and understanding throughout this journey. I am deeply indebted to my parents for always encouraging me to pursue my interests and passions. Finally, I cannot begin to express my thanks to Paul for his unwavering support and patience, particularly in the final months of writing.

Table of Contents

Chapter 1: Literature review and introduction to thesis... 17

1.1 Introduction 17

1.2 Defining and categorising spatial ability and spatial cognition 21

1.2.1 Introduction 21

1.2.2 Uttal et al. (2013): 2 x 2 typology (see also Newcombe & Shipley, 2015)................... 21

1.2.2.1 Evidence supporting the model... 28

1.2.2.2 Object-based transformations (intrinsic-dynamic) vs perspective taking (extrinsic-dynamic) within the model. 29

1.2.3 Previous typologies 33

1.2.3.1 Carroll (1993): psychometric model 33

1.2.3.2 Linn and Petersen (1985) 35

1.2.3.3 Summary of previous typologies in relation to Uttal et al (2013) model 37

1.2.4 Other spatial factors.. 38

1.2.4.1 Visuospatial memory 38

1.2.4.2 Multiple object dynamic spatial ability.. 39

1.2.5 Interim summary 41

1.2.6 The structure of spatial cognition in childhood 42

1.2.7 The development of spatial skills 44

1.2.8 Interim summary 45

1.2.9 Spatial thinking tasks used in the thesis 46

1.2.10 Spatial tools 48

1.2.11 Embodied cognition.. 49

1.3 Overview of science.. 52

1.3.1 Approaches to assessing science knowledge, skills and understanding within the thesis. 53

1.4 Mental models and the nature of conceptual understanding in science 55

1.4.1 Mental models: Johnson-Laird (1980) 56

1.4.2 Mental models: Gentner & Stevens (1983) 57

1.4.3 Framework theory: Vosniadou (1994) 59

1.4.4 Knowledge in pieces: diSessa (1988).. 61

1.4.5 Mental simulation and mental animation within mental models.................... 63

1.4.6 Mental models summary 64

1.5 Approaches to science instruction in the thesis.. 65

1.6 Existing research on the relationship between spatial ability and science, technology, engineering and mathematics (STEM) 66

1.6.1 Spatial ability and STEM: longitudinal evidence.. 67

1.6.2 Spatial ability and science: evidence from adults and adolescents ... 68
1.6.2.1 Chemistry ... 68
1.6.2.2 Biology/Medicine ... 70
1.6.2.3 Physics/Engineering/Geoscience ... 72
1.6.3 Spatial ability and science: evidence from younger children ... 75
1.6.4 Summary of possible mechanisms which may explain the relationship between spatial thinking skills and science ... 77
1.6.5 Spatial tools and science learning in children ... 78
1.7 Overall summary and rationale for thesis ... 78
Chapter 2: The latent structure and development of spatial skills in middle childhood ... 82
2.1 Introduction ... 82
2.1.1 The 2x2 typology of spatial thinking ... 82
2.1.2 The latent structure of spatial cognition in childhood ... 83
2.1.3 The development of spatial skills in middle childhood ... 87
2.1.4 Current study ... 90
2.2 Methods ... 91
2.2.1 Participants ... 91
2.2.2 Materials ... 92
2.2.2.1 2D mental rotation (intrinsic-dynamic sub-domain) ... 92
2.2.2.2 Mental folding task (intrinsic-dynamic sub-domain) ... 93
2.2.2.3 Children’s Embedded Figures Task -CEFT (intrinsic-static sub-domain). ... 94
2.2.2.4 Perspective Taking Task for Children- (extrinsic-dynamic sub-domain) ... 95
2.2.2.5 Spatial scaling (extrinsic-static sub-domain) ... 96
2.2.3 Procedure ... 97
2.3 Results ... 97
2.3.1 Development of spatial skills ... 97
2.3.1.1 Analysis strategy ... 97
2.3.1.2 Descriptive statistics ... 99
2.3.1.3 Main analysis of development of spatial skills ... 102
2.3.2 Latent structure of spatial skills ... 104
2.3.2.1 Analysis strategy ... 104
2.3.2.2 Correlation analysis ... 106
2.3.2.3 Confirmatory factor analysis ... 107
2.4 Discussion ... 110

Chapter 3: The contribution of spatial ability to science achievement in middle childhood ***118***

3.1 Introduction 118

3.1.1 Spatial ability and science learning 118

3.1.2 Spatial ability and adult's science learning 119

3.1.3 Spatial ability and younger children's science learning 119

3.1.4 Changes in the relationship between spatial ability and science at different stages of learning 121

3.1.5 Science assessment approach 121

3.1.6 Current study 122

3.2 Methods 122

3.2.1 Participants 122

3.2.2 Measures 123

3.2.2.1 Spatial measures 123

3.2.2.2 Science assessment 123

3.2.2.3 Control variables 127

3.2.3 Procedure 127

3.3 Results 127

3.3.1 Descriptive statistics 127

3.3.2 Correlation analysis 127

3.3.3 Regression analysis 128

3.4 Discussion 133

Chapter 4: The contribution of spatial ability to children's understanding of sound propagation ***139***

4.1 Introduction 139

4.1.1 Summary of aims and goals for chapter 139

4.1.2 Overview in relation to findings from previous chapter 139

4.1.3 Children's understanding of sound across development 143

4.1.4 Current study and hypotheses 146

4.2 Methods 149

4.2.1 Participants 149

4.2.2 Materials 150

4.2.2.1 Spatial measures 150

4.2.2.2 Control measures 153

4.2.2.3 Science lesson 153

4.2.2.4 Prior knowledge assessment 155

4.2.2.5 Post-test ... 156
4.2.3 Procedure ... 163
4.3 Results ... 163
4.3.1 Descriptive statistics ... 163
4.3.2 Confirmatory Factor Analysis ... 164
4.3.3 Preliminary analysis of school class group and gender ... 167
4.3.4 Correlation analysis ... 168
4.3.5 Regression analysis ... 169
4.4 Discussion ... 179
Chapter 5: The role of gesture as a spatial tool for learning about magnetism ... 187
5.1 Introduction ... 187
5.1.1 Overview in relation to findings from previous chapters ... 187
5.1.2 Gesture overview and theoretical rationale ... 189
5.1.3 Prior research on the role of gesture in relation to learning ... 191
5.1.4 Current studies ... 194
5.2 Study 1 ... 198
5.2.1 Hypotheses ... 198
5.2.2 Method ... 199
5.2.2.1 Participants ... 199
5.2.2.2 Design ... 200
5.2.2.3 Materials and procedure ... 200
5.2.3 Results ... 213
5.2.3.1 Descriptive statistics and correlations ... 213
5.2.3.2 Analysis strategy ... 214
5.2.3.3 Consolidation question analysis ... 214
5.2.3.4 Near transfer question analysis ... 215
5.2.3.5 Intermediate transfer question analysis ... 216
5.2.3.6 Post-test gesture production analysis ... 217
5.2.4 Discussion ... 218
5.3 Study 2 ... 220
5.3.1 Overview ... 220
5.3.2 Method ... 221
5.3.2.1 Participants ... 221
5.3.2.2 Materials and procedure ... 221
5.3.2.3 Pre-teaching familiarisation activity ... 221
5.3.3 Results ... 222
5.3.3.1 Descriptive statistics ... 222
5.3.3.2 Consolidation question analysis ... 223

5.3.3.3 Near transfer question analysis 224
5.3.3.4 Intermediate transfer question analysis 225
5.3.3.5 Post-test gesture production analysis 226
5.3.4 Discussion 226
5.4 Additional analysis for study 1 and 2 228
5.5 Overall discussion 230
Chapter 6: Discussion 236
6.1 Introduction 236
6.2 The structure of spatial thinking skills in childhood 237
6.2.1 General empirical and theoretical critique of Uttal et al. (2013) model in light of available evidence. 243
6.3 The relationship between individual differences in children's spatial thinking skills and science learning 244
6.3.1 Review of findings from Chapters 3 and 4 245
6.3.1.1 Intrinsic-dynamic thinking and mental models 250
6.3.1.2 Summary of findings from Chapters 3 and 4 254
6.4 The role of gesture as a spatial tool to support science learning 255
6.5 Overall summary thesis findings 259
6.6 Overall evaluation of main lines of evidence in current thesis 260
6.7 Implications for education 263
6.8 Limitations and future work 265
6.9 Concluding remarks 268
References 269
Appendices 289
8.1 Example problem-solving question, as presented to children (Chapter 4) 289
8.2 Near transfer questions (magnetism assessment; Chapter 5) 290
8.3 Intermediate transfer questions (magnetism assessment; Chapter 5) 294

List of Tables

Table 1: Summary of Uttal et al.'s 2x2 model, in relation to factor-analysis models. Source Uttal et al. (2013). 37
Table 2: Summary of spatial thinking skills based on Uttal et al. (2013) typology. ..42
Table 3: Spatial tasks included in Mix et al. (2018)....................85
Table 4: Demographic information of participants across age groups....................92
Table 5: Summary of tasks included in relation to Uttal et al. (2013) model 92
Table 6: Percentage accuracy for males and females. Means and standard deviations in parentheses. 99
Table 7: Descriptive statistics, by age group. Means and standard deviations in parentheses. Percentage accuracy is reported....................100
Table 8: Zero order correlations (upper triangle) and partial correlations controlling for age in months and scores on remaining spatial tasks (lower triangle) between spatial tasks....................107
Table 9: Fit indicators for CFA for whole sample. 108
Table 10: Summary of sub-topics included in the science assessment 124
Table 11: Descriptive statistics for science total scores, BPVS raw scores and spatial measures. Maximum possible score in parentheses. Percentage accuracy is reported....................128
Table 12: Zero-order correlations between study variables (upper triangle) and partial correlations (lower triangle). 130
Table 13: Multiple regression analysis predicting science total score. 131
Table 14: Multiple regression analysis predicting biology score. 131
Table 15: Multiple regression analysis predicting chemistry score. 131
Table 16: Multiple regression analysis predicting physics score 132
Table 17: Summary of study hypotheses 149
Table 18: Factual knowledge questions 157
Table 19: Coding scheme for prior knowledge assessment, with example responses. 157
Table 20: Summary of problem-solving questions (lower level of difficulty)....................159
Table 21: Summary of problem-solving questions (higher level of difficulty)....................160
Table 22: Summary of question one....................161
Table 23: Example responses for prediction element of question one 162

Table 24: Example responses for explanation element of question one 162
Table 25: Descriptive statistics (percentage accuracy) 164
Table 26: Fit indicators for one and two factor models. 166
Table 27: Bivariate and partial correlations between the study variables. 170
Table 28: Multiple regression analysis predicting factual recall score 171
Table 29: Multiple regression analysis predicting total problem-solving score 172
Table 30: Multiple regression analysis predicting total problem-solving score 172
Table 31: Multiple regression analysis predicting total problem-solving score, excluding matrix reasoning 173
Table 32: Multiple regression analysis predicting total problem-solving score, excluding matrix reasoning 174
Table 33: Multiple regression analysis predicting overall prediction total. 174
Table 34: Multiple regression analysis predicting overall prediction total, excluding matrix reasoning. 175
Table 35: Multiple regression analysis predicting overall explanation total 175
Table 36: Multiple regression analysis predicting overall explanation total, excluding matrix reasoning. 176
Table 37: Summary of predictors in multinomial logistic regression model 177
Table 38: Summary of predictors in relation to categorical outcome. 178
Table 39: Summary of findings in relation to study hypotheses 180
Table 40: Summary of study hypotheses. 199
Table 41: Summary of teaching content for all conditions 202
Table 42: Example responses for example near transfer verbal item. 210
Table 43: Example responses for example intermediate transfer verbal item. 211
Table 44: Summary of descriptive statistics by condition. 214
Table 45: descriptive statistics for mean percentage accuracy on first attempt of consolidation questions by condition. 215
Table 46: Frequency of iconic gestures produced overall during post-test by condition. 218
Table 47: Descriptive statistics for study 2 223
Table 48: Descriptive statistics for percentage accuracy on first attempt, by condition. 223
Table 49: Frequency of iconic gestures produced during post-test, by condition. .. 226

List of Figures

Figure 1: Example of a hidden/embedded figures trial 23
Figure 2: Example of a snowy figures trial 23
Figure 3: Example of a trial from the identical pictures test ………………………..23
Figure 4: Example of a paper folding trial 24
Figure 5: Example of a surface development trial...24
Figure 6: Example of a form board trial..24
Figure 7: Example of a flag rotation trial 25
Figure 8: Example of a card rotation trial 25
Figure 9: Example of a 3D mental rotation trial...25
Figure 10: Localisation spatial scaling task trial 26
Figure 11: Water-level task trial...26
Figure 12: Example trial from the Guilford-Zimmerman test of spatial orientation..27
Figure 13: Kozhevnikov & Hegarty (2001) perspective taking task trial 28
Figure 14: 135° anti-clockwise 2D mental rotation trial 93
Figure 15: Mental folding trial 94
Figure 16: Perspective taking, 90°, three object trial 96
Figure 17: Spatial scaling trial at a scaling factor of 1:4...97
Figure 18: Mental folding and mental rotation accuracy by age group
Figure 19: CEFT accuracy by age group...101
Figure 20: Spatial scaling accuracy by age group
Figure 21: Perspective taking accuracy by age group...101
Figure 22: Model one, one-factor, baseline model..109
Figure 23: Model two, two-factor, intrinsic-extrinsic model 109
Figure 24: Model three, two-factor, static-dynamic model.....................................110
Figure 25: Biology item from paper 1. Conceptual understanding focus. 125
Figure 26. Biology item from paper 1. Factual knowledge focus...........................126
Figure 27: Mental rotation trial 151
Figure 28: Arrival judgment task sequence..152
Figure 29: First video clip presented in science lesson 154
Figure 30: Second video clip presented in science lesson.......................................154
Figure 31: Third video clip presented in science lesson..154
Figure 32: Image shown to children for prior knowledge assessment 156

Figure 33: Model 1 (one-factor) ... 166
Figure 34: Model 2 (two-factor, intrinsic-extrinsic) ... 167
Figure 35: Model 3 (two-factor, static-dynamic) ... 167
Figure 36: Prior knowledge questions. ... 201
Figure 37: Bar magnets used in study 1 and study 2 ... 201
Figure 38: Example slide from magnetism teaching presentation ... 202
Figure 39: Gestures representing repelling south poles.. ... 205
Figure 40: Gestures representing repelling north poles. ... 205
Figure 41: Gestures representing attracting poles.. ... 205
Figure 42: Gestures representing changing from repelling to attracting poles by rotating hand 180º ... 206
Figure 43: Gesture representing holding a magnetic object (metal disc). ... 206
Figure 44: Gesture representing holding a non-magnetic object (plastic cup). ... 206
Figure 45: Concrete model/gesture consolidation question ... 208
Figure 46: Verbal-diagram consolidation question ... 208
Figure 47: Word problem slide ... 208
Figure 48: Diagrammatic and verbal item example from near transfer question set. ... 210
Figure 49: Diagrammatic and verbal item example from intermediate transfer question set. ... 211
Figure 50: Engagement questions ... 213
Figure 51: Mean near transfer score by condition ... 216
Figure 52: Mean intermediate transfer score by condition ... 217
Figure 53: Mean near transfer score by condition ... 224
Figure 54: Scatter plot of near transfer performance against mental folding scores, by condition. ... 225
Figure 55: Intermediate transfer performance by condition ... 225
Figure 25: Biology item from paper 1. Conceptual understanding focuss. ... 297
Figure 26. Biology item from paper 1. Factual knowledge focus. Source:. ... 297
Figure 36: Prior knowledge questions. ... 297
Figure 39: Gestures representing repelling south poles. Participants moved hands to represent the resistance felt. ... 297
Figure 40: Gestures representing repelling north poles. ... 297

Figure 41: Gestures representing attracting poles. Participants moved hands together to represent the force. 297
Figure 42: Gestures representing changing from repelling to attracting poles by rotating hand 180º 297
Figure 43: Gesture representing holding a magnetic object (metal disc). 297
Figure 44: Gesture representing holding a non-magnetic object (plastic cup).........297
Figure 48: Diagrammatic and verbal item example from near transfer question set 297
Figure 49: Diagrammatic and verbal item example from intermediate transfer question set 297

Chapter 1

Literature review and introduction to thesis

1.1 Introduction

Science plays a crucial role in society at a personal, social and economic level. For instance, scientific literacy, the ability to apply knowledge of scientific concepts within our day-to-day lives, is important for decision making about societal issues. For example, one factor which affects the likelihood of behaviour change is the level of knowledge an individual holds about the issue in question (Gifford, Kormos, & McIntyre, 2011). Taking climate change as an example, whereas air pollution remains in the atmosphere for a few hours or days, carbon dioxide remains for hundreds of years. However, in a recent study, participants did not differentiate between the two. The most frequently reported category, for both carbon dioxide and air pollution, was 'decades', a dramatic underestimation of the residence time of carbon dioxide (Dryden, Morgan, Bostrom, & Bruine de Bruin, 2018). Such an underestimation may impact beliefs about the reversibility of climate change, and ultimately, the perceived urgency with which personal behaviour change is needed.

At an economic level, science is also a significant contributor to the United Kingdom. For example, life sciences alone contribute £30.4bn to the UK Gross Domestic Product (GDP) and also support 482,000 jobs (PricewaterhouseCoopers, 2017). However, there are also concerns surrounding the shortfall in skilled workers within many science disciplines. Taking STEM (science, technology, engineering and mathematics) as a whole, a recent study estimated that there is currently a STEM shortfall of approximately 173,000 workers (Baker, 2018). Moreover, the most up-to-date UK shortage occupations lists includes close to 50 engineering occupations, and 13 occupations related to physics, geology and meteorology ("Shortage Occupation List," 2018). Beyond industry, there are also significant difficulties in recruiting specialist science teachers to teach in secondary schools (Sibieta, 2018). Consequently, outside of London, only 17% of secondary school physics teachers hold a relevant degree (Sibieta, 2018). What factors contribute to the shortage of specialist science teachers to teach in schools, and suitably qualified scientists and engineers to work professionally within industry?

Cognitive processes, abilities and skills are one factor that play a role in educational success. Focusing on the learner in science education, then, would seem to be a crucial way of addressing the STEM gap. On the one hand, the importance of science education is reflected in science being a ‘core’ curriculum subject, along with mathematics and English (Department for Education, 2013b). Yet, compared to mathematics and literacy, the psychological, and particularly cognitive factors, that influence science learning are less well understood. This is particularly the case for younger children (Tolmie, Ghazali, & Morris, 2016).

This thesis focuses on spatial thinking as one contributor to primary-school aged children’s successful learning and achievement in science. We frequently use spatial thinking in our daily lives; for example, when we pack a suitcase or use a map to navigate around an unfamiliar environment. Spatial thinking is a ubiquitous part of learning and doing science. Historically, many famous scientific discoveries have been attributed to spatial visualisation. For example, using Rosalind Franklin’s 2D X-ray images, James Watson and Francis Crick were able to infer the 3D structure of the DNA double helix structure. More generally, spatial thinking is used within science through the use of diagrams, animations, models and other spatial representations; processes which have cross-cutting importance in all science domains. At a more specific level, spatial thinking is important in biology, for instance, when we learn about the structure of organs within the body, and, in physics, when we reason about the spatial relational structure of the solar system. Scientists also choose to represent non-spatial information in a spatial format. For example, a central element of the scientific process is the dissemination of research findings. Research may be shared with peers for critique, or with the wider public, to inform. The presentation of non-spatial information (e.g., numerical data) in a spatial format (e.g., tables, graphs) is commonplace in science.

Large-scale longitudinal studies (e.g. Wai, Lubinski, & Benbow, 2009) spanning the past 50 years have also demonstrated convincingly that individual differences in spatial thinking skills in adolescence are predictive of STEM achievement in adulthood. However, little research has addressed the specific relationship between spatial thinking and science learning, in primary-school aged children. Science in the primary school years lays the foundation for future science learning at all levels. Having a more in-depth understanding of the relationship between spatial thinking and science learning at this earlier point in development has the potential to support spatial thinking, and science, within the primary-school years, and beyond. For instance, a deeper understanding of which spatial thinking skills relate to different aspects of children's science learning could inform training and interventions. These interventions, at an earlier stage, could improve engagement and achievement in science throughout adolescence, and ultimately encourage more students to pursue further qualifications, and careers, in science. The primary goal of the current thesis is, therefore, to comprehensively investigate the relationship between spatial thinking skills and science learning in the primary-school years (Chapters 3, 4 and 5).

Research to date investigating the relationship between spatial thinking and science, with learners of all ages, has been hampered by inconsistencies between, and limitations of, prior theoretical models of spatial thinking and approaches to defining spatial cognition. For example, historically, most prior typologies have been based largely on bottom-up, psychometric data, such as exploratory factor analysis (EFA) evidence, and have also not been sufficiently driven by theory and recent developments in neuroscience. There has also been a heavy focus on factors such as object-based visualisation, or on skills such as mental rotation. Many prior spatial skill typologies have therefore been limited in their scope. As a result, much prior research investigating the connection between spatial thinking and science learning in the primary school years has adopted composite measures of spatial thinking, or, has focused on a restricted range of spatial skills.

To address the inconsistencies and restricted focus of prior spatial typologies, Uttal et al., (2013) proposed a novel spatial typology. In contrast to prior models, the model is theory-driven and draws on cognitive, linguistic and neuroscientific evidence, rather than being exclusively driven by 'bottom-up' psychometric data. However, to date, there have been very few direct empirical tests of this typology, and only one with children. The secondary goal of the thesis is to therefore empirically assess the model in childhood, using several strands of inter-related evidence. One line of evidence presented within the thesis is the differential predictiveness of spatial skills in relation to science learning (Chapters 3, 4 and 5). That is, to what extent does the pattern in which different spatial skills predict science learning lend support to the typology? A second strand of evaluation presented is psychometric modelling through confirmatory factor analysis (Chapters 2 and 4). The final source of evidence presented is the developmental trajectories of various spatial skills in middle childhood, and the extent to which they map onto or support the Uttal et al. (2013) typology (Chapter 2).

Spatial tools, symbolic representations of spatial information, such as diagrams, gestures, concrete models and spatial language, are another component of spatial cognition. However, spatial tools have been neglected within the dominant spatial typologies to date. Spatial tools are useful in science learning. For example, diagrams provide a structured spatial representation of scientific processes and structures. Concrete models afford the opportunity to physically interact with science concepts. Additionally, hand gestures can also support understanding of the spatial-relational elements within scientific processes (e.g., a hand moving in a circular fashion, to represent the orbit of a planet). Spatial tools may be independently useful for children as instructional tools within science, regardless of their level of spatial thinking skill. However, spatial tools may also have a moderating relationship with spatial thinking ability. For instance, hand gestures may be particularly useful for children with lower spatial ability, because they may provide an external spatial scaffold, or they may promote the mental simulation of physical actions in science processes. Similarly, spatial ability may be particularly important for learning with diagrams in science, because of the need to visualise and animate elements within a static image. To date, there has been no research that investigates the effectiveness of gesture as a spatial tool to support children's science learning, and no investigation of how spatial thinking skills may moderate this effect. The final goal of the thesis is to therefore investigate the effectiveness of a particular spatial learning tool, gesture, in relation to other spatial tools, and how gesture interacts with spatial thinking skills, in the context of science instruction and learning (Chapter 5).

The remainder of this introductory chapter presents a review of the pertinent literature, beginning with a discussion of the Uttal et al. (2013) model of spatial cognition and how it relates to previous dominant models and typologies. Next, there is a review of science skills and models of conceptual development. Following this, there is a review of prior studies focusing on the relationship between these domains. Finally, drawing on this review, the main aims and goals of the thesis are reviewed.

1.2 Defining and categorising spatial ability and spatial cognition

1.2.1 Introduction

Individual differences in spatial thinking skills, spatial ability, relates to our knowledge of, and ability to represent and transform, "the location of objects, their shapes, their relation to each other, and the paths they take as they move" (Newcombe, 2010, p30). Spatial ability has long been recognised as an ability at least partly independent of general intelligence, reasoning and verbal ability (Hegarty, 2014; Thurstone, 1948). Furthermore, although some authors have described spatial ability as a unitary concept, there are thousands of spatial ability tests, measuring a wide range of spatial skills (Elliot & Smith, 1983). Various psychometric (e.g., Carroll, 1993) and more theoretical (e.g., Uttal et al., 2013) approaches to categorising the breadth of these skills exist, with differing emphases. These models, described below, generally have in common, however, the underlying assumption that skills falling within different categories measure partially distinct but correlated spatial abilities.

The section that follows begins with a discussion of the Uttal et al. (2013) spatial typology, the model that will be evaluated within the thesis. This model is then compared with other previous, dominant spatial typologies (Carroll, 1993; Linn and Petersen, 1985). Following this, there is a discussion of additional spatial factors that are either not directly included within these models or the placement of these factors within the proposed sub-domains is unclear.

1.2.2 Uttal et al. (2013): 2 x 2 typology (see also Newcombe & Shipley, 2015)

Definitions and categorisations of spatial abilities developed from psychometric approaches, which adopted an exploratory factor analysis methodology, throughout the 20th century (see discussion of Carroll, 1993; section 1.2.3.1). Exploratory factor analysis explores the relationship between observed variables (e.g., scores on spatial thinking tasks) to determine if they measure a common underlying (unobservable) latent variable (e.g., a type of spatial skill). Exploratory factor analysis studies with adults have typically identified several distinct but correlated spatial factors. However, the specific emerging structure of spatial thought differs across studies. This reflects the tasks included in the analyses, and in some cases, the subjectivity involved at various points of the analysis process. This, in addition to inconsistent use of terminology across authors, explains some of the confusion in definition and categorisations evident in the literature (Hegarty, 2014; Newcombe & Shipley, 2015; D'Oliveira, 2004). The number of spatial factors identified in previous factor analysis studies with adults ranges from two to ten (D'Oliveira, 2004). Unlike confirmatory factor analysis, exploratory factor analyses can be run without a theory in mind.

Building on suggestions and concerns raised in a review by Hegarty and Waller (2005), Uttal et al., (2013) and Newcombe and Shipley (2015) argued that the field of spatial cognition needed a new approach to defining spatial abilities. They further suggest that this approach should be based on the top-down understanding of spatial skills, rather than exclusively being from bottom-up inductive analysis, as in the exploratory factor analysis-tradition. Uttal et al. (2013) argue that the model should draw upon developments in cognitive psychology and cognitive neuroscience. Such an approach would address the issues with definitions and categorisations outlined above.

The authors, therefore, propose a '2 x 2 model', distinguishing between skills as being intrinsic and extrinsic, along one dimension, and static or dynamic, on the other. *Intrinsic skills* are within-object focused skills. *Extrinsic skills* focus on the relationship between objects, or, relate to the relationship between an object and a frame of reference, or to the wider environment. Within this model, as well as being either intrinsic or extrinsic, skills are classified as either being *static* or *dynamic*. Dynamic skills involve movement or transformation (e.g., an imagined rotation), whereas static skills do not, and may involve representation only. In addition, movement can either be mentally operated, as in the mental rotation or externally created (Uttal et al., 2013). In many cases, a static skill is a pre-requisite to a dynamic skill (e.g., it is necessary to encode a shape before mentally rotating it), and therefore it may be difficult to clearly distinguish between skills along this additional dimension.

The proposed 2 x 2 typology results in 4 spatial sub-domains: intrinsic-dynamic; intrinsic-static; extrinsic-dynamic; extrinsic-static. Within this model, *intrinsic-static skills* involve the processing of objects or shapes, or parts of objects or shapes, without further transformation. Examples of spatial thinking tasks that would measure *intrinsic-static* spatial thinking skills are: the hidden/embedded figures task; the snowy figures task and the identical pictures task. The hidden/embedded figures task (Figure 1) requires participants to locate a particular 2D shape from another, distracting, shape made up of joined other 2D shape. The snowy figures test (Figure 2) involves searching for an unknown shape within a 'snowy' background.

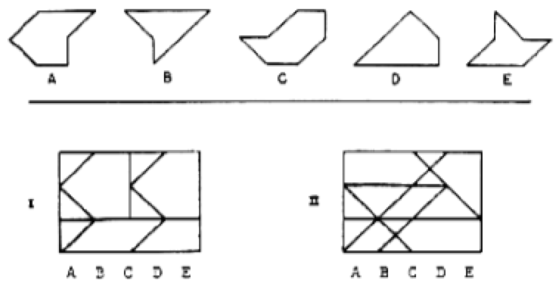


Figure 1: Example of a hidden/embedded figures trial (source Hegarty & Waller, 2005)

Figure 2: Example of a snowy figures trial (source Hegarty & Waller, 2005)

Figure 3: Example of a trial from the identical pictures test (source Hegarty & Waller, 2005)

Finally, the identical pictures task (Figure 3) requires a participant to determine which of five images on the right is identical to a given pattern on the left.

Within the model, intrinsic-dynamic thinking is said to involve the processing and manipulation or transformation of objects or shapes. Examples of spatial thinking tasks that measure *intrinsic-dynamic* spatial thinking skills are: mental paper folding; surface development tasks; the form board task; and 2D and 3D mental rotation. In the *paper folding task* (e.g., Ekstrom, Dermen, & Harman, 1976; Figure 4) participants firstly see an image of a sheet of paper, presented with an indication of a fold line. Participants are required to imagine the paper being folded in this way, and then, to visualise a hole being punched through the folded sheet. Participants then choose the correct image, of five images of the unfolded paper, which correctly shows the location of the punched holes. *Surface development tasks* (Figure 5), a different variation of a folding task, involve participants visualising the folding and unfolding the net of 3D shapes. The example in Figure 5 requires participants to visualise which of the four 3D shapes on the right would be constructed with the unfolded net on the left. *The form board task,* Figure 6, requires the mental construction of a 2D shape, using a number of other smaller, 2D shapes. In the example shown in Figure 6, participants select which of the five shapes together form the rectangle above.

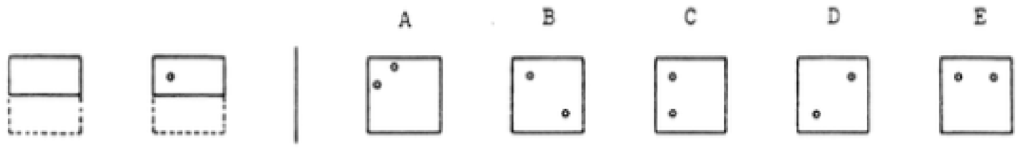


Figure 4: Example of a paper folding trial (source Hegarty & Waller, 2005)

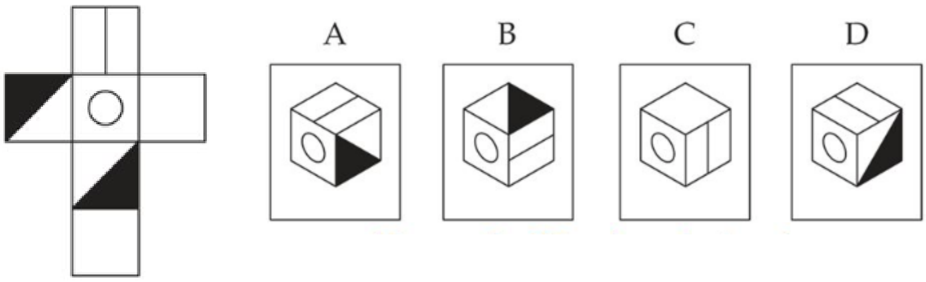


Figure 5: Example of a surface development trial (source Hegarty & Waller, 2005)

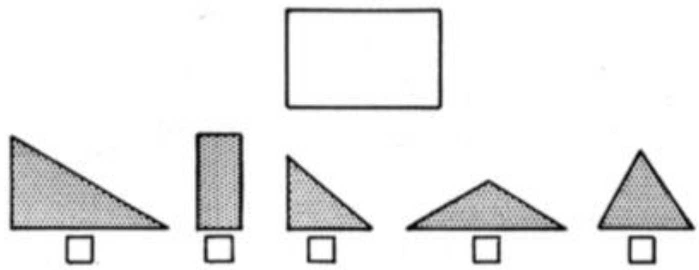


Figure 6: Example of a form board trial (source Hegarty & Waller, 2005)

2D and 3D mental rotation are the most well-known and researched skills falling within the intrinsic-dynamic domain. Considering first 2D rotation, in the *flag* *rotation test* (Figure 7) participants rotate a 2D target shape, and then determine whether the other figure, to the right of the target, is the same as the target, but rotated, or is a mirror version of the target (Carroll, 1993). The *card rotation task* (Figure 8), has four response options, rather than one.

Figure 7: Example of a flag rotation trial (source: Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001)

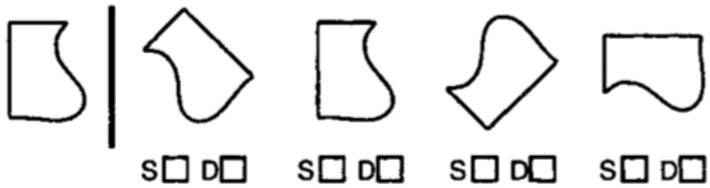


Figure 8: Example of a card rotation trial (source Miyake et al., 2001)

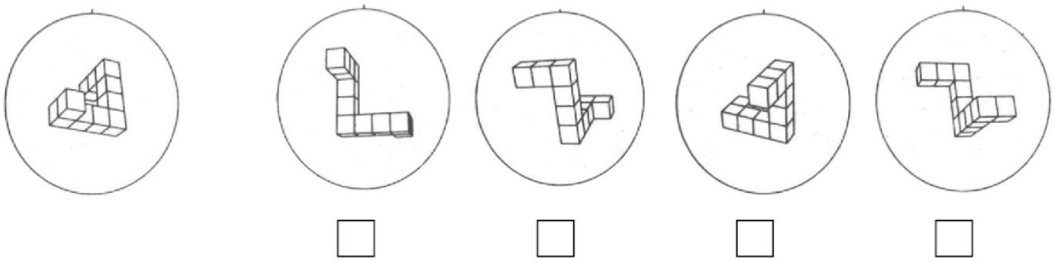


Figure 9: Example of a 3D mental rotation trial.

Finally, Figure 9 illustrates an example of a 3D mental rotation task (Vandenburg & Kuse Mental Rotation task [Peters et al., 1995]). Participants are asked to choose one of two response options from the group of four on the right, which match the target on the left.

Newcombe and Shipley (2015), in their final categorisation and updated version of the Uttal et al. (2013) model, further break down intrinsic-dynamic skills into *rigid and non-rigid transformations*. In a rigid transformation, the spatial relations within the object are preserved, whereas in a non-rigid transformation, spatial relations are not preserved. For example, mental rotation is rigid because the distances are preserved and objects remain whole, whereas mental folding is a non-rigid transformation.

Extrinsic-static thinking involves the coding of the locations and spatial relations between objects, and between objects and other landmarks. Within the extrinsic-static category, the authors refer to alignment, which relates to reasoning about temporal and spatial correspondence. An example of an extrinsic-static skill, based on spatial correspondence, is the ability to find corresponding locations between shapes of equal proportions but differing sizes: spatial scaling. For instance, in a localisation scaling task, Figure 10 (Frick & Newcombe, 2012), the participant is shown a map with an object at a target. The participant is then shown another map, which is identical, other than it being scaled by a specific scaling factor. The participant is asked to place an object on the new map, in the corresponding location. Spatial perception tasks are also used to measure the extrinsic-static dimension. For example, the water level task (Piaget & Inhelder, 1956) requires participants to draw the level of a liquid in a bottle, shown at different orientations (Figure 11).

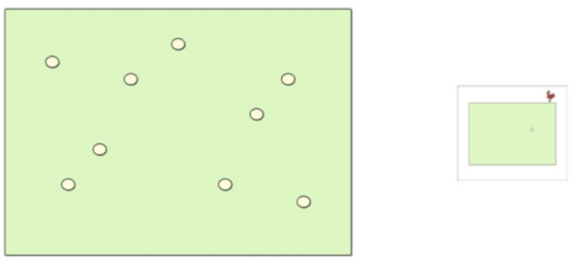


Figure 10: Localisation spatial scaling task trial

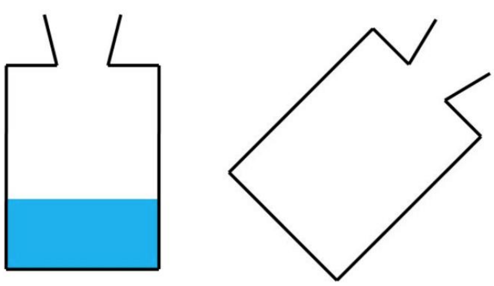


Figure 11: Water-level task trial

Extrinsic-dynamic thinking involves the transformation of the relationship between objects, or the relationship between objects and frames of reference. Perspective taking and spatial orientation are the main examples of extrinsic-dynamic spatial thinking skill, within the model. Perspective taking and spatial orientation are used as terms within the literature to refer to similar types of spatial skill. In these tasks, a participant visualises how an object or scene might look from another vantage point or viewing angle.

For example, in the Guilford-Zimmerman test of Spatial Orientation (Guilford & Zimmerman, 1948; Figure 12), a participant is shown an image of a boat on a lake, from the perspective of someone piloting the boat. A second image shows how the scene appears after the boat changes direction, still from the perspective of the pilot. The participant is provided with a selection of aerial views which represent the change in direction from the first to the second image. Participants are asked to choose which one aerial view best represents the change in direction.

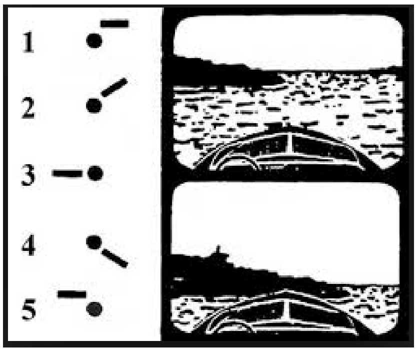


Figure 12: Example trial from the Guilford-Zimmerman test of spatial orientation

Another variation of a perspective taking task was later presented by Kozhevnikov & Hegarty (2001). In this task (Figure 13), participants were shown a set of two-dimensional objects and were directly asked to imagine that they were facing a particular way, within the array. Participants then determined what direction another object would be, based on this new perspective.

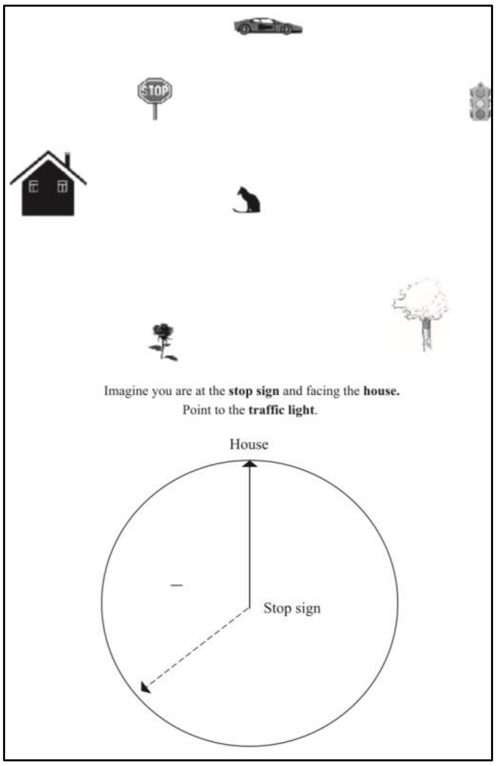


Figure 13: Kozhevnikov & Hegarty (2001) perspective taking task trial

1.2.2.1 Evidence supporting the model

Although there has been very little direct evaluation of the model to date, there is a broad range of supportive evidence for the dimensions of the model, and for specific sub-domains of the model. As one evolutionary basis for the broad *intrinsic/extrinsic* dimension, Newcombe (2018) states that humans have two primary spatial functions, tool use and navigation, which have a distinct evolutionary and neural basis. Tool use involves *intrinsic*, object-based spatial relations, whereas navigation involves the *extrinsic* coding of relations between objects and frames of reference. In addition, connecting perspective taking and navigation within the extrinsic spatial domain: when navigating, it may be the case that information from multiple views or perspectives are integrated (Newcombe, 2018).

In support of the *static-dynamic distinction*, Kozhevnikov, Kosslyn, and Shephard (2005) reported a distinction between object-visualisers, who tend to excel at intrinsic-*static* skills, and spatial-visualisers, who tend to excel at intrinsic-*dynamic* skills. Moreover, this study found that artists were more likely to be object-visualisers, and scientists were more likely to be spatial-visualisers.

There is also neuroscientific evidence to support the model. First, for the distinction between intrinsic-static spatial skills, and other sub-domains in the model. For example, object comparison (intrinsic-static) and location comparison (extrinsic-static) tasks result in different patterns of brain activation (Haxby et al., 1994). Farah, Levine, and Calvanio (1988) also reported a case of a brain-damaged patient with selective impairment in the processing of visual (intrinsic) information (shape and colour), but not spatial information (object transformation and location information). More generally, evidence also exists for the static/dynamic distinction; for example, in the macaque, part of the dorsal stream, the medial temporal and medial superior temporal areas, contain neurons that are only sensitive to motion (Chatterjee, 2008).

1.2.2.2 Object-based transformations (intrinsic-dynamic) vs perspective taking (extrinsic-dynamic) within the model.

The distinction between spatial orientation/perspective taking (extrinsic-dynamic skills) and object-transformation, e.g., mental rotation (intrinsic-dynamic skills) is particularly central evidence for the intrinsic-extrinsic dimension within the model. This distinction is supported by, and links to, egocentric and allocentric encoding of space. Egocentric and allocentric encoding refers to specific frames of reference in space (Klatzky, 1998). Egocentric frames of references refer to understanding spatial locations and representations in relation to the self, and are based on body-based coordinates (self-to-object encoding). Allocentric encoding relates to a spatial frame of reference that is external to the person, and more specifically, is based on the relationship between objects and landmarks within the environment (object-to-object encoding). In order to perform object-based transformations (e.g., intrinsic-dynamic skills, such as mental rotation), the body coordinates of the observer are often used as a reference point (i.e., egocentrically) (Newcombe, 2018). For example, an object may be mentally rotated 90^{o}, in relation to the current orientation of the self. During perspective taking tasks, egocentric encoding is often also used but is updated, in order to adopt a different spatial perspective. However, in perspective taking tasks, and not mental rotation or object-based mental transformation tasks, it is necessary to also utilise allocentric encoding (i.e., the mapping between environmental frameworks and landmarks). Thus, egocentric encoding is common to both intrinsic and extrinsic skills, within the Uttal et al. (2013) model, but allocentric encoding and manipulation is an additional requirement of mature perspective taking performance (Newcombe, 2018).

It is important to note, however, that there has not been consistent support for the intrinsic-dynamic (e.g., mental rotation) vs extrinsic-dynamic (e.g., spatial perspective taking) within the prior literature. One reason for this is that many extrinsic-dynamic perspective taking tasks can be solved using various rotation-related methods, or through methods which use spatial encoding methods linked more with mental rotation. Indeed, Newcombe, Uttal, and Sauter (2013) argue that the crossover between these domains is that, for some spatial orientation/perspective taking tasks, participants could choose to mentally transform the object itself, rather than change their perspective. However, as will be discussed below, the extent to which this is the case varies depending on the type of perspective taking or spatial orientation task, and also the instructions that are given to participants. The following paragraphs review the evidence to date on this distinction.

One approach to studying the dissociation between mental rotation and perspective taking has been to use psychometric approaches, such as factor analysis, to distinguish between these skills. For instance, in a meta-analysis of 90 exploratory factor analysis studies, Carroll (1993) was unable to find consistent evidence for a psychometrically dissociated spatial orientation/perspective taking ability, and therefore, included it with his visualisation factor (Carroll's [1993] model is discussed more fully in section 1.2.3.1). The majority of studies included in this analysis used the Guilford-Zimmerman test of Spatial Orientation, described previously in section 1.2.1. However, more recently, a psychometric distinction between mental rotation and spatial perspective taking has been reported (Kozhevnikov & Hegarty, 2001; Hegarty & Waller, 2005). These studies used the previously described perspective taking task (Figure 13). Confirmatory factor analyses within these two studies resulted in a dissociation between this variation of perspective taking task, and measures of mental rotation.

Self-report data provides some indication of the discrepancy between these findings. Barratt (1953), for example, found that the majority of participants reported solving the Guilford-Zimmerman test of Spatial Orientation task by mentally imagining the movement of the boat (i.e., via object-based visualisation), not by imagining themselves moving. In addition, in this task, participants are not directly asked to adopt a different perspective. However, in the task used by Kozhevnikov &

Hegarty (2001), the majority of participants self-reported changing their spatial perspective, rather than rotating the array. In this task, participants are more explicitly asked to imagine facing a particular way and face a new direction. Much of the subsequent research comparing perspective taking and mental rotation, within the experimental psychology and cognitive psychology tradition, has adopted similarly more careful and specific task paradigms to allow a clearer distinction between the two processes.

For example, brain imaging and neuropsychological data from adults suggest that mental rotation and perspective taking are linked to dissociable but overlapping systems (Zacks, Vettel, & Michelon, 2003; Wraga, Shephard, Church, Inati, & Kosslyn, 2005; Lambrey, Doeller, Berthoz, & Burgess, 2012). Within these studies, participants were explicitly asked to either rotate an array, or to update their viewpoint to the one adopted by another person. For instance, fMRI research indicates that perspective taking (extrinsic-dynamic skills) show patterns of brain activation more in line with navigation (e.g., retrosplenial cortex and hippocampus). However, brain activation for array rotation (intrinsic-dynamic skills) is more similar to the rotation of objects, i.e., activation of the right intraparietal sulcus (Lambrey, Doeller, Berthoz, & Burgess, 2012).

As well as this psychometric and neuropsychological distinction, more recent adult based behavioural studies indicate that mental rotation and spatial perspective taking tasks often have different chronometric profiles (Crescentini, Fabbro, & Urgesi, 2014). That is, mental rotation tasks frequently demonstrate a linear increase in response time with increasing angle of rotation (e.g., Shepard & Metzler, 1971). This is thought to be because mental rotation of objects involves participants visualising the rotation of the object through each angle. The larger the angle of rotation required, the longer the imagined rotation would take, as would be the case in a physical rotation. Chronometric data from perspective taking tasks, however, often do not show this linear pattern of response time (Habacha, Moreau, Jarraya, Lejeune-Poutrain, & Molinaro, 2018). Response times often remain low for low levels of angular disparity, but, increase suddenly for higher angular disparities (Keehner, Guerin, Miller, Turk, & Hegarty, 2006; Michelon & Zacks, 2006; Kessler & Thomson, 2010; Zacks & Michelon, 2005;). This suggests that, at least for lower angles of perspective change, adults do not necessarily perform a continuous, imagined rotation of the self around the array. It may be that participants utilise the allocentric framework to a greater extent, to perform a more direct 'blink' transformation, to another location (Wraga, Creem, & Proffitt, 2000). As with the brain imaging data described above, there were also attempts in these studies to reduce the intrinsic/extrinsic strategy crossover. For example, in Keehner et al (2006), the spatial tasks were designed in such a way that attempts to solve a spatial perspective taking trial with a rotation based strategy, or vice-a-versa, would result in an incorrect response, at least some of the time.

In addition to task features, a second reason why a dissociation is not always evident between these skills relates to developmental factors. Within perspective taking tasks, younger children's response times do sometimes demonstrate a linear relationship with angular disparity (Roberts & Aman, 1993), in contrast to adults. This suggests that younger children may sometimes use a more egocentric strategy (i.e., a graduated, imagined rotation of the self around the array), and not an allocentric strategy (i.e., utilising landmarks). It may also be the case that younger children choose to rotate the array. Prior research has shown that allocentric encoding of space develops later than egocentric encoding of space, and is not fully mature until 10 years (Bullens, Iglói, Berthoz, Postma, & Rondi-Reig, 2010). This will be revisited in sections 1.2.6 and 1.2.7.

To summarise, there is strong evidence for a dissociation between intrinsic-dynamic and extrinsic-dynamic spatial skills, behaviourally, psychometrically and neuroscientifically. However, the extent to which this is evident is dependent on careful task designs and instructions. Of course, it is still possible within some of these tasks that participants choose to use a different strategy, even though they have been instructed to use another. This may still add variability and noise to analyses which may reduce the likelihood that a psychometric dissociation is evident. In addition, developmentally, younger children may be less likely to use an allocentric/extrinsic strategy for perspective taking tasks. This may also affect the extent to which an intrinsic-extrinsic dissociation is found throughout development. Therefore, although the intrinsic-extrinsic dimension with the Uttal et al. (2013) model has a reasonable basis, these caveats mean that there is a need to further evaluate it empirically. As outlined in section 1.1, a goal of the current thesis is to therefore evaluate the Uttal et al. (2013) model, including the intrinsic-extrinsic dimension, using several lines of evidence.

1.2.3 Previous typologies

In the sections that follow, the Uttal et al (2013) is compared to prior, dominant spatial typologies: Carroll (1993), and Linn and Petersen (1985). As outlined above, Carroll's (1993) was rooted in the exploratory factor analysis tradition. Linn an Petersen (1985) largely based their model on exploratory factor analysis evidence, although also included some cognitive rationale for the typology.

1.2.3.1 Carroll (1993): psychometric model

Following the reanalysis of over 90 exploratory factor-analysis studies with adult samples, Carroll (1993) concluded that spatial ability consisted of five factors: spatial visualisation, spatial relations, closure speed, flexibility of closure and perceptual speed. Carroll also considered visual memory within his overall analysis of human ability; however, this fell in a different category to these spatial factors. In his analysis, Carroll firstly found strong evidence for spatial visualisation, which emphasises "power in solving increasingly difficult problems involving spatial forms" (Carroll, 1993, p315). Visualisation is the most frequently measured spatial factor and involves the mental transformation of shapes and forms. Moreover, the complexity associated with the spatial transformations in these tasks means that they are not linked to speed (Höffler, 2010; Hegarty & Kozhevnikov, 1999). The *visualisation* factor is frequently measured by the mental paper folding, surface development and form board tasks (Hegarty, 2014; Hegarty & Waller, 2005). As stated previously, Carroll was unable to find evidence for a separate spatial orientation or perspective taking ability, and so included it within this visualisation category, often measured by the Guildford-Zimmerman test of spatial orientation (Hegarty & Waller, 2005).

In contrast to tests of spatial visualisation, which often include a sequence of transformations, tests of th*e spatial relations* factor, in Carroll's typology, typically involve the rotation of 2D objects in a short period of time (e.g., the previously described card rotation and flag rotation tasks). Tasks in this factor are less cognitively demanding, and are often speeded. The spatial relations factor is sometimes dissociated from the spatial visualisation factor (Hegarty & Waller, 2005). However, Carroll found only inconsistent evidence for this dissociation, perhaps because, in some cases, the main difference between the factors is the complexity of the transformation involved. Thus, they may be quantitatively, rather than qualitatively different, in terms of cognitive demand (Hegarty, 2014; Höffler, 2010; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). This has more recently been supported by cognitive analyses of the factors which indicate that they differ mainly in the extent that they place demand on working memory (Hegarty, Shah, & Miyake, 2000; Shah & Miyake, 1996; Miyake et al., 2001; see section 1.2.4).

The visualisation factor and the spatial relations factor with the Carroll (1993) model map onto the intrinsic-dynamic category, within the Uttal et al. (2013) model. Because spatial perspective taking is included within the visualisation factor within the Carroll (1993) model, this also therefore maps onto the extrinsic-dynamic category, within the Uttal et al. (2013) model.

Carroll's third and fourth factors, *closure speed* and *flexibility of closure* are related to the speed and skill at 'apprehending and identifying a visual pattern, often in the presence of distracting stimuli' (Miyake et al., 2001, p638). In flexibility of closure tasks, participants know prior to taking the test what the pattern is that they are searching for. An example of a flexibility of closure task is the embedded/hidden figures task, previously described in section 1.2.2. In tests of closure speed, participants are not informed of the pattern they are searching for (Carroll, 1993). The snowy figures task, described in section 1.2.2, is an example of a closure speed task.

In the case of flexibility of closure, although Carroll's categorisation places it separately from visualisation, there are some similarities between these skills. In both visualisation and flexibility of closure, there is a requirement to maintain a mental image and to counteract the distracting stimuli. In visualisation tasks, there is often an additional step of performing a transformation of the image. This cross-over is supported by evidence that variance from spatial visualisation and relations, is often present in tests of flexibility of closure (Carroll, 1993).

The final factor, *perceptual speed,* concerns the 'speed in comparing figures or symbols, scanning to find figures or symbols, or carrying out very simple tasks involving visual perception' (French, 1951). Tests that load on the perceptual speed factor do not involve transformation or require effortful maintenance of images; rather, they require rapid matching of patterns (Miyake et al., 2001; Carroll, 1993). The previously described identical picture task falls within Carroll's (1993) perceptual speed category. Closure speed, flexibility of closure and perceptual speed map onto the intrinsic-static sub-domain, within the Uttal et al. (2013) model.

Caroll's model, which summarises the factor analysis approach, has delineated key aspects of spatial ability overall, the reliance on exploratory factor analysis, which can be used without any a priori understanding of the variables being tested, has led to inconsistent results (Hegarty & Waller, 2005). Furthermore, a factor analysis is only able to identify factors based on the quality and inclusion of the measures to identify that construct (Hegarty & Waller, 2005). Thus, for example, although Carroll was unable to find a clear distinction for a separate spatial orientation/perspective taking ability, this may have been due to the lack of sensitivity of the measure traditionally used. In addition, tasks such as spatial scaling, which fall into the extrinsic-static domain within the Uttal et al. (2013) model, do not feature within this model. This can similarly be attributed to the fact that the majority of research and task development in the domain of spatial scaling occurred after the development of the Carroll (1993) model.

1.2.3.2 Linn and Petersen (1985)

Linn and Petersen's (1985) spatial skill classification has also been useful and informative in the study of spatial ability. This meta-analysis primarily investigated the degree of sex differences in spatial ability. However, within the study, Linn and Peterson also provide a categorisation of abilities, to guide the analysis. It has subsequently been used as a more general spatial typology. Although not a factor-analysis, or a direct analysis of other factor-analysis studies in itself, Linn and Peterson's paper refers to psychometric research in the rationale for the categorisation. In addition, the authors refer to other cognitive factors, and propose that grouping of skills should also be carried out with reference to the similarity of spatial tasks, rather than exclusively on psychometrically loaded factors.

The major categories of Linn and Petersen's (1985) model are *spatial visualisation*, *mental rotation* and *spatial perception*. As with Carroll's (1993) classification, the model includes a *spatial visualisation* category, which also encompasses tests such as the form-board and mental paper folding tasks. However, their second category, *mental rotation*, includes both two and three-dimensional rotation tasks; neither of which are included within Linn & Petersons visualisation category. These tasks do not necessarily emphasise speed, as was the case with Carroll's (1993) spatial relations category. The classification proposed that mental rotation as a skill is distinct from tasks in the visualisation category. The cognitive rationale for the distinction was based on the finding that visualisation tasks often require multiple steps, and, it is often necessary that participants choose appropriate strategies, which may vary from trial to trial (see also discussion of working memory/executive function below). *Spatial visualisation* and *mental rotation* within Linn and Petersen's (1985) model map onto intrinsic-dynamic skills within the Uttal et al. (2013) model.

Another key difference is the inclusion of a *spatial perception* category. Tasks in this category require participants to separate the frame of reference of an object, from its surroundings (Hegarty, 2014). The water level task, described in section 1.2.2, falls into this category. *Spatial perception* fits within the extrinsic-static category, in the Uttal *et al.* (2013) model. The model includes flexibility and speed of closure tasks within the visualisation category, rather than as separate factors (as was the case within Carroll's [1993] and the Uttal et al [2013] model). Finally, perspective taking/spatial orientation are not included within Linn and Petersen's (1985) model.

Table 1 links the Uttal et al. (2013) typology to previously mentioned factor-analysis based categorisations and to commonly cited spatial abilities, discussed in previous sections.

Table 1: Summary of Uttal et al.’s 2x2 model, in relation to factor-analysis models. Source Uttal et al. (2013).

Spatial skills described by 2 x 2 classification	Description	Examples of measures	Linn & Peterson (1985)	Carroll (1993)
Intrinsic and Static	Perceiving objects, paths or spatial configurations amidst distracting background information	Embedded Figures Task, Flexibility of Closure	Spatial Visualisation	Flexibility of Closure, Speed of Closure, Visuospatial perceptual speed
Intrinsic and Dynamic	Piecing together objects into more complex configurations, visualising and mentally transforming objects. Rotating 2D or 3D objects.	Form board, Block Design, Paper Folding, Mental rotation, Card Rotation	Spatial visualisation Mental rotation	Spatial Visualisation Spatial relations
Extrinsic and static	Understanding abstract spatial principles such as horizontal invariance or verticality	Water-level test Spatial scaling	Spatial perception	Not included
Extrinsic and dynamic	Visualising an environment from a different perspective	Guilford-Zimmerman spatial orientation	Not included	Spatial visualisation

1.2.3.3 Summary of previous typologies in relation to Uttal et al *(2013) model*

To summarise, Carroll’s (1993) model was based entirely on data from exploratory factor analysis studies, whereas the Uttal et al. (2013) draws on data and theory from a range of sources. Linn and Petersen (1985)’s model was based on psychometric and cognitive data, and therefore bears some resemblance to the Uttal et al. (2013) model. However, Linn and Petersen’s (1985) model is still relatively fragmented, and does not provide an overarching theoretical framework, as such. In the ‘extrinsic’ domain, spatial perspective taking/orientation are omitted as separate factors the Carroll (1993) model, and not included all, in the Linn and Petersen (1985) model. However, as was outlined in section 1.2.2.2, there is evidence of a dissociation between mental rotation and perspective taking, within specific task situations. In addition, neither the Carroll (1993) model or the Linn and Petersen (1985) model include a category to include skills such as spatial scaling.

Of all of the spatial categories, the ‘intrinsic-dynamic’ (i.e., object-based transformation/manipulation) category is the most difficult to group together in a single way; perhaps, because it has received the most research, and authors have conceived it in a variety of ways. In Carroll's (1993) model, the distinction is between more complex visualisation, and, relatively straightforward, speeded spatial relations tasks. Linn and Petersen (1985), however, include a separate category for mental rotation versus spatial visualisation, based more on the similarity of tasks. Linn and Petersen's (1985) rotation category also includes complex 3D rotation. In their final updated version of the Uttal et al. (2013) model, Newcombe & Shipley (2015) distinguish between rigid and non-rigid transformation, within the intrinsic-dynamic sub-domain.

1.2.4 Other spatial factors

In the sections that follow, two additional elements of spatial cognition are discussed: visuospatial memory and multiple object dynamic spatial ability. These factors are either omitted from previous models, or their placement within categories in the models is unclear.

1.2.4.1 Visuospatial memory

As mentioned, Carroll additionally identified visual memory as a factor; however, this was included in a factor of 'general memory and learning' (Carroll, 1993). Furthermore, Baddeley and Hitch's (1974) working memory model includes the visuospatial sketchpad, a slave system specialised for the maintenance and manipulation of visual and spatial representations (Alloway, Gathercole, & Pickering, 2006).

A primary distinction within the visuospatial sketchpad, is between the *visual* aspect of working memory, which focuses on shape and colour, and the *spatial* aspect, which focuses on paths, rotation and location (Logie, 1995). Prior research into visuospatial working memory also suggests a distinction between simple storage, which involves the retention of visual and spatial information, and complex storage, which also involves the processing of visual and spatial information. The former have been termed visuospatial short-term memory span tasks, whilst the latter tasks, which involve simultaneous processing as well as storage, as visuospatial working memory span tasks (Miyake et al., 2001; Alloway et al., 2006).

Miyake et al. (2001) examined the latent relationship between visuospatial short term/working memory, executive functioning and spatial ability. The study primarily adopted Carroll's (1993) spatial skills model, and therefore, included spatial visualisation, spatial relations and perceptual speed tasks. The study also included four short-term/working memory tasks. In addition, two measures of general executive functioning were included. The central executive, an additional component of Baddeley and Hitch's (1974) model, is a general-purpose function, which is said to be responsible for controlling and regulating information processing. It is also associated with the control of goal directed behaviour.

The results demonstrated that executive functioning significantly predicted visualisation, spatial relations and perceptual speed. However, executive functioning was most strongly related to spatial visualisation, followed by spatial relations, and then perceptual speed. Visuospatial short-term/working memory, however, only uniquely predicted perceptual speed. The finding that visuospatial short-term/working memory predicted perceptual speed only, independently of executive function, is likely to have be due to the lack of transformation, or maintenance, involved in these tasks.

Thus, this analysis provides some cognitive support to Carroll's (1993) model, and also partly supports Linn and Petersen's (1985) model. That is, the results suggested that the observed psychometric dissociation between visualisation and spatial relations is partly due to the increased executive function demands with the former. The authors chose to place visuospatial working/short-term memory as latent predictors of the spatial thinking tasks. However, other authors define visuospatial memory tasks as measures of spatial thinking, in their own right (e.g., Mix et al., 2016; Burton & Fogarty, 2003). An alternative model might, therefore, place spatial memory tasks alongside the other observed spatial tasks, but, retain the central executive task as latent predictor.

1.2.4.2 Multiple object dynamic spatial ability

Dynamic spatial ability tasks measure the ability to reason about moving objects or stimuli (Hunt, Pellegrino, Frick, Farr, & Alderton, 1988). Larson (1996) notes that static paper-and-pencil tests correspond poorly to the real world, which is three-dimensional, dynamic and interactive. However, computerised testing has allowed greater emphasis to be placed on dynamic spatial ability as a construct (Hegarty & Waller, 2005).

Hunt et al., (1988) provided initial evidence for a separate spatial ability which is used to reason about moving objects. In particular, they proposed the construct of multiple object dynamic spatial ability. In Hunt et al. (1988), participants completed computer-administered tests of spatial ability involving static stimuli (i.e., traditional measures of spatial ability), as well as tests which involved dynamic stimuli. The dynamic, computer-administered tests measured skills such as being able to extrapolate movement trajectories and judge velocities. In one task, for example (an intercept judgment task), participants saw a target moving horizontally across a computer screen and were required to press a button to launch a second target, which moved vertically, timed at the correct moment, so that the two intercepted. The tasks, therefore, required participants to reason across space and time, and, also to spatially update the relationship between objects. An exploratory factor analysis (EFA) revealed that the computer-administered, 'static spatial ability' tasks (i.e., traditional spatial thinking measures) loaded onto separate factors, to the computer-administered, multiple object, dynamic spatial ability tests. The authors, therefore, concluded that spatial reasoning about moving stimuli is distinct from spatial reasoning about static stimuli. Contreras, Colom, Hernández and Santacreu (2003) and D'Oliveira (2004) subsequently conducted similar EFA and confirmatory factor analyses (CFA). Taken together, these two studies also suggested that arrival judgment and intercept judgment, dynamic spatial thinking tasks, are distinct from traditional paper-and-pencil tasks, i.e., static tasks.

An issue with the conclusions of Hunt et al., (1988), and also Contreras et al., (2003) and D'Oliveira (2004), is that their testing batteries did not include a task such as mental rotation, in a dynamic/moving format. Larson (1996), however, investigated whether spatial tasks involving dynamic stimuli, per se, were distinct from static stimuli, or if the distinct dynamic spatial factor related specifically to speed-space judgments. Larson administered mental rotation tasks in both a static (i.e., traditional format-2 objects, same mirror judgment) and dynamic format. In the dynamic format, each of the objects to be mentally rotated also moved around the computer screen, in their own circular path. The results of the study showed very high correlations between static and dynamic versions of the rotation task (.80 to .90). The authors concluded that a separate 'dynamic' spatial ability therefore relates to specialised space-speed-time linked tasks (e.g., involving arrival judgment), rather than spatial thinking performed on moving spatial stimuli, per se. This study, however, was limited somewhat by the use of zero-order correlation only, as opposed to either factor analysis or multiple regression, and, the lack of any other judgment-type arrival tasks, for comparison.

Finally, it should also be noted that these dynamic spatial measures (intercept and arrival judgment) also differ from many other traditional paper-and-pencil spatial ability tasks, in that they focused on the relationship between two or more objects, rather than on the spatial relations that make up a single object. Moreover, the comparison paper-and-pencil tasks were typically single objected focused tasks. Relating back to the Uttal et al. (2013) model, multiple object dynamic spatial ability tasks could fit within the extrinsic-dynamic category. Indeed, Newcombe and Shipley (2015) include the updating movement of objects through space, within the extrinsic-dynamic category. However, no research to date has investigated whether dynamic spatial ability measures might psychometrically group with other 'extrinsic' tasks, such as perspective taking or spatial scaling. This will form part of the evaluation of the Uttal et al. (2013) model. Multiple-object, dynamic spatial ability and its relation to other types of spatial ability will also be discussed further in Chapter 4.

1.2.5 Interim summary

As outlined, the Uttal et al. model is adopted and evaluated in the current thesis. Table 2 is a proposed summary of spatial thinking skills, drawing together and based on the prior typology of Atit et al., (2013), Uttal et al., (2013) and proposals by Atit et al., (2013). This summary also refers to dynamic spatial abilities. Visuospatial memory is not included separately in Uttal et al., (2013) typology. However, these working memory abilities could also fit into the 2x2 model (see Mix, Hambrick, Satyam, Burgoyne, & Levine, 2018). The placement of some of these tasks is currently theoretical (e.g., multiple-object dynamic spatial ability) because no psychometric data exist.

Table 2: Summary of spatial thinking skills based on Uttal et al. (2013) typology.

Uttal et al. (2013) 2 x 2 classification	Outcome category (From Uttal et al., 2013, unless specified)	Description	Examples of measures
Intrinsic and Static abilities	**Disembedding**	Perceiving objects, paths or spatial configurations amidst distracting background information	Embedded Figures Task, Flexibility of Closure,
Intrinsic and Dynamic abilities	**Rigid Intrinsic Transformation** Atit et al. (2013)	Transformations on objects in which spatial relations are preserved	Card rotation, flags, 3D cube rotation
	Non-rigid Intrinsic Transformations Atit et al. (2013)	Non-rigid/non-brittle: Transformations where spatial relations are not preserved and distances among points change continuously	Mental folding test, surface development test. Form board Block design
Extrinsic and static abilities	**Spatial Perception**	Understanding abstract spatial principles such as horizontal invariance or verticality	Water-level test
	Alignment (Newcombe & Shipley, 2015)	Reasoning about spatial and temporal correspondence	Spatial scaling
Extrinsic and dynamic abilities	**Perspective taking**	Visualising an environment from a different perspective	Guilford-Zimerman spatial orientation (1948) Hegarty & Waller (2004) spatial perspective taking test
	Dynamic multiple object spatial ability (Hunt et al., 1988)	Visualising the movement of an object relative to other objects; spatiotemporal thinking.	Dynamic spatial ability tests (e.g. Hunt et al, 1998)

1.2.6 The structure of spatial cognition in childhood

Recently, research has also begun to investigate the inter-relational structure of spatial cognition in childhood (Vander Heyden, Huizinga, Kan, & Jolles, 2016; Mix et al, 2018; Heil, 2018). Vander Heyden et al. (2016) investigated the factor structure of spatial thinking skills, developmentally, in a sample of 8-12 year olds, grouped into three age groups: 7.5-9 year olds, 9-10.5 year olds, 10.5-12 year olds. Children completed three intrinsic-dynamic spatial tasks, typically administered to adults (Vandenburg & Kuse Mental Rotation task [Peters et al., 1995]; Ekstrom paper folding task, [Ekstrom et al., 1976]; Wechsler Adult Intelligence Scale, Block Design task [Wechsler, 2003]). In addition, they completed two novel extrinsic spatial thinking tasks. One task required the rebuilding of a layout, following a change of perspective;

the other, involved navigation through a route, after a change of perspective. The results of the factor analysis revealed that a two-factor structure (i.e., intrinsic versus extrinsic transformation) was shown only between the ages of 10.5-12 years, whereas a one-factor structure fit the data better for the younger age groups. However, the study used tasks typically administered to adults. It is possible that a dissociation might have been revealed earlier in development, if more developmentally appropriate tasks were used.

In a related study, Heil (2018) also found that intrinsic-dynamic spatial thinking skills (i.e., object transformation) were dissociated from extrinsic-dynamic (i.e., perspective-taking) skills, with a sample of 10.5 year olds. More specifically, as with the study above, a CFA showed that a two-factor model fitted the data better than a one-factor model. The findings of Vander Hayden et al. (2016) and Heil (2018) are thus broadly in line with the previously outlined adult findings, from Hegarty and Waller (2004) (i.e., a dissociation between object transformation and perspective transformation)

As outlined above, prior research suggests that allocentric encoding of space is not fully mature until 10 years (Bullens et al., 2010). Thus, the finding of a lack of a significant dissociation between intrinsic-dynamic and extrinsic dynamic skills for younger children in Vander Hayden et al. (2016), may reflect either lack of allocentric skills, the tendency not to spontaneously use allocentric frameworks, or, a difficulty in shifting between egocentric and allocentric frameworks. Similarly, given a lack of allocentric skills, some of the younger children may have chosen to use an object-transformation strategy, rather than a perspective change strategy (i.e., move the object, not the self).

Only one study, with adults or children, includes a direct psychometric test of the whole Uttal et al. (2013) 2x2 model. Mix et al. (2018) used confirmatory factor analysis to test the 2x2 model, with children, aged 5-11. The findings of the study did not provide any support for the full 2x2 model (i.e., a four factor model), or, for a static/dynamic model. However, for the two younger age groups (6 year olds; 9 year olds), there was support for an intrinsic/extrinsic, two-factor model. For the 12 year olds, however, a one-factor model fit the data best. Yet, even the one-factor model did not fit the data well for the 10-11 year olds, and, mental rotation had a low factor loading on this single factor. This suggests that a 2-factor model may be needed for the older children, but, the two factor model that the authors tested was not suitable.

Moreover, testing of the full 2x2 (i.e., 4 factor) model was limited within this study, because the CFA included only two spatial measures, for each of the sub-domains. Thus, overall, there was support predominantly for the intrinsic-extrinsic spatial dimension, but only for younger children in the sample. This study is discussed further in Chapter 2.

To review, research to date is somewhat mixed regarding the structure of spatial cognition throughout development. When the analyses include only intrinsic-dynamic (object transformation) and extrinsic-dynamic (perspective taking) skills (i.e., Vander Heyden et al., 2016; Heil, 2018), a dissociation is evident for older children (aged approximately 10). The findings of Vander Heyden et al. (2016) also indicate that these skills become more dissociated throughout development. Yet, when a range of skills, covering the whole 2x2 model is included in the analysis, a dissociation is evident for younger, but not older children. More specifically, when the two-factor, intrinsic/extrinsic model includes both static and dynamic skills, no dissociation between intrinsic-extrinsic skills is evident for older children. The latter finding is somewhat surprising, given evidence that object-based transformations and perspective transformations are indeed dissociated in adults. It would seem more likely that spatial skills become more, and not less dissociated, as children develop. Further analysis of the structure of spatial thought in childhood is important particularly in relation to the Uttal et al. (2013) model, to attempt to address some of these inconsistencies.

1.2.7 The development of spatial skills

In addition to the possible psychometric dissociations reported above, there is also some evidence of differences in developmental trajectories for some spatial thinking skills. In particular, there is some evidence that mental rotation develops earlier within middle childhood than spatial perspective taking (Crescentini et al., 2014). This finding map onto the finding that egocentric encoding of space develops earlier than allocentric encoding of space (Bullens et al., 2010). However, beyond this, there has been little detailed research comparing the developmental trajectories of a range of spatial skills in middle childhood. Comparing developmental trajectories provides another source of evidence in evaluating the Uttal et al. (2013) model. Further detail on the developmental trajectories of spatial skills in middle childhood is given in Chapter 2.

1.2.8 Interim summary

To summarise, exploratory factor-analysis research throughout the 20th century, most thoroughly summarised by Carroll's (1993), began to delineate the various dimensions of spatial ability, yet, at times, suffered from a restricted focus. The heavy emphasis on object-based transformations, and on bottom-up exploratory factor analysis methods, led to a narrow definition of spatial ability in certain areas of research. These analyses neglect other skills such as multiple object spatial ability, spatial scaling, map use and perspective taking.

The model by Uttal et al. (2013) presents an appealing alternative approach, because it is theory-driven, addresses both object-focused skills, and spatial skills linked to spatial relations, and, has a neuroscientific basis. However, to date there has been little research that has directly evaluated the model. The intrinsic-dynamic (rotation) vs extrinsic-dynamic (perspective taking) distinction is given as strong evidence for the intrinsic-extrinsic dimension. However, while the evidence for this distinction is indeed convincing, the caveats outlined in section 1.2.2.2 justify further evaluation of the model. In particular, features of task design are important for perspective taking tasks. In addition, developmental factors may affect the intrinsic-extrinsic dissociation; however, the limited developmental data that does exist is contradictory. Evaluation of this model in the primary school years is therefore timely.

A further possible issue with the intrinsic-dynamic and extrinsic-dynamic categories is the extent to which they adequately describe apparently homogenous tasks. For example, both spatial perspective taking and arrival judgement could fit into the extrinsic-dynamic categories, and both mental folding and mental rotation are categorised as intrinsic-dynamic skills. However, there are key differences between these tasks, within their respective sub-domains. The thesis also provides an opportunity for a nuanced and fine-grained evaluation of these tasks, skills and categorisations.

1.2.9 Spatial thinking tasks used in the thesis

As outlined above, one of the goals of the current thesis is to further evaluate the Uttal et al. (2013) model in childhood, using several inter-related sources of evidence. One of these lines of evidence is the relative predictiveness of different spatial skills as predictors of science learning in the primary school years. In the current thesis, within Chapters 2, 3 and 4, individual tasks were therefore included to assess and evaluate each of Uttal et al.'s (2013) spatial sub-domains: intrinsic-static, intrinsic-dynamic, extrinsic-static and extrinsic-static domains. Detailed descriptions of the task procedures are given in the respective chapters, and a summary is provided below.

Considering the theorised 'intrinsic-static' domain, in Chapters 2 and 3, a version of the embedded figures task suitable for children (The Children's Embedded Figures Task; CEFT; Witkin, Otman, Raskin, & Karp, 1971) is used. This task is used because it requires children to locate a geometric shape, without additional transformation of those shapes. The adult version of the task is widely used to measure geometric form processing, and Uttal et al. (2013) propose this task as a measure of intrinsic-static thinking, within the description of the model. In Chapter 4, a visual discrimination task is used (Gardner, 1996). In this task, children view a 2D geometric shape, and then find another 2D geometric shape that matches exactly. Compared with the CEFT, the visual discrimination task has the advantage that it does not involve the additional 'disembedding' process from the distracting background, in addition to the static shape processing. The ability to 'disembed' a target from a complex background also draws on 'global-local' processing skills (e.g., Nilsson Jobs, Falck-Ytter, & Bölte, 2018).

Two types of 'intrinsic-dynamic' tasks are used in the thesis, reflecting the possible sub-division of this category within later conceptualisations of the Uttal et al. (2013) model. Both mental rotation and mental folding tasks were included to assess intrinsic-dynamic spatial thinking. Mental folding, whilst also involving intrinsic-dynamic spatial skills, differs from mental rotation in the precise skills that it taps into (Atit et al., 2013), meaning that the inclusion of both types of task is beneficial. A distinction between different types of object-based transformation skill is also reflected in other prior spatial typologies. Therefore, including both types of task permits comparison with these models. Considering 'rigid' intrinsic-dynamic skills, in Chapter 2 and 3, a 2D mental rotation task is used (Broadbent, Farran, and Tolmie, 2014).

Children are asked to determine which presented animal is the same as a target animal. The 'same' item is a rotated version of the target animal. In Chapter 4, another 2D mental rotation task is used, with more abstract stimuli (Thurstone, 1948). This task is used to address some of the potential limitations with the animal rotation task used in Chapters 2 and 3. To measure 'non-rigid' intrinsic-dynamic skills, a mental paper folding task (Harris et al., 2013) is used in Chapters 2, 3, 4 and 5. In this task, children are shown an image of a piece of paper, and an arrow indicating a fold. Children are asked to determine what the piece of paper would look like after being folded, by choosing from one of four response options. As will be outlined in section 1.4.6, non-rigid mental folding ability is of particular interest, given the complex and flexible nature of the transformation and the ways in which this may support reasoning about physical processes.

A spatial scaling task is used in Chapters 2, 3 and 4 to measure the theorised extrinsic-static domain. As outlined in section 1.2.2, spatial scaling can be considered an extrinsic-static spatial skill within the Uttal et al. (2013) model, due to the need to align and encode locations between two scaled spaces. A discrimination scaling task is used, whereby children are shown a larger printed map, and four smaller maps on a laptop computer. The smaller maps are scaled in size, relative to the larger printed map. Children are asked to choose which of the four smaller maps shows a target located in the same location as shown on larger printed map.

For the extrinsic-dynamic domain, a perspective taking task (Frick, Möhring & Newcombe, 2014b) is used in Chapter 2 and 3. Children are shown a character holding a camera, who is taking photographs of various combinations of objects and other characters. The character is sometimes shown to be taking photographs from a different spatial perspective and viewpoint than the child. Children are asked to determine what the photographs taken by the character the camera would look like. In Chapter 4, a multiple-object dynamic spatial ability task (see section 1.2.4.2) is used (Hunt et al, 1988). In this arrival judgement task, children are shown four objects in a race, on a laptop computer. The objects are shown at varying distances from a starting line, which travel at different speeds within the race. Children are asked to determine which of the four moving objects they predict would win the race. Including both types of skill allows closer for investigation of the extrinsic-dynamic domain. In particular, although they both might be categorised as extrinsic-dynamic, there are clear differences between them. For instance, the perspective taking task requires an updating of ego-centric perspectives, whereas the arrival judgment task has an additional speed-time processing element. In addition, using the arrival judgment as an alternative task in Chapter 5 in part addresses some of the possible limitations with using a perspective taking task. For instance, although this task is intended to measure extrinsic-dynamic skills, as outlined in section 1.2.2.2, it is possible to use a mental rotation (i.e., intrinsic) strategy to solve the task.

1.2.10 Spatial tools

In defining spatial cognition, Newcombe (2018) notes that, in addition to the intrinsic and extrinsic spatial thinking skills discussed above, humans also possess the ability to spatialise: to use symbolic, spatial tools. As outlined in the introduction, other prior spatial typologies have neglected spatial tools. It has been proposed that there is a continuum of spatial tools, from action to abstraction, with direct action at one end, gesture towards the middle, and language at the other end (Novack, Congdon, Hemani-Lopez, & Goldin-Meadow, 2014).

For example, visualisations such as *models and diagrams* represent spatial relational information, with varying degrees of abstraction. For instance, in the context of learning, a diagram might be highly realistic and isomorphic to an external referent. However, a diagram could also include specific symbols, or, use conventions specific to the discipline in question; this is often the case in science. Moreover, there are individual differences in diagram comprehension ability (Hegarty & Steinhoff, 1997). In the classroom, for example, some students fail to attend to elements of diagrams properly, may not understand the conventions within diagrams (e.g., use of arrows in science diagrams), or, may not integrate individual elements within a diagram, into a larger mental model (Bergey, Cromley, Kirchgessner, & Newcombe, 2015). In addition to comprehending diagrams, we are also able to produce diagrams of our own.

Spatial language is a symbolic spatial tool which verbally describes, and represents, spatial relationships (e.g., between, above) and object properties (e.g., thin, bendy). There are individual differences in skill at both production and comprehension of spatial language terms between the age of 5 to 10 (Pruden, Levine, & Huttenlocher, 2011). Moreover, there is also variation in the use of parents' spontaneous use of spatial language with children, and, in children's spontaneous spatial language use (Pruden, Levine, & Huttenlocher, 2011).

Finally, *gestures* are hand movements which occur in three-dimensional space and dynamically represent spatial-relational, spatio-motoric, and abstract, information (McNeill, 1992). Therefore, like spatial language, gestures can represent spatial relationships, such as above and below. However, the gestures which represent above and below are clearly more concrete than the corresponding, spatial language terms, in terms of the spatial similarity to the represented concept. Gestures are also a type of action, and prominent theories of gesture see gestures and actions as emerging a common action generator (e.g., Hostetter & Alibali, 2008). Indeed, many gestures are very close to the actions they represent; for instance, producing a gesture of pouring water out of a glass. However, because gestures are only ever representational, they are more abstract than actions. Therefore, gestures both allow individuals to have spatial and motor representational experiences related to a concept, but, are also somewhat removed from action itself.

In contrast to prior spatial typologies, and in line with Newcombe (2018), the definition of spatial cognition within this thesis therefore also expands to include the use of symbolic spatial tools, and, in particular, the use of gesture in the context of science instruction. In addition, there is also a focus on the connection between spatial thinking skills and spatial tools, and how spatial thinking skills such as mental folding ability may be more important for using particular spatial tools (e.g., diagrams) than others (e.g., concrete models). Gesture as a spatial tool, theories of gesture, and the relation between gesture, cognition and learning, will be considered in more detail, in Chapter 5.

1.2.11 Embodied cognition

As well as both falling within the theoretical scope of spatial cognition, gesture and spatial thinking skills, such as mental rotation, can both be theoretically linked within the embodied cognition approach. Theories of embodied cognition (Gallese & Lakoff, 2005; Barsalou, 1999; Damasio, 1989; Meyer & Damasio, 2009) challenge the view of the mind as an abstract information processor. Within an embodied cognition approach, psychological processes are proposed to be influenced and affected by our bodies: sensory systems in different modalities, motor processes, and emotions (Glenberg, 2010).

Considering gesture, one specific idea stemming from an embodied approach to cognition is that our actions affect the representations we have about objects and processes. It is theorised that this occurs through the connections representations have with the sensorimotor experiences we have had, or are having, with the object itself; that is, actions affect thought. Barsalou's perceptual symbols systems approach (Barsalou, 1999), for example, proposes that the current representations we hold of events and objects are rooted in the neural states that were active when the objects or events were encountered and perceived (Goldin-Meadow & Beilock, 2010). Representations are therefore said to be multimodal traces, based on our prior experiences with objects and events. This theory is supported by the finding that individuals who have had prior experiences with an object, or event or type of action, have different representations or show different brain patterns, than those who have not had these experiences (Goldin-Meadow & Beilock, 2010; Calvo-Merino, Glaser, Grèzes, Passingham, & Haggard, 2004).

As stated in the previous section, gestures are a unique kind of action, rooted in an action-based system, which have representational qualities, but do not directly affect the world in the same way that a physical action on an object does (Goldin-Meadow & Beilock, 2010; Trofatter, Kontra, Beilock, & Goldin-Meadow, 2015). Gestures affect thought and conceptual representations (see Goldin-Meadow & Beilock, 2010; Trofatter, Kontra, Beilock, & Goldin-Meadow, 2015; section 5.1.3). Thus, from an embodied cognition perspective, gesture, as a special form of action, affects thought.

Considering now spatial thinking skills, one theoretical approach suggests that visualisation-based skills such as mental rotation occur in a depictive fashion, through a 'mental simulation' of action (Kosslyn & Pomerantz, 1977; Kosslyn, 1994). That is, when asked to rotate a shape, a participant does so in an analog fashion, by 'imagining' the rotation of the shape as they would do with a real object, but in the 'mind's eye'. However, a mental simulation of such a process is not necessarily visual in nature, i.e., as though a camera had recorded a video of the process (Rapp, 2005). Instead, aspects of the model may be imagistic in nature (Kosslyn, 1994), meaning that visual characteristics and spatial relationships are maintained. From this perspective, they are not exact replicas of the external process or system; rather, they simplify or schematise the process, whilst preserving important information (Rapp, 2005).

Relatedly, another embodied cognition framework, the convergence-divergence zone framework (Damasio, 1989; Meyer & Damasio, 2009) specifically proposes that mental images, such as the ones activated during mental rotation, occur through activation of the early sensory cortices, both during actual perception of the object, and also, during simulated recall of the object. Thus, a mental simulation is a recreation of the neural states which were involved in physically performing or witnessing an action or object, in the absence of the physical action or object. In support of this, motor, somatosensory and sensorimotor areas are activated during mental rotation tasks (Parsons et al., 1995).

In addition, linking gesture and mental transformations: having more experience with rotating real objects and gesturing about rotating objects in a training paradigm leads to stronger mental rotation skills in young children (Levine, Goldin-Meadow, Carlson, & Hemani-Lopez, 2018). Having more *physical experience* of an action therefore improves the subsequent *mental simulation* of that action. Thus, gesture as a type of action produced by the body impacts upon thought processes and representations, including mental simulations of physical actions, such as mental rotation. Similarly, in the context of science learning, producing gestures may therefore enhance the mental simulation of physical processes, therefore, supporting reasoning about that process. This will be addressed in the study in Chapter 5.

In summary, an embodied cognition approach suggests that our conceptual representations and thought processes, are affected by and link with our bodies. Intrinsic-dynamic spatial thinking skills such as mental rotation and mental folding may involve an imagistic simulation of action, with associated motor and sensory activation. In addition, this activation, particularly the motor activation, is rooted in prior experiences of these processes. That is, the ability to visualise and simulate a mental fold stems from experiences of folding in the 'real world'. Gestures are a type of action, and are therefore fundamentally linked to the body, and gestures impact upon thought.

1.3 Overview of science

As outlined, the primary goal of the thesis was to investigate the relationship between spatial thinking skills and science learning, in the primary school years. We now therefore turn to approaches to defining the knowledge and skills involved in learning and doing science.

In its most general sense, the goal of science is to extend our knowledge of the world. As such, ‘science’ refers to both the existing body of knowledge we have about the world, and, the activities and processes by which this knowledge comes about (Zimmerman, 2000). Engaging in science therefore partly involves understanding and using the factual knowledge and conceptual understanding we have, relating to the phenomena around us. In addition to this, it involves specific reasoning, strategies and skills which are directed towards the discovery of, and changes to, the theories we have about the world (Zimmerman, 2000). In the latter case, this would include, for example, being able to manipulate experimental variables, generate and test hypotheses, and analyse or interpret data.

Wellington (1988) similarly proposed a trio of types of knowledge central to science: ‘knowledge that’ (facts, happenings and phenomena), ‘knowledge why’ (explanations, models, frameworks, theories) and ‘knowledge of how to’ (skills, processes and abilities). The former two categories of knowledge map broadly onto factual and conceptual understanding, respectively, whereas the last maps more onto procedural knowledge/skills and scientific investigation skills.

The above division is also apparent in the UK primary school science curriculum (Department for Education, 2013a). Indeed, the two broad curriculum aims relate to ‘scientific knowledge and conceptual understanding’ (mapping onto the first two of Wellington’s domains), and, ‘the nature, processes and methods of science’ (mapping onto the third area). Within the curriculum, scientific knowledge and conceptual understanding are grouped around the disciplines of biology, chemistry and physics, and further, into sub-topics, within these domains.

Although a useful heuristic, the distinction between ‘procedural/investigative’ and ‘knowledge’-related science domains is, in some ways, overly simplistic. The artificial distinction between process and knowledge-based science skills is reflected in a reference within the national curriculum. In particular, it emphasises that ‘working scientifically…must always be taught through and clearly related to substantive science content in the programme of study' (Department for Education, 2013a, page 5). That is, there should be an integration of scientific investigative skills with conceptual understanding. In addition, in 'real world' science, the two areas are rarely separate. A biologist conducting an experiment uses investigative skills, but these skills also draw upon existing conceptual understanding of the area (for example, to enable theory-based hypotheses to be made). In addition, their conceptual understanding of the subject area may be developed as a result of the experiment. Ideally, this development would also contribute to the broader knowledge base of the discipline.

Tolmie et al., (2016) further argue that 'conceptual' skills and 'investigative' skills are themselves not singular or unitary constructs, and instead propose three 'core' science skills, which encompass both areas. First, *accurate observation,* which includes an awareness of covariation and the explicit description of observations (e.g., when the drum is hit with a small force, it makes a quiet sound; when the drum is hit with a large force, it makes a loud sound). Second, the ability to *extract and reason about causal connections*, involves drawing conclusions from data, or verbally explaining trends in available data (e.g. the harder the drum is struck, the louder the sound). Finally, the ability to provide *explanations of the mechanisms* that explain inferred connections (e.g., when a drum is hit with a greater force, a louder sound is produced, because…). The last skill maps onto Wellington's (1998) 'know why' domain, and relates to understanding and applying conceptual models and theories.

1.3.1 Approaches to assessing science knowledge, skills and understanding within the thesis.

In the current thesis, several different but complementary approaches were adopted to assess knowledge and skills in science. First, the study reported in Chapter 3 adopts a curriculum-based approach to assessing science achievement. In particular, children completed questions designed to assess the achievement of science curriculum objectives, appropriate to the age range. As outlined above, the UK science curriculum includes both conceptual knowledge and understanding, and investigative skills. Other than the reasons given above, an advantage of adopting a curriculum-based approach to defining and assessing science is that it broadly reflects the activities and knowledge children actually engage in and with, in school. Therefore, adopting this approach to science provides insights into the extent to which spatial thinking skills are predictive of the broad outcomes of science learning as aligned with the realities of the classroom and science curriculum.

As will become clear in the subsequent literature review, a core hypothesis of the thesis is that spatial thinking specifically relates to conceptual understanding of conceptual science content. For example, spatial thinking may be particularly important for children's understanding of the solar system. In order to investigate this hypothesis in more depth, the approach to science assessment in Chapters 3 and 4 therefore focus specifically on conceptual knowledge and understanding of scientific processes.

The study reported in Chapter 4 focuses on a single conceptual physics topic: sound. The advantage of focusing on a single topic is that it enables specific hypotheses to be made about the relationship between spatial thinking and science understanding, and for in-depth analyses to be conducted. One of the goals of Chapter 4 is to provide evidence for the hypothesis that spatial thinking specifically relates to *conceptual understanding* of science content, rather than *factual recall or retention* of content (i.e., the contrast between Wellington's [1998] 'know that' and 'know why' domains). This study therefore includes science assessment measures which assess both of these broad domains. Retention involves retrieving remembered information from long-term memory, in the same or a similar format in which it was taught. The study reported in Chapter 4 therefore included a measure of scientific factual knowledge to measure retention. To measure conceptual understanding, the study includes a measure of transfer, in the form of problem-solving questions. Transfer involves the ability to use what has been learnt, and to then apply it to new problems or to answer new questions, in different contexts (Barnett & Ceci, 2002; Mayer, 2002a). The inclusion of both of these types of measure means it is possible to compare the extent to which spatial thinking predicts both recall of scientific knowledge and conceptual understanding.

In addition, within each problem-solving question (i.e., each question assessing conceptual understanding), children were asked to make a *prediction* and then to provide an *explanation* of their prediction, using their conceptual understanding. Prediction and explanation are two important science skills (Tolmie et al., 2016), and are also two specific cognitive processes associated generally with the transfer of learning and conceptual understanding (Mayer, 2002a), although explanations often require more conceptual understanding than predictions (Mayer, 2002a). Including prediction and explanations as a further sub-division of transfer allows a comparison to be made between the predictive power of spatial thinking within these skills.

The study in Chapter 5 also focuses on a single conceptual physics sub-domain: magnetism. Because spatial thinking was not found to be predictive of factual knowledge recall in Chapter 4, the study in Chapter 5 focuses only on conceptual understanding. However, the science assessment within Chapter 5 divides conceptual understanding along different dimensions to Chapter 4. The assessment separates transfer into *near transfer* and *intermediate transfer*. *Near transfer*, refers to the application of conceptual understanding to very similar contexts than taught originally. *Intermediate* transfer refers to the application of conceptual understanding to a different context, but still within the same domain to that which was taught originally (Barnett & Ceci, 2002). Within the study reported in Chapter 5, the near transfer questions assess the understanding of the taught content (bar magnets). The intermediate transfer questions assess understanding of untaught content (e.g. disc magnets). Intermediate transfer requires the learner to apply their conceptual understanding to a more abstract context. Including both types of transfer in this study meant that it was possible to determine whether gesture-based instruction, which is more abstract, was particularly useful for intermediate-level transfer compared with near transfer.

Given that the latter half of the thesis focuses primarily on conceptual understanding in science in relation to spatial thinking skills and spatial tools, the following sections go into more detail on theories of the nature of conceptual understanding, and the source of conceptual explanations and models.

1.4 Mental models and the nature of conceptual understanding in science

One of the major goals of science education is to gain a conceptual grasp or understanding of the world around us. For example, how does the heart work? Why does the sun appear to move across the sky? To answer questions such as these, the learner must build and use some form of mental, conceptual, representation, often termed ‘mental models’ (Mayer, 2002b). There is consensus in the literature that the structure of this type of knowledge is more likely to be domain-specific (e.g., physics, biology), than domain-general (e.g., related to one of Piaget’s stages of development; Mayer, 2002). Yet, what is not clear, or agreed upon theoretically, is the structure of these mental models. Some theories of conceptual understanding see mental models as being structured, coherent and organised (Johnson-Laird, 1980; Gentner & Gentner, 1983; Vosniadou, 1994). Other theories, most notably DiSessa (1988), 'knowledge in pieces', see representations as being fragmented and lacking a strong organisational structure. The following sections discuss each of these theoretical approaches in turn.

1.4.1 Mental models: Johnson-Laird (1980)

Prior to contemporary mental model theories, in the late 19th century Boltzmann made the general proposal that 'all our ideas and concepts are only internal pictures' (Johnson-Laird, Girotto, & Legrenzi, 1998). In addition, Craik (1943) made subsequent reference to individuals carrying an internal, 'small-scale model' of external reality in the mind, which allows individuals to generate and test predictions, and also, to process future events, based on prior knowledge.

Building on these ideas, Johnson-Laird, Girotto and Legrenzi (1998) later characterised mental models as psychological representations of people, entities, processes and complex systems. Furthermore, representations can be based on real events or entities, or alternatively, they might be hypothetical or imaginary (Johnson-Laird, Girotto & Legrenzi, 1998; Johnson-Laird & Khemlani, 2017). A further central aspect of Johnson-Laird's notion of mental models is that the individual parts of a mental model, and the relations between their parts, correspond to the external reality being represented: iconicity (Johnson-Laird, 1980; Johnson-Laird & Khemlani, 2017). In this way, mental models are therefore analogous to the physical models used by a scientist to represent some physical reality (Rapp, 2005). Within this approach, mental models are therefore based on truth – that is, a representation of what is possible, given a premise (Johnson-Laird & Khemlani, 2017).

A study by Ehrlich, Mani and Johnson-Laird (1979) provides one example of supportive evidence for mental models. Participants in one condition heard three sentences about the spatial relations between objects: 'The knife is in front of the spoon. The spoon is on the left of the glass. The glass is behind the dish.' In a second condition, participants read the same sentences, in a non-successive order. Participants were then asked to draw the layout of the objects as they remembered, based on the sentences. The results of the study showed that participants who read the ordered sentences were more likely to draw the correct layout of the objects. The authors proposed that, in the ordered example, as participants heard each of the sentences, they constructed a single mental model, representing the entire layout, because the model could be successively built (i.e., knife, spoon, spoon, glass, glass, dish). In the non-ordered example, participants constructed at least two models, rather than one. Having to encode, maintain and recall two models was more challenging than one, leading to poorer layout recall.

Although the general notion of a mental model is useful, this specific approach is less relevant to the development of models of explanatory frameworks. Johnson-Laird's (1980) approach focuses more on mental models as being temporary and transient. Moreover, research within this approach which has provided support for mental models has typically focused on the construction of mental models in response to propositional inference and logical reasoning tasks (e.g., García-Madruga, Moreno, Carriedo, Gutiérrez, & Johnson-Laird, 2001; Cherubini & Johnson-Laird, 2004). However, there is still a place for this short-term mental model construction for scientific reasoning, particularly when considering the more spatial relational and causal representations. For instance, when addressing problems and scenarios, learners may construct short-term models to represent aspects of the problem in hand. Moreover, in line with Gentner & Stevens (1983), discussed below, Johnson-Laird acknowledges that the ability to predict and understand physical phenomena depends on the construction and use of a mental model.

1.4.2 Mental models: Gentner & Stevens (1983)

Whereas Johnson-Laird's mental models are more transient, and focus on reasoning within tasks and situations, Gentner and Steven's (1983) notion of mental models is that they are based on understanding and knowledge acquisition. Gentner (2001) suggests that we develop simplified, causal mental models of physical systems, based on qualitative relations, which allow predictions and inferences to be made about that system. In addition, Gentner (2001) relatedly argues that mental models allow individuals to run mental simulations of the physical system or process. An individual is also said to hold one of several different models for a physical phenomenon or dynamic system. The specific model that a person holds affects their reasoning about that phenomenon (Gentner & Stevens, 1983). For instance, there are two dominant models people hold for understanding a thermostat, and the model they possess impacts upon their reasoning about that system (Gentner & Stevens, 1983).

Gentner and Gentner (1983) more specifically provide evidence that mental models about physical systems are also often based on analogies: for example ‘the hydrogen atom is like the solar system’. The background to this approach to mental models is the structure-mapping theory of analogy (Gentner, 1983). This approach suggests that an analogy maps the structural relations between two domains. The *base domain* is the familiar domain (in the example above, the solar system), whereas the novel domain (i.e., the new domain being learnt) is the *target domain (*in the example above, the hydrogen atom). The basis of the analogical mapping is that the learner has a strong understanding of the base domain, and in particular, the structural relations among objects in the base domain (i.e., sun, planet). Central to the understanding of the base domain in the example given are relational predicates such as ‘attracts’ and ‘revolves around’. When learning about the new, target domain, the relational predicates from the known base domain are mapped onto the new entities (i.e., nucleus, electron) in the target domain.

As a further example, analogies are often used to aid the understanding of electricity (Gentner and Gentner, 1983). The *flowing water analogy* represents water flowing through wires in a circuit in a manner analogous to water flowing through a pipe. Within the analogy, a resistor is a narrow pipe. The *moving crowd analogy* portrays electricity as the rate of movement of people through a room, while a resistor is a door to the next room. Both models, however, have drawbacks in their ability to map particular aspects of the concept of electricity. The flowing water analogy easily maps voltage: the voltage of the circuit corresponds to the number of pumps that are pumping water through the circuit (Gentner & Gentner, 1983). Yet, voltage is less easily mapped in the crowd model. Furthermore, gates in the crowd model easily explain and map onto resistors, but map less well onto the constricted pipes in the flowing water analogy.

Gentner and Gentner (1983) provide evidence for mental models of electricity. The authors propose that if analogies are part of, or linked to people’s causal models of electricity, then reasoning about electricity should be linked to the particular analogical model someone uses. As predicted, participants with the flowing water analogy-based model were more accurate when reasoning about combinations of batteries, whereas participants with the moving crowd/objects model reasoned more accurately about combinations of resistors.

Thus, the existing analogical models individuals hold are in some ways supportive of conceptual understanding. Providing learners with analogies for scientific concepts, in some cases, improves conceptual understanding (Jaeger, Taylor, & Wiley, 2016; Donnelly & McDaniel, 1993; Bean, Searles, Singer, & Cowen, 1990), although not universally (e.g., McDaniel & Donnelly, 1996). In the case of analogies linked to everyday experiences, because they are highly familiar, it is perhaps easier to simulate these experiences and generate inferences. In addition, the idea that models are based upon and linked to specific experiences implies that there is a somewhat imagistic, spatial or motor element, and links to theories of embodied cognition (see section 1.2.11). Yet, as is shown above, there are limitations to these models in terms of their explanatory power. Moreover, in some cases (as will be described in Chapter 4 in the context of sound) building models based on analogies from everyday experience can lead to erroneous conceptions. Importantly, overall, Gentner's view is that we hold a single, dominant mental model, to explain a system or process.

1.4.3 Framework theory: Vosniadou (1994)

Vosniadou (1994) also proposes that learners seek to construct a coherent explanatory framework, or mental model, to explain a process or a system. Learners build mental models by integrating information from science lessons and their own experiences. This approach further proposes that knowledge is organised and embedded in larger theoretical structures. Vosniadou distinguishes between the *naïve framework theory of physics* and *specific theories.*

The *naive framework theory* is based on entrenched presuppositions or constraints, unavailable to conscious awareness (Vosniadou, 1994). These constraints are domain-specific, most prominently in the domain of physics. For instance, Spelke (1991) proposed that infants are aware of five physical constraints early on in life: continuity, solidity, no action at a distance, gravity and inertia. This framework theory of physics, present early on in life, is said to constrain children's subsequent understanding of the physical world. *Specific theories* include a set of beliefs which describe the behaviour of physical objects or processes. Beliefs are built from everyday experience and observations, and also from socially transmitted knowledge (e.g., teaching), to explain phenomena. These beliefs are also influenced, and constrained by, the presuppositions from the naïve framework theory. A single, coherent mental model of the physical process is constructed from this set of beliefs.

Vosniadou and Brewer (1992) provide the example of children's understanding of the spherical earth. Children categorise the earth as a physical object, and thus apply to it some of the physical constraints presuppositions of objects. In particular, Vosinadou argues that two major presuppositions are important. First, that space is organised in terms of up and down, in relation to a flat ground. Second, that objects fall in a downward direction. The study reports that children demonstrate one of five mental models of the earth, constrained by these presuppositions. An example of children's initial, alternative, mental model of the earth, in the study, is of the earth represented as being a flat, stable object, with the sky above.

Eventually, children acquire the belief that the earth is spherical, through socially-transmitted knowledge. However, this introduced belief is difficult to integrate into the model, because it contradicts the presuppositions of the framework theory. One way in which a model can be changed is to construct a synthetic model which integrates the new belief within the constraints of the framework theory. A synthetic model children demonstrated is of a spherical earth in which humans live on a flat ground, within the sphere. This model meets both the presuppositions of the framework theory, and also the new socially-transmitted spherical belief. The authors propose that children can only fully understand the earth as being spherical if the physical presuppositions, which constrain the model, are challenged.

Another example was reported by Ioannides and Vosniadou (2002), who investigated force understanding, in a sample of 5-14 year olds. The results of this study showed that 88.6% of the participant's responses fell into one of seven consistent models of force (internal force, internal force affected by movement, internal and acquired push, acquire force and force of push/pull, force of push/pull, and gravitational and other forces). Moreover, the predictions and explanations that participants made consistently aligned with the model, regardless of the context.

To summarise, Vosniadou states that children hold a set of related beliefs about physical processes which are closely tied to physical presuppositions. As a result, at any one time, children possess a single, coherent theory or mental model, of a physical process. Thus, the predictions and explanations children provide for a phenomenon should be consistent, regardless of the situation or problem they are given.

1.4.4 Knowledge in pieces: diSessa (1988)

diSessa's view on concept formation and development contrasts with the previous accounts. diSessa argues that knowledge of physical phenomena is not organised in a stable, coherent, systematic way (i.e., in a single theory), but, that knowledge is 'in pieces'. That is, learners possess many fragments of knowledge which are loosely organised into a system of knowledge. These pieces of knowledge are known as p-prims (phenomenological primitives). P-prims are simple, self-explanatory, shallow interpretations of physical processes, which are constructed from every day, common experiences. Within this theory, conceptual change occurs through the reorganisation and integration of p-prims, rather than through replacement or deletion of models. diSessa further asserts that given the fragmented nature of knowledge, the explanations people provide for the same phenomena should vary depending on the context in which they are given. Note the contrast to previous approaches, which suggest that we hold a single coherent model, which is applied similarly in different contexts.

Each physical phenomenon may involve the organization and combination of different p-prims, and each p-prim is involved in the explanation of more than one phenomena. diSessa (2014) provides the example of the explanations individuals provide to explain a vacuum cleaner motor speeding up when the end of the tube is covered. Many people believe that the motor speeds up to 'overcome' the resistance which is generated by the end being covered. In reality, the motor speeds up because there is less effort needed, due to the decreased air resistance when the end is covered. For instance, in this example, there may be the application of the 'overcoming' p-prim, which is based on the notion of a greater force overcoming the weaker force. The 'overcoming' p-prim is said to originate from prior experiences people have had in their everyday lives, when reacting to physical resistance (i.e., a tug of war). Although p-prims in this instance are unhelpful (i.e., they lead to a non-normative explanation), diSessa asserts that they can be used in a positive way. For example, the ohm's law p-prim (increased effort leads to increased results) supports student's understanding of thermal equilibration (diSessa, 2014).

DiSessa (1988) provides supportive evidence for the fragmented nature of knowledge, relating to circular motion. From a 'knowledge as theory' perspective, circular motion is explained by a branch of impetus theory: circular impetus. The authors used variations of Piaget's sling problem with younger and older children, and physics naïve adults. In the basic version of the sling problem, a participant is asked 'what happens to a ball twirling around on the end of a string, if the string is cut?'. For example, one of the variations of the problem, the ball was pushed so that instead moved around a circular tube. diSessa's argues that if people hold a single, coherent, impetus theory, then their answers to the problems should be consistent, regardless of the changes between them. However, participants in fact gave very different predictions and explanations across the different variations of the problem. In fact, none of the participants maintained a circular impetus theory throughout.

Another study (diSessa, Gillespie, & Esterly, 2004) addressed the previously described research by Ionnides and Vosiandou (2002), relating to forces. The study included a quasi-replication of Ionnides and Vosniadou (2002), and an extension. The study failed to replicate the prior findings. In particular, they reported that less than 50% could be classified into one model of force only (compared to the 88.6% in Ioannides and Vosniadou (2002). Moreover, an additional extension part of the study showed that conceptual understanding of force was more context-dependent than had been shown by Ionnides and Vosniadou (2002). For example, even the colour of an object affected the explanations of force given by the youngest children.

diSessa does, however, also acknowledge the existence of mental model type structures (diSessa, 1996; diSessa, 2002). He describes a mental model as a typically instructed, compound knowledge form, which allow for extended explanations, problem-solving and hypothetical reasoning (diSessa, 1996). However, diSessa argues that mental models typically involve a small number of key causal principles, which may, in some cases, be p-prims. He also suggests that mental models depend on an underlying substrate system, such as spatial reasoning (diSessa, 2002). For example, a mental model of gear movement could involve a small set of p-prims (e.g., contact conveys motion), along with spatial reasoning about gears diSessa (2002). Moreover, mental models are dependent on p-prims, and during instruction, the development of mental models occurs in interaction with p-prims (diSessa, 1996). The implication is that models are still less centrally coherent and organised than the other theories suggest.

1.4.5 Mental simulation and mental animation within mental models

Several of the theories above (e.g., Gentner & Stevens, 1983) refer to mental simulation and an imagistic, dynamic, element to mental models (e.g., spatial reasoning about gears). Such an approach links to the theories of embodied cognition described in section 1.2.11 (Gallese & Lakoff, 2005; Barsalou, 1999; Damasio, 1989; Meyer & Damasio, 2009).

Evidence for an analog, simulation-type process to scientific reasoning and mental models exists in the domain of mechanics. For example, research has investigated the process of 'mental animation', whereby individuals infer the kinematics of a mechanical system based on a static representation of the system. Hegarty (1992), for example, used reaction time and eye-fixation data to determine how inferences are made from static diagrams of a pulley system. An important finding of the study was that participants performed better when they made inferences in the direction of causality (i.e., from the beginning of the chain of causal events), than when they made inferences backwards, from the result of an event. Hegarty concluded that the mental animation process occurred in a piecemeal fashion. That is, the units of the system were decomposed into smaller elements, and participants 'animated' each component in the causal sequence. Thus, she suggests that it was not a full 'animation' of the entire causal process.

Schwartz and Black (1996) also investigated analog simulation as a strategy for inferring mechanical behaviour. They were asked to determine if the groove and knob on two gears, presented as static diagrams, would meet, if the gears were rotated. A linear relationship between response time and angular disparity was found, which the authors suggest point towards the use of analog imagery to solve the task. Hegarty (2004) proposes that inferences such as these involve *spatial* simulations (i.e., piecemeal, sketchy, schematic), and not *visual* simulations (detailed, holistic).

Finally, some studies have shown that directing people to visualise, simulate or imagine, leads to greater learning or more accurate reasoning. For example, in the context of a biological process, Leopold and Mayer (2015) found that asking participants to visualise a process (e.g., visualise the steps in the nervous system when the brain sends a signal to the diaphragm and rib muscles) improved participants' explanations of the process (i.e., the accuracy of their mental model). Moreover, other research, with pre-schoolers, has investigated the role of directed visualisation in relation to a gravity bias, typically observed at this age. Pre-schoolers typically erroneously predict that an unsupported object will fall straight down, even if the path is constrained by a curved tube. However, asking children to imagine the ball rolling down the tube, before making their prediction, leads to increased prediction accuracy, relative to control conditions (Palmquist, Keen, & Jaswal, 2017; Joh, Jaswal, & Keen, 2011).

1.4.6 Mental models summary

To summarise, in order to understand and explain a scientific process, the learner must hold some form of mental representation of that process. The theoretical approach of this thesis is that, through everyday experiences and science instruction, learners generate internal mental models of scientific processes. Based on the theories and literature above, mental models are thus defined within the current thesis as an internal representation of a scientific process or system which allow predictions, inferences and explanations to be generated.

Mental models are viewed as being loosely isomorphic to the external referent. That is, they preserve or reflect the spatial-relational or schematic nature of the process or system in question. Furthermore, mental simulation type processes may be involved when developing and reasoning within these models, and, some of these simulations link to, or are based upon, in an analogical sense, to everyday experiences. However, this simulation is likely to be piecemeal and involve individual interactions between small numbers of elements. Mental models may be stored as long-term explanatory structures, or, may be more transient, in response to on-line demands of problems or questions.

Moreover, as will be discussed further in the thesis, spatial thinking, particularly intrinsic-dynamic, spatial visualisation skills, are theorised to play a supportive role in the development of, and reasoning within, these models. In particular, non-rigid transformation skills such as mental folding ability, may be particularly useful for mental simulations within these models. In a non-rigid transformation the distance among the points change (Atit et al., 2013). For instance, visualising the bonnet of a car before and after a collision would involve a non-rigid transformation, because the distance among the points change as the metal of the bonnet bend (Atit et al, 2013). Thus, non-rigid transformations enable complex and flexible spatial visualization. Non-rigid transformations may be useful when reasoning and mentally simulating about physical processes within mental models of scientific concepts. For example, a non-rigid transformation ability such as mental folding may be particularly useful when visualising the ‘flowing water’ analogy of electricity, within the models outlined in section 1.4.2.

However, this definition does not necessarily present mental models as being completely structured and theory-like, and, it does not preclude the possibility that the expression of some aspects of the model may vary somewhat by context (e.g., depending on the given scenario). An accurate or complete mental model of a process involves several interconnected elements. However, this set of elements may be coherently organised around key beliefs or physical presuppositions, or, may represent a collection of loosely connected, piecemeal, fragments of knowledge. Prior evidence, to date, is inconclusive on this matter. Finally, part of the mental model may include the record of fragmented reactivation needed to simulate prior experiences with the concept. This definition will be particularly relevant in Chapter 4.

1.5 Approaches to science instruction in the thesis

As well as adopting several complementary approaches to the assessment of science knowledge, understanding and skills, the studies in the current thesis include a number of different approaches to science instruction.

One distinction within the thesis is between whole-class science instruction versus individual, or one-to-one science instruction. Chapter 4 takes a whole-class instructional approach, and focuses on how spatial thinking skills relate to the learning outcomes of a typical primary school science lesson. The whole-class approach is typically used by teachers in schools, and therefore has the advantage of ecological validity and practical value. There has also been a limited amount of research which has adopted this instructional approach in relation to cognitive skills and abilities to date. Chapter 5, in contrast, focuses on individual or one-to-one instruction in the context of learning about magnetism. This approach is more suited to the two studies in this chapter, which take a more fine-grained, experimental approach to comparing instructional approaches based on different spatial tools (see below). Therefore, the one-to-one context provides an amount of control and precision within an experimental paradigm.

A second distinction is between the types of spatial tools used within the instructional methods. As stated above, Chapter 5 focuses on gesture as a spatial tool for supporting children's learning about magnetism, and how this approach compares to other spatial tools: concrete models and diagrams. As stated in section 12.2.10, both gestures and actions performed with objects are types of embodied action. Including the diagram-based instruction method therefore allows these embodied approaches to be compared to a non-embodied, control condition. However, as discussed previously, gestures are a special type of action, which differ from actions on objects in terms of their representational qualities. Therefore, including concrete models as an alternative instructional approach to gesture also means these two methods can be compared. If gesture does indeed have special qualities that distinguish it from actions with objects, then a difference in magnetism learning outcomes between the approaches should be evident. In addition, by including these three contrasting instructional approaches, it is possible to also compare the extent to which spatial thinking skills (i.e. mental folding ability) differentially predict learning with different spatial tools.

1.6 Existing research on the relationship between spatial ability and science, technology, engineering and mathematics (STEM)

The following section presents the current evidence base on the relationship between spatial thinking and science learning. This will provide the background to the main goal of the thesis, to investigate the relationship between spatial thinking skills and science in the primary school years. This section begins with an overview of the longitudinal evidence connecting spatial ability to STEM broadly. The main sections that follow present a summary of existing research relating spatial ability specifically to science learning and achievement. Evidence is discussed in relation to key scientific disciplines (chemistry, biology, physics). However, there is a crossover in domain-general skills that may be relevant to each of these different disciplines, which will be addressed in the summary of mechanisms. The section on physics and engineering also includes studies that have specifically investigated the role of spatial ability in mental model construction, linking to the previous section. Finally, the spatial ability and science review begins with research focused on adults and adolescents, and then, addresses research studies which are based on samples of younger children.

This age-based distinction is important for several reasons. For instance, correlations between spatial ability and science may be stronger for students learning science at an earlier stage, than those at more advanced stages. For example, Hambrick et al., (2012) found that spatial ability was more predictive of performance on a geologic structure task at lower levels of geologic knowledge. However, for those with more domain-specific knowledge, spatial skills were less important. Thus, as students progress through to advanced study, domain-specific strategies and conceptual knowledge may be more important, than more domain-general spatial abilities (Uttal & Cohen, 2012). The importance of focusing on children's science learning, in particular, will be discussed in subsequent sections.

1.6.1 Spatial ability and STEM: longitudinal evidence

Longitudinal studies over the past 50 years, sampling thousands of participants, provide strong evidence of a link between spatial ability and STEM achievement; both in academic and career outcomes (Lubinski & Benbow, 2006; Wai et al., 2009). For example, Wai et al., 2009 ('The Project Talent study'), first tested the spatial abilities (mechanical reasoning, surface development and abstract reasoning) of a representative sample of 400,000 general American secondary school students, aged 14-18 years, in 1960. The educational outcomes of the sample were then followed up 11 years later. The results indicated that the likelihood of earning an advanced degree in a STEM subject increased as a function of the level of a participant's spatial ability. For example, 45% of individuals with a STEM PhD at follow-up had been in the top 4% of spatial ability in adolescence, and, nearly 90% were in the top 7% for spatial ability. Controlling for mathematical and verbal ability, spatial ability accounted for an additional 4% of the variance in predicting the likelihood of having studied for either a STEM vs Non-STEM degree at follow up. In addition, these effects remain even after controlling for interest in STEM subjects, in high-school (Hedges cited in Newcombe, Uttal & Sauter, 2013).

Another longitudinal study, reported by Lubinski and Benbow (2006), involved five cohorts identified at different time points, based upon the Study of Mathematically Precocious Youth. Participants in the sample were identified by age 13 as intellectually talented. The most crucial finding of the study was that, controlling for mathematical and verbal SAT scores, spatial ability at age 13 significantly predicted the likelihood of earning undergraduate and graduate STEM degrees, and, being in a STEM career, 20 years later.

Assessing the longitudinal evidence overall, there is therefore a strong indication that individuals with stronger spatial ability in adolescence are more likely to pursue STEM-related degrees and careers.

1.6.2 Spatial ability and science: evidence from adults and adolescents

1.6.2.1 Chemistry

A number of studies have found associations between spatial ability and chemistry exam performance in university-level students (e.g., Baker & Talley, 1972; Bodner & McMillen, 1986; Pribyl & Bodner, 1987). Bodner & McMillan (1986), for example, found that spatial visualisation (intrinsic-dynamic ability) and flexibility of closure (the Embedded Figures Task, intrinsic-static spatial ability) were correlated with students' scores in an introductory organic chemistry course. Correlations were found between spatial skills and questions requiring problem-solving which were spatially represented, such as identifying crystal structures, but also, with apparently non-spatial items, such as stoichiometry questions. The authors argued that the non-spatial stoichiometry questions utilised spatial thinking because they required students to reorganise information in the question. That is, in order to solve these problems, students 'restructured' the problem information, a process which they consider to be related to the skills that a visualisation task draws on, i.e., transformation. Alternatively, it is also possible that the correlations reflect general cognitive ability; the research design did not enable this to be discounted as an explanation.

Pribyl and Bodner (1987) conducted another study involving undergraduate organic chemistry students, using the same spatial ability measures as above. A main finding of the study was that students with high spatial ability often produced drawings when solving problems which involved molecular structures, a strategy which authors argue contributed to their success in problem-solving. Individuals with poorer spatial abilities were less likely to draw diagrams, and those who did, produced drawings which were incorrect. The authors proposed that producing drawings allowed individuals with stronger spatial ability to organise and reconstruct the information in the question. However, the drawing analysis was anecdotal and not systematically coded. Moreover, such a conclusion would ideally be based on a mediational analysis (i.e., frequency / quality of drawings as a mediator of the relationship between spatial ability and learning outcomes).

Individuals with high spatial ability have also been shown to produce diagrams for mathematics problems which, on the surface, do not appear 'spatial' (Fennema & Tartre, 1985). This finding is also reinforced by research by Hegarty & Kozhevnikov (1999) who found that individuals with higher spatial visualisation ability tended to produce schematic, rather than visual, diagrams when solving non-spatial maths problems. Overall, this perhaps suggests that individuals with stronger spatial abilities tend to spontaneously think in a 'spatial' way, which in turn contributes to problem-solving; a claim previously made by (Lohman, 1993); although, see Fiorella and Mayer (2017) in relation to biology (section 1.6.2.2).

A number of studies have also addressed more specific conceptual aspects of chemistry. For example, a specific difficulty for chemistry students that has been identified is being able work with, translate and relate different types of representations of molecules. In Stull, Hegarty, Dixon and Stieff (2012), spatial ability, as measured through 3D visualisation, correlated (r = .28 to .32) with students' ability to translate between different diagrammatic representations of structures in an organic chemistry class.

Similarly Keig & Rubba (1993) examined the role of spatial ability (3D visualisation), logical reasoning ability and knowledge of representations, on secondary school student's (aged 15-17 years) ability, to translate between different types of representations. Although reasoning and knowledge predicted performance, in contrast to the study above with adults (Stull et al., 2012), spatial ability did significantly predict performance, after controlling for reasoning and knowledge. The inclusion of a measure of general reasoning ability may have accounted for the lack of unique role for spatial ability. That is, spatial ability did not predict outcomes, once general reasoning was accounted for, which would therefore suggest there was shared variance between the two abilities. However, it may have been due to the low power of the analysis, given the sample size was relatively small for the number of predictors in the regression analysis (n = 42, four predictors).

Finally, in a study with 12-13 year old pupils, Rhodes et al., (2016) investigated the relationship between several components of working memory and student's learning about acids and alkalis. Regression analyses identified visuo-spatial working memory as the only independent predictor of performance, when receptive vocabulary (an estimate of general IQ), and other executive function measures, were controlled for. This relationship was only significant for conceptual questions, which required a deeper knowledge and problem-solving, in contrast to the factual questions.

1.6.2.2 Biology/Medicine

Spatial ability has been linked with success at learning anatomical structures and with general academic performance in anatomy learning (Garg, Norman, Spero, & Maheshwari, 1999; Garg, Norman, & Sperotable, 2001; Rochford, 1985; Guillot, Champely, Batier, Thiriet, & Collet, 2007; Lufler, Zumwalt, Romney, & Hoagland, 2012). In a study involving second-year medical students' examinations, a composite spatial battery made up from tests of spatial visualisation, was predictive of performance in spatial multiple-choice test items and practical anatomy exams, but not non-spatial anatomical knowledge essays or multiple-choice questions (Rochford, 1985). Spatial questions (multiple choice and practical exams) related to the three-dimensional properties, characteristics and relative positions of the organs and systems in the human body. Lufler et al. (2012) more recently found similar results in a study with first-year medical students. Those enrolled in a gross anatomy course who were in the highest quartile of 3D mental rotation performance were 2.2 times more likely to score 90% on practical examinations, and 2.1 times more likely to score 90% or above on written examinations. Therefore, overall, individuals with stronger spatial abilities appear to be better able to process and learn anatomical structures, which are heavily spatial-relational in nature, and depend on processing of spatial configurations more so, than those with lower spatial abilities.

Spatial ability also relates to identifying and drawing cross-sections of three-dimensional objects, a type of penetrative thinking (Kali & Orion, 1996) and a skill which has implications for anatomy learning (Cohen & Hegarty, 2007). For example, Cohen and Hegarty (2007) demonstrated that perspective taking ability (extrinsic-dynamic spatial ability) moderately correlated with success at drawing cross-sections of novel (non-anatomical) three-dimensional objects. In addition to being better able to learn anatomical structures in terms of relative spatial configurations, spatial ability therefore also supports penetrative thinking in determining cross-sections of individual parts of these systems.

Spatial ability has also been associated with those working in the surgical discipline and with proficiency at learning and performing surgical procedures (Anastakis, Hamstra, & Matsumoto, 2000; Wanzel, Hamstra, Anastakis, Matsumoto, & Cusimano, 2002). Wanzel et al. (2002), for example, administered a range of spatial tasks to surgical trainees and subsequently taught them a spatially complex surgical procedure (a z-plasty). The results of the study indicated that performance on the form-board and the mental-rotation tasks, but not the snowy pictures (see section 1.2.1), the gestalt completion, shape memory or the cube comparison tests, correlated with performance on the surgical procedure.

Related to the previous study by Pribyl & Bodner (1987) in the chemistry section (section 1.6.2.1), Fiorella and Mayer (2017) investigated the spontaneous spatial strategy used by undergraduates when learning about the human respiratory system. The results of the study revealed that the total number of spontaneous spatial learning strategies used during learning (mind-map or drawing) significantly predicted overall learning outcomes. However, the use of these spatial strategies did not mediate the relationship between spatial ability and learning outcomes. That is, there was no evidence that one of the mechanisms through which participants with higher spatial ability achieved greater learning outcomes was accounted for through the greater likelihood of spontaneously using spatial learning strategies.

Two studies in the biology domain, based on samples of secondary school students, investigated different topics and reported contrasting outcomes. First, Klein & Koroghlanian, (2004) investigated the association between spatial ability (paper folding) and learning about mitosis and meiosis. The study, failed to find any association between spatial ability level, based on a median split, and post-test conceptual understanding (a mixture of multiple-choice and free-response questions). The authors suggest that a reason for the lack of relationship, compared to other similar studies, may have been that their achievement assessment measured comprehension only, and did not measure problem-solving or transfer.

Rhodes et al. (2014) assessed 12-13 year old student's visual-spatial working memory, and other aspects of executive function, in relation to their understanding of DNA. Regression analyses revealed that visuospatial working memory predicted performance on the conceptual, but not factual, sections of the assessment. However, unlike the previously described study by Rhodes (2016), which focused on chemistry and secondary school students, there was no measure of, or control for, general ability. Therefore, it is not clear whether the observed correlations might reflect a more general cognitive ability, such as reasoning or IQ.

Interestingly, performance in the mitosis/meiosis lesson was not related to spatial thinking related skills, whereas the lesson on DNA was. This may be a reflection of the different levels of spatial thinking required in the conceptual understanding of each topic. DNA has a spatial link for example, through the helix structure and pairing of bases.

1.6.2.3 Physics/Engineering/Geoscience

Spatial ability has been linked with learning and understanding specific conceptual aspects of physic, such as electricity. In a study with third-year undergraduate electrical engineering students, for example, mental rotation and mental cutting ability, were predictive of test items related to physical aspects of electric circuits (r = .505 and r = .527 respectively) but not with other sections of the same test on energy, current and voltage (Duffy, Sorby, & Bowe, 2016). Questions on the physical aspects of electric circuits required students to match images of circuits with the correct diagrammatic version. The spatial aspects of the question include being able to establish the correspondence between the two representations, by encoding both images, and, in particular, being able to transform (i.e., rotate) aspects of the image, such as the battery, to make effective comparisons.

Similarly, Delialioğlu & Askar (1999), investigated the relationship between 15 and 16 year old's spatial ability and performance on a test relating to electrostatic and electric current. Controlling for mathematical ability, a composite spatial ability measure (intrinsic-dynamic) accounted for an additional 9.6% variation in physics achievement. The previously discussed study by Duffy et al. (2016) with adults did not find a correlation between questions relating to electric current. This possibly adds weight to the argument that spatial abilities are more predictive at early stages of science learning, although it is difficult to make direct comparisons between studies due to differences in assessment types.

Research has also revealed an association between spatial ability and mechanics problem-solving. In a study involving undergraduate students, Kozhevnikov, Hegarty and Mayer (2002) reported that accuracy on a composite spatial ability measure (this included paper folding, card rotation, cube comparison and the form board task) correlated with performance on mechanics problems which involved extrapolation of motion from graphs. Participants were also asked to think aloud whilst they solved the problems, and were also asked specific further questions afterwards, about their problem-solving strategy. Analysis of performance on these qualitative mechanics tasks indicated that participants with low spatial ability tended to interpret graphs as literal pictures of events, and used mostly visual imagery. In contrast, high spatial ability participants were more likely to construct schematic images and use spatial manipulation to solve problems.

Further subsequent evidence of the relationship between spatial ability and mechanics problem-solving was reported by Kozhevnikov and Thornton (2006). This study investigated the relationship between spatial ability (the same composite measure as used in the previous study) and performance on questions relating to force and motion, with 92 physics-naïve undergraduate students. The results of this study indicated correlations in the range $r = .28$ to $r = .32$ between and the mechanics tasks. The authors proposed that the extent to which spatial ability impacts and predicts performance on mechanics tasks partly depends on the number of moving objects that the mechanical system contains. Multidimensional problems are more likely to require spatial resources. Participants with low spatial ability were more likely to choose answers which failed to simultaneously take into account all dimensions in the problem (e.g., the original direction and velocity of a moving object, and also, the effect of an additional force on the object). In addition, both Kozhenviokov, Hegarty and Mayer (2002) and Kozhenikov and Thornton (2006) ruled out general intelligence as a third variable by controlling for standardised assessments (Scholastic Aptitude Tests [SAT]) of mathematics and verbal ability.

Similarly, Isaak and Just (1995) found that individuals with poorer spatial ability were more susceptible to the cycloid illusion, a physics misconception, which relates to understanding rolling motion. The authors hypothesised that those with poorer 3D spatial visualisation skills made more incorrect judgments, and were more susceptible to the illusion, because the problem required simultaneous rotation and translation, thus requiring significant spatial resources, as in the above example.

Other lines of research propose that individuals with higher spatial ability reason more accurately about physical systems and processes because they develop more accurate and complete mental models during instruction. These findings are particularly relevant to a study in the current thesis, reported in Chapter 4.

For example, undergraduate students in Narayanan and Hegarty (2002) first read a text about how a flushing cistern works. Participants in the higher spatial ability (paper folding) median group scored more highly on a written description of the causal chain, and also on transfer questions (troubleshooting questions: i.e., suggesting what might be wrong in the system). However, spatial ability did not predict performance on basic comprehension/function questions. Similarly, undergraduates in Hegarty and Just (1993) learnt about the kinematics and configuration of a pulley system through studying a text with diagrams. The results of the study suggested that participants with lower spatial ability (as measured through a mechanical reasoning task) had greater difficulty in constructing a mental model. In particular, they performed less well on comprehension questions about the system, when learning from text and diagrams. Moreover, they were more likely to look more frequently at the diagrams, re-read clauses in the text, and, spent more time, overall, reading the descriptions. These findings suggest that spatial ability may support mental model construction, and the ability to explain and infer about a system, using the model.

Within the domain of geoscience/meteorology, Sanchez and Wiley (2014) found that multiple object, dynamic spatial ability predicted undergraduate's ability to explain the process of plate tectonics, which had previously been taught to them. Thus, this type of spatial ability may have supported mental model construction. An additional result not reported above was that there was also an interaction with the mode of teaching and dynamic spatial ability. In particular, participants with lower dynamic spatial ability performed worse, and thus perhaps struggled to construct a mental model, only when the illustrations in the text were static, and not when they were animated. This suggests that dynamic spatial ability is particularly important for 'animating' static diagrams into dynamic mental models.

Finally, undergraduates in Jaeger et al., (2016) learnt about the El Nino weather phenomena through reading a science text. Participants with higher paper folding scores were provided a larger number of accurate causal points in an essay which explained how the phenomena occurred. However, the paper folding scores did not significantly predict performance on the true or false concept verification questions. This therefore again perhaps supports the theory that spatial thinking skills particularly support mental model construction of physical processes, and in particular, utilisation of the whole model in explanation and inference tasks. Specifically, participants with stronger mental folding skills may have constructed a more accurate mental model, and also, have been more able to manipulate the model. When participants produced the written explanation of the entire process, they were required to sequentially link each of the stages of the weather phenomena. Thus, they may have been more so required to access and manipulate the entire mental model. The concept verification questions also required a degree of understanding, and thus the ability to access the model. However, because the concept verification question addressed only one element of the process, less co-ordination within the model may have been required. Thus, mental folding was a stronger predictor of the explanations, than the concept verification questions.

1.6.3 Spatial ability and science: evidence from younger children

Little research to date has examined the relationship between spatial ability and science in younger children. In addition to the reasons previously outlined, investigating this relationship is also important because it has implications for early curriculum design and also informs the development of spatial training interventions. The downstream effects of early interventions also have the potential to support learners when they are at more advanced stages of science education.

Jarvis and Gathercole (2003) focused on science achievement in UK national standardised tests in science at age 11 years, in relation to working memory. A key finding of the study was that visuospatial working memory was the only unique predictor of science achievement, when controlling for the other working memory measures. St Clair-Thompson & Gathercole (2006) found similar results, again with a sample of 11 year olds, also based on science achievement as measured through a general UK national standardised assessment. The study revealed a similar domain-specific association between visuospatial working memory and general science achievement. A limitation of visuospatial working memory measures, however, is that, they often, from a spatial thinking perspective, measure more than one ability. For example, the odd-one-out task, used in St-Clair-Thompson and Gathercole (2006) requires visual discrimination of forms (intrinsic-static skills) and also memory for location (extrinsic-static skills). More fine-grained research is therefore needed to investigate different aspects of spatial thinking in relation to science learning in the primary-school years.

Tracy (1990) also assessed science performance using a general standardised assessment, but in a sample of 10 and 11 year old American students. The sample was split into high and low spatial ability, as measured through a composite spatial ability test. The study revealed a significant difference in science performance between those in the high versus low spatial ability grouping. However, this study did not include any measure of IQ or other cognitive factors, and so provides generally weak evidence because it does not attempt to discount general ability as an alternative explanation.

In a more recent and methodologically sound study, Mayer, Sodian, Koerber, and Schwippert (2014) specifically investigated scientific reasoning ability or investigative skills, in relation to spatial ability, in children. A sample of 10 year olds completed a paper-and-pencil scientific reasoning assessment, which covered questions about the nature of science, understanding theories, designing experiments and interpreting data. Spatial ability was measured through a composite of 2D mental rotation and 3D visualisation tasks (i.e., intrinsic-dynamic skills). Controlling for various covariates (inhibition, fluid reasoning, problem-solving/central executive function and reading ability), spatial ability added 5% unique variance to the model in predicting overall scores. Spatial ability may have contributed particularly to the data interpretation questions, which included graphs and tables, but may have also assisted in mentally representing relationships between variables. However, the study did not consider the impact of the different kinds of spatial ability, due to the use of a composite measure, and, the composite measure focused only on intrinsic-dynamic spatial skills. There was also no examination if different elements of the science assessment associated more closely with spatial ability than others.

Finally, Harris, George, Hirsh-Pasek, and Newcombe (2018) investigated 6-year olds' and adults' force and motion understanding in relation to spatial ability. Participants were asked to predict and infer the effect of forces that were acting upon objects, through a computerised game. Participants also completed a mental folding and mental rotation task, as well as a measure of receptive vocabulary, which acted as a proxy control for general ability. The results of the study showed that, for children, mental folding, but not mental rotation, predicted performance. For adults, however, neither mental folding or mental rotation significantly predicted performance on the task. Although these findings appear to provide some support for the age-related change in the role of spatial thinking discussed previously, it should be noted that the adult findings in this study contrast to findings previously reported in the adult physics literature above (e.g., Kozhevnikov & Thornton, 2006) where it was reported that this type of spatial ability did indeed predict adult's performance on similar force and motion problems.

1.6.4 Summary of possible mechanisms which may explain the relationship between spatial thinking skills and science

Several inferences about possible mechanisms can be drawn from the literature as a whole explaining the relationship between spatial ability and different aspects of science learning. First, on a more domain-specific level, prior research suggests that spatial ability (mostly intrinsic/dynamic transformation skills, but, some intrinsic/static skills), supports the understanding and learning of conceptual topics that are particularly spatial in nature, such as mechanics and anatomy. For example, individuals with stronger spatial abilities are more able to learn and problem solve about topics which require co-ordination between multiple spatial aspects or components. This co-ordination might involve, for example: physical systems with more than one component; transforming and coordinating between multiple representations of the same molecule; and, supporting the processing of the spatial-relations of anatomical structures.

The ability to generate well-structured mental images, i.e., visualise, may also support mental model constructions, and therefore conceptual explanations, particularly of physical processes. For example, when explaining a scientific process, a learner might run a previously established mental model (see section 1.4.6) in order to support the reasoning process. The evidence broadly points to the application of conceptual understanding, explanations and transfer, rather than factual recall, being more closely linked to spatial ability.

Second, some of the mechanisms appear to be more domain-general. For example, spatial ability may be involved, or support, problem-solving generally, regardless of topics. The capacity to represent information mentally in a spatial format may support the construction of mental models of individual problems in assessments. Furthermore, the transformational aspects of intrinsic-dynamic spatial skills may support restructuring the information that is given in a question in order to solve it.

Spatial thinking skills may thus support thinking about non-spatial phenomena, or in solving apparently non-spatial problems; this may be supported through strategies, such as drawing (i.e., by employing a spatial tool). Third, and related to this point, some domain-general scientific reasoning and investigation skills appear to draw upon spatial thinking skills, such as graph and table use and interpretation, and the general ability to use diagrams (i.e., the use of spatial tools).

1.6.5 Spatial tools and science learning in children

As well as there being little research on the relationship between spatial ability and science learning in younger children, there is currently also limited comparative evidence concerning the role of spatial tools. Separate lines of research exist concerning the role of physical and concrete objects (e.g., Evangelou & Kotsis, 2019; Triona & Klahr, 2003) and diagrams (e.g., Herrlinger, Höffler, Opfermann, & Leutner, 2017) in children's science learning. However, no research to date has investigated the role of gesture as a spatial learning tool to support children's science learning. In addition, no prior work has compared these three types of spatial tools and instructional approaches in the context of learning about a scientific concept. Moreover, there has yet to be an investigation into the extent to which spatial thinking skills (e.g., mental folding) may moderate the impact of spatial tools in science, or vice-a-versa. For example, gestures and concrete objects may be particularly useful for children with weaker spatial thinking skills. In the context of instruction, a child with weaker intrinsic-dynamic visualisation skills could be directed to produce a gesture in relation to a science concept, which could in turn activate motor areas in the brain. When the child is later required to reason about that science concept in the absence of the object, and therefore mentally simulate the action, the prior gesture-based experience theoretically may enhance that mental simulation. This aspect of the thesis, which relates to the third goal outlined in section 1.1, is outlined in more depth in Chapter 6.

1.7 Overall summary and rationale for thesis

To review, the main the thesis is to examine the relationship between spatial thinking skills and science learning in childhood. A strong body of literature exists which examines the association between spatial ability and science achievement and learning with adults and more experienced learners, but much less is known about this relationship in younger children. Prior research suggests that spatial ability and spatial skills may be more important for younger and less experienced learners, with less formal knowledge, than older or more experienced learners. Also, given the degree of malleability of spatial skills at a younger age (Heckman, 2006) research with younger children may provide avenues for training and intervention. Early intervention and training may also have a long term benefit for later STEM education. Additionally, no studies to date have considered how the importance of spatial thinking may vary across throughout the primary-school years.

No research to date, including research in the adult literature, has considered the broader model of spatial skills proposed by Uttal et al. (2013) in relation to science learning; the vast majority of research to date has focused only on object-based/intrinsic skills. For example, the role of spatial perspective taking and spatial scaling as predictors of children's science learning is unclear. In addition, the majority of studies with younger children also focus only on science achievement generally, but do not provide any indication of the disciplines of science or science skills, which may show a stronger or weaker relationship with spatial skills. Relatedly, research has yet to address whether spatial skills are particularly related to skills linked to the development of mental models (i.e., explanations), compared with other skills. The current thesis therefore includes several complementary approaches to assessing science learning and how this relates to spatial thinking skills. In Chapter 3, a curriculum-based science assessment provides a broad measure of learning outcomes which align with school-based expectations. In contrast, Chapter 4 and 5 focus on specific physics sub-topics. Chapter 4 includes science assessments which separately assess factual knowledge recall and conceptual understanding, therefore allowing a comparison to be made between these science skills. In addition, conceptual understanding is further sub-divided into predictions and explanations, thus providing a more fine-grained analysis, and evidence in relation to the relationship between spatial thinking and mental models. Chapter 5 also sub-divides conceptual understanding, but does so along a dimension of near and intermediate transfer.

The second aim of the current thesis is to examine the structure of spatial cognition in middle childhood, particularly in relation to the Uttal et al (2013) model. A considerable amount of research to date has focused on the structure of spatial cognition in adults. This has resulted in the proposal of several prominent theoretical models, including the more recent 2x2 theoretical model (Uttal et al, 2013). However, there has been little direct evaluation of this model, and even less research in childhood. Furthermore, the research in childhood that does exist provides a somewhat inconclusive picture. There are also several general caveats to the model which suggest that further evaluation is timely. Some of these issues include, for example, the breadth of the intrinsic-dynamic and extrinsic-dynamic sub-domains, and the dissociation object-based transformation from perspective taking spatial orientation depending on features of task design.

One approach to investigating the structure of spatial cognition in the thesis is through psychometric modelling through confirmatory factor analysis (Chapter 1 and Chapter 4). In addition, the evaluation of the model will also be supported through the analysis of developmental data (Chapter 1). Although research indicates that spatial skills develop within middle childhood, no research to date has compared the development of a broad range of spatial skills, chosen using this theoretical framework. The conclusions from both of these sources of data will be considered together. For example, a psychometric dissociation between two skills may be reflected in different developmental trajectories. In addition, the data relating to the predictive role of different spatial thinking skills in relation to science learning, described in the aim above (Chapters 3 and 4), will be integrated into the evaluation of the model. For instance, the finding that spatial skills falling into different sub-domains within the model uniquely predict science learning could lend support to the model.

The final aim of the thesis is to examine the effectiveness of gesture as a spatial tool in supporting children's conceptual learning in science, and how gesture compares to learning with other spatial tools, an area previously unaddressed. An additional aspect of this final aim was to investigate the possible moderating role of spatial thinking skills in the role of gesture, concrete models and diagrams, in children's science learning (Chapter 5). In some prior studies with adults, the benefit of spatial thinking tools to science learning is independent of spatial thinking skills. For example, the spontaneous use of spatial strategies by learners, such as mind maps, is positively correlated with learning outcomes (e.g., Fiorella & Mayer, 2017). In other cases, spatial thinking skills have a moderating effect on their use. For instance, dynamic spatial ability is important for adults' learning with static diagrams, but not with dynamic illustrations (e.g., Sanchez & Wiley, 2014). In addition, spatial analogies are particularly helpful for learners with lower spatial ability (Jaeger et al., 2016). The study in Chapter 5 therefore connects the findings relating to spatial thinking skills and science learning in Chapters 2 and 3, to the use of spatial tools, in the context of instruction.

Chapter 2

The latent structure and development of spatial skills in middle childhood

2.1 Introduction

Spatial cognition includes a range of skills requiring, to varying degrees, the processing, representation, comparison and transformation of spatial information. Attempts at categorising spatial thinking skills have been approached from both psychometric and theoretical perspectives, which has led to several contrasting typologies (Linn & Peterson, 1985; Caroll, 1993; Uttal et al., 2013). Uttal et al. (2013) proposed a theoretically driven model that goes beyond purely psychometric data and links to evidence from cognitive neuroscience. However, the model has received little direct evaluation to date. The goal of the current study was therefore to investigate the developmental trajectories of spatial skills in middle childhood, to psychometrically evaluate the degree of relationship between these skills, and then use these analyses as evidence to further evaluate the model.

To achieve this, a sample of 6-11 year olds completed a selection of spatial skills which were chosen to measure each of Uttal et al's (2013) four spatial sub-domains. The developmental trajectories of each spatial task were compared to determine if any similarities or differences mapped onto the 2 x 2 model. For example, a difference in the developmental trajectories of mental rotation (intrinsic-dynamic skill) and perspective taking (extrinsic-dynamic skill) would support the theorised distinction between these sub-domains within the model. In addition, confirmatory factor analysis was used to test the intrinsic-extrinsic and static-dynamic divisions within the model. The following sections review the background to the 2 x 2 model, the existing evidence on the structure of spatial cognition in childhood, and current evidence on the development of spatial skills.

2.1.1 The 2x2 typology of spatial thinking

Previous spatial typologies have been criticised for not being sufficiently theory-driven (Hegarty & Waller, 2005). In response, Uttal et al. (2013) and Newcombe and Shipley (2015) proposed a two by two classification of spatial thinking, which categorises skills as being intrinsic or extrinsic along one dimension, and static or dynamic, along the other. In combination, this two by two classification gives rise to four spatial sub-domains: intrinsic-static, intrinsic-dynamic, extrinsic-static and extrinsic-dynamic.

There is good evidence to support an intrinsic vs. extrinsic distinction between spatial skills. First, there is an evolutionary basis for the intrinsic/extrinsic dimension, based on the distinct spatial functions of tool use and navigation, respectively (Newcombe, 2018). Considering intrinsic-dynamic vs extrinsic-dynamic skills, psychometric studies in adults indicate that mental rotation (intrinsic skills) and perspective taking (extrinsic skills) are dissociated, psychometrically (Kozheninikov and Hegarty, 2001; Hegarty & Waller, 2004). However, the extent to which an intrinsic-dynamic vs extrinsic-dynamic dissociation is observed psychometrically depends upon features of perspective taking task design, and between-participant strategy choices. Brain imaging data also indicates that mental rotation and perspective taking are also linked to dissociable but overlapping systems (Zacks, Vettel, & Michelon, 2003; Wraga, Shephard, Church, Inati, & Kosslyn, 2005). Finally, response time data from mental rotation and spatial perspective taking tasks frequently demonstrate different chronometric profiles (Crescentini et al., 2014, Habacha et al., 2018, Keehner et al., 2006; Michelon & Zacks, 2006).

Less direct evidence exists for the dissociation between dynamic and static spatial skills. In many cases, a static skill is a pre-requisite to a dynamic skill (e.g., it is necessary to encode a shape before mentally rotating it), and therefore it may be difficult to clearly distinguish between skills along this additional dimension (Newcombe, 2018). There is, however, some evidence of a distinction between object-visualisers (intrinsic-static) and spatial-visualisers (intrinsic-dynamic) (Kozhevnikov et al., 2005). However, though there is supportive evidence for elements of the Uttal et al. (2013) model, there has been very little direct empirical evaluation of the model.

2.1.2 The latent structure of spatial cognition in childhood

A limited number of studies have investigated the structure of spatial cognition in childhood, although the results are inconclusive. Some studies have compared individual domains of the model specifically. In particular, some evidence suggests that object-based transformations (intrinsic-dynamic skills) and spatial perspective taking (extrinsic-dynamic skills) appear to be dissociated for children at the end of middle childhood, aged 10.5-12 year olds (Vander Hayden et al., 2016; Heil, 2018).

These skills may not, however, be dissociated for younger children, aged 7.5 to 10.5 (Vander Hayden et al., 2016).

These findings can also be interpreted in the context of egocentric and allocentric encoding of space (Klatzky, 1998). Egocentric encoding is common to both intrinsic transformations (i.e., object-based transformations, such as mental rotation) and extrinsic transformations (i.e., perspective taking). However, during perspective taking tasks, it is optimal to also utilise allocentric encoding (i.e., environmental frameworks), and to also update egocentric encoding. In line with this, the response times adults show in perspective taking tasks often do not show a linear trend with angular disparity (Kessler & Thomson, 2010; Zacks & Michelon, 2005). Adults may therefore use an allocentric framework to perform a direct 'blink' transformation to another location (Wraga et al., 2000). In contrast, children's response times do sometimes demonstrate a linear relationship with angular disparity (Roberts & Aman, 1993). Younger children may therefore sometimes use a more egocentric strategy such as a graduated rotation of the self. Prior research has shown that allocentric encoding of space is not fully mature until 10 years (Bullens et al., 2010). The lack of a significant dissociation between intrinsic-dynamic and extrinsic dynamic skills for the younger children in Vander Hayden et al. (2016), may reflect: a lack of allocentric skills; the tendency not to spontaneously use allocentric frameworks, or difficulty in shifting between egocentric and allocentric frameworks. As a result of the above possibilities, it may be that younger children choose to use an object-based strategy or a more ego-centric strategy.

However, the only study to directly test the whole of the Uttal et al. (2013) two by two model, with adults or children (Mix et al., 2018), found somewhat contradictory evidence to the findings above. Given this is the only comprehensive assessment of the Uttal et al. (2013) model, it is therefore the central study of comparison in the current study. The study included children divided into three age groups: 6 year olds (n = 251), 9 year olds (n = 246) and 12 year olds (n = 241). Children completed a selection of spatial thinking tasks. A summary of the tasks for each sub-domain is provided in Table 3. For some tasks, variations of the same type of spatial task were used with different age groups. For example, a 3D mental rotation task was used with the 12 year olds whereas a 2D mental rotation task was used with the 6 and 9 year olds.

Table 3: Spatial tasks included in Mix et al. (2018)

	2x2 sub-domain	Spatial tasks	Description/comment
6 year olds **9 year olds**	**Intrinsic-static**	Visuo-spatial memory	Children recalled the locations of previously presented dots
		Figure copying	Children copied a line drawing of a geometric shape
	Intrinsic-dynamic	2D Mental rotation	
		WISC block design	Children constructed specified figures using small cubes
	Extrinsic-static	Map reading	No map rotation involved
		Proportional reasoning	
	Extrinsic-dynamic	Map reading	Involved map rotation
		Spatial perspective taking	Task involved PlayMobil (toy) characters
12 year olds	**Intrinsic-static**	Visuo-spatial memory	Children recalled the locations of previously presented dots
		Figure copying	Children copied a line drawing of a geometric shape
	Intrinsic-dynamic	3D Mental rotation	
		WISC block design	Children constructed specified figures using small cubes
	Extrinsic-static	Proportional reasoning	
	Extrinsic-dynamic	Map reading	Involved map rotation
		Spatial perspective taking	Kozhevnikov & Hegarty, 2001 task; described in section 1.2.2

The authors used confirmatory factor analysis to test three hypothesised models: a two-factor intrinsic vs extrinsic model (which included both static and dynamic tasks), a two-factor static vs dynamic model (which included both intrinsic and dynamic tasks), and a four-factor model, in which each of the four factors represented one of the four sub-domains. These factors were compared to a one-factor model, in which spatial thinking was presented as a unitary construct. Because the 12 year olds completed only one measure for the extrinsic-static domain, it was not possible to test the four-factor model with this age group.

For the younger two age groups (6 and 9 year olds), Mix et al. (2018) found that an intrinsic-extrinsic, two-factor model fitted the data better than a one-factor model. This therefore supported the intrinsic-extrinsic dimension. For the 12 year olds, neither of the two-factor models fitted the data better than the one-factor model. Therefore, a one-factor model fitted the data best; however, a one-factor model did not fit the data well, according to fit statistics. This indicates that a factor structure more complex than a single dimension may exist for 12 year olds, but was not adequately captured by the tasks and pre-specified task loadings in this study. It was only possible to test the full four-factor model with the 6 and 9 year olds, and, in both cases, a four factor-model was inadmissible. Moreover, there was no significant support for a 2-factor, static-dynamic model, at any age group.

Therefore, Mix et al.'s (2018) findings suggest less, and not more, of an intrinsic-extrinsic dissociation with age. As discussed, research with adults indicates that intrinsic-dynamic and extrinsic-dynamic skills are dissociated. Taken together, the authors suggest that there may be a non-linear developmental pattern of dissociation, whereby dissociation exists for younger children, decreases in late childhood, and is then apparent again for adults. However, the findings of Mix et al (2018) contrast to Vander Hayden et al.'s (2016) and Heil's (2018) findings above, who both report an intrinsic-extrinsic dissociation in late childhood.

It should be noted that there are issues with task choices in the Mix et al. (2018) study which may have affected the extent to which a psychometric intrinsic-extrinsic dissociation was evident for 12 year olds. For example, linking to section 1.2.2.2, the 'dynamic' map reading task was considered an as *extrinsic*-dynamic task. However, participants were required to rotate the map in order to determine the correct location. This rotation could have been made intrinsically. The non-rotation, more extrinsic ('static') version of the map reading task was not available for the 12 year olds. It may have therefore been the case that, on balance, the extrinsic dimension was somewhat more 'intrinsic' for the 12 year olds, leading to noise in the design. In addition, the study included the minimum number of tasks per dimension, two, to test a four factor design.

Thus, overall, a limited amount of research currently exists which tests the dimensions of the 2x2 model, and, even less exists with children. Moreover, there are inconsistencies in the data that do exist. Further research on the latent structure of spatial cognition in childhood is therefore timely.

2.1.3 The development of spatial skills in middle childhood

The theory that psychometric dissociations between spatial skills may vary throughout development is potentially supported by the fact that spatial thinking skills develop significantly from childhood to adulthood. As such, another aim of the study was to detail the developmental trajectories of each spatial sub-domain, as this will be informative to our understanding of the relational structure of these skills in childhood. The spatial tasks included in this study were selected based on Uttal et al.'s (2013) theoretical framework of spatial cognition (see also Newcombe & Shipley, 2015). The design of the task battery enabled comparisons of intrinsic vs. extrinsic spatial skills, and static vs. dynamic spatial skills developmentally. Presented below is a brief review of existing research on the development of these skills in middle childhood, following from section 1.2.7.

Intrinsic-dynamic spatial thinking involves visualising and mentally transforming 2D and 3D objects. Children begin to show evidence of using mental rotation strategies in 2D rotation paradigms, from approximately the age of four (Estes, 1998). Crescentini et al. (2014) found that basic 2D rotation skills further improve significantly between the ages of 7 and 8 years, with no significant improvement from 8 to 11 years. The finding that 2D rotation skills do not develop significantly beyond the age of 8 was also supported by Marmor (1975), who found that 8 year old's 2D rotation speed was at near adult levels. With 3D rotation tasks more typically used with adults, however, performance continues to improve until late childhood. Further development is evident between the ages of 8 and 9 years (Vander Heyden et al., 2016), but, no development is evident between 9 and 10 years (Titze, Jansen, & Heil, 2010).

Intrinsic-static spatial thinking is the "coding of spatial features of objects, including their size and the arrangement of their parts (e.g., to identify objects as members of categories)." (Newcombe & Shipley, 2015, p.6). The Children's Embedded Figures Task (CEFT) was used as a measure of intrinsic-static spatial thinking in the current study. The embedded figures task is often used as a measure of intrinsic-static spatial thinking. This task requires perception and processing of the spatial configuration of an object. Specifically, individuals are required to locate a geometric shape, embedded or hidden within a more complex image. A version of the embedded figures task suitable for children is the Children's Embedded Figures Task (CEFT). Children can complete pre-school versions of the CEFT by 3 years, and, performance improves in early childhood, from 3 to 5 years (Busch, Watson, Brinkley, Howard, & Nelson, 1993; Witkin, Otman, Raskin, & Karp, 1971). In addition, within middle childhood, significant development between the ages of 7 and 8 years, and also between 8 and 9 years, is also evident (Bigelow, 1971).

Extrinsic-dynamic spatial thinking involves the processing of the spatial relations between objects, with additional movement or transformation of the objects or elements involved. Extrinsic-dynamic spatial skills were assessed in the current study using a spatial perspective taking task. A core extrinsic-dynamic spatial thinking skill, spatial perspective taking, involves visualising an environment in its entirety from a different position (Uttal et al., 2013). Developmentally, level 1 perspective taking requires a child to *know that* another person may see an object which they currently cannot (Flavell, Everett, Croft, & Flavell, 1981). Three year olds demonstrate level 1 perspective taking knowledge (Flavell et al., 1981), and this basic skill may be present in 14-month olds (Sodian, Thoermer, & Metz, 2007). Level 2 perspective taking, in contrast, is the ability to determine *exactly how* another person would perceive an object from a different perspective. Level 2 perspective taking emerges around 4 or 5 years of age (Pillow & Flavell, 1986). Frick, Möhring, and Newcombe (2014b) report that further development in level 2 perspective taking occurs within middle childhood, mainly between the ages of 7 and 8 years.

However, Crestectini et al. (2014) found that perspective taking develops later, predominantly between the ages of 8 and 9 years. Relating back to the previously reported object-focused skills from Crestectini et al. (2014), their findings together suggest that rotation and perspective taking have different developmental trajectories, with perspective taking showing development later than mental rotation. In line with this, Xistouri and Pitta-Pantazi (2006) also found that perspective taking skills continue to develop between the ages of 9 and 12 years. The difference in developmental trajectory is likely due to the difficulties younger children have in inhibiting their own ego-centric perspective (i.e., in not choosing the response option which represents their own viewpoint). This is possibly linked to the extended development of executive function (e.g., inhibitory control) skills, throughout middle childhood (Diamond, 1990; Gathercole, Pickering, Ambridge & Wearing, 2004).

Extrinsic-static spatial thinking involves the coding of object locations in relation to other objects, spatial frameworks or landmarks. Extrinsic-static skills were measured in the current study using a spatial scaling task. As outlined in section 1.2.1, spatial scaling ability has been used as a measure of extrinsic-static spatial thinking. Scaling tasks can be described as extrinsic and static, within the Uttal et al. (2013) model, because of the requirement to align between two different sized spaces, and then map locations and distances from one space to another. In terms of the development of spatial scaling in naturalistic environments, Vasilyeva and Huttenlocher (2004) found that 90% of 5-year-old children could successfully place objects on a rectangular rug, using a scaled, two-dimensional map. However, only 60% of 4 year olds were successful at the same task. Other paradigms have used smaller, table-top-sized materials. For instance, when participants are shown a location on a map, and are asked to find the corresponding position in a scaled space, no significant development is evident above the age of 6 years, compared to adults (Frick & Newcombe, 2012).

In summary, the development of spatial skills varies somewhat between sub-domains, and studies. Furthermore, there is some mixed evidence that intrinsic skills, particularly mental rotation, and extrinsic skills, specifically perspective taking skills, have different developmental trajectories. Specifically, perspective taking appears to develop later than mental rotation, i.e., reaches an adult-level of competence at a later age. Such a finding maps onto the previously described psychometric findings, whereby a dissociation between this type of skill does not emerge until later childhood. That is, it may be that the optimal strategies (i.e. allocentric encoding) for effective performance on extrinsic-dynamic tasks, or the cognitive processes supporting these strategies, do not emerge until later in middle childhood. It is currently not known, however, how other types of intrinsic-dynamic transformation (e.g., mental folding) develop in middle childhood, and, if they show a similar or dissimilar trajectory to mental rotation, or to extrinsic spatial skills.

Other than studies in which mental rotation and perspective taking have been measured simultaneously, the development of the spatial skills discussed thus far have typically been addressed in isolation, in distinct participant samples. The use of such distinct samples across studies makes it difficult to make direct comparisons across the developmental trajectories of each skill. Additionally, no known study assesses the development of a range of spatial skills in consecutive age groups, throughout middle childhood. This fine-grained approach is needed to provide more detail as to spatial skills development in this age range, which will in turn provide a source of evidence to further evaluate the Uttal et al. (2013) model.

2.1.4 Current study

To review, the current study had two aims. First, it aimed to further investigate the inter-relational psychometric structure of spatial skills in middle childhood, based on an assessment of the two dimensions of the Uttal et al. (2013) model. Second, it aimed to investigate the development of these skills, linking this to the observed psychometric structure. Only one study (Mix et al., 2018) to date has directly tested both dimensions of the Uttal et al. (2013) model, and the findings are difficult to reconcile with the wider literature. Moreover, in terms of developmental trajectories, research has to date only directly compared mental rotation (intrinsic-dynamic) and perspective taking (extrinsic-dynamic) skills. By including measures from each of the Uttal et al. (2013) sub-domains in a cross-sectional sample, throughout middle childhood, it was possible to address these areas in the current study. Furthermore, by simultaneously addressing both the latent structure and development of the skills, it was also possible to use each area to mutually inform our understanding of the other. The combined developmental and psychometric data provide a source of evidence to further evaluate the Uttal et al. (2013) model.

It was predicted that there would be significant age-based increases in accuracy within middle childhood, and, based on prior research, that this would be predominantly between the ages of 7 and 9 years. However, it was also predicted that there would be some differences in the developmental trajectories between intrinsic and extrinsic tasks. Based on the evidence to date, a difference in the developmental trajectories was expected between mental rotation and mental folding (intrinsic-dynamic), compared with spatial perspective taking. More specifically, it was expected that perspective taking (extrinsic-dynamic) would show significant developments in accuracy later than mental rotation and mental folding skills. Furthermore, it was expected, based on prior research, that, if a latent dissociation existed within this age range, it would be along the intrinsic and extrinsic dimension. Given that there is less supportive evidence and less research generally for the static-dynamic distinction, a clear dissociation was not expected to emerge along this dimension. Due to the conflicting data, there were no firm predictions regarding whether a dissociation was more likely in earlier or later, within middle childhood.

2.2 Methods

2.2.1 Participants

Participants were sampled from a large, culturally diverse primary school in London. The eligibility for free school meals was 19%, slightly above the national average (Department for Education, 2017). The initial sample consisted of 185 children, across six age groups (6 years to 11 years). Due to missing data caused by technical failure, 8 participants did not have a full set of scores available for analysis. Seven of these participants were missing data from one task only, and to maximise statistical power, their missing scores (two British Picture Vocabulary Scale-III scores, two mental folding scores, two perspective taking scores, one CEFT score) were estimated by calculating the mean for their respective year group and replacing their missing score with the mean value. The eighth participant was missing several variables and was excluded from the analysis. The final sample therefore consisted of 184 participants. A summary of the age and gender of participants in each age group is given in Table 4.

Table 4: Demographic information of participants across age groups

Age group	N	% male	Age years (mean ± SD)
6 years	30	53.3	6.00 ± 0.34
7 years	31	41.9	6.99 ± 0.29
8 years	32	56.3	8.03 ± 0.28
9 years	31	45.2	8.97 ± 0.32
10 years	31	51.6	9.95 ± 0.33
11 years	29	58.6	11.00 ± 0.30

2.2.2 Materials

A summary of the spatial tasks used in this study, and in Chapter 3, are provided in Table 5.

Table 5: Summary of tasks included in relation to Uttal et al. (2013) model

Spatial Task	Dimension 1		Dimension 2	
	Intrinsic	Extrinsic	Static	Dynamic
Mental Rotation Task	✓			✓
Mental Folding Task	✓			✓
CEFT	✓		✓	
Perspective Taking Task		✓		✓
Scaling Task		✓	✓	

2.2.2.1 2D mental rotation (intrinsic-dynamic sub-domain)

This task was modified from Broadbent, Farran, and Tolmie, 2014. In this task, participants were asked to identify which of two monkeys above a horizontal line, matched the target monkey below the horizontal line (see Figure 14). One monkey below the line was rotated by a fixed degree, relative to the target monkey and one monkey above the line was a mirror image of the target monkey. Participants indicated their answer by pressing one of two pre-selected computer keys. Participants first completed four trials at 0°, i.e., where the monkey below the horizontal line was not rotated. These practice trials were checked by the experimenter, and only participants achieving at least 50% on the practice trials continued to the experimental trials. Following the practice trials, participants completed 36 experimental trials, 8 x 45°

trials, 8 x 90° trials, 8 x 135° trials, and 8 x 180° trials and 4 trials at 0°). Equal numbers of clockwise and anticlockwise rotations were included.

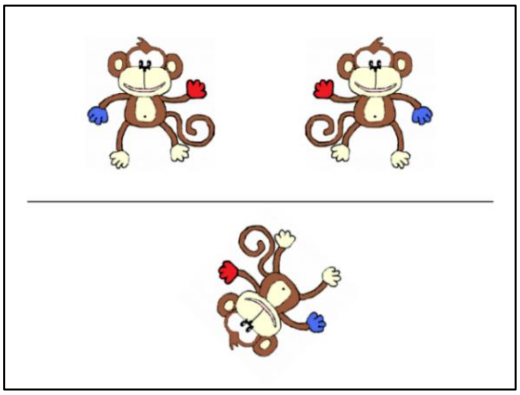


Figure 14: 135° anti-clockwise 2D mental rotation trial
(correct answer is left monkey)

2.2.2.2 Mental folding task (intrinsic-dynamic sub-domain)

In the mental folding task (Harris, Newcombe & Hirsh-Pasek, 2013), children imagined folds made to a piece of paper, without the physical representation of the fold itself (a physical piece of paper). Participants were shown a green shape at the top of a computer screen (see Figure 15). Across all trials, the shape included a dotted line, which represented the folding line, an and arrow, which indicated the direction and distance of the required fold. Below this shape were four possible response options, each of which showed a possible outcome following the directed fold. Only one of these four response options (the target item), showed the outcome of the fold correctly. The other three options were systematic error types. Participants first completed two practice items, in the first of which they were given a physical card version to check their answer. Answers to each practice question were checked by the researcher and if a child indicated an incorrect option they were given one further attempt to answer the item. The majority of participants answered the practice item correctly on the first attempt and all participants answered correctly by the second attempt. Following the practice items, participants completed 14 experimental items.

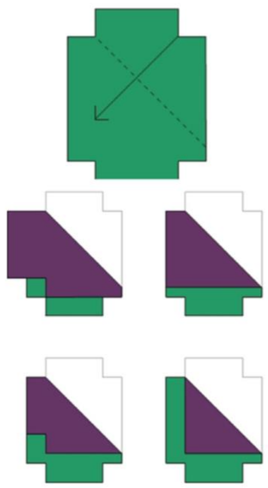


Figure 15: Mental folding trial (correct answer is C)

2.2.2.3 Children's Embedded Figures Task -CEFT (intrinsic-static sub-domain).

The Children's Embedded Figures Task (Witkin et al, 1971) requires participants to locate a target shape within a more complex meaningful figure. The task was administered as per the administration guidelines (Witkin et al., 1971). The task contained a maximum of two blocks, presented in a fixed order (A-B). Participants were first shown the target shape (a 'tent' shape for block A, and a 'house' shape for block B). Before the practice and experimental trials, participants were familiarised to the shape through four discrimination trials, where they were required to identify the target shape from a selection of other shapes. Participants repeated the discrimination trials until two items were answered correctly in succession. After this, participants completed either two (block A) or one (block B) practice trials, where they located a target shape hidden within a more complex image. Participants outlined the shape with their finger to indicate their answer. Participants were required to successfully locate the target shape in each practice item, before progressing to the experimental trials. Participants repeated the practice trials until they located the target item, and all children eventually passed the target trials. Following the practice trials, participants completed either 11 (block A) or 14 (block B) experimental trials, in which they again were required to locate the target shape hidden with a larger more complex image. Participants progressed from block A to block B, provided they correctly located the target shape for at least one trial within block A.

2.2.2.4 *Perspective Taking Task for Children- (extrinsic-dynamic sub-domain)*

This task (Frick, Möhring & Newcombe, 2014b) involved spatial perspective taking whereby participants visualised what photographs would look like, when taken from cameras placed at different angles and positions, relative to their viewpoint (Figure 16). Participants first completed four practice questions involving physical Playmobil characters. The experimenter placed two Playmobil characters, each holding a camera, next to two objects, in a specified arrangement on a table. For the first practice question, the participant was shown four photographs of the objects, which were taken from the perspective of one of the characters. The participant was then asked which of the two characters could have taken the photograph. The characters were rearranged, and the question was asked again with new photographs, for the next practice question, and the process repeated. Participants were able to check their answers by standing behind the photographer. Following this, participants completed an additional practice question on a laptop computer, which showed a Playmobil character taking a photograph of two objects from the same perspective as the child (0° angular difference trial).

On passing the practice questions, the task then continued with the main trials. These varied per the number of objects in the layout (1, 2 or 3) and the angular difference between the photographer's and the child's perspective (0°, 90° or 180°). The task consisted of two blocks of 9 trials; each of the three angular differences were presented once for 1, 2 and 3 object trials. The first block progressed with all 1 object trials first, followed by 2 object trials and finally all 3 object trials. The second block was reversed such that it began with 3 objects, working back to 1 object only. Accuracy was recorded on the computer through the child's touch screen response to each item.

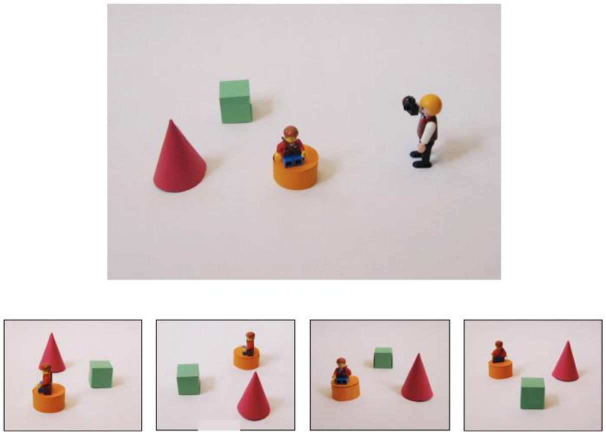


Figure 16: Perspective taking, 90°, three object trial (correct answer is A)

2.2.2.5 *Spatial scaling (extrinsic-static sub-domain)*

In this spatial scaling task (Gilligan, Hodgkiss, Thomas, & Farran, 2018) children were required to find equivalent corresponding locations on two maps, when one was varied in size relative to the other. Participants were presented with four treasure maps on a touch screen computer, which each contained a single black square (the treasure location), positioned at different locations for each map (Figure 17). To the left of the computer, children were presented with a printed treasure map, mounted in an A3 ring bound pad. The child's task was to determine which of the four maps on the computer screen had the treasure (i.e., the black square) positioned in the same place as the larger printed map. Only one of the four computer maps displayed the treasure in exactly the same position as the printed map. The other three, incorrect, options were created uniformly, for each trial.

Printed maps were either unscaled (1:1; 7cm x 7cm), or, were scaled, to either a ratio of either 1:2 (14cm x 14cm) or 1:4 (28cm x 28cm), relative to the maps on the computer (which were 7cm x 7cm each). Half of the items contained grids which separated the maps into 6 x 6 (larger) grid sections, and therefore required gross level acuity, while the other 9 items contained grids which separated the map into 10 x 10 (smaller) sections, and therefore required fine level acuity.

Participants first completed two unscaled practice items, after which, they completed the main 18 trials of the test. If participants did not answer correctly, they were given verbal feedback, and one further chance to complete the practice item. Only participants who correctly answered 50% or more of the practice items on their first attempt correctly continued to the main trials. All participants answered 50% or more correctly on their first attempt. Following the practice items, participants completed 18 experimental trials, which included six items presented at each scale factor.

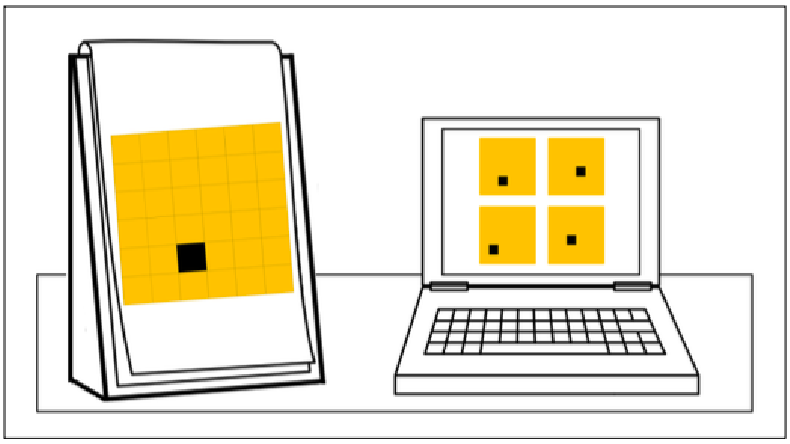


Figure 17: Spatial scaling trial at a scaling factor of 1:4
(correct answer is A)

2.2.3 Procedure

The mental rotation task and mental folding task were completed by participants in a group testing session, lasting approximately 35 minutes, involving groups of eight children. Each group testing session was supervised by at least two researchers. Within the group testing session, children aged 6-10 years completed additional mathematics tasks not reported here. The CEFT, the perspective taking task and the scaling task were then completed in an individual testing session, lasting approximately 45 minutes. Within the individual testing session, children aged 6-10 completed an additional spatial language measure, not reported here.

2.3 Results

2.3.1 Development of spatial skills

2.3.1.1 Analysis strategy

Percentage accuracy was used as the dependent variable for the spatial tasks, for all analyses, which was converted into a z-score, across the whole sample. To investigate the development of spatial thinking across age groups, a separate one-way analysis of variance (ANOVA) was conducted for each spatial task. For each spatial task, a preliminary ANOVA was conducted, which included gender (2 levels: male and female) and age group (6 levels: 6,7,8,9,10,11 years) as between-participant factors. If there was no interaction with gender, it was then removed from the analysis. Homogeneity of variance was determined through Levene's test. For measures in which the homogeneity of variance was violated, Games-Howell post-hoc tests were used to investigate significant main effects. Otherwise, post-hoc Bonferroni adjusted t-tests were used. Gender was not included as a variable within the main ANOVA because Games-Howel post-hoc tests were not available for factorial designs.

To further analyse the developmental trends across the age groups, polynomial contrasts were calculated, for each spatial task. The aim was to determine if the pattern of development for that task was only linear, or if there was also a significant higher-order trend; for example, a quadratic trend (i.e., indicating one bend or inflection in the trend). Linear contrasts were first inspected, across the age groups. If this trend was significant, the deviation statistic was examined, which indicated if it is necessary to further investigate the higher-order (i.e., quadratic trend). Where a non-linear trend was found, this was also followed up with a curve estimation, using age as a continuous variable. The variance explained by the linear and more complex trend were compared, using an extra-sum-of squares test. An extra-sum-of-squares test produces an F-value, which takes into account the improvement in model fit, along with the complexity of the models. A significant F-value indicates that the more complex trend is a better fit. Normality of the dependent variables was determined through inspection of normality tests. Normality was acceptable for all dependent variables and age groups, other than for mental rotation performance for 10 and 11 year olds. However, the groups were of sufficient size for the central limit theorem to apply; therefore, parametric tests were used.

A power analysis conducted in G*Power (Power = 0.8, $\alpha = 0.05$) indicated that a sample size of 212 was needed to detect a medium-sized effect (.25) for the most complex analysis (the ANOVA interaction between gender and age group). A medium effect size was chosen based on findings in previous research (e.g., Crestectini et al., 2014). The achieved sample (N=185) was therefore short of this total.

2.3.1.2 Descriptive statistics

Descriptive statistics for each spatial task are presented in Table 7. Percentage accuracy is reported for ease of comparison. There were no floor or ceiling effects in the data for any task, for each age group. Table 6 also includes a summary of the means and standard deviations for males and females for each task. Details of significant gender differences are reported in the respective section below.

Table 6: Percentage accuracy for males and females. Means and standard deviations in parentheses.

Task	Males	Females
Mental Rotation	77.78 (15.24)	75.25(19.06)
Mental Folding	55.47(22.82)	60.76(22.49)*
CEFT	50.45(18.82)*	44.47(18.25)
Spatial Scaling	59.87(20.87)*	53.82(20.09)
Perspective Taking	60.75(21.73)	60.68(21.94)

*Note: * indicates significantly higher mean accuracy for gender group*

Table 7: Descriptive statistics, by age group. Means and standard deviations in parentheses. Percentage accuracy is reported

	Max	**Range**	**6 years**	**7 years**	**8 years**	**9 years**	**10 years**	**11 years**
Mental Rotation	100	15.00-100.00	58.84 (11.90)	69.19 (17.36)	80.23(15.19)	82.95(11.60)	80.00(18.33)	88.10 (9.70)
Mental Folding	100	0-100.00	37.20 (14.82)	43.32(20.94)	66.29(15.63)	61.52(23.45)	66.36 (17.90)	73.73 (18.81)
CEFT	92	4.00-92.00	30.62(12.49)	35.87 (13.37)	50.88(16.45)	50.71 (15.95)	56.52 (17.94)	60.70 (16.53)
Spatial Scaling	100	11.11-100.00	37.78(13.95)	46.24 (19.53)	56.77(19.69)	64.34 (15.89)	68.46(14.80)	68.00 (19.73)
Perspective Taking	100	16.68-100.00	43.68 (13.82)	48.75 (16.34)	57.99(17.79)	66.48 (3.8)	71.15 (20.36)	76.82 (21.12)

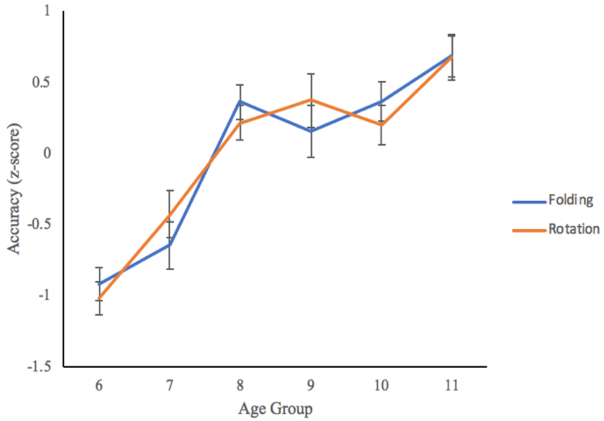


Figure 18: Mental folding and mental rotation accuracy by age group

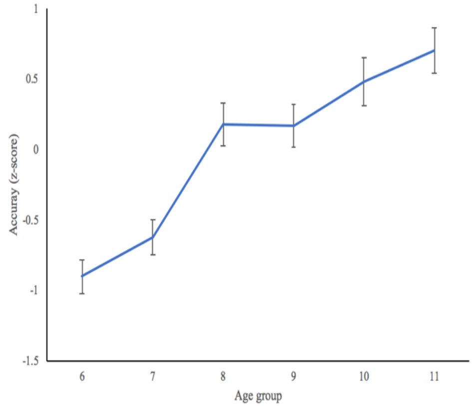


Figure 19: CEFT accuracy by age group

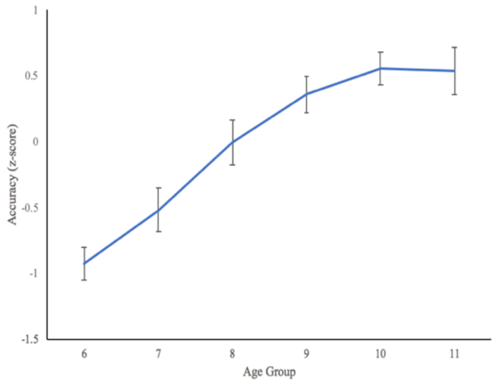


Figure 20: Spatial scaling accuracy by age group

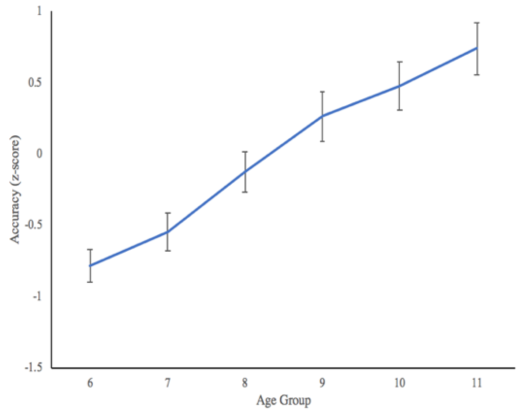


Figure 21: Perspective taking accuracy by

2.3.1.3 Main analysis of development of spatial skills

Beginning with intrinsic-dynamic skills; for the mental rotation task, the preliminary ANOVA with gender revealed no significant main effect of gender (p = .372), or interaction between gender and age group (p = .115), as such gender was removed from further analysis. For the main ANOVA, there was a main effect of age, F (5,178) = 16.31, p <.001, η_p^2 = .314. Games-Howell post-hoc tests revealed that the 8 year olds showed significantly higher accuracy than the 6 year olds (p < .001); see Figure 18. However, the performance of 7 year olds was not significantly higher than the 6 year olds (p = .087), or significantly lower than the 8 year olds (p = .094). There were no further significant improvements in accuracy beyond the age of 8 years. Polynomial contrasts revealed that the effect of age on mental rotation performance was described by a linear trend, F (1,178) = 67.04, p < .001. The deviation statistic indicated the existence of a more complex trend. Further analysis also revealed a significant quadratic trend, F (1,178) = 8.07, p = .005. Curve estimation using age in months revealed that a linear and quadratic trend accounted for a similar amount of variance (linear r^2 = .28, quadratic r^2 = .30), but, the quadratic trend was the better fitting model, F (1,181) = 7.65, p < .001).

For the mental folding task, the preliminary ANOVA with gender revealed a main effect of gender F (1,172) = 6.36, p = .013, η_p^2 = .040; the average accuracy of females was higher than of males (see Table 6). The interaction between gender and age group was not significant (p = .786), which justifies removing gender from the analysis (i.e. any effects of gender were uniform across age). The ANOVA revealed a main effect of age, Figure 18, F (5,178) = 17.82, p <.001, η_p^2 = .33. Bonferroni post-hoc tests further revealed there was a significant increase in accuracy between the ages of 7 and 8 years (p <.001). However, there were no significant differences between 6 and 7 year olds, or any further significant improvements in accuracy beyond the age of 8 years. As with mental rotation, a linear trend, F (1,178) = 73.50, p <.001, and quadratic trend F (1,178) = 4.73, p = .031, described the development of mental folding across the age groups. Curve estimation, revealed that although a linear and quadratic trend accounted for a similar amount of variance (linear r^2 = .30, quadratic r^2 = .32), the quadratic trend was the better fitting model, F(1,181) = 5.23, p = .023).

For the CEFT (intrinsic-static skill), the preliminary ANOVA revealed a main effect of gender *F* (1,172) = 4.55, *p* = .034, whereby the average accuracy of males was higher than of females. The interaction between and age group was not significant (p = .970), and thus gender was removed from the analysis. The main ANOVA revealed a main effect of age, *F* (5,178) = 17.29, *p* <.001, η_p^2 = .327, Figure 19. Bonferroni corrected post-hoc tests revealed that there was a significant improvement in performance between 7 and 8 years (*p* < .001). There were no significant differences between the ages of 6 and 7 (*p* > .99), and no significant improvement in performance accuracy beyond the age of 8. The effect of age had a significant linear trend only, *F* (1,178) = 79.42, *p* < .001.

For the spatial scaling task (extrinsic-static spatial skill) the preliminary ANOVA revealed a main effect of gender *F* (1,172) = 4.638, *p* = .033, η_p^2 = .026, whereby the accuracy for males was higher than the females. The interaction between gender and year group was not significant (p = .706) and thus gender was removed from the analysis. The main analysis revealed an effect of age group, *F* (5,178) =15.73, *p* <.001, η_p^2 = .306, see Figure 20. Bonferroni corrected post-hoc tests revealed that although there were no significant increases in accuracy between consecutive age groups, 8 year olds outperformed the 6 year olds (*p* < .001). There were no significant improvements between 8-year olds and older groups (8-9, *p* > .99; 8-10, *p* = .125; 8-11, *p* = .188). Increases in accuracy described by a significant linear trend only, *F* (1,178) = 72.10, *p* < .001).

For the preliminary analysis of performance on the perspective taking task, there were no significant main effect of gender (*p* = .764) and the interaction between gender and age group was also not significant (*p* = .972), and thus gender was removed from the analysis. The main ANOVA revealed a main effect of age group, *F* (5,178), = 14.67, *p* < . 001, η_p^2 = .89, Figure 21. Games-Howell corrected post-hoc tests revealed that performance improved between ages of 6 and 8 years. There were no significant differences in performance accuracy between 6 and 7 (*p* > .99), or 7 and 8 year olds (*p* = .800). Unlike the other spatial tasks, there was also significant improvement in accuracy beyond the age of 8, whereby the 11 year olds significantly outperformed the 8 year olds (*p* = .005). The effect of age was described by a significant linear trend only, *F*(1,178) = 72.46, *p* <.001.

To summarise, performance on the CEFT, mental folding, spatial scaling and mental rotation tasks showed significant development until the age of 8 years, with no significant changes thereafter. For the CEFT and mental folding tasks, there was a significant improvement in performance specifically between the ages of 7 and 8 years. The perspective taking task also showed development in the initial 6 to -8 year age window, but, also continued to show significant development beyond the age of 8, and therefore showed a significant amount of continued later development than the other skills. A linear pattern of development best characterised the data for all tasks other than for mental rotation and mental folding, where a quadratic trend best fit the data. This appeared to be driven by linear increases in development in early middle childhood, and a significant tapering off of performance increases at the age of 8 years.

2.3.2 Latent structure of spatial skills

2.3.2.1 Analysis strategy

Percentage accuracy was used as the dependent variable for all analyses. To analyse the relationship between observed spatial skills, Pearson correlations were first completed between all measures. Partial correlations were also calculated, between each skill, controlling for age in months as well as for the scores of the remaining spatial tasks. This provided an initial indication of unique variance between pairs of tasks.

To analyse the latent structure of the data, a Confirmatory Factor Analysis was conducted. This approach was used over an Exploratory Factor Analysis, as there were a priori predictions about the theorised latent structure of the observed data (see below). Three models of spatial thinking were tested, with a priori predictions that Model 2 would be the best fit, given the aforementioned evidence supporting it. Model 1, included all of the spatial tasks loaded onto one spatial factor, i.e., a model in which all of the spatial tasks were tapping the same underlying construct. This one-factor model acted as a ‘null’ model. In the main hypothesised model, Model 2, a two-factor structure of spatial thinking was tested, in which mental folding, mental rotation and the CEFT loaded onto an ‘intrinsic’ factor, and perspective taking and spatial scaling loaded onto an ‘extrinsic’ factor. In Model 3, an alternative hypothesised model was tested, where the CEFT and spatial scaling loaded onto a ‘static’ factor and mental rotation, mental folding and perspective taking loaded onto a ‘dynamic’ factor. Due to the number of tasks per sub-domain in the current study, it was not possible to test the four-factor model.

A χ^2 difference test used to compare the fit of the hypothesised models (see below), against this one-factor null model. In line with the convention within the CFA literature (Schermelleh-Engel, Moosbrugger, & Müller, 2003), if the difference in fit between the hypothesised models (i.e., model 2 and 3) and the null model (model 1) was not significant, the more parsimonious model was deemed the best fitting model (i.e., the one factor, model 1).

Each model used maximum likelihood estimation, with standardised latent factors allowing free estimation for all factor loadings. Several indicators of model fit were used, as is typical in CFA (Brown, 2014). The χ^2 statistic provides an indication of model fit, where a significant χ^2 statistic indicates a poor fit, contrary to typical hypothesis testing. The Standardised Root Mean Square Error Approximation (RMSEA) divides the total model error by the degrees of freedom and also adjusts by the sample size included. RMSEA is a 'badness of fit' measure and therefore smaller values indicate better fit. MacCallum, Browne, & Sugawara (1996) suggest the following criteria: .01 excellent fit, .05 good fit, .08 and above poor fit. The final fit indicator used was the comparative fit index (CFI). The CFI indicates the improvement of model fit over a baseline model in which covariances are zero. A CFI of at least .9 indicates a fair fit, and .95 indicates a good fit between the model and the data (Kline, 2015). Discriminant validity in CFA represents the degree to which identified factors are distinct. The inter-factor correlation provides an indication of the degree of discriminant validity between factors in the model. Very high correlations indicate that the factors in the model are not distinct constructs, and therefore demonstrate poor discriminant validity. Furthermore, a correlation of 1 or more indicates that the factors are indistinguishable. For this reason, within this analysis, where an inter-factor correlation value was 1 or greater, the model was considered inadmissible.

In addition, prior to the main analysis, to check for possible developmental differences in factor structure, measurement invariance was used. Measurement invariance indicates whether differences in factor structure exist between groups. This analysis involved splitting the sample into a younger (6-8 year olds) and older (9-11 year olds) group. Using measurement invariance, configural model invariance was tested across the groups, i.e., whether the factor structure fit consistently across the age groups. If the configural invariance analysis suggested group equivalence, the CFA was run on the whole sample, because there was no justification for separately analysing by age group. However, age was still controlled for in all models by including age in months as a covariate. This was achieved by fitting a structural equation model which included the CFA component, but also age in months scores. The main CFA models were fit using the residualised covariance matrix after partialing out age in months.

For CFA, simulation data reveals that 150 participants is the minimum required sample size for simple models with a small number of observed variables (Muthén & Muthén, 2002). Thus, the sample of 185 children in the current study is sufficient for the CFA analysis including the entire sample, given the small number of observed variables. For multi-group modelling, the rule-of-thumb is 100 participants per group (Kline, 2015). Therefore, for the models in which measurement invariance was used and the two age groupings were considered, the sample sizes were slightly less than the minimum required number of participants.

2.3.2.2 Correlation analysis

A summary of Pearson's correlations and partial correlations are given in Table 8. Other than perspective taking and paper folding, the correlations between the tasks were small to moderate, which suggests that there is a degree of shared processing, with some unique aspects to each skill. The lower triangle of the table also presents partial correlations between each skill, controlling for age in months as well as for the scores of the remaining spatial tasks. This provides an indication of unique variance between pairs of tasks. As is evident in the table, there were significant partial correlations, and therefore shared variance, between the paper folding task and the embedded figures, scaling task, and the perspective taking tasks. In addition, there were significant partial correlations between the mental rotation and scaling task, and also significant partial correlations between the scaling and perspective taking task.

Table 8: Zero order correlations (upper triangle) and partial correlations controlling for age in months and scores on remaining spatial tasks (lower triangle) between spatial tasks

	Rotation	Folding	CEFT	Scaling	Perspective Taking
Rotation		.50**	.37**	.56**	.46**
Folding	.17*		.57**	.58**	.61**
CEFT	-.05	.24**		.52**	.52**
Scaling	.20*	.20**	.15*		.58**
Perspective Taking	.07	.27**	.12	.23*	

* < .05 ** < . 001

2.3.2.3 Confirmatory factor analysis

Prior to the main CFA analysis, measurement invariance was analysed in relation to developmental differences across age. The configural model variance analysis indicated no significant difference in factor structure between the 6-8 and 9-11 year olds, when analysed with any of the three main models. Specifically, each model was tested with the age grouping variable included. For all three models, this resulted in a good fit (model 1- χ^2, [χ^2 (10) = 11.486], CFI: .99, RMSEA: .04; model 2: χ^2, [χ^2 (8) = 9.373], CFI: .99, RMSEA: .04; model 3: χ^2, [χ^2 (8) = 10.34], CFI: .99, RMSEA: .056), indicating that there was no evidence the factor structured differed significantly between the two age groups. Therefore, the main analysis includes the whole sample, and developmental differences in factor structure were not considered further.

A summary of the fit indicators for the main CFA models, for the whole sample, is provided in Table 9.

Table 9: Fit indicators for CFA for whole sample.

Model	χ^2	df	χ^2 *p*-value	CFI	RMSEA	χ^2_{diff}	χ^2_{diff} *p*-value
1-factor	4.62	5	.464	1.00	.000	-	-
2-factor (intrinsic-extrinsic)	4.45	4	.348	.999	.025	.17	.682
2-factor (static-dynamic)	4.61	4	.400	.999	.029	.01	.923

The one-factor and two-factor models all produced nonsignificant χ^2 values, indicating that the models did not significantly deviate from the data. The CFI values for the one-factor and two-factor models were > .95, indicating a good fit to the data. RMSEA values indicated an excellent fit for all models.

χ^2 difference tests revealed that the two-factor, intrinsic-extrinsic model (Figure 23) did not fit the data better (p = .682) than the one-factor model (Figure 22). Similarly, an χ^2 difference test revealed that the two-factor, static-dynamic model (Figure 24) did not fit the data better (p = .924) than the one-factor model (Figure 22). Therefore, the more parsimonious one-factor model was not rejected, and was thus the better fit.

Although the one-factor model provided the best fit the data, this single spatial factor did not account for 100% of the variance in any spatial task. The highest factor loading was mental folding (.59). The spatial factor accounted for 35% of the variance in mental folding scores.

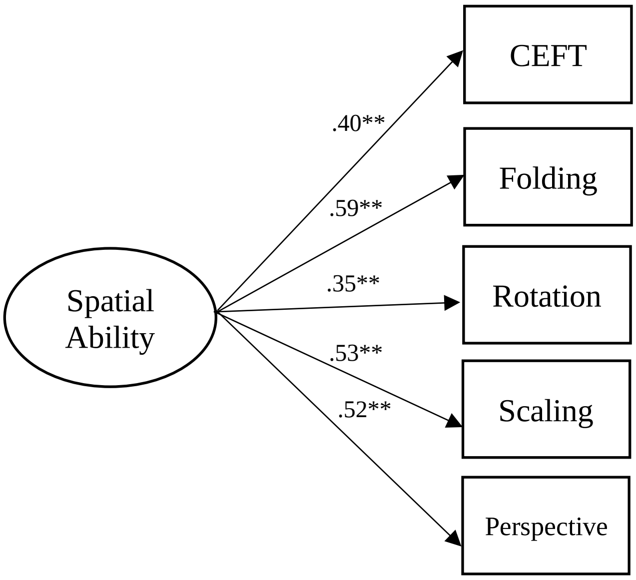


Figure 22: Model one, one-factor, baseline model.

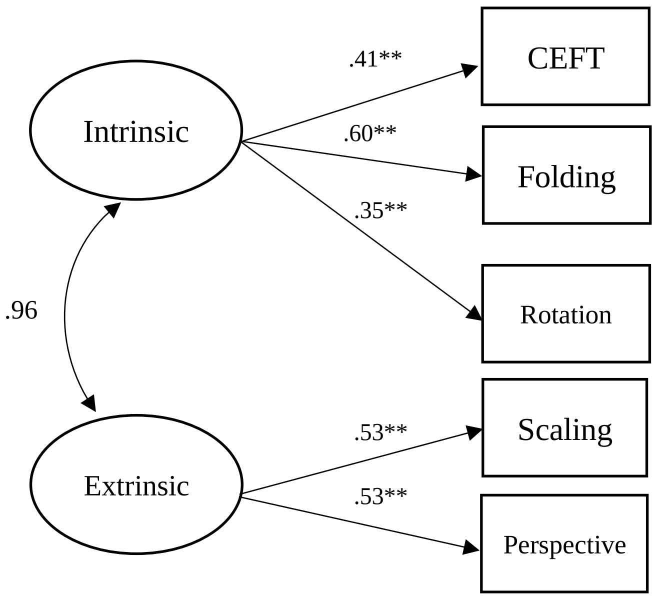


Figure 23: Model two, two-factor, intrinsic-extrinsic model

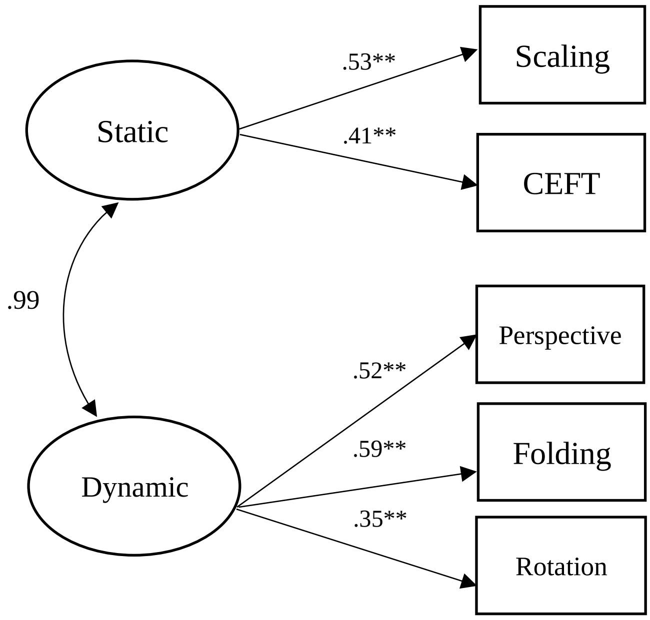


Figure 24: Model three, two-factor, static-dynamic model.

2.4 Discussion

The current chapter had two aims. The first aim was to investigate the developmental trajectories of a range of spatial skills throughout middle childhood, based on the theoretically-based Uttal et al. (2013) model. The second aim was to investigate the inter-relational, latent structure of these skills. The goal was to use the findings from each aim to inform each other, and to provide a novel source of evidence to further evaluate the Uttal et al. (2013) model.

In terms of developmental similarities, through pairwise tests, mental folding, mental rotation, the CEFT and spatial scaling all showed a significant increases in accuracy between the ages of 6 and 8 years, and no significant improvements from 8 to 11 years. Considering intrinsic-dynamic skills, the finding that 2D mental rotation showed development in this 6 to 8 year developmental window, with no significant development following until 11 years, is broadly in line with prior research by Crescentini et al. (2014) and Marmor (1975). However, it should be noted that Marmor (1975) had an adult control group for comparison. It is therefore not possible here, and below, to conclude from this data set that development ends at age 11.

A novel finding, however, was that mental folding, an intrinsic-dynamic mental transformation only investigated in children previously up to the age of 7 years (Harris et al., 2013), followed a similar developmental pattern to mental rotation. This suggests that both types of intrinsic-dynamic abilities broadly follow a similar developmental trajectory in middle childhood, despite differences between them (e.g., rigid versus non-rigid transformation types; but, see also latent structure discussion below).

However, development was apparent on the spatial scaling task, a discrimination paradigm, to a later age than has been previously shown in localisation paradigms (e.g., Frick & Newcombe, 2012). Within a localisation scaling paradigm, participants are shown a location on a map, and then asked to indicate the corresponding location, in a single scaled referent space. Frick and Newcombe (2012), for example, reported that 6 year old children demonstrated adult-level performance within a localisation paradigm. Thus, the later scaling development observed in the current study may reflect the increased demands of discriminating between several referent spaces, compared with the localisation paradigms, used in previous studies. Finally, the current findings contrast somewhat to previous research findings on intrinsic-static skills, assessed using the CEFT, which indicated that development continues until 9-10 years (Bigelow, 1971; Witkin, Otman, Raskin, & Karp, 1971).

There were, however, two main developmental differences between skills. First, only in the case of mental folding and the CEFT was there significant development between two consecutive age groups; specifically between the ages of 7 and 8 years. This suggests that, for these tasks, there is a steeper trajectory in the earlier part of the age range. Second, age group differences for the mental rotation task and the mental folding task were better described by a quadratic trend, although the relative difference in effect between a linear and quadratic fit was small. This suggests that for these tasks, there is an initially linear increase, followed by a significant attenuation of performance increases at the age of 8 years. This contrasts to the other spatial skills where development was described by a linear trend only.

For the perspective taking (extrinsic-dynamic spatial thinking) task, there were improvements in accuracy between the ages of 6-8 years, as with the other spatial tasks. However, in contrast to the other spatial skills assessed, there were significant group difference performance improvements beyond 8 years, to 11 years.

By amalgamating the above findings, one possible conclusion is that the intrinsic spatial thinking tasks tended to show more rapid development in the early part of middle childhood, whereas the extrinsic tasks show more steady development throughout, or in the case of perspective taking, a larger amount of development into later, middle childhood. In terms of the Uttal et al (2013) model, this developmental analysis therefore supports the intrinsic vs extrinsic dimension. However, this conclusion is based on some disparate findings. For example, although the CEFT did show significant development specifically between the ages of 7-8, and no significant development between the ages of 8 to 11, the trend analysis did not reveal a significant deviation from linearity. At the very least, there are differences in the development of the intrinsic-dynamic object-based transformations, and extrinsic-dynamic, perspective taking ability. Because this was one of the main predictions, it will be discussed further below.

The finding that the developmental trajectories of perspective taking and mental rotation, in particular, differed somewhat, is in line with previous research by Crescentini et al. (2014), who found that mental rotation developed earlier (from age 7 years) than perspective taking (from age 8 years). The finding that perspective taking continues to develop into late childhood also supports prior findings from Xistouri and Pitta-Pantazi (2006). As outlined above, the finding that mental folding, specifically, also differed from perspective taking, is a novel finding.

These findings can be explained within the spatial frames of reference theory outlined in the introduction. Object-based transformation (intrinsic-dynamic tasks) require the imagined manipulation of an object, from a maintained egocentric perspective. However, when performing the spatial perspective taking task in the current study, children were required to allocentrically encode the array of objects, encode the character with the camera, and, also, the spatial relationship between them. As outlined in the introduction, in contrast to data from mental rotation, the response times adults show within perspective taking tasks typically do not show a linear trend with orientation difference, at least for angles up to 90^{o} (Kessler & Thomson, 2010; Zacks & Michelon, 2005). Therefore, for lower angles of perspective change, adults do not necessarily perform a continuous rotation of the self, but, may rely more so on an allocentric framework to jump to another location (‘a blink’; Wraga et al., 2000). Data from younger children do sometimes indicate a linear relationship between angular disparity and response time (Roberts & Aman, 1993), thus suggesting that they may use a strategy such as graduated mental rotation of the self, around the array (i.e., a more egocentric strategy). However, even in this case, in order to determine how the array might look, the participant must update their egocentric perspective, but, within the encoded allocentric framework. The implication of these findings is that more mature performance involves greater utilisation of an allocentric framework.

Effective performance on the perspective taking task, more so than the mental rotation task, depends on the utilisation of an allocentric framework, and the ability to switch between and update egocentric and allocentric frameworks. As discussed, the spontaneous use of allocentric strategies increases progressively throughout middle childhood (Bullens et al., 2010). Moreover, this cognitively effortful process also requires the maintenance and updating of information in mind, and thus additional executive function demands, such as working memory. In addition, executive functions (e.g., working memory) continue to develop throughout middle childhood and into adolescence (Gathercole, Pickering, Ambridge & Wering, 2004). Within the perspective taking task in the current study, participants were also required to directly inhibit their fixed egocentric perspective, when responding, because one possible response option was an image of the array, taken from their own egocentric perspective. In line with this, prior research with 6 year olds found that inhibitory control, in particular, predicted perspective taking performance, controlling for covariates, including verbal IQ (Frick & Baumeler, 2017). Notably, however, inhibitory control was not predictive of performance on any other spatial task (e.g., mental rotation). Thus, the more linear development of perspective taking, and continued development of accuracy beyond the age of 8 years, may be a reflection of an increased ability in later childhood, to use allocentric frameworks, and also to switch from, and inhibit, an egocentric frame of reference.

The analysis of the latent structure of spatial skills across the two age groupings provides additional insight into the validity of the two by two model. The results of the CFA revealed that neither of the two-factor models fit the data significantly better than a one-factor model, and there was no evidence that this pattern varied significantly across development. There was therefore no significant evidence of the predicted distinction between intrinsic and extrinsic spatial skills, in the current study. However, the single spatial factor did not account for more than 35% of the variance in any spatial task. Thus, the spatial tasks must measure additional capacities beyond a common spatial skill.

These findings contradict previous research by Vander Heyden et al. (2016), found a significant dissociation between intrinsic-dynamic and extrinsic-dynamic skills for 10.5 year olds, and not younger children, aged 7.5-10.5. The findings in the current study are also not in line with the reported findings of Mix et al. (2018) who found a dissociation along the intrinsic/extrinsic dimension (which also included both static and dynamic skills), for 9 year olds.

This discrepancy between studies may be due to methodological factors. For example, Mix et al. (2018) had a larger sample size, and more spatial tasks loaded were loaded onto each factor. It may have been the case that there was greater statistical power to detect a difference between the one and two factor models in Mix et al. (2018). A second reason for the discrepancy between the results relates to strategy use. As outlined in section 1.2.2.2, it is possible to solve spatial thinking task, particularly mental rotation and perspective taking tasks, using several different strategies (Schultz, 1991). It is possible that tasks in the current study which were considered ‘intrinsic’ may have sometimes been solved using an ‘extrinsic’ strategy, or vice-a-versa. For example, perhaps participants rotated the array, rather than adopted a change in perspective, for the perspective taking task. Given that children’s allocentric skills develop until the age of 11, is possible that some children in the age range of this study were more confident using this type of object-based strategy. Future work might more closely investigate children’s spatial strategy use in a range of spatial thinking tasks.

A remaining issue concerns mental rotation. Although mental rotation and mental folding showed similar developmental trajectories, the CFA suggested that they were measuring partly distinct skills. Mental folding loaded more strongly within the CFA models. In addition the spatial factor, or factors, accounted for 13% of the variance in mental rotation scores, compared to 35% of the variance in mental folding. Given the low factor loading, it is possible that this type of rotation task might load onto a separate factor altogether, given a larger number of similar tasks (as listed in ‘spatial relations’, within the Carroll [1993] model). The mental rotation task itself is discussed further in the next chapter.

It is also worth noting that the developmental results were not consistent with all previous studies. In particular, Vander Heyden et al. (2016) found that the development of intrinsic-dynamic spatial skills occurred somewhat later (i.e., mental rotation 8-9/10 years; mental folding: 11-12 years) than observed in the current study.

However, they measured these skills using tasks typically administered to adults, which require more complex transformations and stimuli, i.e., performing a sequence of folds; rotating abstract 3D cubes. It could be argued that their tasks measured a more sophisticated level of mental transformation skill. This is reminiscent of the apparently contradictory findings of mental rotation skills between infants and young children (Frick, Möhring, & Newcombe, 2014a). That is, infants often show surprisingly good mental rotation skills (e.g., Moore & Johnson, 2008), whereas very young children perform poorly (e.g., 3 year olds in Frick, Hansen, & Newcombe, 2013). This has been taken to reflect the differences in the paradigms used for the two age groups, and perhaps different levels of rotation skill. Specifically, the demands of the rotation tasks used within infants are quite distinct from the demands of the tasks used with 3 to 4 year olds.

The current data also revealed gender effects, with small effect sizes, for the CEFT and the spatial scaling task in favour of males, and the folding task, in favour of females. However, there were no gender effects for the perspective taking or mental rotation tasks. This adds to the mixed picture of gender effects on spatial skills in childhood, which largely show either small, significant effects in favour of males (e.g., Neuburger, Jansen, Heil, & Quaiser-Pohl, 2011) or no significant differences between males and females (e.g., 9 year olds in Titze, Jansen, & Heil (2010); 5-7 year olds in Harris et al., 2013). To the best of our knowledge, this is the first study to assess gender differences in mental folding ability throughout middle childhood. The finding that no sex differences emerged in favour of males, and in fact, accuracy was significantly higher for females, fits with adult data which suggests that sex differences for mental folding in favour of males, typically do not exist (Linn & Peterson, 1985).

In conclusion, the current study showed that although spatial skills develop in middle childhood broadly within the 6-8 year age range, there were some differences between skills. In particular, the intrinsic tasks (2D rotation, mental folding and disembedding) showed particularly rapid early development in this range. In contrast, perspective taking showed further performance gains between the ages of 8 to 11, and thus continued to develop into the end of middle childhood. This is likely to be a reflection of the additional demands involved with this task. In particular, the need to update and switch between egocentric and allocentric spatial frameworks. There was also some non-linearity evident for the intrinsic-dynamic tasks, which appears to be a reflection of a significant tapering off of development from 8 years. Despite somewhat different trajectories, there was no evidence for a clear psychometric dissociation between intrinsic and extrinsic skills in this age range. In addition, the mental rotation task had a low factor loading on both factors, for the older children. Thus, although a one-factor model provided the best fit, the data suggests that a two or more factor structure might be evident given a wider range of tasks (e.g., several extrinsic dynamic tasks) and or, a larger sample.

In terms of the Uttal et al. (2013) model, the developmental data therefore provides support for the intrinsic vs extrinsic dimension, but not the static vs dynamic, or the full 2x2 model. However, the psychometric data did not support any of the Uttal et al. (2013) dimensions. The latter finding should be considered in light of the methodological considerations outlined above and below.

With respect to limitations of the current study, the use of a longitudinal design over the time period would have provided a more precise indication of the development of each skill for individual children. Additionally, including an adult control group in the analysis would have meant that conclusions could have been made regarding whether development 'ends' at the age of 11. Considering the CFA, a larger sample size may have revealed a significant factor dissociation. Future research might also use a larger number of spatial skills, which may result in a clearer factor structure. In particular, it was not possible to test the full four-factor model; however, in the Mix et al. (2018) study, such a model was consistently inadmissible. It is worth noting that it may not even be practical to test the 2x2 model, particularly with younger children. In order to adequately test the model, 3-4 spatial thinking measures would be needed for each sub-domain, resulting in a test battery of 12-16 measures.

These findings have implications for the training of spatial thinking skills. First. the possibility of training spatial skills is generally supported by the degree of individual variation in performance shown throughout development in this study, for the majority of the spatial tasks. Furthermore, periods of strong development may be when skills are most malleable and affected by external input, and therefore, better time windows for training (Zelazo & Carlson, 2012). The pattern of development shown in the current data specifically suggest that training of mental folding and disembedding skills, via the CEFT, between the age of 7 and 8 years, may lead to the greatest gains within middle childhood. Performance increased significantly for both of these tasks, between these consecutive ages. Furthermore, mental folding was also the strongest loading spatial task on the identified spatial factor, and thus training of this skill is likely to reap the highest rewards with respect to improved spatial thinking

In the next chapter, the role of these spatial skills are considered as a predictor of science achievement within middle childhood. Although the data in the current chapter suggests that a one-factor model provides the best fit, the skills are considered separately. The single spatial factor accounts for between 19%-35% of the variance in the various spatial tasks. There may therefore be unique aspects to certain tasks that might make them stronger predictors of science achievement, despite the tasks as a whole sharing a large amount of variance

Chapter 3

The contribution of spatial ability to science achievement in middle childhood

3.1 Introduction

3.1.1 Spatial ability and science learning

Large-scale longitudinal studies spanning the past 50 years provide convincing evidence that spatial ability in adolescence predicts later science, technology, engineering and mathematics (STEM) achievement (Lubinski & Benbow, 2006; Wai, et al., 2009). In addition to often-cited examples of scientific discoveries resulting from creative spatial thought, a growing body of research with adults and adolescents highlights a more specific link between spatial ability and various aspects of science learning (e.g., Delialioğlu & Askar, 1999). However, the relationship between spatial ability and science learning in younger children has been largely neglected. A deeper understanding of this relationship at an earlier stage of development is important because it has implications for early curriculum design, informs the development of spatial training interventions, and has the potential to support learners when they are at more advanced stages of science education. The focus of this study was therefore on the relationship different spatial ability and science achievement in middle childhood.

The first specific aim was to investigate whether spatial thinking skills were generally predictive of science achievement in 8-11 year olds. The second aim was to determine which specific spatial thinking skills were the strongest predictors of science achievement for 8-11 year olds. To achieve these aims, the 8, 9, 10 and 11 year olds from Chapter 1 completed a paper-based, curriculum linked science assessment. The various spatial skills reported in Chapter 1 were then considered as predictors of performance on this science assessment.

The study also had the additional aims of investigating whether the relationship between spatial thinking and science varied by age, and whether this relationship varied by conceptual domain of science. In relation to the first of these, the four sampled age groups meant that it was possible to investigate possible interactions between spatial thinking skills and age group. In relation to the second of these aims, the science assessment contained content spanning biology, chemistry and physics, meaning it was also therefore possible to determine whether the predictive power of the spatial tasks varied by conceptual science domain.

The following sections review the existing literature on the relationship between spatial ability and science learning in adults and children, and then review the approach to science assessment used in the current study.

3.1.2 Spatial ability and adult's science learning

As outlined in Chapter 1, spatial skills may particularly support learning, problem-solving and reasoning within conceptual science areas that have a clear spatial-relational basis (e.g, astronomy and mechanics). Most prior research with adults points to spatial visualisation skills as being related to science learning. Spatial visualisation involves mentally transforming object-based spatial information, and is assessed through intrinsic-dynamic spatial skills such as mental rotation. Existing research with adults suggests a link between intrinsic-dynamic spatial skills and conceptual understanding in aspects of biology (Rochford, 1985; Lufler et al., 2012), chemistry (Stull, et al., 2012, Baker & Talley, 1972; Bodner & McMillan, 1986), and physics (Kozhevnikov & Thornton, 2006; Kozhenviokov, et al., 2002). For example, in Stull et al. (2012) spatial ability, as measured through 3D object visualisation, correlated with undergraduate students' ability to translate between different diagrammatic representations of chemical structures. There is also some evidence linking adults' chemistry performance to disembedding (intrinsic-static) spatial skills (Bodner & McMillen, 1986) and undergraduate's geoscience understanding to multiple-object (extrinsic-dynamic) spatial skills (Sanchez & Wiley, 2014). However, no research to date has addressed other skills, such as spatial scaling ability, in relation to science learning.

3.1.3 Spatial ability and younger children's science learning

Research relating spatial ability and science learning in younger children is sparse, and some studies that have addressed this have done so only in relation to visual-spatial working memory, or a limited range of spatial skills. Two studies (Jarvis & Gathercole, 2003; St-Clair Thompson & Gathercole, 2006) focused on 11-year olds' achievement in UK national standardised science tests in relation to working memory. The findings of both studies pointed towards the visual-spatial working memory task as being predictive of performance in science. However, because these tasks are designed to test both visual and spatial aspects of spatial thinking, complex working memory span tasks often confound object/visual, and location/spatial skills. In the odd-one-out task (St-Clair Thompson & Gathercole, 2006), for example, participants distinguish between shapes based on visual form (intrinsic-static) but also remember the location of the odd item, relative to the other two items (extrinsic-static).

A few studies to date have examined children's science performance and learning in relation to other spatial skills (e.g., Harris et al., 2018; Mayer, et al., 2014; Tracy, 1990). Tracy (1990), for example, found that 10-11 year olds in a higher spatial ability grouping outperformed those in a lower spatial ability grouping on a standardised science measure. However, this study did not include any other non-spatial cognitive measures, and therefore did not discount such cognitive factors as an alternative explanation. It also used a composite spatial measure.

One more recent study did compare different spatial ability measures (Harris et al, 2018), in relation to children's science learning. This key study is particularly important because it includes more than one type of spatial thinking measure and compares between them, and because the study included receptive vocabulary as a proxy control of general ability. The study involved 5.5 year olds and adults completing a game which involved making judgments about force and motion. Participants also completed age-appropriate versions of a mental folding and a mental rotation task (2D for children; 3D for adults). The results revealed that, controlling for receptive vocabulary, mental folding moderately predicted children's ability to make judgments about force and motion, whereas mental rotation did not. For adults, neither mental rotation or mental folding significantly correlated with performance. The study therefore provides some initial indication of a connection between children's spatial thinking and reasoning within the domain of physics, in particular. Although the study included two measures of spatial thinking, it is still limited in including only intrinsic-dynamic transformation skills. This study is also important because it emphasises the need to consider the distinctiveness of mental rotation and mental folding within the intrinsic-dynamic domain.

3.1.4 Changes in the relationship between spatial ability and science at different stages of learning

Spatial skills may be more important for individuals at an earlier stage of learning than those in later stages (Uttal & Cohen, 2012). During initial learning, or for individuals with lower levels of domain-specific knowledge, a learner may use spatial processing to establish mental maps and models, or to problem solve (Mix et al., 2016). In line with this, for example, Hambrick et al. (2012) found that spatial ability interacted with adults' level of geological knowledge in a geology task in which participants inferred the geologic structure of a mountain range. Specifically, spatial ability was more predictive of performance for participants who had lower levels of geologic knowledge, whereas for those with more domain-specific knowledge, spatial skills were less important.

Developmentally, this hypothesis is also supported by the finding, described above, that mental folding ability, an intrinsic-dynamic skill, predicts children's, but not adult's, understanding of forces (Harris et al., 2018). One possible interpretation of this finding is that younger children must actively visualise the effects of forces to make predictions, whereas adults rely more on knowledge of forces and their effects, which has accumulated over time. The above findings suggest that spatial skills may therefore play a more important role in science achievement for younger compared with older children; however, this has yet to be addressed empirically.

3.1.5 Science assessment approach

In the current chapter, a curriculum-based approach to science assessment was adopted. The UK science curriculum includes the previously outlined aspects of factual knowledge, conceptual understanding and scientific investigation (Department for Education, 2013a). Thus, this approach also aligns with Wellington's definition of science knowledge (1988). Within the UK science curriculum, objectives related to 'factual knowledge and conceptual understanding' are grouped into sub-topics (e.g., plants), which are linked to one of the broader conceptual domains of biology, chemistry or physics. Within the curriculum, 'working scientifically' includes skills such as asking questions and making predictions, designing controlled experiments, making observations and taking measurements, recording findings and data and drawing conclusions. As described in Chapter 1, the curriculum also emphasises that opportunities to work scientifically should be integrated within the conceptual topic areas. Science achievement in the current study was therefore assessed using a composite assessment of factual knowledge, conceptual understanding and investigation skills taught in the age range of interest. A curriculum-based approach has the advantage that it covers the breadth of knowledge and skills children learn in the classroom. Such an approach has also been successfully adopted in the past, for example, in studies investigating the role of executive functions on children's performance in national science assessments (Jarvis & Gathercole, 2003; St Clair-Thompson & Gathercole, 2006).

3.1.6 Current study

The aim of the current study was to examine the relationship between 8-11 year olds' spatial skills and their performance in a science assessment, which covered aspects of biology, chemistry and physics knowledge as well as scientific investigation skills within these areas. School year groups in the UK are further grouped into larger curriculum-linked 'key stages'. Children in Years 3 to Year 6 (aged 7-11)[1] are grouped into 'Key Stage 2'. Children from each year group within Key Stage 2 were sampled, which meant that the children in the sample were working towards the same overall curriculum objectives. By using a range of ages, the aim was to determine if this relationship was moderated by age. Given the dearth of literature on the relationship between children's spatial skills and science reasoning, it is difficult to make specific predictions. Based on the findings of Harris et al. (2018), it was predicted that, minimally, intrinsic-dynamic skills would be related to science performance, and, this relationship may be stronger for younger children.

3.2 Methods

3.2.1 Participants

The participants in the current study were the 8, 9, 10 and 11 year olds from the previous chapter. There was no additional missing data to that reported in the previous

[1] Although the widest age range for Years 3 to 6 is 7-11 years, at the time of data collection, the average age of children in years 3-6, was 8,9,10 and 11 years. Hence, in the previous chapter and this chapter, for consistency, these age groupings are referred to. However, they also map onto Years 3-6, as curriculum year groups.

chapter. The sample therefore consisted of 123 children. A power analysis conducted in G*Power (Power = 0.8, α = 0.05) indicated that a sample size of 127 was needed to detect a small to medium effect (f^2 = .12) for the basic analysis (the multiple linear regression analysis with 7 possible predictors). A small to medium-sized effect was chosen rather than a medium effect size because of the number of spatial variables included and the shared variance between them. For the models with the additional age-based interactions, 165 participants were needed to detect a small to medium effect. The achieved sample was therefore sufficient for the main analyses, but was underpowered for the additional age-based interaction analyses.

3.2.2 Measures

3.2.2.1 Spatial measures

The same five spatial measures (mental rotation, mental folding, CEFT, perspective taking, spatial scaling) were used as reported in Chapter 2.

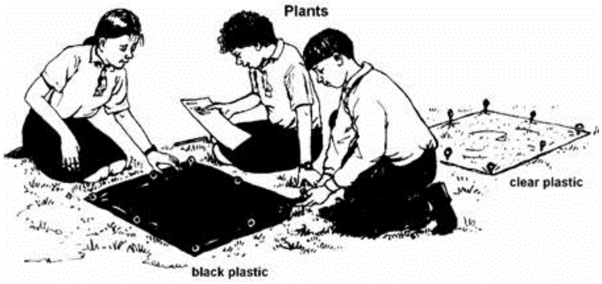
3.2.2.2 Science assessment

The science assessment consisted of two paper-based tests, which children completed in two sessions, in class groups, under the supervision of the researcher. All questions were read to participants by the researcher. The assessment was a composite, curriculum-based measure, and questions were taken from a selection of past science UK standardised ('SATS') test papers designed to assess science achievement in this age range (e.g., Qualifications and Curriculum Authority [QCA], 1996, 2009). The test included approximately equal numbers of biology, chemistry and physics focused questions on topics appropriate to this curriculum stage ('Key Stage 2': age 7-11).

Each paper had a total possible score of 50 marks leading to a total science mark of 100. The assessment included questions which varied in difficulty. The difficulty level of each question was determined by the categorisation given in the testing materials, which is linked to curriculum target descriptors. Paper one contained questions of low to medium demand and paper two contained questions of high demand. Paper one contained 11 questions and paper two contained 10 questions. Each question focused on one sub-topic, e.g., magnets (see Table 10 for topics). Each question was divided into several sub-items (approximately 4 per question). Some items were more factual/recall based (e.g. what is the function of the roots of a plant?), others required more conceptual understanding (explain why the bigger sail makes the boat go faster) or were more problem-solving-based. Some items in the context of hypothetical experiments, related to the sub-topic, required investigation skills (e.g. make a prediction…). There was a mixture of free response and multiple-choice items throughout. An example of an item requiring conceptual understanding (Figure 25) and factual recall (Figure 26), in the context of a hypothetical experiment, are provided below. The two papers had good levels of internal consistency as measured by Cronbach's $\alpha = .841$ (paper 1) and $\alpha = .794$ (paper 2), across all items. A second coder scored a random 10% of the first and second papers and demonstrated a high degree of inter-rater reliability with the first coder ($r = .99$, $p < .001$).

Table 10: Summary of sub-topics included in the science assessment

Biology	Chemistry	Physics
Plants (functions of parts, seed dispersal, life cycle)	Properties of materials	Light (shadows, reflections)
Human skeleton	Changing state (condensation, melting and evaporation)	Sun, earth and moon
Human growth and development	Reversible and non-reversible changes	Gravity and forces
Classifying and sorting animals	Rocks	Electricity



The children left two squares of plastic on healthy grass for five days.

The chart shows what they found when they lifted the plastic.

type of plastic	**colour of grass**
black	yellow
clear	green

Why was the grass **yellow** where the black plastic had been?

....
....

....
....

Figure 25: Biology item from paper 1. Conceptual understanding focus. Source: QCA (1996).

What do all plants need to grow well?

Tick **TWO** boxes.

pots	☐	insects	☐	worms	☐
water	☐	rocks	☐	air	☐

Figure 26. Biology item from paper 1. Factual knowledge focus. Source: QCA (1996).

3.2.2.3 Control variables

Vocabulary is highly correlated with overall general intellectual ability (Sattler, 1992); therefore, the British Picture Vocabulary Scale-III (Dunn & Dunn, 2009) was included as a measure of verbal ability, but also serves as an estimate of general intelligence. The experimenter read a word to the child, who then matched it to one of four pictures. The words became increasingly difficult and testing was discontinued when the child made 8 errors within one set.

3.2.3 Procedure

Children completed the two paper-based science assessments, in two separate sessions. Sessions lasted approximately 45 minutes each. Science assessments were administered by the researcher in whole class groups, within the child's own classroom. The science assessments were completed before the spatial data, reported in Chapter 2, was collected.

3.3 Results

3.3.1 Descriptive statistics

A total science score was calculated by totalling the participants' scores across both paper 1 and paper 2. A total for biology, chemistry and physics questions across both papers was also calculated. Percentage accuracy was used for all spatial tasks (as reported in Chapter 2). Descriptive statistics are presented in Table 11.

3.3.2 Correlation analysis

Bivariate Pearson's correlations were also analysed between the predictive variables (BPVS, age and spatial ability measures) and the dependent variables (total science score and biology, chemistry and physics sub-scores), which are reported in Table 12. Partial correlations, controlling for age and BPVS raw scores, between each of the spatial measures and each of the science totals are reported in the lower triangle of Table 12.

Controlling for these covariates, mental rotation did not correlate with any science variable. The mental folding task, the embedded figures task and the scaling task had small to moderately sized partial correlations (range: $.211 < r < .384$) with total science scores and biology, chemistry and physics scores. Perspective taking scores also had small to moderately sized positive partial correlations (range: $.229 < r < .295$) with all science variables other than chemistry scores, where there was no significant correlation.

Table 11: Descriptive statistics for science total scores, BPVS raw scores and spatial measures. Maximum possible score in parentheses. Percentage accuracy is reported.

Measure	M	SD	Range
Correct overall science score (100)	43.97	14.60	7-75
Correct overall science score, Age 8 (100)	35.75	10.87	7-51
Correct overall science score, Age 9 (100)	41.42	14.78	14-72
Correct overall science score, Age 10 (100)	47.26	14.31	18-71
Correct overall science score, Age 11 (100)	52.24	13.31	21-75
Correct overall biology score (36)	18.63	6.17	3-33
Correct overall chemistry score (32)	13.11	5.03	1-26
Correct overall physics score (32)	12.91	5.56	2-29

3.3.3 Regression analysis

Regression analyses were run for overall science scores and for biology, chemistry and physics scores. Assumptions for linear regression were met (normality of residuals, no significant collinearity, homoscedasticity of residuals, no significant influential cases). There were no significant gender differences in any science scores ($p > .05$ for all); therefore, participants were treated as one group in the subsequent regression analyses. A hierarchical and stepwise approach was taken to determine the amount of variance in science outcomes that was accounted for by participants' spatial ability, taking into account the covariates (age and BPVS raw score). In all regression models, covariates were added hierarchically first. Betas reported refer to the final models (Tables 13-16).

Entered in the first step of each model, age in months significantly predicted overall scores and scores for individual science areas. Age remained a significant predictor in the final model for overall science scores and physics scores. However, age was not significant in the final model for biology or chemistry. Participants' BPVS raw score was entered in the second step of each model and was a significant predictor of all science outcomes. BPVS scores remained a significant predictor in all of the final models.

Following entry of age and BPVS scores, the predictive role of the spatial ability measures were then considered. All spatial predictors found to be significantly associated with the respective science score in the prior partial correlation analysis were entered together as a block using forward step-wise entry. Forward step-wise entry was used due to the inter-relatedness of the spatial variables, and because there were no strong theoretical predictions about the basis for a hierarchical ordering of variables within this block.

Table 12: Zero-order correlations between study variables (upper triangle) and partial correlations (lower triangle).

	1	2	3	4	5	6	7	9	10	11	12
1. BPVS raw score	-	44**	.75**	.64**	.66**	.63**	.27**	.29**	.20*	.40**	.42**
2. Age	-	-	.45**	.33**	.42**	.53**	.20*	.20*	.26**	.26**	.37**
3. Science overall total	-	-	-	.88**	.87**	.88**	.29**	.47**	.37**	.51**	.50**
4. Biology total	-	-	-	-	.76**	.71**	.26**	.42**	.31**	.46**	.48**
5. Chemistry total	-	-	-	-	-	.71**	.24**	.35**	.32**	.43**	.42**
6. Physics total	-	-	-	-	-	-	.28**	.40**	.35**	.43**	.47**
7. Mental Rotation	-	-	.12	.12	.07	.12	-	.29**	.07	.22*	.42**
9. Mental Folding	-	-	.38**	.31**	.21*	.28*	-	-	.41**	.41**	.46**
10. CEFT	-	-	.31**	.23*	.23*	.25**	-	-	-	.31**	.36**
11. Spatial Scaling	-	-	.33**	.29**	.23*	.22*	-	-	-	-	.52**
12. Perspective Taking	-	-	.28*	.30**	.18	.23*	-	-	-	-	-

Note. *$p < 0.05$, ** $p < 0.01$. Upper triangle shows zero-order correlations and lower triangle shows partial correlations between spatial measures and the science total score, controlling for BPVS raw score and age in months.

Table 13: Multiple regression analysis predicting science total score.

Predictor	*b*	ß	*p*	Δ F	Sig Δ *F*	R^2	$R^2\Delta$
Step 1) Age (months)	.13	.12	.044	31.3	<.001	.21	.21
Step 2) BPVS raw score	.41	.57	<.001	106.2	<.001	.58	.37
Step 3) Mental Folding	1.14	.21	.001	20.6	<.001	.64	.06
Step 4) Spatial Scaling	.74	.16	.010	6.8	.010	.66	.02

Table 14: Multiple regression analysis predicting biology score.

Predictor	*b*	ß	*p*	Δ F	Sig Δ *F*	R^2	$R^2\Delta$
Step 1) Age (months)	.02	.03	.648	15.10	<.001	.11	.11
Step 2) BPVS raw score	.15	.50	<.001	60.38	<.001	.41	.30
Step 3) Mental Folding	.45	.20	.008	12.77	.001	.47	.06
Step 4) Spatial Scaling	.33	.17	.025	5.13	.025	.49	.02

Table 15: Multiple regression analysis predicting chemistry score.

Predictor	*b*	ß	*p*	Δ F	Sig Δ *F*	R^2	$R^2\Delta$
Step 1) Age (months)	.05	.12	.103	26.09	<.001	.18	.18
Step 2) BPVS raw score	.13	.52	<.001	60.52	<.001	.45	.28
Step 3) CEFT	.17	.14	.046	6.47	.012	.48	.03
Step 4) Spatial Scaling	.23	.15	.049	3.95	.049	.50	.02

Note. Beta values refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model.

Table 16: Multiple regression analysis predicting physics score

Predictor	*b*	*ß*	*p*	Δ F	Sig Δ *F*	R^2	$R^2\Delta$
Step 1) Age (months)	.12	.30	<.001	47.28	<.001	.28	.28
Step 2) BPVS raw score	.12	.44	<.001	44.98	<.001	.48	.20
Step 3) Mental Folding	.43	.21	.002	9.78	.002	.52	.04

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model

The forward entry of spatial measures predicting overall science score retained mental folding and spatial scaling. Mental folding accounted for an additional 6% of the variance in total science score, ΔF (1,119) = 20.62, $p < .001$, and spatial scaling then accounted for a further 2% of the variance in total science scores, ΔF (1,118) = 6.79, $p = .010$, above the covariates. In the final model, which accounted for 65% of the variance in total science scores (adjusted r^2), mental folding was a stronger predictor (*ß* =.21) than spatial scaling (*ß*= .16).

Forward entry of the spatial measures predicting biology scores also retained mental folding and spatial scaling. After step 2, mental folding accounted for an additional 6% of the variance in biology scores, ΔF (1,119) = 12.77, $p = .001$ and the spatial scaling task accounted for an additional 2% of the variance in biology scores ΔF (1,118) = 5.13, $p = .025$. The overall model accounted for 47% of the variance in biology science scores (adjusted r^2). Mental folding was a stronger predictor (*ß* =.20) than scaling (*ß*= .17) in the final model.

The CEFT was retained as a significant spatial predictor of chemistry scores accounting for a further 3% of the variance in chemistry scores, ΔF (1, 119) = 6.47, $p = .012$, above the covariates. In addition, the spatial scaling task was also retained as a predictor of chemistry scores, which accounted for an additional 2% of the variance, ΔF (1, 118) = 3.95, $p = .049$. The final model accounted for 48% of the variance in participants' chemistry total score (adjusted r^2). The two spatial skills in this model had similarly sized *ß* coefficients: embedded figures, *ß* =.14; scaling *ß*= .15. Mental folding was the only retained predictor of the physics scores. It was entered in step 3 and it accounted for an additional 4% of the variance in physics scores, $\Delta F(1,119) = 9.78$, $p = .002$. The final model accounted for 51% of the variance in physics scores (adjusted r^2).

To determine if age interacted with any of the spatial ability measures, and therefore if this pattern varied across the age groups, a further four models were constructed in which the covariates were again entered in step 1, followed by the spatial ability measures found to be significant for that science score, followed by an interaction term (age in months * spatial measure). No significant age interactions were found ($p > .05$ for all).

3.4 Discussion

The aim of the current study was to examine the contribution of spatial skills to primary-school children's performance in a curriculum-based science assessment. The study revealed overall that spatial ability is a predictor of 7-11 year olds' science achievement. After controlling for receptive vocabulary, which provided an estimate of general intelligence, spatial ability accounted for an additional 8% of the variance in total science scores. This builds upon longitudinal research linking spatial ability to STEM outcomes in adults (Wai et al., 2009; Lubinski & Benbow, 2006) as well as correlational research associating spatial ability to various aspects of science learning in adults (e.g., physics problem-solving: Kozhevnikov & Thornton, 2006). It also builds on research linking visual-spatial working memory to general science performance in 11 year olds (Jarvis & Gathercole, 2003; St-Clair Thompson & Gathercole, 2006) and spatial skills to 5 year olds' force and motion understanding (Harris et al., 2018) in two main ways. First, it investigated a broader range of spatial skills and science topic areas. Second, it sampled a wider age range of children within one study to investigate possible developmental changes.

It is first interesting to note that both an intrinsic and an extrinsic spatial skill, within the Uttal et al. (2013) model, uniquely predicted overall science scores. This suggests that both within-object and between-object spatial skills support children's science reasoning, and supports the broad dissociation between intrinsic and extrinsic spatial skills (Hegarty, Montello, Richardson, Ishikawa, & Lovelace, 2006). Considering the role of specific spatial skills, the results revealed that mental folding, an intrinsic-dynamic spatial skill, was the strongest spatial predictor of total science scores. This general finding builds on past research linking mental folding ability to adult science outcomes (e.g., Baker & Talley, 1972).

Mental folding also emerged as the strongest spatial predictor of biology scores. This is the first study to date linking mental folding ability to biology with children. The ability to flexibly visualise, maintain and manipulate spatial information may be related to mental model construction and utilisation (Lohman, 1988). A mental model (Johnson-Laird, 1980; Gentner & Gentner, 1983) is a structural analog that contains spatial and conceptual relations of a process or situation. Children may construct spatially-grounded mental models of problem-solving questions, which include relational aspects of the problem, and then manipulate these mental models to solve them. This has been proposed in mathematics research with children (e.g., Rasmussen & Bisanz, 2005). Additionally, the representations children have for domain-specific concepts within biology may be spatially-grounded. For example, many of the plant-related questions involve knowledge and understanding of plant anatomy and function, which may be related to one another in mental model format. When recalling the function of roots, children may recall a spatial mental model of a plant, which includes spatial-relational information about the location and structure of different parts of the plant.

Mental folding also predicted physics scores, a finding which builds on the work of Harris et al. (2018), who found that mental folding predicted 5 year olds' force and motion understanding. Recall that the mental folding task requires non-rigid, dynamic visualisation. The spatial skills required to accurately visualise paper folds may support children in, for example, visualising and predicting the dynamic effects of forces acting on objects, or the general dynamic transfer of energy, which is central to physics topics. More specifically, spatial visualisation skills may enable children to mentally simulate actions and processes, such as reasoning about the way two magnets react to each other (see Chapter 4).

After controlling for BPVS scores, mental rotation was not a predictor of science achievement, despite it falling into the same Uttal et al. (2013) category as mental folding; this was also found by Harris et al. (2018) in relation to children's force and motion understanding in 5 year olds. There are several plausible reasons for this. First, as previously described, rotation is a rigid transformation and folding is a non-rigid transformation. In contrast to rotation, where the relationship between all points of the object are preserved, folding creates two separate areas and the spatial-relations between these areas must be maintained as the shape is folded. It is plausible that the additional spatial requirements of the folding task supported more complex visualisation between multiple elements in the science assessment. In addition, there are also possible limitations with the rotation task itself. The task uses the same monkey stimuli throughout, with the choice stimuli having the same pattern of blue and red hands, rather than using a range of animals, as is the case with other 2D rotation tasks (e.g., Neuburger, Jansen, Heil & Quaiser-Pohl, 2011). It is possible that this resulted in children of this age range using a rule-based strategy (i.e., if the monkey's right hand is red in one stimuli, then it will appear to be on the left side on the rotated version), rather than an analog, rotation-based strategy. Finally, research to date with adults and adolescents linking mental rotation to science achievement uses abstract 3D cube mental rotation, in contrast to the 2D animal stimuli used in the current study. Although children up to the age of 10 have difficulty with 3D rotation in its traditional format (Jansen, Schmelter, Quaiser-Pohl, Neuburger, & Heil, 2013), a 3D mental rotation task with tangible objects has more recently been developed which is suitable from 4 years (Hawes, LeFevre, Xu, & Bruce, 2015). Future work could further investigate the possible influence of stimuli type and test format. The mental rotation task was also the lowest-loading spatial task in the CFA, with this group of children, reported in the previous chapter. This will be considered further in the overall discussion.

Spatial scaling, an extrinsic/static skill, also emerged as a predictor of total scores, biology scores and chemistry scores. To our knowledge, this is the first study to link extrinsic static spatial skills with science achievement. The National Research Council's report 'A Framework for K-12 Science Education' (National Research Council, 2012), also identifies scaling within the core theme 'scale, proportion and quantity'. It emphasises that understanding relative magnitude and scale is essential for science; for instance, children must learn to appreciate how systems and processes vary significantly in size (e.g., a cell versus an organism). Taking a chemistry topic example from the current study, when understanding states of matter, children link how a liquid behaves at the observable *macroscopic* scale with the molecular processes at the *microscopic* scale. The report also identifies that children need to confidently move back and forth between representational models of different scales (e.g., for biology: a diagrammatic representation and a life-sized human skeleton model).

Switching between scaled models is a central component of the scaling task used in the current study.

The embedded figures task, an intrinsic-static spatial skill, was a significant predictor of chemistry scores only. This builds on prior work which found a relationship between this task and adults' chemistry performance (Bodner & McMillan, 1986). Intrinsic-static spatial skills relate to form perception and the processing of objects without further transformation. Several of the chemistry items include diagrams which require processing sub-parts of objects (e.g., 3 beakers, each with 4 ice cubes, which either have 1, 2 or 3 layers of insulation). The visual discrimination between the diagrams may support problem-solving needed for this type of question.

Interestingly, biology emerged as the discipline area which was most strongly predicted by spatial ability generally, despite the fact that it is not generally thought of as a spatially demanding area, relative to physics, for example. Although there are examples of spatial ability being related to biology learning in adults (e.g., learning anatomy: Lufler et al., 2012), in the Wai et al. (2009) longitudinal study, spatial ability in adolescence was predictive of outcomes in physics, engineering and chemistry, but not biology. Although biological concepts may not immediately appear as spatial as other areas, the abstract spatial representations used to organise and classify (e.g. classification keys: binomial, branching tree diagrams used to identify species) may be spatially demanding. It is possible that there is greater utilisation of these kinds of spatial representations for children than for adults.

Although the discussion above has been mostly considered by conceptual domains, there are also other possible cross-cutting, more domain-general mechanisms. For example, the link between spatial scaling and scores could be linked to spatial representations more generally. More specifically, it could be that extrinsic-static skills (i.e., the ability to align between two objects or coordinate between two or more spatial elements) are useful for using certain types of spatial representations. When using a table, for example, in order to locate a specific piece of information, it is necessary to co-ordinate between a column heading and row identifier, in order to identify the specific cell needed. Tables were used throughout the assessment, in more conceptual items, but also in more scientific investigation-related items (e.g., tables of results).

Models predicting overall science scores and performance in each area of science were consistent across development. It had been predicted that spatial skills may contribute more to science performance for younger children, suggesting that as domain-specific knowledge increases, spatial abilities play less of a role in science (e.g., Hambrick et al., 2012); however, this was not upheld in the data. Such a hypothesis is based on the idea that older or more experienced learners can apply knowledge more readily without having to process spatially. For example, this prediction would suggest that spatial visualisation would not be a strong predictor of questions where children determined the direction of a force acting on an object because they would simply 'know' the answer, without having to visualise it. However, this was not the case. The assessment covered a wide range of topics and it may be that, although the older children were indeed more experienced in science, their in-depth knowledge (i.e., knowledge they could recall at the time of doing the assessment) may have been restricted to the topic or topics they have recently covered in class, for example. Furthermore, the children were all in the same academic Key Stage; with a wider age range, above 12 years possibly, developmental changes may have been observed.

There are also limitations with the study. First, although the BPVS was included as a measure of verbal ability, a measure of non-verbal reasoning ability was not included. It is possible that the relationships observed may be partly accounted for by aspects of the task that involve fluid intelligence or non-verbal reasoning, in addition to the spatial skill measured. Second, the nature of the composite science assessment used includes aspects of factual knowledge, conceptual understanding and problem-solving. Dividing outcome measures into these sub-skills is a possibility for future research.

Relatedly, items also differed in the extent to which they required participants to use overtly spatial representations, such as diagrams. The observed relationship between spatial skills and science achievement may be driven by items which included spatial representations such as these. This is supported by a prior study demonstrating the effectiveness of a science curriculum which included spatial skills training in the form of diagram reading instruction (Cromley et al., 2016). The training was most effective for science post-test items in which interpretation of the diagram was particularly important in answering the question because the diagrams had been used to relate novel curriculum content. That is, the students had not been exposed to the topic or diagram previously in class and the question answer could therefore be derived from interpretation of the diagram alone. Many diagrams in the current study also had a degree of novelty because they were often included to accompany previously unseen problems and scenarios. Future research could compare the contribution of spatial skills to performance on items which rely on diagrams to varying degrees.

It should also be noted that the study was underpowered to detect a small to medium effect when the additional age-based interactions were included within the regression models. Future research is therefore needed to confirm these findings.

The results observed in the current study have implications for interventions to support children's science learning. Given evidence that spatial skills are malleable (Uttal et al., 2013), the finding that spatial scaling, mental folding and disembedding predict children's science achievement suggests that they are good candidates for spatial training. Long-term interventions involving the training of multiple spatial skills, embedded within the curriculum, may be a particularly effective approach (see Hawes, Moss, Caswell, Naqvi, & MacKinnon (2017) for a mathematics example). Furthermore, interventions to support children's spatial thinking skills could lead to additional long-term benefits for science achievement and engagement.

Chapter 4

The contribution of spatial ability to children's understanding of sound propagation

4.1 Introduction

4.1.1 Summary of aims and goals for chapter

The goal of the current chapter was to investigate the relationship between spatial thinking skills and children's science learning in further depth, to build on the previous chapter. The first specific aim was to investigate the relationship in the context of learning about a physics topic, sound, in the classroom, in contrast to the context of a curriculum-based assessment in the previous chapter. The second specific aim was to investigate whether the relationship varied by different types of science skills, such as factual knowledge recall, compared with problem-solving. A secondary aim of the chapter was to further investigate the structure of spatial thought, and more specifically the validity of the Uttal et al. (2013) model.

To achieve these goals, five classes of children in Year 5 (aged 10) took part in a science lesson on sound propagation, and then completed a science assessment, which targeted several science skills. Children also completed five spatial thinking tasks which mapped onto the proposed Uttal et al. (2013) model. The predictive power of these spatial skills was explored in relation to science skills assessed following the lesson. The main prediction was that spatial thinking skills would be predictive of children's ability to apply conceptual understanding of sound to novel problems. It was hypothesised this would be the case because spatial skills would contribute to the construction and manipulation of a vibrational mental model of sound propagation. Confirmatory factor analyses were also conducted to test the intrinsic-extrinsic, and static-dynamic dimensions of the Uttal et al. (2013) model.

The following sections provide more detail on the rationale for this chapter in relation to previous findings, both within this chapter and in the wider literature.

4.1.2 Overview in relation to findings from previous chapter

The findings in the previous chapter revealed that spatial skills are a significant predictor of 7-11 year olds' general science achievement. Mental folding and spatial scaling emerged as the best spatial predictors of overall science achievement, assessed using a curriculum-based science measure. Furthermore, mental folding was the only unique spatial predictor of physics achievement. This chapter moves on to address the latter relationship in more depth, and also with a number of methodological variations.

As outlined in Chapter 1, a curriculum-based science assessment, as used in the previous study, has the advantage of covering a broad range of procedural and conceptual science skills, meaning that the findings have wide educational applicability and validity. However, a disadvantage of this approach is that it is difficult to draw more specific conclusions about the observed relationship between children's spatial skills and science learning. Furthermore, as discussed in Chapter 1, it was hypothesised that a primary role of spatial thinking in science relates to learning and understanding conceptual topics. With this in mind, procedural skills were not included in the current study. Instead, the focus was specifically on one conceptual, physics sub-topic: sound.

Within the domain of knowing and understanding scientific phenomena, the current study further distinguished between different science skills. Specifically, it distinguished between children's ability to retrieve scientific factual knowledge, and, their ability to provide conceptual predictions and explanations, in response to novel problem-solving questions. Through assessing predictions and explanations the aim was to determine if, and how, spatial ability relates to children's ability to construct and use mental models. Predictions and explanations are commonly used in science-related research as a measure of aspects of conceptual understanding (e.g., Siegler ,1976; Glauert, 2009; Messer, Pine, & Butler, 2008). Factual knowledge, which is less dependent on conceptual understanding, was assessed to provide a contrast to predictions and explanations. The aim was to determine if children's spatial skills were generally more or less predictive of some of these aspects of knowledge than other skills. For example, based on the literature outlined in study one (e.g., Jaeger et al., 2016) spatial thinking may be more related to explanations, than to factual recall or predictions.

In the classroom, children typically learn about a topic in-depth in the context of a science lesson. In the current study, groups of 9-10 year olds therefore participated in a whole-class, researcher-delivered science lesson on the topic, prior to completing the science assessment. The researcher was also a qualified teacher for this age group. The primary reason for this was to ensure that children had a similar level of instruction on the topic. Furthermore, it meant that, for all children, there was the same amount of time between instruction, and, completing the post-test. Moreover, having the same researcher deliver the lesson to all classes, also reduced the variability that might exist across teachers. Finally, it also increased the ecological validity of our findings, by linking spatial skills to learning activities encountered by children, in the classroom.

There were two other main, methodological variations, compared to the previous study. First, in terms of IQ, a measure of fluid reasoning ability was included, in addition to a measure of vocabulary. A limitation of the previous study was that it was unclear whether some of the observed relationship between spatial ability and science achievement was accounted for or shared with a more general, fluid reasoning ability. One possibility is that some of the variance in science learning is uniquely accounted for by fluid reasoning, or, that some variance is uniquely accounted for by spatial ability. A further possibility is that some, or all, of the variance is shared between both types of tasks. For example, in the case of matrix reasoning type fluid reasoning tasks and spatial mental folding tasks, skills such as the ability to process shapes, or to maintain goal-directed behaviour in the context of novel problem-solving questions, might be common across these tasks.

As outlined previously, spatial ability is recognised as being at least partly distinct from reasoning ability (Thurstone, 1948). Indeed, Carroll's (1993) psychometric model of human ability, includes fluid reasoning (Gf) as a separate aspect of general intelligence (g) to visual processing (Gv) (i.e., spatial thinking). Spatial factors also have high loadings on fluid reasoning, with some uniqueness (Carroll, 1993). In line with this, it was demonstrated in an fMRI study (Ebisch et al., 2012) that induction (i.e., fluid reasoning), and two of Carroll's (1993) spatial factors (visualisation and spatial relations), activated both shared and distinct areas of the brain. Fluid reasoning, spatial visualisation and spatial relations showed shared patterns of activation linked to frontoparietal networks, areas linked to the preparation of goal-directed behaviour, the implementation of complex task rules, and, to working memory (see discussion of executive function section 1.2.4.1). In contrast, spatial visualisation, more so than inductive reasoning and spatial relations, showed increased activation of bilateral supramarginal gyrus, an area which has been linked in previous research to the manipulation of visual material.

Second, the current study also included some variations to the spatial tasks compared with the previous study. First, a different 2D mental rotation task was administered, for the intrinsic-dynamic category, due to the possible limitations with the task used in Chapter 2 and 3, previously discussed. Second, instead of the CEFT, a visual discrimination task was administered for the intrinsic-static category. The basis of this change was that, in addition to processing the form of a shape, the CEFT also involves additional processing (i.e., the ability to separate a shape from a complex background). The visual discrimination task used in the current chapter is more focused on individual visual forms, thus, a better measure of the intrinsic-static category.

Most significantly, a different task was administered for the extrinsic-dynamic category. In the previous chapter, it was predicted that the extrinsic-dynamic (perspective taking) task would be a unique predictor of science achievement. This was particularly expected to be the case for physics, because physics often involves the dynamic interaction between multiple elements. However, perspective taking was not a unique predictor, either for overall scores, or for physics specifically. One possibility is that, although perspective taking involves the dynamic interaction between objects and the frame of reference/person's perspective, it does not directly require the manipulation and perception of the relationship between multiple objects.

In order to include a dynamic task that involved direct manipulation of multiple objects, the current study used a measure of multiple-object dynamic spatial ability (see Chapter 1). To review, prior research indicates that these types of tasks are dissociated from object-focused transformation tasks, i.e., intrinsic-dynamic tasks (Hunt et al., 1988; D'Oliveria, 2004). Thus, within this study, these tasks are categorised as being extrinsic and dynamic, within the Uttal et al. model, because they require the perception and updating of the spatial relations between multiple, dynamic elements. 'Dynamic' here is used to refer mainly to the on-screen movement of stimuli, rather than the visualisation of movement, as in mental rotation or perspective taking. However, there may be a degree of 'imagined' dynamic visualisation; for example, when making predictions about the outcomes of future events. Moreover, Larson (1996) reported very high correlations between standard mental rotation tasks, and rotation tasks which involve the additional movement of the rotation stimuli on screen. Multiple object dynamic spatial ability tasks also align with one of three types of extrinsic-dynamic skills outlined by Newcombe and Shipley (2015), labelled as 'updating movement through space'.

As discussed in the previous spatial skills review (see Chapter 1), there has been limited investigation to date of this type of spatial ability (Hegarty & Waller, 2005); and, only one published study has investigated this type of task, in relation to science learning or achievement (Sanchez & Wiley, 2014).

Due to the fact that the spatial tasks to be used are somewhat different to from Chapter 1, a further confirmatory factor analysis will be run, to further test the intrinsic-extrinsic and static-dynamic dimensions of the Uttal et al. (2013) model. The findings of the chapter may therefore also add to the findings on the structure of spatial cognition from Chapter 1.

Given the conceptual topic of focus in the current study was sound propagation, the following section provides an overview of the dominant models and misconceptions children have about sound.

4.1.3 Children's understanding of sound across development

The scientifically accepted explanation of sound states that sounds are produced by a source of energy striking an object which results in the object vibrating. As the object vibrates, it pushes the particles of the medium (e.g., air particles), and the particles in the air vibrate. These air particles push on adjacent air particles, which also vibrate, and the vibrations propagate through the medium. Sound is therefore a process of energy transmission; it has the properties of a physical process, and not the properties of a physical object. However, a dominant misconception, which constrains younger children's understanding of sound, is that sound has materialistic or object-like properties (Mazens & Lautrey, 2003).

Piaget (1971) reported that 6-year-olds believe that a sound 'lives' inside an object, and travels to them when it makes a sound. Beyond these early findings, Mazens and Lautrey (2003) provide the only developmental analysis of children's understanding of sound. The focus of the current study is on whether variation in with spatial thinking skills children's are associated with children's understanding of sound. The study by Mazens and Lautrey (2003) is therefore of particular importance to the current study because it provides an insight into the typical mental models children hold of sound at different ages, and the extent to which there is variation at different points in development, including the age of children in the current study.

The study involved 89 French children, aged 6-10 years (pre-school aged: mean age 6 years; second grade: mean age 7 years; fourth grade: mean age 10 years). The study considered three properties of physical objects, which children may attribute to sound: solidity, weight and permanence. Results of the study showed that the properties of matter children attributed to sound were abandoned in a hierarchical way. In particular, across development, children abandoned the property of weight first, then permanence, followed by solidity.

These results can be interpreted in light of the dominant models of conceptual understanding. The inconsistencies in attributions might be linked to a 'knowledge in pieces' perspective' (diSessa, 1988). In particular, because at different points in development children attribute some, but, not all, of the properties of matter to sound, it might indicate fragmentation of knowledge. Different contexts (i.e., the different scenarios used in the study) might have activated different p-prims. The authors argue, however, that the hierarchical and synchronic manner in which the attributes of matter cease to be attributed to sound, suggest consistency and coherence. Their theoretical interpretation is therefore in line with Vosniadou's (1994) framework theory account. Specifically, they suggest that, at each point in development, children's overall concept of sound is based on relatively stable physical presuppositions (i.e., properties of objects). This is similar to Vosniadou's previously discussed arguments that children's mental models of the earth are constrained by the properties of physical objects. However, an additional point is that there was also decreased variability across age, in terms of the number of physical attributes demonstrated. This could perhaps still be linked with a knowledge in pieces account (diSessa, 1988). It may be that knowledge is initially quite fragmented, but becomes more coherent over time.

As outlined, solidity was more resistant to change, which suggests that this property is more central to children's concept of sound. Solidity was assessed by asking children how they were able to hear a sound through a wall or a door. In line with Vosniadou's approach, the authors propose five mental models of sound. However, these models are based only on the solidity responses. Although, in the discussion, the authors suggest that the abandonment of the other two object properties is reflected within the first three models, they do not suggest how exactly.

First, they report a model in which sound is completely solid, and cannot pass through objects at all. This was observed for only some 6 year olds. In the second model, a sound can be heard through a wall because it travels through holes in an object. This explanation dominated among the 6 year olds children (79%) but decreased at age 7 (37%) and further decreased for the 10 year olds (26%). The authors argue that everyday experiences with sand or water may provide an analogical (see structure mapping theory; Gentner (1983), Chapter 1) framework for this belief. Third, explanations which suggested that the properties of the material (i.e., the strength or hardness), determine whether a sound can be heard or not. For instance, sound travels through cardboard more easily than metal. Again, everyday experiences (e.g. a pencil piercing a piece of paper) may inform this model. This type of explanation emerged at age 7 (30%) and decreased at age 10 (15%).

The fourth model involved explanations that were based on sound being immaterial. For example, sound is like a ghost, which can travel through objects. This is a 'transitional' belief, which does not seem to be constrained by solidity, but, does not yet meet the criteria for the final model (a vibrational model). Explanations based on immateriality were present for a small number of the pre-school children (17%), increasingly only slightly for second grade children (18%), and increasing significantly for fourth grade children (29.5%). Finally, explanations which referred to the vibrations of objects. References to vibrations were typically in terms of adjacency, in a sequential process. Vibration explanations increased from 4% to 15% between pre-school to second grade, with a large jump in fourth grade (29.5%).

Therefore, by the age of 10, around the age of children in the current study, although the majority of children (50%) in Mazens and Lautrey (2003) no longer applied the properties of objects to sound, 39% continued to apply at least one property. Of the children who did not attribute solid properties to sound, only half refer to vibrations or resonance. Many of the remaining children provided immaterial ('sound is like a ghost'), explanations. Overall this generally suggests that many children struggle with conceptual understanding of sound, even into later childhood. The aim of the current study is to consider whether individual differences in children's spatial thinking skills are a significant factor affecting 9-10 year old's ability to apply conceptual understanding of sound, to novel scenarios.

As outlined, the emerging structure of children's concept of sound is unclear from Mazens and Lautrey (2003). In particular, although the authors interpret the findings within Vosniadou's framework, as outlined above, it also possible to interpret aspects of the data more in line with diSessa's approach (i.e., knowledge as being fragmented). However, there does, at the very least, appear to be greater consistency and coherence for the older children. The term ‘model’ is therefore adopted from this point forward, but, with the theoretical caveats outlined in Chapter 1.

4.1.4 Current study and hypotheses

To summarise, the current study aimed to assess the role of different spatial skills as predictors of children’s scientific recall of, and reasoning about, the concept of sound. As outlined above, the focus was on distinguishing between children’s ability to retrieve scientific factual information, as well as their ability to provide predictions and explanations, in the context of sound problem-solving questions. To ensure that all children had a similar level of exposure to formal teaching on the topic, and also to increase the ecological validity of our findings, children also took part in a one-hour science lesson, within their classes, prior to the assessment. The focus of the study was on children aged 9-10 years. This age range was chosen, firstly, because, based on the spatial development data described in section 2.1.3, children showed little development in skills beyond the age of 8. Therefore, this meant that the spatial developmental variability within the sample was reduced. Moreover, as outlined above, prior research suggests that children of this age are at a point of transition in their conceptual understanding of sound. That is, some but not all, children appear to have adopted a vibrational model, by this age. A secondary aim was to investigate the structure of spatial cognition in childhood, by evaluating the intrinsic-extrinsic and static-dynamic dimensions of the Uttal et al. (2013) model.

Hypothesis 1: It was expected that spatial ability would not be predictive of performance on the factual recall questions, or, that this relationship would be smaller than the relationship with the sound problem-solving questions. The basis of this prediction was that factual knowledge is less dependent on conceptual understanding (Mayer, 2002a), and thus, perhaps less likely to be stored as, or within, a mental model. Therefore, the role of spatial skills (for model construction, manipulation, and so on) would not necessarily apply.

Hypothesis 2: It also was expected that, controlling for verbal and fluid reasoning ability, spatial ability would predict overall performance on the sound problem-solving questions, following the lesson. It was theorised that spatial skills would support the construction or manipulation of a mental model of sound propagation. Spatial skills may enable children to construct representations of elements of scientific processes, and, represent relations between these elements, within the mental model (Gentner & Stevens, 1983; Johnson-Laird, 1980). For example, for this topic, representing the interrelations between an object vibrating, and the vibrating particles that make up a solid object, resulting in a sound being heard. In this case, it was expected that construction or manipulation of the internal mental model would be supported by the external models presented during the lesson, presented verbally and through static and dynamic visualisations. When required to make inferences within the problem-solving questions, it is proposed that children would 'run' aspects of this internal mental model. This may be supported by a process of mental simulation (section 1.4.5). As discussed in section 1.4.5, this would most likely involve the simulation of piecemeal fragments of the process.

Children with higher spatial ability, who are more able to create and manipulate mental models, should, hypothetically, be more able to generate accurate predictions and explanations, when problem-solving. As discussed in Chapter 1, research exists with adults to support the notion that spatial ability supports model construction of physical processes (Narayanan & Hegarty, 2002; Jaeger, Taylor & Wiley, 2016; Sanchez & Wiley, 2014; Hegarty & Just 1993). However, no research to date has addressed the role of different spatial skills as being supportive of mental model construction, in children. Moreover, no research has compared the role of a range of spatial thinking skills as predictors of children's factual knowledge, predictions and conceptual explanations, relating to a specific topic.

Hypothesis 3: It was also expected that the 'dynamic' spatial skills (mental folding, mental rotation and the arrival judgment task) would be stronger predictors of problem-solving scores, than the static spatial skills (spatial scaling task and the discrimination task). In addition, it was expected that the dynamic spatial skills would play a larger role in the development of the vibrational model of sound, which is a dynamic process. In particular, that the arrival time judgment task, an extrinsic-dynamic task, would be a particularly strong predictor of problem-solving scores, because it involves dynamic multiple elements, which may be particularly relevant to grasping the vibrational model of sound propagation. This may be through the use of dynamic visualisations and models used in the lesson to represent the vibrational model. In addition, intrinsic-dynamic spatial thinking skills, as assessed through mental folding and mental rotation, may be supportive of piecemeal mental simulation, within the model (Hegarty, 2004).

[Hypothesis 4]: It was expected that although spatial ability would be beneficial to predictions and explanations, there would be a stronger relationship with explanations, overall, than predictions. When generating an explanation of a process, it is necessary to coordinate between various aspects or elements of the process. Explanations may involve a more complete running or accessing of the mental model. In contrast, predictions are more descriptive, relate to one or a smaller number of aspects of the model (for example, an outcome), and, are less dependent on integrating elements of the model (cf. Sanchez & Wiley, 2014).

[Hypothesis 5]: Using confirmatory factor analysis of the spatial tasks, it was expected that a two-factor intrinsic-extrinsic factor model, based on the 2x2 Uttal et al. (2013) typology, would fit better than a one-factor model. The intrinsic factor will include mental rotation, mental folding and visual discrimination, and the extrinsic factor will include spatial scaling and the arrival judgment task.

[Hypothesis 6]: Finally, also using confirmatory factor analysis of the spatial tasks, it was also expected that a two-factor static-dynamic factor model, also mapping onto the Uttal et al. (2013) typology, would also fit better than a one-factor model. The static factor will include spatial scaling and visual discrimination mental rotation, and the dynamic factor will include mental folding, mental rotation and the arrival judgment task.

A summary of the main hypotheses is provided in Table 17.

Table 17: Summary of study hypotheses

Hypothesis 1	Spatial ability not predictive of factual scores, or, relationship weaker than with problem-solving scores (predictions plus explanations).
Hypothesis 2	Spatial ability predictive of overall performance on the sound problems solving scores.
Hypothesis 3	Dynamic spatial skills stronger predictors of problem solving scores overall than static spatial skills.
Hypothesis 4	Spatial ability stronger prediction of explanations overall than predictions.
Hypothesis 5	Two-factor intrinsic-extrinsic model better fit than one-factor model
Hypothesis 6	Two-factor static-dynamic model better fit than one-factor model

4.2 Methods

4.2.1 Participants

Five ‘Year 5’ (aged 9-10) class groups of participants were initially recruited in total, from three London schools, totalling 130 participants. Two classes each were recruited from two schools, and one class was recruited in a further school. An opt-out consent procedure was used to maximise the number of participants in the study, which was approved by the UCL, Institute of Education, Research Ethics Committee. One whole class, 28 children, was excluded, due to a significant (45 minute) interruption in the middle of the science lesson, and then a subsequent overnight delay, before completion of part of the post-test. The final sample therefore consisted of 102 participants (mean age: 9.6 years; SD: .33). The proportion of females (58.5%) was higher than the proportion of males (41.5%).

A power analysis, conducted in G*Power indicated that a sample size of 113 was needed to detect a small to medium effect (f^2 =.12) for the main analysis (the multiple linear regression analysis with 5 possible predictors). (Power = 0.8, α = 0.05). A small to medium-sized effect was chosen due to the number of spatial measures included, and the likelihood of shared variance between them. The achieved sample was therefore slightly smaller than the desired sample.

4.2.2 Materials

4.2.2.1 Spatial measures

4.2.2.1.1 Visual discrimination task (intrinsic-static sub-domain)

Participants completed the discrimination subtest of the Test of Visual Perceptual Skills, which was administered according to the manual (TVPS; Gardner, 1996), which includes 16 items of increasing difficulty. In this task, the participant was shown a 2D, black and white geometric design at the top of a page of a ring-bound pad. They were also shown five similar-looking geometric patterns at the bottom of the page. The participant was asked to choose one shape from the bottom set of shapes that exactly matched the target shape. Participants responded by verbally indicating the answer to the experimenter. The participant first completed two practice items. If they failed on a first attempt, they were given a maximum of one further attempt to answer each practice item. All children completed the practice items after a maximum of two attempts at each item. Upon completing the first two practice items, they progressed to the main 16 trials. Accuracy was recorded for each trial.

4.2.2.1.2 Mental folding (intrinsic-dynamic sub-domain)

This task (Harris et al., 2013) was the same as was administered in the previous chapter. However, for this study, the task was coded in Psychtoolbox for Matlab, and presented on a 13" Macbook laptop. The task was also administered individually, rather than in groups. Accuracy was recorded through the child's mouse response to each item.

4.2.2.1.3 2D Mental rotation (intrinsic-dynamic sub-domain)

Mental rotation ability was tested with the rotation subset of Thurstone's Primary Mental Abilities Test (Thurstone, 1948). The task was coded in Psychtoolbox for Matlab, and presented on a 13" Macbook laptop. Children were shown a black outlined shape on the left of the screen (Figure 27). The researcher asked the child to choose which one of four shapes on the right, would fit together to form a square with the shape on the left (see Figure 27). As is evident in the sample below, the three distractor items differed slightly in shape to the target item (item 2). Children completed a shortened version (Gunderson, Ramirez, Beilock, & Levine, 2012), due to time constraints, which included 8 items.

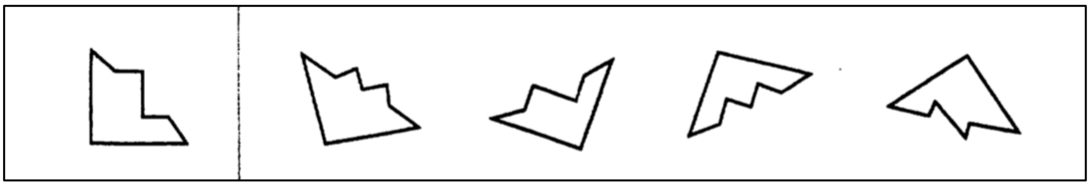


Figure 27: Mental rotation trial (correct answer is b)

4.2.2.1.4 Spatial scaling (extrinsic-static sub-domain)

This task (Gilligan et al., 2018) was the same task used in the previous chapter. However, the task was coded in Psychtoolbox for Matlab, and presented on a 13” Macbook laptop.

4.2.2.1.5 Arrival judgement (extrinsic-dynamic sub-domain)

This task was developed for the purposes of the current study. As outlined, it was adapted for use with children from a paradigm developed and tested on adults by Hunt et al. (1988) and required judgment of relative motion. It was developed and presented in Matlab using Psychtoolbox, and presented on a 13” Macbook laptop. The task was presented to children as representing a race (Figure 28). Children were presented with a black screen, which contained two vertical white lines on the left and right side (the start and finishing line). At the start of each trial, four different coloured squares appeared vertically, at varying distances from the left starting line. Across all trials, the colour of each square was consistent. The squares then moved to the right of the screen, at varying speeds. The objects disappeared when they reached the middle of the screen. If they had continued at the same speed, three of the squares would have arrived at the finishing line at the same time, and one would arrive before the other three. Children were asked to make a judgment about which of the four coloured squares would have arrived first (i.e., won the race). Children pressed a key on the keyboard to indicate which coloured square they believed would finish first. Accuracy was recorded. The participant first completed two practice items. If they failed on a first attempt, they were given a maximum of one further attempt to answer each practice item. All children completed the practice items after a maximum of two attempts at each item.

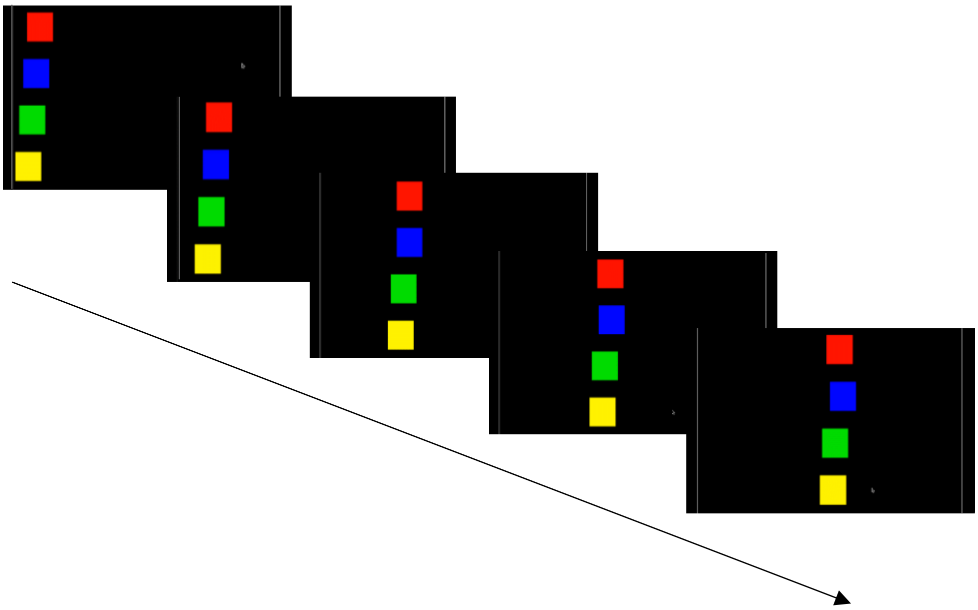


Figure 28: Arrival judgment task sequence. Example trial from item 1, level 1.

Participants completed 20 experimental trials of increasing difficulty, split into 5 levels. The difficulty levels were varied by decreasing the difference in arrival time between the 'winning' square and the other squares, by varying the speed and starting points of the objects. The difference in the first level was larger, and so the 'winning' square won by a larger margin, meaning the judgment was easier. The difference in the final level was smaller, so judging the future winner was more difficult.

The starting positions of the shapes were varied between trials. There were four different configurations of starting points, which were randomised within the four trials of the level. The starting points and winners were also considered together, so that within each level, a winning object started at one of the four positions. For example, within each level, there was always one winning square which started in first place. The speed of the shapes were systematically varied within each trial, based on the starting positions, so that, if they had continued, the winning object would have arrived first, and the other three objects would have arrived, simultaneously, afterwards. The winning square was varied randomly so that at each level, one of the squares was the winner. The trials were presented in order of increasing difficulty, but, not as separate levels. However, there was a short break halfway through the task.

4.2.2.2 Control measures

4.2.2.2.1 Matrix reasoning (fluid reasoning ability)

Fluid/abstract reasoning was assessed using the Matrix reasoning subtest of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999). The matrix reasoning test involves pattern completion. In this task, the participant inspects an incomplete series or matrix, and then selects the option that best completes the series or matrix. Participants completed a maximum of 30 items. Testing was discontinued after three incorrect responses.

4.2.2.2.2 Vocabulary

Vocabulary was assessed using the vocabulary subtest of the Wechsler Abbreviated Scale of Intelligence (Weschler, 1999), which was administered according to the manual. Participants first completed three picture items, where the participant named an object presented visually. The remaining 28 items were verbal, and required the participant to define words that were presented visually and orally. The words were presented in order of increasing difficulty, and included nouns, verbs and adjectives. Participants were given a score of 0, 1 or 2, depending on the accuracy and completeness of the definition. Testing was discontinued after three consecutive scores of zero.

4.2.2.3 Science lesson

The researcher (who is also a qualified primary school teacher) delivered the science lesson to each class. It followed a pre-prepared plan, which was the same for all classes. The lesson was designed to include activities which children would typically encounter in this age range. With this in mind, the plan was discussed with a teacher of this age range, prior to the study. It lasted approximately 1 hour, with a short break in the middle. The first part of the lesson covered the general vibrational model of sound, and the second part focused on how the model applies specifically to sound traveling through solids, liquids and gases. The lesson was delivered via a combination of verbal teaching, as well as videos (BBC, 2007; Boyles, 2017), text and images presented via an interactive whiteboard.

Figure 29: First video clip presented in science lesson (BBC, 2007)

Figure 30: Second video clip presented in science lesson (BBC, 2007)

Figure 31: Third video clip presented in science lesson (Boyles, 2017).

The start of part 1 began with a whole-class discussion of vibrations. During this, children were invited to feel vibrations from their vocal chords, as well as carefully vibrate a ruler on the table. Children then watched a short video clip, which explained how air vibrations allow us to hear sounds. This video included two visual models. The first showed air molecules represented as balls next to each other, each hanging from their own rope, hitting each other (Figure 29). The second showed a line of people, each representing a molecule. The person at the back pushed on the person in front, and the pattern continued down the line (Figure 30). A final video provided a demonstration using a speaker and candles (Figure 31). The presenter placed candles on the floor next to the speaker and explained how the flickering of the candles showed the vibration in response to the vibration of the speaker. The researcher finally explained the ways in which sound travels, including why sounds get fainter the further they are away from the source and how sound can vibrate through, and also sometimes 'bounce off', objects. After this, children completed a short group activity, which involved approximately four children. Each group was provided with images and short text descriptions, each showing parts of the process of how we hear a sound, from a speaker. Children were asked to sequence the image/text pairs.

At the start of the second part of the session, the researcher asked the children to name the three states of matter. The researcher then asked children whether they thought sound travelled best/fastest through a solid, a liquid or a gas, and to give some reasons. After this, the children were shown a short video clip (1 minute 30 seconds), which explained how sounds travel differently through a solid, a liquid and a gas. The video compared the different particle structures of each medium and the effect this has. Children were then finally shown a short clip (1 minute) about sound in a vacuum, which showed an alarm clock slowly becoming silent, after air is removed.

4.2.2.4 Prior knowledge assessment

Prior knowledge was assessed before children's participation in the science lesson, but within the same session. Children were shown an illustration of a girl listening with her hand to her ear, in a room, and a ringing bell (Figure 32). Children were asked to explain in their own words, on a piece of paper how they thought she heard the sound. Children were given a score of 0-3, with scores corresponding to, respectively: no understanding, partial understanding, established understanding or good understanding. Table 19 provides examples of each type of answer. A second coder scored a random 20% of the pre-tests. Because the inter-rater reliability assessment was based on a single text response and coding scheme, Cohen's Kappa was calculated. This analysis indicated high inter-rater reliability (Kappa = .952).

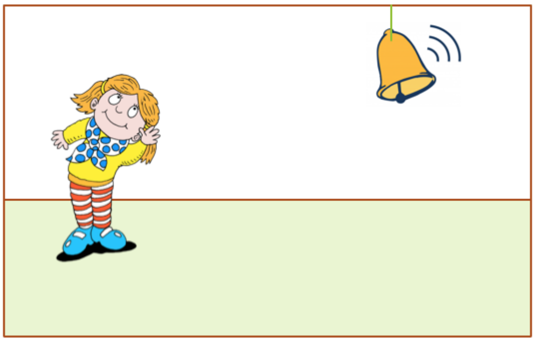


Figure 32: Image shown to children for prior knowledge assessment

4.2.2.5 Post-test

The science measure was made up of two paper and pencil assessments: an assessment consisting of factual knowledge questions, and an assessment consisting of problem-solving questions.

4.2.2.5.1 Factual knowledge

There were six factual questions. Five questions had a maximum score of one, apart one question, which had a possible score of two, resulting in a maximum score of 7. The questions required basic factual knowledge about sound, and did not require an in-depth understanding or application. Table 18 presents each of the factual knowledge questions.

Table 18: Factual knowledge questions

1) As you move further away from an object making a sound, the sound____________________.
   ☐ Becomes quieter ☐ Becomes louder ☐ Stays as loud as it was
2) Which of the following type of wave travels the fastest?

   ☐ Light ☐ Sound
3) Tick all of the following types of medium which sound **can** travel through

   ☐ Solid ☐ Liquid ☐ Gas ☐ Vacuum
4) Sounds are made when objects_______________.
5) Another name for a sound wave is a _________________ wave.
   ☐ Compression ☐ Surface ☐ Transverse
6) Give one example of a use for echo________________________.

Table 19: Coding scheme for prior knowledge assessment, with example responses.

Score and description	Example response
Score 0: No evidence of relevant prior knowledge. Irrelevant / entirely incorrect response/ no mechanisms described.	She hears a loud noise because the bell is tolling
Score 1: Some evidence of prior knowledge. No reference to what is vibrating.	I think the sound is caused by vibrations
Score 2: Evidence of established prior knowledge. Reference to what is vibrating.	Because when the little ball hits the bell, the bell vibrates
Score 3: Evidence of good prior knowledge. Reference to what is vibrating, and at least one additional causal element.	The bell rings and vibrates, and then the sound travels to her ears, and there are small hairs inside her ears which vibrate and send a message

4.2.2.5.2 Problem-solving questions

Participants also completed eight problem-solving questions. Each problem-solving question centered on a scenario involving sound traveling in some way. To ensure that the problems were challenging for all children, problem-solving questions of two levels of difficulty were included. The first four problem-solving question (Table 20) were closer to the content learnt in the lesson, but still unfamiliar. The remaining four problem-solving questions (Table 21) required the application of knowledge and understanding to more novel situations. Participants were presented with a written scenario, and a picture to accompany the scenario (see Appendix, section 8.1). For each problem-solving question, participants were first asked to make a written prediction (e.g. what do you think would happen if…). They were then asked to provide an explanation of their prediction, using their theoretical knowledge and understanding about sound. Four of the problem-solving question (see Q1,2,3,8) had explanation responses which scored up to 3. Four problem-solving questions (see Q4,5,6,7) had explanation responses broken down into two parts, scoring a maximum of two each and four overall.

Table 20: Summary of problem-solving questions (lower level of difficulty)

Question	**Scenario**	**Prediction**	**Explanation**
1	Josh stood in his classroom next to the door. Michael then closed the door. Josh then shouted to Michael on the other side of the door.	Imagine you were Michael. Predict what happened when Josh shouted with the door closed.	Explain as fully as you can how this happened, starting from the moment Josh shouts.
2	Sam is standing on the side of the swimming pool. There are metal steps which lead into the swimming pool. Sam taps the metal stairs with a hammer. His friends are in the pool with their heads underwater.	Imagine you were one of Sam's friends underwater. Predict what would happen when Sam taps the steps.	Explain as fully as you can how this happened, starting from the moment Sam hits the stairs with the metal object.
3	Daniel makes a string telephone as shown below. He stands on one end of the room and speaks quietly into one end of the telephone. His friend holds the other end of the telephone to her ear.	Imagine you were Daniel's friend. Predict what would happen.	Explain as fully as you can how this happened, starting from the moment Daniel speaks into his end of the telephone.
4	Reis taps an empty pipe in one part of his house and his friend listens to the pipe in another part of the house.	Predict whether you think sound would travel faster in pipe 1, faster in pipe 2, or whether it would be as fast/ the same, in both:	Explain as fully as you can how and why this would be the case. Write about both pipes.

Table 21: Summary of problem-solving questions (higher level of difficulty)

Question	Scenario	Prediction	Explanation
5	Lydia wants to soundproof her room so that her neighbours, who live next door, are not disturbed by her music. She first covers the wall between her and the neighbours in carpet tiles, plays her music and then asks her neighbour to measure the sound. After that, she removes the carpet tiles and then covers the wall with wooden tiles instead, and her neighbour measures the sound again.	Predict which material would be the best at soundproofing (the best at blocking the sound and reducing how much sound travelled to her neighbours). Do you think wooden tiles would be better, carpet tiles would be better, or they would both be as good as each other, for soundproofing?	Use your understanding of how sound travels through different mediums to explain why you think this, as fully as you can. Write about both types of material.
6	Sally shouts normally. Joanne then also shouts, but uses her hands in a cup shape over her mouth.	Predict what would be different about the sound when Joanne shouts with her hands in a cup shape.	Use your understanding of how sound waves travel to explain why you think this, as fully as you can.
7	Hasan has a double glazed window. This means there is one glass layer on the outside of the house, a middle layer which is air, and another glass layer on the inside. The pictures show what the window looks like from a side angle. John has a single glazed window. This means there is one layer of glass only. They are same total depth overall.	Predict whether you think sound would travel easiest through the double glazed window, the single glazed window, or, whether it would be the same.	Using your understanding of how sound travels, explain why you think this. Write about both kinds of window in your explanation.
8	Benedict walks through an open field and shouts. He then stands in a narrow cave and shouts towards a wall. He shouts equally loudly in both places.	Predict what might be different about the sound Benedict makes in the cave compared to the open field.	Use your understanding of how sound travels to explain as fully as you can, how this happens.

4.2.2.5.3 Scoring of problem-solving questions

The predictions element of each scenario was coded as either correct or incorrect, using a pre-selected range of correct and acceptable answers. Explanations were coded depending on the depth of explanation and conceptual understanding demonstrated, using a pre-defined scheme for each scenario. Table 23 and 24 display example prediction and explanation responses for question one (see Table 22 for summary of question). A second coder scored a random 20% of the problem-solving assessments (predictions plus explanations). Because the problem-solving questions were coded based on the overall score for the problem-solving assessment, Pearson's correlations were calculated between the coder's scores. This analysis resulted in an overall high inter-rater reliability ($r = .98$, $p < .001$). Cronbach's alpha values were also calculated, to assess the reliability and internal consistency of the measures. Cronbach's alpha reliability of the problem-solving questions, with the prediction and explanation for each problem-solving question included as separate items, was good ($\alpha = .810$). The reliability for predictions alone was $\alpha = .556$, and the explanations alone was $\alpha = .795$. The reliability of the factual test was $\alpha = .529$.

Table 22: Summary of question one.

Scenario	Josh stood in his classroom next to the door. Michael then closed the door. Josh then shouted to Michael on the other side of the door
Prediction	Imagine you were Michael. Predict what happened when Josh shouted with the door closed
Explanation	Explain as fully as you can how this happened, starting from the moment Josh shouts.

Table 23: Example responses for prediction element of question one

Score	Example response
0	‘it would not have opened’
1	‘Michael could still hear Josh’

Table 24: Example responses for explanation element of question one

Score	Description	Example responses
0	Irrelevant answer; inaccurate or incomplete response/lacking detail.	‘Because something is blocking it.’
1	One point (with vibrational mechanism).	‘because the door vibrates’ ‘the particles in the door vibrate on the other side’
	Two points (path and no clear vibrational mechanism).	‘the sound travels through the air and then through the door’
2	Two points (with vibrational mechanism).	‘the vibration travelled through the door and then vibrates in Michael’s ear’
	One point (with vibrational mechanism), with one or more additional points referring to path (without vibrational mechanism).	‘the particles vibrate and the sound travels to Michael’s ears’
3	Three points (with vibrational mechanism).	‘The air particles vibrate, hitting more air particles next to it. The sound then travels through the solid door. The air particles vibrate around his ear and then his eardrum vibrates and sends a message to the brain’

4.2.3 Procedure

Children first completed the cognitive testing (matrix reasoning and vocabulary task, and spatial tasks) in a single 1:1 session which lasted approximately 35 minutes. This testing took place in a quiet space in the child's school. Once all of the children in the class had completed this individual cognitive testing (no more than 2 weeks), the science lesson was delivered to the children by the researcher. The prior knowledge task was completed directly before the lesson, and the post-test directly after the lesson. Children completed the prior knowledge task and post-test silently, within their class groups. To reduce the impact of reading (specifically decoding) ability, the researcher read aloud each question, and then gave the whole class approximately 5 minutes to complete each question. The children completed the questions silently and independently. The children were asked not to move to the next question, even if they had finished the previous answer.

4.3 Results

Percentage accuracy was used for all spatial tasks. To evaluate performance on the science tasks, a factual score was first totalled, which had a maximum score of 7. A prediction total was also calculated, which included all predictions across all problem-solving questions, scored out of 8. An explanation score was also calculated, which included all marks gained for all explanations, with a maximum score of 28. Finally, a total problem-solving score was calculated, which was made up of a total of predictions plus explanations, with a maximum score of 36.

4.3.1 Descriptive statistics

Descriptive statistics for the spatial tasks, control variables and science tasks are reported in Table 25. The mean accuracy scores for factual knowledge (66%) suggests that children generally were able to recall at least some of the core information from the lesson. As expected, children's mean scores on the problem-solving questions were lower (24% accuracy). Within problem-solving scores overall, although mean prediction scores were high (60% accuracy), explanation scores were poor overall (14% accuracy). Children therefore, on average, made approximately 4 valid explanation points, across the problem-solving questions. This is not surprising, given the difficulty many children have in understanding sound, as outlined in the literature review. Moreover, the problem-solving questions required children to not only to generate an explanation, from the conceptual model of sound, but to apply it to specific, novel situations. Recall that the goal was not to determine if children could master the application of conceptual explanations, but to see how individual differences in spatial ability predict the degree to which children could apply their understanding. The distribution of explanation scores was partly addressed by running an additional categorical analysis, discussed below in the analysis strategy.

Table 25: Descriptive statistics (percentage accuracy)

Measure	M	SD	Range
Prior knowledge	33	30	0-100
Factual total	66	23	0 - 100
Total problem-solving	24	15	0 - 81
Predictions	60	22	0 - 100
Explanations	14	15	0-75
WASI Matrix reasoning total	42	15	17 - 90
WASI Vocabulary total	51	13	23 - 89
I-D (mental rotation task accuracy)	80	18	13 - 100
I-D (mental folding task accuracy)	61	20	21 - 100
I-S (discrimination task-accuracy)	61	15	31 - 94
E-S (scaling task - accuracy)	69	15	33 - 94
E-D (arrival judgment task-accuracy)	64	13	15- 95

Note. I-D: intrinsic-dynamic; I-S: intrinsic-static; E-S: extrinsic-static; E-D: extrinsic- dynamic.

4.3.2 Confirmatory Factor Analysis

Before the main analyses, the latent structure of the spatial skills was investigated, as in Chapter 2. The choice of spatial tasks had changed from Chapters 2 and 3, but the task selection, and hypotheses, were still based on the Uttal et al. (2013) model. The aim was to determine if the same latent pattern was observed as in Chapter 2 and to evaluate any resulting theoretical implications. Three models were tested, with the same factor structure as in Chapter 2. The intrinsic/extrinsic model now included mental folding, mental rotation and visual discrimination on the intrinsic factor, and the arrival judgment and scaling task on the extrinsic factor. The static/dynamic model included visual discrimination and spatial scaling on the static factor, and mental rotation, mental folding and arrival judgment on the dynamic factor. The process of model evaluation mirrored Chapter 2. To increase statistical power, these models included all 131 children. This is because the spatial data from the excluded class was not affected by the lesson / post-test interruption.

The same analysis process and model fit indicators were used as in Chapter 2. To review, the χ^2 statistic indicates of model fit, and a significant χ^2 statistic indicates a poor fit. For the Standardised Root Mean Square Error Approximation (RMSEA), smaller values indicate better fit (0.1 excellent fit, .05 good fit, .08 and above poor fit [MacCallum, Browne and Sugawara, 1996]). The CFI indicates the improvement of model fit over a baseline model; a CFI value of .9 indicates a fair fit, and .95 indicates a good fit (Kline, 2015).

Table 26 provides a summary of the models along with details of model fit. As can be seen in Table 26, all three models produced a non-significant χ^2 value ($p > .05$ for all). The CFI and RMEA values indicated an excellent and good fit, respectively, for all models.

The χ^2 value for the baseline model (Figure 33), and the intrinsic-extrinsic model (Figure 34) were very similar. By χ^2 value, the static-dynamic model (Figure 35) was a better fit than one-factor baseline model. However, the χ^2 values did not differ significantly ($p > .05$). Therefore, although there was some support for the static-dynamic distinction, the data best supported the one-factor model. However, the arrival judgement task loaded poorly onto the one-factor model. The one-factor spatial ability model accounted for 17% of the variance in arrival judgment scores. The finding that neither two-factor model fitted the data than a one-factor model mirrors the findings from Chapter 2.

Table 26: Fit indicators for one and two factor models.

Model	χ^2	df	χ^2 *p*-value	CFI	RMSEA	χ^2_{diff}	χ^2_{diff} *p*-value
1-factor	3.22	5	.667	1	0.00	-	-
2-factor (intrinsic-extrinsic)	3.19	4	.527	1	0.00	.029	.863
2-factor (static-dynamic)	1.10	4	.894	1	0.00	2.21	.144

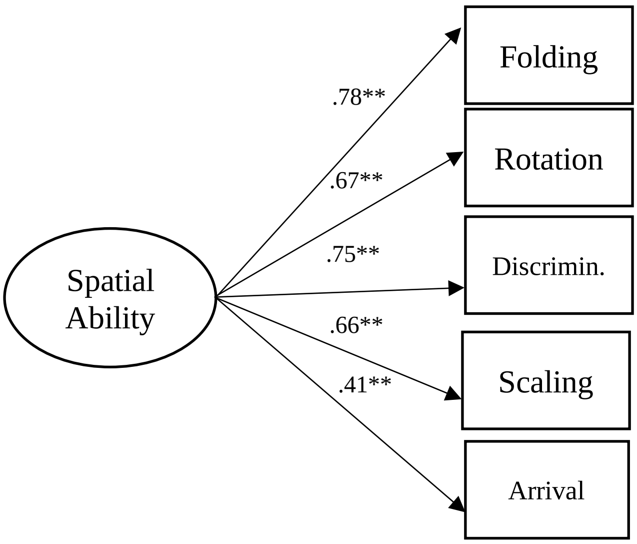


Figure 33: Model 1 (one-factor)

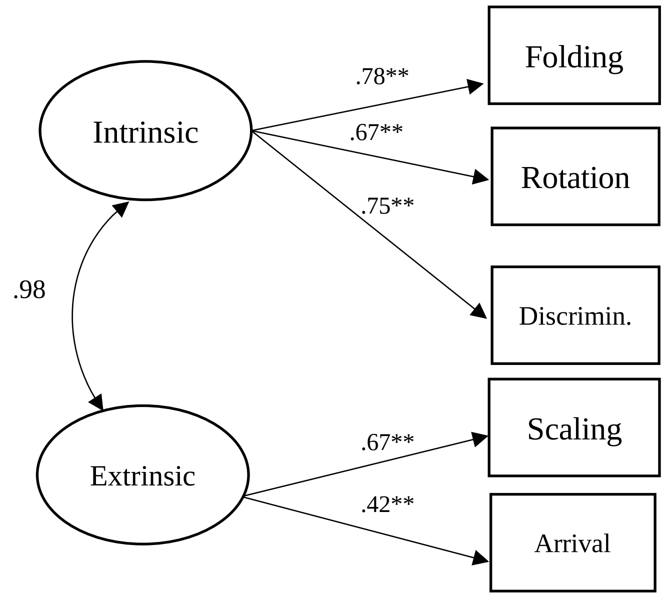


Figure 34: Model 2 (two-factor, intrinsic-extrinsic)

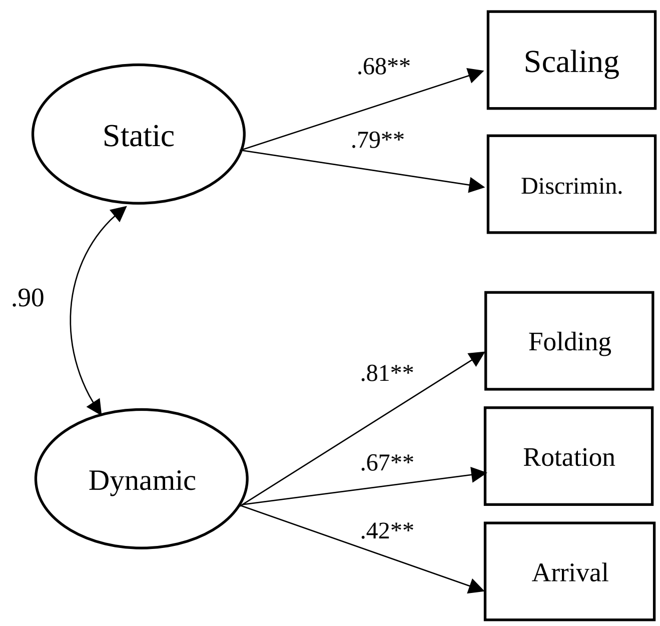


Figure 35: Model 3 (two-factor, static-dynamic)

4.3.3 Preliminary analysis of school class group and gender

A preliminary analysis was conducted to determine the effect of class grouping (i.e., school class) and gender. Separate ANOVAs were run with factual knowledge score and problem-solving score as dependent variables. For each dependent variable, a separate ANOVA was conducted with gender and the class group as independent factors (i.e., four models in total).

One-way ANOVAs revealed that there were no significant differences between class groups for total problem-solving scores and factual scores, and therefore class-grouping was not considered further in the analysis. Likewise, one-way ANOVAs also revealed no significant gender differences across total problem-solving score and factual scores, and so gender was also not considered further in the analysis. The assumptions for parametric testing were not met for total problem-solving scores, for the class grouping analysis (outliers) and gender analysis (outliers and homogeneity of variance). However, non-parametric analyses (Kruskal-Wallis for class grouping, and Mann-Whitney for gender), resulted in the same outcomes ($p > .05$ for both).

4.3.4 Correlation analysis

As a preliminary investigation of the relationship between the study variables, bivariate Pearson's correlations were computed between age, science variables, spatial variables and IQ variables. A summary is provided in Table 27. To avoid overfitting the regression models that followed, which was particularly important given the reduced number of participants, and therefore power in the analysis, an initial partial correlation analysis was also run. The partial correlations were computed between the science variables and each spatial task, controlling for the matrix reasoning and vocabulary tasks. Only variables which had significant partial correlations in this analysis were then entered into the regression models.

The partial correlation analysis (controlling for matrix reasoning and vocabulary, see Table 27) revealed that none of the spatial tasks demonstrated significant partial correlations with factual scores. However, mental folding had moderate, significant partial correlations with total problem-solving scores and explanations. Mental rotation had a small, significant partial correlation with prediction scores and total problem-solving scores. Finally, the arrival judgment task had a small, significant partial correlation with explanations, and with total problem-solving scores.

4.3.5 Regression analysis

A regression analysis was conducted to determine the amount of variance contributed by the spatial tasks to each science score (factual knowledge, total problem-solving score, predictions, explanations), taking into account the control variables. In each model, the vocabulary and matrix reasoning (IQ) tasks were entered together in Step 1. Age in months was not also entered in step one, because it did not significantly correlate with any of the science variables. Adding age into the model would have decreased the power of the model, which was, as described above, important to avoid, given the number of variables already included within the model, and the reduced sample size. In step 2, the participant's prior knowledge score was entered. In step 3, the spatial tasks were added (using the criteria described above). To determine if spatial ability and prior knowledge were independent predictors, additional models were included, for total problem-solving scores, in which step 2 and step 3 were reversed.

Additionally, less conservative models were run, which excluded the matrix reasoning task. This is discussed in detail in section 4.3.5 below. Because of the positively skewed distribution of the explanation scores, an additional categorical analysis was conducted, with a multinomial logistic regression. This is also discussed in detail in section 4.3.5. The regression assumptions (normality of residuals, homoscedasticity of residuals, no significant influential cases, no significant collinearity) were met for all models.

Table 27: Bivariate and partial correlations between the study variables.

	Age	PriorKn	Fact.	Predi.	Explan.	Tran.Tot.	Matrix	Vocab	Folding	Ment.Rot	Vis.Disc	Scal.	Arrival
1. Age (months)		.02	.14	.05	.17	.15	.01	.054	-.07	-.05	-.08	.02	-.05
2. Prior knowledge			.37**	.32**	.48**	.48**	.28**	.33**	.18	.09	.21*	.19	.28**
3. Factual				.47**	.48**	.52**	.21*	.49**	.14	.17	.11	.23*	.23*
4. Predictions					.57**	.78**	.25*	.40**	.19	.26**	.17	.10	.25*
5. Explanations						.96**	.35**	.48**	.36**	.26**	.33**	.30**	.31**
6. Problem-solving total							.36**	.50**	.35**	.29**	.31**	.26**	.32**
7. Matrix reasoning								.28**	.44**	.34**	.41**	.51**	.42**
8. Vocabulary									.04	.08	.20*	.20*	.13
9. Mental folding			.11	.14	.33*	.31*				.45**	.63**	.45**	.37**
10. Mental rotation			.13	.21*	.18	.22*					.51**	.44**	.18
11. Visual discrimination			-.03	.05	.19	.16						.57**	.21*
12. Spatial scaling			.12	-.07	.13	.07							.27**
13. Arrival judgement			.17	.17	.20*	.22*							

Note. * $p < 0.05$, ** $p < 0.01$. Upper triangle shows zero-order correlations and lower triangle shows partial correlations between spatial measures and the science scores, controlling for matrix reasoning and vocabulary scores.

For factual knowledge scores (Table 28) the vocabulary and matrix reasoning task entered into step 1 accounted for 25% of the variability in scores, F (2,99) = 16.39, $p < .001$. In the final model, vocabulary was the only unique IQ predictor, with a large beta value ($\beta = .41$). In step 2, prior knowledge accounted for an additional 5% of the variability in factual scores, ΔF (1,99) = 7.27, $p = .008$; $\beta = .41$). Step 3 was not included, because the prior partial correlation analysis indicated that the spatial ability measures did not significantly correlate with the factual scores, after accounting for IQ. The model overall accounted for 30% of the variability in factual scores.

Table 28: Multiple regression analysis predicting factual recall score

Predictor	*b*	ß	*p*	Δ F	Sig Δ *F*	r^2	$r^2\Delta$
Step 1) Vocabulary	.11	.41	< .001				
Matrix reasoning	.11	.03	.754	16.39	< .001	.25	.25
Step 2) Prior knowledge	.45	.25	.008	7.27	.008	.30	.05

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model.

For total problem-solving scores (Table 29), the vocabulary and matrix reasoning task entered into step 1 accounted for 31% of the variance in scores, F (2,99) = 21.74, $p < 0.001$. In the final model, vocabulary was again the only significant IQ predictor ($\beta = .38$). Prior knowledge entered in step 2 accounted for an additional 8% of the variance in scores, F (1,98) = 13.23, p < 001, and was a unique predictor in the final model $\beta = .28$. The mental folding task, mental rotation task and arrival judgment task were then simultaneously entered in step 3, F (3,95) = 4.19, $p = .009$, and accounted collectively for an additional 7% variance in total problem-solving scores. In the final model, only mental folding was a significant, unique predictor, $\beta = .19$. The final model accounted for 46% of the variance in total problem-solving scores.

Table 29: Multiple regression analysis predicting total problem-solving score

Predictor	*b*	ß	*p*	Δ *F*	Sig Δ *F*	r^2	$R^2\Delta$
Step 1) IQ(Vocabulary)	.35	.38	< .001				
IQ(Matrix reasoning)	.01	.01	.922	21.7	< .001	.31	.31
Step 2) Prior knowledge	1.7	.28	.001	13.2	< .001	.39	.08
Step 3) Mental folding (I-D)	.35	.19	.047				
Mental rotation (I-D)	.46	.13	.145				
Arrival judgment (E-D)	.21	.10	.242	4.19	.009	.46	.07

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the *p* value of the change in *F* for each step of the regression model

Table 30 presents the regression analysis for total problem-solving scores, with steps 2 and 3 reversed. Entering the spatial skills in step 2 accounted for 9% additional variance, compared with 7% when entered in step 3, the previous model. Therefore, 2% of the variance in total problem-solving scores were shared between prior knowledge and spatial ability.

Table 30: Multiple regression analysis predicting total problem-solving score

Predictor	*b*	ß	*p*	Δ F	Sig Δ *F*	R^2	$r^2\Delta$
Step 1) IQ(Vocabulary)	.35	.38	<.001				
IQ(Matrix reasoning)	.01	.01	.922	21.7	< .001	.31	.31
Step 2) Mental folding (I-D)	.35	.19	.047				
Mental rotation (I-D)	.46	.13	.145				
Arrival judgment (E-D)	.21	.10	.242	4.86	.003	.40	.09
Step 3) Prior knowledge	1.65	.28	.001	11.01	< .001	.46	.06

A second, less conservative model was run, which excluded the matrix reasoning task. The matrix reasoning task was not a unique predictor in the initial model (Table 31). This suggests that the more inductive aspects of the matrix task, which are not shared with the spatial tasks, did not correlate with performance overall, on the problem-solving questions. The variance was therefore shared with the spatial tasks (as outlined in the introduction). Controlling for performance on the matrix reasoning task may therefore be overly conservative.

Table 31: Multiple regression analysis predicting total problem-solving score, excluding matrix reasoning

Predictor	*b*	*ß*	*p*	Δ *F*	Sig Δ *F*	r^2	$r^2\Delta$
Step 1) IQ(Vocabulary)	.35	.38	<.001	33.97	< .001	.25	.25
Step 2) Prior knowledge	1.65	.28	.001	13.52	< .001	.36	.11
Step 3) Mental folding (I-D)	.35	.18	.038				
Mental rotation (I-D)	.46	.13	.132				
Arrival judgment (E-D)	.21	.10	.212	4.11	.009	.46	.10

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model

Table 31 presents the model for total problem-solving scores, but, excludes the matrix reasoning task. In this updated model, the combined spatial skills now accounted for 10% additional variance when added, in step 3. Notably, the overall model r^2 was the same as the previous model. Therefore, excluding the matrix reasoning task did not reduce the amount of variance explained, overall.

Table 32 presents the model for total problem-solving scores, again excluding the matrix reasoning task, but with steps 2 and 3 reversed. Entering the spatial tasks into step 2 accounts for 14% of the variance in total problem-solving scores, compared with 10% in the previous model. Spatial ability and prior knowledge therefore account for 4% shared variance in total problem-solving scores.

Table 32: Multiple regression analysis predicting total problem-solving score, excluding matrix reasoning

Predictor	b	ß	p	Δ F	Sig Δ F	r^2	r^2Δ
Step 1) IQ(Vocabulary)	.35	.38	< .001	33.97	< .001	.25	.25
Step 2) Mental folding (I-D)	.35	.18	.038				
Mental rotation (I-D)	.46	.13	.132				
Arrival judgment (E-D)	.21	.10	.212	7.59	< .001	.40	.14
Step 3) Prior knowledge	1.65	.28	.001	11.32	.001	.46	.06

For prediction scores (which were part of the composite problem-solving score above) the vocabulary and matrix reasoning task entered into step 1 accounted for 18% of the variance in scores, F (2,99) = 10.84, p < 0.001. In the final model (Table 33) only vocabulary was again the only significant IQ predictor, β = .31. Prior knowledge entered in step 2 accounted for an additional 3% of the variance in scores, F (1,98) = 3.84; it did not significantly improve the model, p = .053, but was marginally significant in the final model, β =.19, p = .048. This can be accounted for by the addition of the spatial task in the final model removing some of the error or noise in the data. The mental rotation task was then entered in step 3, F (3,97) = 4.105, p = .009, and accounted for an additional 4% variance in prediction scores. The final model accounted for 25% of the variance in prediction scores. A parallel model which, as above, excluded the matrix task is presented in Table 34. Within this model, there was a small increase in the variance explained by mental rotation (4%, rising to 5%).

Table 33: Multiple regression analysis predicting overall prediction total.

Predictor	b	ß	p	Δ F	Sig Δ F	r^2	r^2Δ
Step 1) IQ(Vocabulary)	.10	.31	.002				
IQ (Matrix reasoning)	.20	.05	.648	10.84	< .001	.180	.18
Step 2) Prior knowledge	.40	.19	.048	3.83	.053	.210	.03
Step 3) Mental rotation (I-D)	.24	.20	.036	4.51	.036	.246	.04

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ F is the p value of the change in F for each step of the regression model.

Table 34: Multiple regression analysis predicting overall prediction total, excluding matrix reasoning.

Predictor	*b*	ß	*p*	Δ F	Sig Δ *F*	r^2	$r^2\Delta$
Step 1) IQ(Vocabulary)	.10	.31	.001	18.71	< .001	.158	.16
Step 2) Prior knowledge	.40	.20	.034	3.83	.053	.210	.04
Step 3) Mental rotation (I-D)	.26	.21	.018	4.51	.036	.246	.05

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model.

Considering explanation scores (Table 35), the vocabulary and matrix reasoning task entered into step 1 accounted for 29% of the variance in scores, F (2,99) = 19.70, p = < 0.001. In the final model, only vocabulary was again the only significant IQ predictor, β = .36, p < 001. Prior knowledge entered in step 2 accounted for an additional 9% of the variance in scores, F (1,98) = 13.79; p < .001, and was a significant predictor in the final model, β=.29, p < .001. Mental folding and the arrival task were entered in step 3, accounting for an additional 7% of the variance in explanation scores, F (2,96) = 5.568, p = .005. Only mental folding was a unique predictor in the final model, β = .26, p = .005. The final model accounted for 44% of the variance in explanation scores. Table 36 presents the model without the matrix reasoning task. Within this model, the variance explained by spatial skills overall increased from 7% to 9%.

Table 35: Multiple regression analysis predicting overall explanation total

Predictor	*b*	ß	*p*	Δ *F*	Sig Δ *F*	r^2	$r^2\Delta$
Step 1) IQ(Vocabulary)	.25	.36	< .001				
IQ (Matrix reasoning)	.30	.03	.679	19.70	< .001	.29	.29
Step 2) Prior knowledge	1.30	.29	.001	13.79	< .001	.37	.09
Step 3) Mental folding (I-D)	.37	.26	.004				
Arrival task (E-D)	.12	.07	.395	5.568	.005	.44	.07

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model.

Table 36: Multiple regression analysis predicting overall explanation total, excluding matrix reasoning.

Predictor	*b*	*ß*	*p*	Δ *F*	Sig Δ *F*	r^2	$r^2\Delta$
Step 1) IQ(Vocabulary)	.26	.37	< .001	30.13	< .001	.22	.23
Step 2) Prior knowledge	1.31	.29	.001	17.47	< .001	.33	.12
Step 3) Mental folding (I-D)	.39	.27	.001				
Arrival task (E-D)	.13	.08	.327	7.90	.001	.42	.09

Note. Betas refer to values when all predictors are entered into the final model. The Sig Δ *F* is the p value of the change in F for each step of the regression model.

Because the explanation scores were not normally distributed, were low overall, and 22% of children scored zero, and were thus not able to provide a least one accurate explanation point, the were also analysed using a categorical approach. Although the regression models above were valid and the assumptions met (i.e., the residuals of the model had a sufficiently normal distribution), a categorical approach provides additional confirmation of the results.

Children were categorised into the following groups: children who got an overall score of zero on explanations, children who scored 1-3 on explanations, children who scored 4-8 on explanations, and children who scored 9 or more on explanations. This was based on dividing the frequencies, such that approximately 65% of the scores fell into the middle two categories (Pasta, 2009). Children who scored zero were used as a reference category, in a multinomial logistic regression model. The aim was to determine if spatial ability related to the likelihood of being in the higher scoring explanation groups, relative to the children who were not able to provide at least one accurate conceptual explanation point (i.e., children in the zero score group). In the regression model, the two IQ variables and pre-test measures were included, via a forced entry method. Rather than enter spatial scores on the basis of the prior partial correlation analysis, given the change from a continuous to a categorical approach, all of the spatial measures were entered, but, forward stepwise entry was used to select the best spatial predictors. In this type of model , the variable which would lead to the greatest model improvement is added first, and then further variables are only added into the if they significantly improve the model.

The overall model was significant, χ^2 (12, N = 102) = 62.98. The Nagelkerke r^2 value (pseudo r^2 statistic) was 49.8. Details of the predictors in the model are given in Table 35. As is shown in Table 37, and echoing the previous analysis, vocabulary and prior knowledge made unique contributions to the model. In addition, also in line with the previous analysis, the mental folding task was retained as the only significant spatial predictor.

Table 37: Summary of predictors in multinomial logistic regression model

Predictor	χ^2	Df	*p*
Vocabulary	9 .52	3	.023
Matrix Reasoning	4.23	3	.231
Prior Knowledge	18.34	3	< .001
Folding	9.64	3	.022

Table 38: Summary of predictors in relation to categorical outcome.

	*b (*SE)	Odds ratio	*p*
Zero score vs 1-3 score			
Matrix reasoning	.033(.09)	1.03	.718
Vocabulary	.117 (.06)	1.13	.086
Prior Knowledge	1.28 (.53)	3.60	.016
Folding	.131(.12)	1.14	.276
Zero score vs 4-8 score			
Matrix reasoning	.052(.10)	.95	.589
Vocabulary	.16 (.07)	1.18	.019
Prior Knowledge	1.73 (.55)	5.64	.002
Folding	.31(.13)	1.36	.017
Zero score vs 9+score			
Matrix reasoning	.16 (.16)	1.17	.266
Vocabulary	.286 (.11)	1.33	.009
Prior Knowledge	2.71 (.82)	15.05	.001
Folding	.59 (.26)	1.81	.023

Table 38 displays the specific parameter estimates for each predictor, based on each categorical comparison. Mental folding in the zero score vs. 1-3 score was not a significant predictor. However, folding was significant for the zero score vs. 4-8 score, and the zero score vs. 9+score. Considering first the zero score vs. 4-8 score, for every one unit increase in mental folding, the odds of being in the 4-8 score group, relative to the zero score group, increased by a factor of 1.36. Furthermore, for every one-unit increase in mental folding, the odds of being in the 9+ score group, relative to the zero score group, increased by 1.81.

4.4 Discussion

The overall aim of the current study was to investigate the relationship between children's spatial thinking skills and different aspects of their learning about sound propagation. Specifically, the study focused on children's recall of factual knowledge and conceptual understanding, as assessed through scientific predictions and explanations in the context of novel scenarios.

Although not a central aim of the study, the results provide some further evidence regarding the structure of children's concept of sound. The prior study by Mazens and Lautrey (2013) concluded that a 'knowledge as theory' approach, such as Vosniadou's Framework Theory (1994), best characterises children's concept of sound; they characterise the concept as being coherent and organised. The evidence in the current study provides some further support of this, for children of this age, and following direct instruction. First, there was a high correlation between predictions and explanations overall. This could be interpreted to indicate that both predictions and explanations orientate from the same mental model or models. Furthermore, despite the significant variations in contexts across problem-solving questions, and slight variations in knowledge required, there was a reasonable amount of consistency evident. For example, there was a good level of internal consistency (Chronbach's alpha) for explanation items overall, and, predictions and explanations combined. The rest of the discussion will therefore, tentatively, consider predictions and explanations as originating from the same source of knowledge (i.e. model). The findings below will be discussed again in detail, in the overall discussion, in relation to the theories of conceptual understanding outlined in Chapter 1.

A summary of the main findings in relation to the study hypotheses is provided in Table 39. Considering Hypothesis 1, as predicted, children's spatial thinking skills did not significantly predict performance on the factual recall questions. Likewise, matrix reasoning scores were not a significant predictor of factual scores. However, vocabulary was a strong, significant predictor. This suggests that the scientific factual information in the lesson was stored in a verbal, rather than spatial, format. In particular, the ability to retrieving specific definitions to words is a similar process to retrieving specific scientific factual knowledge. Given the findings for predictions and explanations, discussed below, the findings relating to factual knowledge suggest that spatial skills mostly support science skills which are more dependent on conceptual understanding (i.e., model construction) and application. This partly supports the finding by Rhodes et al. (2016) who found that visuospatial working memory predicted secondary school student's performance on conceptual application, but not factual, questions, in the context of learning about acids and alkalis.

Table 39: Summary of findings in relation to study hypotheses

Hypothesis	Findings in relation to hypothesis
1) Spatial ability not predictive of factual scores, or, relationship weaker than with problem-solving scores (predictions plus explanations).	Hypothesis supported. Spatial ability was not a significant predictor of factual knowledge scores, controlling for covariates.
2) Spatial ability predictive of overall performance on the sound problems.	Hypothesis supported. Controlling for covariates, spatial thinking skills accounted for 7% additional variance in total problem-solving scores.
3) Dynamic spatial skills stronger predictors of problem-solving scores overall than static spatial skills.	Hypothesis supported. Only dynamic spatial skills were significant predictors of problem-solving scores, after controlling for covariates.
4) Spatial ability stronger prediction of explanations overall than predictions.	Hypothesis supported. Spatial thinking skills accounted for 7% additional variance in explanations, versus 4% for predictions.
5) Two-factor intrinsic-extrinsic model fits better than one-factor model.	Hypothesis not supported. Two-factor intrinsic-extrinsic model did not fit better than one-factor model.
6) Two-factor static-dynamic model fits better than one-factor model.	Hypothesis not supported. Two-factor static-dynamic model did not fit better than one-factor model.

With Hypothesis 2, the expectation was that spatial ability, overall, would predict performance on the problem-solving questions. The findings revealed that spatial ability was indeed a significant predictor of total scores on the problem-solving questions, which required children to apply conceptual understanding of sound through the generation of predictions and explanations. Spatial skills overall accounted for between 7% and 10% of the variability in problem-solving scores, after controlling for prior knowledge, vocabulary and, in the more conservative model, fluid/abstract reasoning. Based on our initial theoretical expectations, it is proposed that children with stronger spatial skills were more able to develop a detailed and accurate conceptual mental model of sound. Furthermore, having a more complete and accurate mental model allowed children with stronger spatial skills to generate more accurate predictions and explanations from the scenarios.

These findings are consistent with prior research with adults which links spatial ability to mental model construction of physical processes (Narayanan & Hegarty, 2002; Jaeger, Taylor & Wiley, 2016; Sanchez & Wiley, 2014; Hegarty & Just, 1993). A further re-analysis of several studies, based on the same cistern mechanism described in Narayanan and Hegarty (2002), found that spatial ability was most predictive of learning outcomes when participants were required to integrate diagrams with verbal information (Hegarty & Kriz, 2008). In the current study, it may have similarly been the case that children with stronger spatial skills were more able to integrate information from static diagrams, dynamic animations and verbal description from the lesson, into an integrated mental model.

We now turn to the more specific hypotheses (Hypothesis 3 and 4). Beginning with Hypothesis 3, it was expected that the dynamic spatial thinking skills (within the Uttal et al., 2013 model) would be stronger predictors of problem-solving scores than the static skills. Results of the study showed that, controlling for both IQ measures, the dynamic spatial skills (i.e., skills which involved visualised or actual movement: mental rotation, mental folding and arrival judgement) were indeed more strongly correlated with performance on the problem-solving questions, compared to the static skills (spatial scaling and visual discrimination). This also aligns with the initial expectations. Specifically, dynamic spatial skills seemed more useful for understanding the process of sound, than the comprehension of static, scaled, relative spatial relations, and discriminating static visual forms. For example, having strong visualisation skills (i.e., mental folding), could support mental simulation of aspects of the model.

However, while the pattern of correlations and associations with total problem-solving scores fit with expectations, there was no strong evidence of the static-dynamic distinction, psychometrically (hypothesis 6). There was also no support for the intrinsic-extrinsic dimension of the Uttal et al. (2013) model (hypothesis 5). The static-dynamic, two-factor model did fit the data somewhat better than a one-factor model, but, the difference in model fit with the baseline model was not statistically significant. Moreover, in all models, the arrival judgment task loaded poorly with the other spatial tasks. This suggests, that multiple object dynamic spatial ability tasks, which involve speed-time judgments, are distinct from traditional spatial thinking tasks. This will be discussed further in the overall discussion.

In addition, though the dynamic spatial skills did have a stronger partial correlation with problem-solving scores than the static skills, only mental folding and mental rotation, and not the arrival judgment task, showed a unique relationship with any of the science scores, in the overall regression models. The finding that the arrival judgement task was not a unique predictor of conceptual explanations, contrasts to the results of Sanchez and Wiley (2014), who found an extrinsic-dynamic spatial ability (intercept) task uniquely predicted undergraduates' geoscience understanding, whereas mental folding did not. One possibility for this discrepancy is that the topic of study in the Sanchez and Wiley (2014) study related to plate tectonics and volcanic eruptions, therefore entirely involved concrete interactions between objects. In the current study, although there are causal elements involved in how we hear a sound, the process is less concrete (i.e., invisible vibrations of the air occurring). There is also clearly a large discrepancy in the ages between the samples.

With hypothesis 4, it was expected that children's spatial skills would be more strongly related to explanations than to predictions. The reasoning presented was that generating a prediction related to a given scenario would be less likely to draw on the entire mental model, or at least, require less co-ordination between different aspects or elements of the model. Many of the predictions in the assessment required participants to describe the likely outcome of a series of events, and therefore focused on a single aspect of the model. In contrast, it was suggested that explanations more so require the utilisation of the whole model, and an ability to co-ordinate between different elements within the model. Moreover, explanations involve a sequence of linked points.

In line with expectations, spatial ability accounted for a larger amount of variance in explanation scores than prediction scores. For both predictions and explanations, spatial skills categorised as intrinsic-dynamic in the Uttal et al. (2013) model (i.e., those requiring the spatial transformation and manipulation of objects) were the best predictors. There was, however, a difference in terms of which intrinsic-dynamic skill was a unique predictor. Mental rotation was more closely related to predictions, while mental folding was more closely related to explanations.

Comparing the two tasks, the mental folding task appears to require the more complex transformation of the two, given accuracy was 20% lower overall, than mental rotation. There are also, previously noted (see Chapter 1), differences between the two types of intrinsic-dynamic transformation. According to Atit et al. (2013) and Newcombe and Shipley (2015), mental rotation, is a rigid transformation, because the distances between pairs of points of an object are maintained. In contrast, in mental folding, a non-rigid transformation, the distances change as the transformation occurs. One interpretation is therefore that the stronger relationship between mental folding and explanations may be due to the greater degree of flexibility involved for this type of transformation (i.e., changing of points). This spatial flexibility may be more useful for model manipulation, which is more important for explanations.

These two transformations also fall into different categories within Carroll's (1993) spatial skills model (i.e., mental folding: visualisation, mental rotation: spatial relations). Tasks within these two Carroll categories also differ in their degree of executive function involvement (Miyake et al, 2001). It may be that the more complex transformation skill, mental folding, was more linked to explanations, because explanations are the more complex science skill. As outlined above, accurate and detailed explanations require access to more of the mental model and a sequence of linked points. This interpretation is supported by research into physics learning with adults, whereby spatial ability predicts reasoning when participants are required to co-ordinate between multiple aspects of physical situations, e.g., simultaneously reason about the effects of two forces acting upon an object (Kozhevnikov & Thornton, 2006; Isaak & Just, 1995).

The current finding that mental folding significantly predicts children's physics learning generally is in line with the findings of the previous chapter. The findings in this chapter build on the previous one by providing more detail about which specific types of science skills are related to, and are not related to, this mental transformation. Moreover, this chapter adds that the relationship is not accounted for by a more general reasoning ability. The effect size was similar between the two studies. Specifically, controlling for verbal ability, approximately 8% additional variance was explained in the overall science scores in the previous study, versus 10% additional variance explained in total problem-solving scores, in the current study. This suggests that spatial skills make a similar degree of contribution to reasoning based on long-term retrieval (in a science achievement assessment), to learning a topic in-depth based on more recent instruction.

To review, the mechanisms discussed above centered on spatial skills supporting the construction and use of a conceptual mental model of sound. However, a further explanation is that it may have been the case that children with stronger spatial skills were more able to manage the on-line spatial demands of the problem-solving questions themselves. For example, children were asked to predict what type of sound someone in one location would experience, if an object was struck at another location, given specific information about the different mediums involved. This required coordination of the spatial relations between the various elements in the scenario itself. This also relates to mental model construction, but the on-line construction a model of the problem scenario in question. Relatedly, each problem scenario included a visualisation, in the form of a diagram, to represent the elements of the scenario, and spatial skills also support the comprehension of diagrams (e.g. through 'mental animation' of the static diagram, see Chapter 1).

The two explanations are not mutually exclusive, however. Spatial skills may support the development and manipulation of the conceptual model within the lesson, and then also, within individual problems, spatial skills are used to interpret the specific spatial relations described, and generate a mental model. There may also be a coordination between these two processes. This is in line with suggestions from Vosniadou and Brewer (1992). The authors suggest that mental models are created 'on the spot' for answering questions and dealing with problems (cf. Johnson-Laird, 1980). However, they propose that this 'online' model is also constrained by an underlying model of the concept involved.

There are also limitations with the study to consider. First, an alternative reason why spatial ability was not a significant predictor of factual knowledge scores may have been due to the low internal consistency/reliability of the measure. Nevertheless, there are a number of reasons why the alpha reliability may have been low. First, although the lesson covered one overall topic, there was some diversity in the content, meaning there was a reasonable range of content on the factual knowledge test. This would have reduced the internal consistency score. In addition, an alpha reliability score is lessened by a smaller number of items. Factual knowledge scores did, however, correlate highly with overall problem-solving scores, which provides some additional support for the quality of the measure. Factual scores were also predicted well by vocabulary scores, which suggests they are both, in part, measuring an ability to comprehend and store verbal information. A second limitation is that a measure of classroom engagement was not administered, which could have been used as a covariate in the analysis. Finally, the number of trials in the mental rotation task was low, due to using the shortened version of the task. This decreases the reliability of the measure, although, the longer version of the test has been successfully used in the past and has proven reliability.

To conclude, this study demonstrated that spatial skills, in particular intrinsic-dynamic, transformation-based skills, specifically relate to children's scientific predictions and explanations about sound. This relationship remained when controlling for verbal ability and fluid reasoning, and was not observed in relation to recall of factual knowledge. Thus, it appears that spatial skills are most important for mental model construction and application. In particular, spatial skills support children in using the mental model to draw inferences and explanations, within given scenarios. In terms of implications to the classroom, these findings suggest that children with poorer spatial visualisation skills may benefit from additional support. For example, these children may benefit from additional exposure to concrete models (e.g., a concrete version of the model of the vibrational process shown in one of the videos in the lesson.) Alternatively, children with poorer spatial skills may need more practice at spatial visualisation, in the context of the particular scientific phenomena they are learning. Finally, directing children to visualise physical processes, in the context of learning, may be beneficial (cf. Palmquist et al., 2018; Joh, Jaswal & Keen, 2011).

Future research could focus on additional conceptual topics to determine if the observed relationships exist, when considering other scientific phenomena. Furthermore, manipulation of the specific aspects of the lesson (e.g., the use of static vs. dynamic visualisations) was beyond the scope of the study, but could be addressed in future research. In particular, a meta-analysis revealed that spatial ability is more predictive of learning from static than dynamic visualisations, and is also more predictive of learning from 2D, versus 3D, visualisations (Höffler, 2010). That is, learners with lower spatial ability perform worse with static and 2D visualisations (because they provide less of a spatial scaffold), and learners with higher spatial ability perform better. However, for learners with lower spatial ability, performance is not as affected if they learn with 3D and dynamic visualisations, presumably because they provide more spatial support. Finally, the specific role of mental folding, as a predictor of conceptual understanding of sound propagation, means that it could be targeted in the future as part of a spatial training study.

Chapter 5

The role of gesture as a spatial tool for learning about magnetism

5.1 Introduction

5.1.1 Overview in relation to findings from previous chapters

The findings in the preceding chapters indicated that individual differences in children’s spatial ability are a significant predictor of science learning and achievement. In Chapter 3, mental folding and spatial scaling were shown to be the best predictors of general science achievement. Mental folding skills were also the best predictor of physics scores. Building on the latter finding, Chapter 4 focused specifically on one physics topic, sound. In this study, mental folding and mental rotation were the best spatial predictors of problem-solving. Taken together, the findings suggest that intrinsic-dynamic spatial skills, which involve object-based spatial transformation, play an important role in the development and application of children’s conceptual understanding of physical processes.

A mechanism was proposed whereby children construct mental models which include dynamic, spatial-relational elements, isomorphic to the physics process or concept (see Chapter 1, section 1.4.5). Furthermore, it was proposed that children with stronger spatial transformation skills are able to construct more accurate mental models. As a result, these children are more able to run these models at a later stage, through a process of mental simulation. As discussed in Chapter 1, in line with theories of embodied cognition (see section 1.2.11), mental simulations are a recreation of the neural states which were involved when witnessing an action or event, without the physical action being performed or witnessed (Hostetter & Alibali, 2010). Spatial transformation skills allow children to manipulate the mental simulation, which in turn supports the generation of predictions, explanations and inferences, related to the physics concept. This may be particularly the case for physics topics which are heavily object-based (e.g., forces and motion, magnetism). To summarise, the suggestion is therefore that children who are more able to mentally transform an object (e.g., fold a shape), would also be more able to construct and manipulate mental simulations of physical processes, within mental models.

Based on these findings above and the mental model theory above, one logical method to improve conceptual understanding and reasoning in physics, therefore, would be to train spatial transformation skills (i.e., mental folding, mental rotation). Theoretically, improving spatial transformation skills would improve the accuracy of mental model construction, and the subsequent accuracy of inferences drawn from these models.

Another possibility for improving conceptual understanding and reasoning in physics is to utilise spatial tools (Newcombe, 2018). As outlined in Chapter 1, in defining spatial cognition, Newcombe (2018) states that as well as tool use and navigation (broadly mapping onto intrinsic and extrinsic spatial skills) humans also possess the ability to spatialise. Spatialisation refers to the ability to construct and use *symbolic* spatial tools, such as spatial language, models, diagrams, analogies, and gesture.

Compared to training spatial transformation skills, the implementation of spatial tools within the science curriculum would be fast, cost-effective and straightforward. Although paradigms vary somewhat, training spatial transformation skills would likely involve several training sessions over an extended period, and it may be necessary to work one-to-one with a child, which would be time-consuming and expensive. As outlined in Chapter 1, the focus of this Chapter is on gesture. It would be straightforward to include gestures within science lessons. For example, teachers could include and integrate them within their whole-class science instruction and children could produce them immediately.

To date, there has been very little research investigating the impact that gesture-based science instruction has on science learning in the primary school years, and no research has compared gesture to other types of spatial tools, such as concrete models. The primary aim of the current study was therefore to investigate the effectiveness of gesture as an instructional spatial tool for learning about a spatially demanding physics topic, magnetism, with primary school-aged children, and how this compares with other spatial tools, often used within science instruction.

The use of spatial tools may directly support learning. However, the extent to which spatial tools facilitate science learning may be also be moderated by spatial thinking skills, such as mental rotation. For instance, gestures and concrete models may be particularly helpful for children with weaker mental transformation skills, because they provide an external spatial scaffold, and also may facilitate the generation of mental simulations (see discussion below). Similarly, mental transformation skills may be particularly important when learning with diagrams, due to the need to 'self-animate' and co-ordinate between different elements, when visualising a dynamic science process. The secondary aim of this chapter was therefore to investigate how the use of spatial tools in science may be moderated by spatial thinking skills.

To achieve these aims, across two studies, children were taught about magnetism through either gesture, concrete models, or verbal description and diagrams only. They then completed an assessment of learning at the level of intermediate and far transfer. Children also completed a mental folding task. The following sections present further details about gesture as a spatial tool, and current research connecting gesture and learning, within and beyond science.

5.1.2 Gesture overview and theoretical rationale

Gestures are hand movements which represent relational, spatial and embodied concepts and are a special kind of action. Although there are several notable typologies of gestures (e.g., Efron, 1941), the typology proposed by McNeill (1992) has received significant support within the gesture literature, and provides the most parsimonious categorisation. McNeill (1992) distinguishes between gestures along several dimensions. For example, *dietic gestures* point to existing objects; such as when we point to give someone directions. *Iconic gestures* represent the semantic content of speech. An example of an iconic gesture is using two hands a specific distance apart to communicate the desired length of a piece of string. This chapter focuses primarily on iconic gestures, and their role in learning.

Theoretically, there are good reasons to predict that the use of gesture supports children's conceptual understanding of physical processes. A dominant theoretical model is the 'gesture as simulated action theory' (Hostetter & Alibali, 2008). Within this theory, the authors propose that gestures are produced as an expression of the mental simulations of actions and events. For example, hearing a sentence such as 'I kick the ball' produces an action-based mental simulation through activation of areas of the premotor cortex where these actions are motorically coded (Tettamanti et al., 2005). Hostetter and Alibali (2008) further propose that gesture and action are both produced from a similar action-related system. Thus, when a representational (iconic) gesture is produced, this is also represented neurally with reference to the action that is represented, akin to the mental simulations observed when reading and hearing an action sentence. The mental model proposal at the beginning of the chapter, above, was that children's conceptual models of physical processes are supported by mental simulations. It is therefore feasible that encouraging children to produce gestures, which are linked to mental simulations of actions, would support the construction, maintenance and use of these conceptual models. Furthermore, it may be that by producing gestures, which are close to the simulated actions they represent, simulations are more likely to be facilitated, than through learning through verbal description of those actions.

Kita, Alibali, and Chu's (2017) 'gesture-for-conceptualisation' theory, proposes that gestures have four functions. First, gestures *activate* spatio-motoric information (i.e, information related to spatial concepts or relationships, or to movements and actions). For example, gesturing might maintain an existing spatio-motoric representation which is already active (e.g., thinking about a spatial concept). As a result of this activation, the extent to which the representation decays, is reduced. Second gestures *manipulate* spatio-motoric information. For example, producing a gesture simultaneously along with solving a spatial transformation task, such as mental rotation, assists in manipulating the spatio-motoric information that is central to the problem. Third, gestures *package* spatio-motoric information into smaller units. For example, when gesturing a route, it is likely that an individual might use several different gestures. Gesturing therefore assists in making complex spatio-motoric information more manageable and communicable. Finally, gestures *explore* spatio-motoric information. For instance, gestures assist in searching for ideas, such as generating novel ideas for the uses of an object.

Kita et al. (2017) also suggest that, within a single interaction or event, a gesture may simultaneously serve more than one function. For example, when a speaker explains different possible solutions to a mental transformation, gesturing may activate spatio-motoric (to reduce representational decay), explore spatio-motoric information (to assist in searching for multiple solutions) and also manipulate spatio-motoric information (to simulate the actions of the objects to be transformed). Many rudimentary physics concepts are rooted in spatio-motoric information. It seems reasonable, therefore, that gesture could support children's understanding and reasoning, of these topics, through some or all of the mechanisms above.

Given the central role of physical action in many physics topics at this level, it would also be reasonable to predict that the best pedagogical approach might be to provide concrete, action-based experiences, where it is possible to do so. In support of this, having physical experience of angular momentum improves undergraduate's conceptual understanding of the topic, compared with observation of the same action (Kontra, Lyons, Fischer, & Beilock, 2015). Yet, in some cases, physical experiences are harmful to reasoning. For example, physically holding objects before making judgments about floating and sinking may reinforce naïve misconceptions that weight relates to sinking rate (Castillo, Waltzer, & Kloos, 2017).

Despite the link between gesture and action discussed above, research has shown that gestures are a special kind of action; that is, they are somewhat removed from action, due to their representational qualities (Goldin-Meadow & Beilock, 2010). More specifically, compared to physical actions, gestures do not have an effect on the world. Thus, it is possible that gesture is not only useful for reasoning about physical processes, it may be more effective than action. There are several mechanisms through which gesture may be more effective than action (summarised in relation to the current study below).

5.1.3 Prior research on the role of gesture in relation to learning

In the domain of learning, there is evidence that the gestures a speaker spontaneously produces during a task predict how well they are likely to profit from later instruction on that task. For example, Perry, Church, & Goldin-Meadow (1988) found that some children made speech-gesture mismatches when being interviewed about a Piagetian conservation task. That is, they sometimes conveyed information in their gestures which was different to what was conveyed in their speech. Gesture-speech mismatches were evident, for example, when children spoke about one variable (e.g., height), but also included gestures which reflected another variable (e.g., width). Children who made these gesture-speech mismatches were more likely to subsequently benefit from instruction on conservation, than those who did not show these mismatches. Similarly, Pine, Lufkin, & Messer (2004) found that children who showed gesture-speech mismatches were more likely to show greater learning gains on a balance beam task.

Overall, children who simultaneously produce an idea in speech and a different through gesture are said to be demonstrating cognitive instability. This instability is an indicator that the person is ready to undergo a knowledge transformation (Pine et al., 2004).

More relevant to the current study, there is also some evidence that directing learners to produce gestures supports learning. There is also some mixed evidence for the advantage of directing learners to produce gestures, over directing learners to use concrete objects, in supporting learning. The most notable examples to date with children have been in the domain of mathematics. Novack et al., (2014), for example, taught 8-9 year old children to solve mathematical equivalence problems (e.g., 2+9+4 = _ +4) using either a physical action, concrete gesture or an abstract gesture. The physical action involved children moving digits across the "equals" sign of an equation to make it equal on both sides. Concrete gesture mirrored the action, without physical objects. Finally, abstract gestures involved making a V gesture with two fingers on one side of the equation, pointing to two numbers, and a single finger point on the other side of the equation, pointing to the space with the missing number. Considering performance on the near transfer questions (questions which could be solved using the taught strategy but were not presented during teaching), participants in both gesture conditions scored more highly than participants in the action condition. However, only the abstract condition led to higher performance in the far transfer questions (equivalence problems which required a deeper understanding, and could not be solved directly using the taught strategy).

This effect does, however, appear to be topic-linked, and, perhaps also depends on individual differences of the learner. For instance, on the topic of measurement, only children who demonstrate conceptually more advanced measurement strategies *before* learning, benefit from instruction through gesture and action. However, even for these children with more sophisticated prior knowledge, gesture-based teaching is no more effective than action-based teaching (Congdon, Kwon, & Levine, 2018). Thus, in Congdon et al. (2018), learning through gesture and action were equally as effective.

Relevant directed-gesture research in the science domain to date has only been conducted with adults (e.g. Stieff, Lira, and Scopelitis, 2016; Atit, Gagnier, and Shipley, 2015). These studies are therefore particularly important in the context of the current study, because they provide the closest point of empirical comparison currently available. Stieff, Lira, and Scopelitis (2016) compared the effectiveness of gesture and concrete models as spatial tools for learning about molecular structures. University student's learning outcomes were equally as good in the concrete model and gesture conditions, overall. However, students taught with concrete models performed worse if these models were not present during the post-test. In contrast, for participants in the gesture condition, performance was not affected if the concrete model was not present during the follow-up. The gesture condition also performed better than the concrete model condition, when there was no model available in the post-test. The frequency of gesture production in the post-test did not correlate with learning outcomes. Gesturing perhaps therefore encouraged the formation of a richer internal representation of the structures during instruction, which was not dependent on having models available. The model-trained group perhaps became dependent on the concreteness of the model, and were not forced to construct as rich a representation of the structures.

Atit, Gagnier, and Shipley (2015) investigated the importance of gesture in geoscience learning. The study focused on penetrative thinking, an important skill in geoscience, which involves visualising and reasoning about the interior of a 3D object from what is visible on the surface. In geoscience, penetrative thinking is important for understanding concepts such as rock formations and groundwater flow. In the study, the intervention involved participants providing a geoscience explanation, which involved penetrative thinking, with their hands (gesture allowed), or providing an explanation whilst sitting on their hands (gesture prohibited). Before and after the intervention, participants completed a 'geologic block cross-sectioning test', a measure of penetrative thinking. The results of the study showed that only participants in the 'gesture allowed' condition improved significantly from pre-test to post-test. A possible limitation with this study is that the 'gesture prohibited' group sat on their hands, which may have impacted upon performance due to feeling awkward or uncomfortable, rather than only being because participants were not able to gesture. Such an effect may be exacerbated with children; this is discussed further below in the context of the current study.

5.1.4 Current studies

To summarise, there are strong theoretical reasons to predict that gesture use would be beneficial to children's physics learning. Empirically, a small number of studies have a small number have been conducted, all of which found that that directing learners to gesture can improve learning and reasoning in science domains. However, this has been limited to geoscience and chemistry topics, with adult samples. In addition, only one study has compared the effectiveness of gesture with the effectiveness of concrete models, and again, this was limited to undergraduate students and a biochemistry topic.

The current studies extend the literature by focusing on teaching a physics topic, magnetism, to children. This topic was chosen because it is based heavily on spatio-motoric information. Moreover, there is solid evidence that gesture supports the mental rotation of objects, a mechanism that is strongly associated with physics (e.g., Chu & Kita, 2008). Therefore, there are good reasons to predict that gesture use would also support understanding and reasoning about magnets and magnetism, given the way that children physically manipulate and explore magnets.

The focus was on investigating whether teaching children about magnetism using gesture, and asking children to also use these gestures within the lesson, would lead to better understanding, than children who were taught with magnets only (hereafter referred to as the concrete model condition). These two conditions were compared to a third condition of teaching through diagrams and verbal description only (hereafter referred to as the verbal/diagram condition). The post-tests in both studies included *near transfer* problem-solving items (which related to the taught magnet – bar magnets) and *intermediate transfer* problem-solving items (which related to different, untaught magnets, e.g., disc magnets).

Both the use of gesture and use of concrete models provide grounded, spatio-motoric experiences related to the scientific concept. Theories of embodied cognition predict that both conditions provide learning benefits, relative to teaching through static diagrams. For example, during later problem-solving, in both cases, when accessing the magnetism mental model, there would be a reactivation of brain areas associated with the past sensorimotor movements of either the magnets (concrete model condition) or gestures (gesture condition). This suggestion is in line with the imaging component of a study by Kontra et al., (2015). The findings in this study indicated that the learning benefits associated with physical experience of angular momentum were explained by reactivation of sensorimotor brain regions, when subsequently reasoning about the topic *after* teaching, in the post-test. When children are asked to predict the actions of magnets in the current study, children taught with gesture and children taught with concrete models should be able to link their sensorimotor experiences to the problems, because children in both conditions will be more able to mentally simulate the actions of the magnets. However, it was expected that children in the gesture teaching group would perform better than the concrete model group on the post-test, particularly on questions that require intermediate transfer. There are several reasons for this hypothesis, listed below.

First, gestures are more schematised than concrete objects (Kita et al., 2017). The process of schematisation involves removing elements of a representation, but, maintaining others. Schematisation encourages people to focus on the important elements or aspects of a representation, and not on features that are tied to that specific representation (Goldin-Meadow & Beilock, 2010). By developing a more schematised, representation or mental model, it was predicted that children would be more able to approach novel contexts, and less tied to the examples given in the teaching session. Through the development of a schematised representation, children are therefore less likely to be distracted by the specific, surface level, perceptual features of the taught version (i.e., one end is blue and one end is red), which are absent in the post-test items.

The use of a schematised gesture might encourage spatial analogical reasoning (see Chapter 1), as described in the prior study by Atit et al. (2015). The analogical mapping (see section 1.4.2) between the taught version of the magnet (a bar magnet) and different types of magnets in the post-test (e.g., horseshoe magnets) may be enhanced by schematised gestures which contain spatial relations, or similarities, between both. The poles of the magnet are a common feature between both the bar magnets, and other types of magnets. The schematised gesture retains the poles of the magnet, but strips away other features such as the oblong shape of the bar magnet, which renders the schema suitable for transfer across novel contexts. Thus, an analogical mapping between the bar magnet and a different magnet, based on the common feature of the poles, might be better supported, because the feature is more salient, within the schematised gesture.

Such an explanation relates to a 'contrasting cases' approach to instruction. For example, Schwartz and Bransford (1998) investigated the effect participants analysing contrasting examples, as a method of preparation for subsequent instruction. The results of the studies showed that students who analysed contrasting cases prior to reading a text or receiving a lecture on the same topic produced more accurate conceptual predictions, compared to various control instructional methods. The authors argued that analysing contrasting examples increased students' ability to distinguish between features of the various concepts during learning. In the context of the current study, bar magnet gestures in the gesture-instruction condition were produced during instruction alongside images of bar magnets. Children therefore had the opportunity to contrast the bar magnet with the schematised bar magnet gestures, and as discussed above, this may highlight and distinguish between the key features of each. This, in turn, may result in a deeper understanding of the concept.

By gesturing, children might develop a richer internal representation of the magnets themselves, or, relatedly, gesturing might encourage children to pay attention to the spatial features of the objects. This mechanism is most endorsed by Goldin-Meadow & Beilock (2010) and is supported by gesture research which has used the Tower of London task (e.g., Beilock & Goldin-Meadow, 2010).

For example, in Beilock and Goldin-Meadow (2010), undergraduate students first solved the Tower of London task, and then explained their solution with gesture plus speech. In this initial stage, the weight of the disc correlated with the size of the disc; the heaviest discs were the largest. Following from this, for half of the participants, the weights of the discs were reversed such that the biggest discs were now the lightest, and the smallest discs were the heaviest ('switch group'). The key finding was that being in the 'switch group' worsened participants' performance, in the final solving of the task. The authors argued that by gesturing about their initial solution, participants developed a rich, internal representation of the objects. When the participants gestured in the initial explanation, they gestured about the small object with one hand, because it was the lighter object, which afforded the use of one hand. The participants therefore developed a representation of the small disc as light. They argue that participants performed more poorly in the switch condition, because the new weight/size combinations were incongruent with the representations developed during the initial gesture explanation.

In the current study, the affordances of the objects (i.e., properties such as magnetic qualities, poles of the magnets) were not directly available when gesturing during teaching. Participants were required to actively generate and produce these properties through their gestures. For example, producing accurate gestures corresponding to the information about two magnets repelling. Through a richer internal representation, children may then be more able to visualise, and mentally simulate, the magnets when problem-solving. Relatedly, by paying more attention to the spatial features of the magnet, children may more readily attach labels to individual parts of the magnet.

Finally, it also possible that by encouraging children to gesture in the teaching phase, they may then subsequently use gesture during the post-test items, many of which require the manipulation of spatio-motoric information. For example, gesturing in the post-test may have the function of off-loading spatial reasoning to the hands (i.e., reducing cognitive load – Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001). Moreover, any of the four gesture functions within Kita et al.'s (2017) model could theoretically enhance problem-solving during the post-test. Therefore, participants' gestures during the post-test were video-recorded, to determine if gesture in the production phase correlated with learning outcomes, and, if the frequency of gesture production was higher in the gesture condition, than the other conditions.

The predictive and moderating role of intrinsic-dynamic spatial ability was also explored, which was measured through a mental folding task. Prior findings in the thesis, particularly those in relation to physics topics, suggest that intrinsic-dynamic spatial thinking skills were likely to be a predictor of learning outcomes related to magnetism. Furthermore, the previously reported findings of Sanchez and Wiley (2014) suggest that, in some learning contexts, spatial thinking skills interact with the mode of science instruction. Specifically, this study reported that spatial thinking skills were significantly correlated with geoscience instruction when the mode of instruction was based upon static diagrams, but, not when it was based upon dynamic animations. The authors suggest that in the static diagram condition it was more necessary to self-animate the diagrams in order to understand the processes described, which therefore required spatial thinking skills. When the images were externally animated, spatial thinking skills were less important. Such a finding was also possible in the current study, given the variation in the type and modality of instructional spatial tools.

5.2 Study 1

5.2.1 Hypotheses

In study 1, performance was compared across three conditions: gesture-based teaching, concrete model based teaching; and teaching based on verbal description and diagrams only. In this study, prior to the main teaching, children were also given one minute of non-directed, free play with bar magnets. The rationale here was that, for gesture-based teaching to be effective, it was expected that children of this age would first require initial physical familiarisation with the concrete models, in order to make links between their physical experiences and the less perceptually rich gesture representations. For example, to be able to produce gestures that represent repelling forces, they would first benefit from experience of those repelling forces.

A summary of the study hypotheses is given in Table 40. It was expected that, following one minute of familiarisation with concrete models, children in the concrete model and gesture teaching conditions would outperform the diagram/verbal teaching condition, for the near transfer questions (hypothesis 1 and 2). In both the gesture and the concrete model conditions, children would have additional, sensorimotor experience related to the magnets, beyond the initial familiarisation, in the context of the teaching related to bar magnets, relative to the verbal description condition. It was also expected that the gesture teaching condition would outperform the concrete model teaching group for near and intermediate transfer questions (hypothesis 3 and 4). Gesture may encourage the construction of a richer internal representation, or, facilitate schematisation. However, for the reasons discussed previously (i.e., schematisation) it was expected the effect would be larger for intermediate transfer than near transfer questions. Regarding spatial ability, in line with previous research and findings in the thesis, it was expected that mental folding would be predictive of scores on both the near and intermediate transfer questions (hypothesis 5 and 6). There were no firm hypotheses regarding any possible moderating effects of mental folding on teaching condition.

Table 40: Summary of study hypotheses.

Hypothesis 1	Concrete model teaching > verbal/diagram teaching (near transfer questions)
Hypothesis 2	Gesture teaching > verbal/diagram teaching (near transfer questions)
Hypothesis 3	Gesture teaching > concrete model teaching (near transfer questions; smaller effect than on intermediate transfer questions)
Hypothesis 4	Gesture teaching > concrete model (intermediate transfer questions; larger effect than near transfer questions)
Hypothesis 5	Mental folding predictive of performance on near transfer questions
Hypothesis 6	Mental folding predictive of performance on intermediate transfer questions

5.2.2 Method

5.2.2.1 Participants

75 children from UK Year 4 state primary schools in London participated in this study, mean age: 8.8 years, S.D.: .28. Children of this age were chosen because they would have a sufficient level of scientific skill to meet the cognitive demand of the near and intermediate transfer problems, but, would not have had too great an exposure to the topic. Children were recruited from three schools (school 1: n = 40, school 2: n = 19, school 3: n = 16). Ethical approval was given by the UCL ethics committee, and parents gave informed consent to take part in the study.

A power analysis, conducted in G*Power (Power = 0.8, α = 0.05) indicated that a sample size of 90 was needed to detect a medium effect (f^2 = .15) for the most complex analysis (general linear model with 3 groups, 2 continuous variables and an interaction variable). The achieved sample size of 75 was therefore slightly underpowered to detect a medium-sized effect. A medium-sized effect was chosen as the anticipated effect size, based on previous research (e.g., Stieff et al., 2016).

5.2.2.2 Design

Children were randomly assigned to either the concrete model teaching condition, the gesture teaching condition, or the diagram/verbal teaching condition.

5.2.2.3 Materials and procedure

The researcher worked one-to-one with children in a quiet space in the child's school. The entire procedure took approximately 40 minutes and followed a fixed order, as listed below.

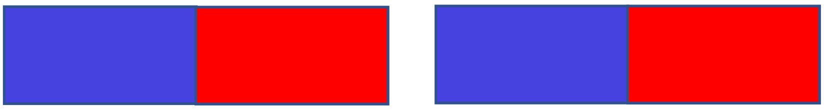
5.2.2.3.1 Mental folding task

Children first individually completed the mental folding task for children (Harris et al., 2013) used in Chapter 2, 3 and 4. As in Chapter 4, the task was presented on a 13" Macbook laptop. This measure of spatial ability was chosen because it was the best predictor of science physics learning in the preceding chapters. In addition, it fit with the hypothesis regarding mental simulation of actions, discussed in the introduction. Finally, it was also the highest loading spatial skill in the confirmatory factor analysis models (Chapters 2 and 4).

5.2.2.3.2 Prior knowledge

Prior knowledge was then screened with three items about magnetism, which covered content relevant to both the near and intermediate transfer questions (Figure 36). Two of the four questions had a possible score of two, and one had a possible score of one, leading to a maximum score of five.

1) Imagine you were holding these two magnets below, as they are shown.



a) Would you expect these magnets to attract, or repel?

☐ Attract ☐ Repel

b) Label (N) north and (S) south on the magnets above

2) The magnets in question 1 are **bar** magnets. Name as many other, **different types** of magnets that you know:

Figure 36: Prior knowledge questions.

5.2.2.3.3 Pre-teaching familiarisation activity

After this, all children completed the pre-teaching, familiarisation activity. In this activity, they were given two bar magnets, shown in Figure 37, and told that for one minute, they could play with, use and look at the magnets. They were told that they did not need to explain or describe their experience.

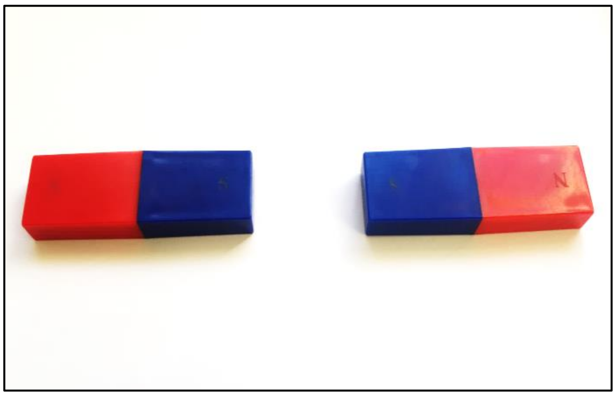


Figure 37: Bar magnets used in study 1 and study 2

5.2.2.3.4 Instruction phase overview

Following this, the researcher delivered the teaching content, which took approximately 10 minutes. The content covered in all three conditions is summarised in Table 41. The same basic PowerPoint slide presentation was used for all three conditions. The verbal teaching content (i.e., written text read by the researcher) referred to magnets generally, but, included diagrams of bar magnets, only. The presentation included written text and schematic diagrams of bar magnets (see Figure 38). The main teaching presentation included 13 slides covering the content in Table 41.

Table 41: Summary of teaching content for all conditions

Definition of a magnet
Definition of magnetism
Magnetic poles and magnetic forces
How magnets attract and repel
Magnetic and non-magnetic objects, with examples

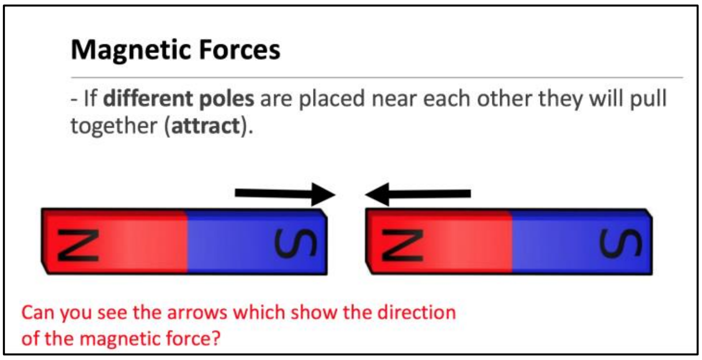


Figure 38: Example slide from magnetism teaching presentation

The following sections describe the instruction phase procedure for the three conditions separately.

5.2.2.3.5 *Instruction phase: concrete-model condition*

The researcher first told participants in the *concrete model group* that they were going to be learning about magnetism. The researcher informed participants that he would be using magnets and other objects to explain, and they would also be using the magnets, and other objects, too. The researcher read the written content on each of the slides, and at pre-specified points, modelled the concepts with the bar magnets shown in Figure 37 (e.g., demonstrating attracting and repelling poles). Immediately after these pre-specified points, the researcher gave the child the magnets, and asked the child to recreate these actions for themselves. This was also the case for the magnetic/non-magnetic object section. The researcher showed the magnetic object attracting to the magnet, and then gave the object and the magnet to the child to try themselves.

5.2.2.3.6 *Instruction phase: gesture condition*

The researcher told participants in the *gesture group* that they would be learning about magnetism, but, that he would be using hand gestures, to represent magnets, to help explain. The researcher explained that the participant would also be using hand gestures. The teaching procedure was identical to the procedure above, except that rather than demonstrating with magnets, the researcher demonstrated through specific gestures, to represent the magnets and movements of the magnets. Likewise, rather than the child then performing the same actions with concrete models, the child reproduced the same gestures, as the researcher.

Both the researcher and the participants used concrete, iconic gestures to represent magnets (Figures 39-42). The gestures co-occurred alongside verbal explanation/description, common to all conditions, as part of the teaching. The hand was placed out flat and the thumb was tucked underneath the hand. Of the four remaining fingers, the index and second finger were labelled as being the north pole of the magnet, and the third and fourth finger were the south pole of the magnet. The opposite hand was also used, as a magnet, and the correspondence between the fingers magnetic poles was used, although symmetrical to the other hand (Figure 39). Moving gestures were also used to describe or show forces. For example, the researcher and child made gestures to represent the repelling force by pushing each hand together and miming the feel of resistance (Figures 39-41). It was also possible to rotate, flip and rearrange the hands in a similar way as would be the case with a concrete magnet (Figure 42). The researcher and participant also produced gestures to represent objects that had been introduced in the teaching materials section. First, they produced a gesture to represent holding the magnetic object, a magnetic disc. The participant and the researcher held their fingers in a 'pinching' type gesture to represent holding the magnetic object (Figure 43). For the non-magnetic object (the plastic cup) – the participant and the researcher produced a static, hand grasping gesture, to representing holding the cup (Figure 44). These object gestures were paired with a single magnet gesture from above (i.e., one hand produced the object gesture, and the other produced the magnet gesture). The researcher and participants also moved the object gesture with the magnet gesture, to represent the object either attracting or not attracting to the poles of the magnet.

The rationale for this element of the gesture instruction was that, within the teaching content, children were taught that a magnet will only attract to another magnet if the poles are not alike, that non-magnetic objects will not attract to either pole of a magnet, but, that magnetic objects will attract to both poles of a magnet. This may be a concept that children confuse. In the transfer test, for example, one item showed a magnet attracting to a metal object, and asked children what would happen if the magnet was turned over, and explain why. It is possible that children may over-extend their understanding of two magnets together, and predict that it would no longer attract (the answer is that it would still attract). The prediction was that, by gesturing about the fact that magnetic objects attract to both poles of a magnet in the teaching, in particular, it would assist in distinguishing this process from the behaviour of two magnets.

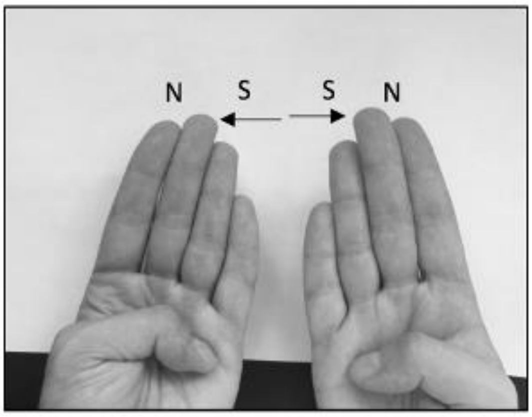


Figure 39: Gestures representing repelling south poles. Participants moved hands to represent the resistance felt.

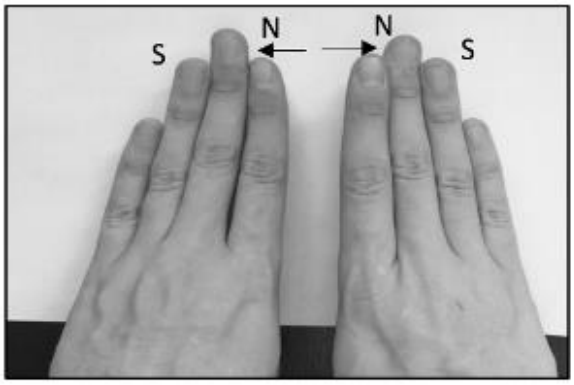


Figure 40: Gestures representing repelling north poles.

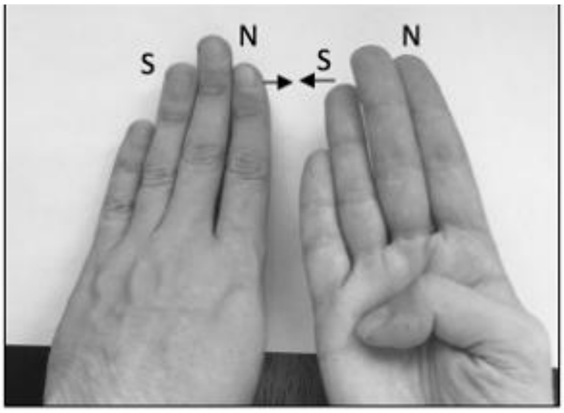


Figure 41: Gestures representing attracting poles. Participants moved hands together to represent the force.

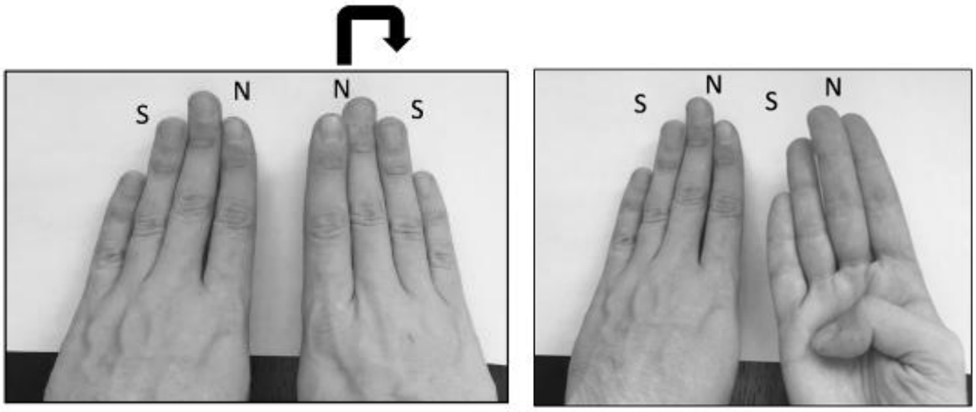


Figure 42: Gestures representing changing from repelling to attracting poles by rotating hand 180º

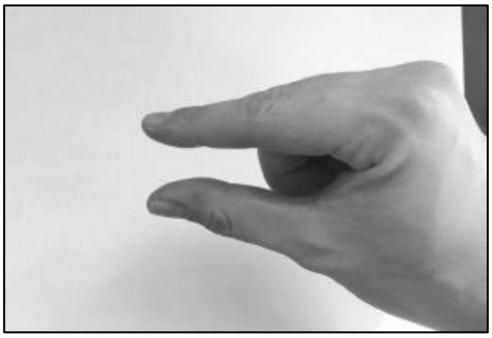


Figure 43: Gesture representing holding a magnetic object (metal disc).

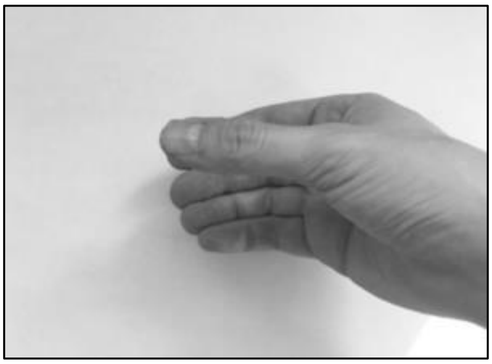


Figure 44: Gesture representing holding a non-magnetic object (plastic cup).

5.2.2.3.7 Instruction phase: verbal-diagram condition

In the *verbal-diagram* condition, the researcher also worked through and read the information given in the slides, which were accompanied by diagrams. There were, however, a number of differences to the magnet and gesture condition. First, recall that, in the gesture and concrete model condition, the researcher would at certain points, model action or gesture, and the child would repeat. In the verbal/diagram condition, at these same specified points, the researcher read the same verbal information which accompanied the gesture/concrete model teaching. However, in this condition, the child also repeated back this same information. Thus, at the points in the teaching where the child was actively engaged in the other two conditions by repeating actions or gestures, in the verbal/diagram condition, they were actively engaged by repeating related pieces of verbal information in the context of the diagrams. The rationale for this was to match the levels of active participation, across all conditions. The correction method for the problem was the same as the concrete model and gesture condition.

5.2.2.3.8 Consolidation phase

Following the Instruction Phase, there were four additional slides, which included consolidation questions relating to the learning that had taken place in the Instruction Phase. For the gesture and concrete model conditions, each question slide included one written question, in which the child was asked to arrange the magnets, or produce magnet gestures, in a particular way (Figure 45). Therefore, in the gesture condition, to show south and south together, they showed the correct gestures for south and south together. In the verbal/diagram condition, these four slides included instead questions which required children to choose one of three sets of magnet diagrams. Each question in the verbal/diagram condition referred to the same magnet combination to the corresponding question in the concrete model /gesture condition, but was reworded, to involve the child choosing the set of diagrams which represented the arrangement (see Figure 46). If a participant made an error on a question, the researcher told them it was not correct, and gave them an opportunity to correct the mistake and make a further attempt. If the answer was still not correct, the researcher told the participant the correct answer. The number of questions correct on the participant's first attempt was used as the consolidation score (see analysis strategy).

The final slide of this phase consisted of a problem related to bar magnets (Figure 47). In the concrete model condition, participants were first asked to make a prediction, without using the magnets, then to explain their answer, using the magnets. In the gesture condition, the child was asked to predict first (without gesturing) and then explain their answer, using the gestures learnt in the teaching. In the verbal/diagram condition, the child gave their prediction and explanation verbally. For this condition only, the problem was printed on an A4 sheet and placed onto a clipboard. To prevent children from gesturing during their explanation, the researcher asked the participant to hold the clipboard throughout their verbal prediction and explanation. For the problem question, the researcher did not tell the child if their answer was correct or incorrect, and did not correct them.

Can you show me two poles attracting, tell me what poles they are, and then change the arrangement to repel?

Figure 45: Concrete model/gesture consolidation question

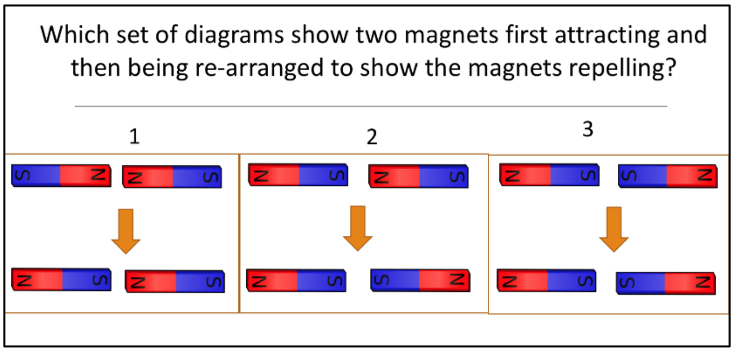


Figure 46: Verbal-diagram consolidation question

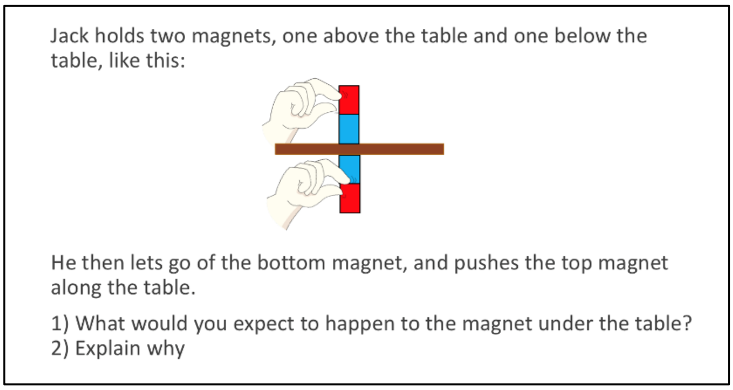


Figure 47: Word problem slide

5.2.2.3.9 Assessment phase

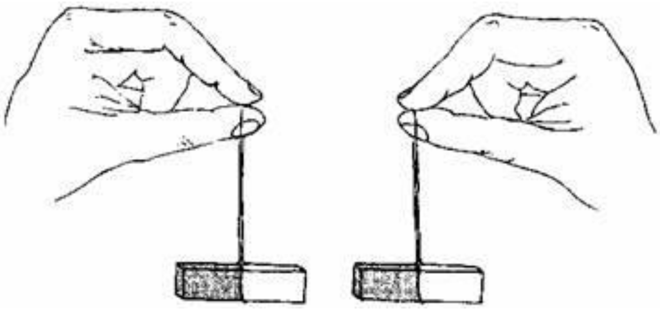
Children in all conditions then completed the assessment of learning, which included near and intermediate transfer questions. Near transfer scenarios required the application of content taught in the teaching section, applied only in the context of the type of magnet discussed or viewed in the teaching, i.e., they related only to bar magnets. Intermediate transfer questions, in contrast, related to other types of magnets which were not taught in the lesson, in any experimental condition (i.e., disc magnets and ball magnets).

Participants completed three near transfer problems and three intermediate transfer problems. The problems were either taken directly from, adapted from, or constructed in a similar style to, questions typically found in past UK national standardised science assessment tests (e.g., Qualifications and Curriculum Authority [QCA], 2002). The aim was to closely align the assessment with the requirements of learning assessments children typically undertake in the classroom, thus increasing the relevance to practice. Items varied in their requirements, including: labelling diagrams; drawing diagrams; making predictions and providing conceptual explanations of predictions. The near and intermediate transfer problems included a mixture of these item types.

Each problem contained several items (ranging from 2-4). Items varied on the number of possible marks available (1 or 2). The total score available for the 3 near transfer questions, combined, was 15, and the score available for the 3 intermediate transfer questions, combined, was also 15. The mark schemes were either taken directly from the original testing material, or, where new problems were constructed, were created in a similar format.

An example of a verbal and diagrammatic item for each question set is provided in Figure 48 (near transfer) and Figure 49 (intermediate transfer). Example responses for the verbal items are provided in Table 42 and Table 43. A full version of the assessment is provided in the appendix (section 8.2 and 8.3).

Wayne has two bar magnets. He hangs each bar magnet from a piece of string. He holds them close together like this:



What happens to the magnets when he holds them like this?

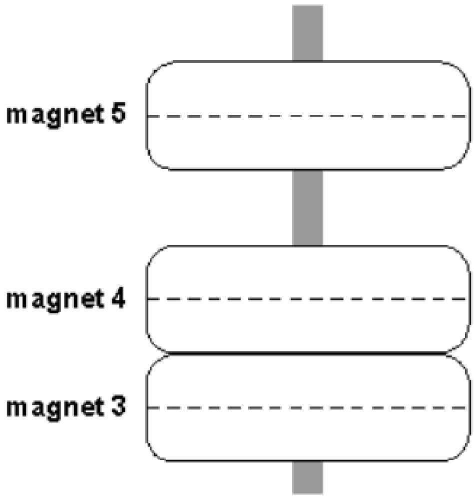
Draw an arrow on each magnet to show the direction of the magnetic forces

Figure 48: Diagrammatic and verbal item example from near transfer question set. Source QCA (2002)

Table 42: Example responses for example near transfer verbal item.

Question	Score	Example response
What happens to the magnets when he holds them like this?	0	'those two will repel'
	1	'the magnets would attract together'

On the diagram below, label 'N' and 'S' to show how the arrangement of the poles of the magnets might be, now, for magnets 3, 4 and 5.



If Wayne wanted magnets 3,4 and 5 **to all float together above magnet 1 and 2,** what would he need to do?

Figure 49: Diagrammatic and verbal item example from intermediate transfer question set. Source QCA (1999)

[note: magnet 1 and 2 were shown to participants in a separate diagram]

Table 43: Example responses for example intermediate transfer verbal item.

Question	Score	Example response
If Wayne wanted magnets 3,4 and 5 to all float together above magnet 1 and 2, what would he need to do?	0	'he would need to put them on the opposite way'
	1	'you would need to turn magnet 5 around'

Participants all completed the near scenarios first, and then, completed intermediate transfer scenarios. Participants always answered the near transfer scenarios first because they naturally progressed from the teaching section. The consolidation questions included a worded problem about bar magnets. The problems in the near transfer test which were about bar magnets. It was decided that having children answer the intermediate transfer problems first, which were about unfamiliar types of magnets, would have been too much of a ‘jump’, from the consolidation phase.

The researcher administered the questions in an interview-type format, which was audio and video recorded. Questions were printed on A4 paper. The researcher read each question, and the child then either gave a verbal answer, or drew or labelled on the paper, depending on the requirements of the question. Verbal answers (e.g., explanations) were scored offline. Drawings and labels were also scored offline. Chronbach’s alpha was calculated, by item, within each question set. The reliability of the transfer questions was α = .67 for near transfer problems and α = .72 for intermediate transfer problems. Although the reliability was slightly lower than expected, this is likely to be a reflection of the varied item types, and also the coverage of sub-topics within the broader magnetism topic (i.e., coverage in Table 40). A second coder scored a random 20% of the transfer questions across study 1 and 2. There was a high correlation between the total transfer scores from the first and second coder (r=.98, p <.001), thus suggesting a high degree of overall inter-rater reliability.

5.2.2.3.10 Engagement Questions

Children’s engagement and interest in the activity was then measured through three questions, shown in Figure 50. Each question featured a 5 choice Likert scale, with a description only at each end of the scale. Children ticked on the scale, indicating their level of engagement in relation to that description. The choice was converted to a score out of 5 for each question. These scores were averaged, leading to an average engagement score.

How easy did you find it to concentrate during this activity?				
☐ Very difficult	☐	☐	☐	☐ Very easy
How interesting did you find this activity?				
☐ Not at all interesting	☐	☐	☐	☐ Very interesting
How confident were you when making your answers?				
☐ Not at all confident	☐	☐	☐	☐ Very confident

Figure 50: Engagement questions

5.2.3 Results

5.2.3.1 Descriptive statistics and correlations

A summary of descriptive statistics is included in Table 44. Four separate ANOVAs were conducted, each with condition as an independent variable. The models, separately, contained age, mental folding score, prior knowledge score and average engagement score as dependent variables. The children in the experimental conditions did not differ significantly by age F (2,72) = .256, p = .775, mental folding score, F (2,72) = 1.515, p = .227, prior knowledge score, F (2,72) = .013, p = .988, or mean engagement score, F (2,72) = .823, p = .443. As outlined in the participants section, there were significantly more females than males in the study. However, the proportion of males and females did not differ significantly across conditions, relative to the overall proportion of males and females, χ^2 (2) = 3.45, p = .175.

Table 44: Summary of descriptive statistics by condition.

	Concrete Model *M* (SD)	Gesture *M* (SD)	Verbal / diagram *M* (SD)
Age (years)	8.85 (.28)	8.8 (.27)	8.8 (.30)
Gender (% female)	52%	76%	56%
Mental Folding	8.28 (2.62)	8.48 (2.74)	7.20 (3.01)
Prior Knowledge	2.24(1.09)	2.28 (.96)	2.28 (1.06)
Mean Engagement	4.38 (.48)	4.3 (.45)	4.2 (.60)
Near Transfer	9.72(2.49)	11.04(2.75)	8.97(2.65)
Intermediate Transfer	7.46 (3.96)	6.92(4.00)	7.14(3.79)

5.2.3.2 Analysis strategy

Parametric testing was used throughout. The suitability of parametric testing was determined through: Levene's test for equal group variances, inspection of z -scores of 3.29 in the positive or negative direction to detect univariate outliers, and sufficient numbers of participants in each group for the central limit theorem to apply in relation to normality. Two preliminary ANOVAs with gender and condition included as an independent variable and each transfer score as dependent variables revealed no significant gender effects (near transfer: $p = .319$; intermediate transfer: $p = .398$), or interactions between gender and condition (near transfer: $p = .548$; intermediate transfer: $p = .269$), and so gender was not considered further in the subsequent analysis. In the subsequent analyses, where appropriate, Bonferroni corrections were applied to pairwise comparisons. Mental folding scores were centered prior to the analysis to reduce collinearity between the mental folding score and the mental folding by condition interaction terms.

5.2.3.3 Consolidation question analysis

Performance on the four consolidation questions (not including the final worded problem) was first examined. A consolidation question accuracy score was calculated, based on the number of consolidation questions children correctly answered on their first attempt, for each condition. Descriptive statistics for consolidation question accuracy is displayed in Table 45. Due to technical failure (video equipment failing to record) consolidation question data was not available for three children. The analysis is therefore based on 72 children.

Table 45: descriptive statistics for mean percentage accuracy on first attempt of consolidation questions by condition.

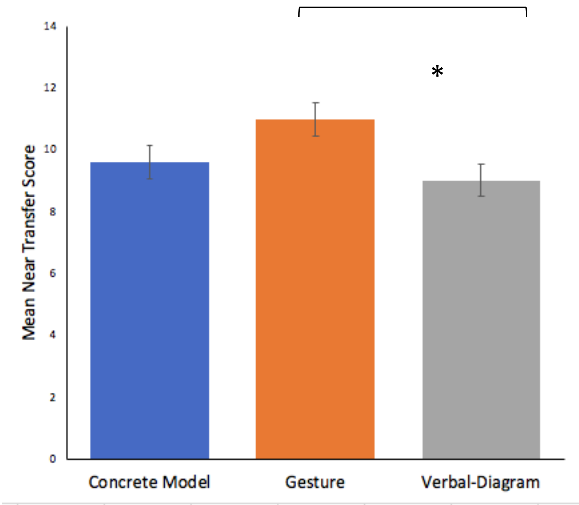
	Concrete Model	Gesture	Verbal/diagram
% accuracy on first attempt (per participant)	Range: .25-100 Mean: .85 SD: .21	Range: .00-100 Mean: .61 SD: .32	Range: .25-100 Mean: .75 SD: .20

A one-way ANOVA with condition as a between-subject factor revealed a main effect of condition, F (2,69) = 5.82, p =.005, η_p^2 = .14. Pairwise tests revealed that consolidation question accuracy was significantly lower for the gesture group than the concrete model group, p = .004. All other pairwise tests were not statistically significant (concrete model vs verbal/diagram: p = .435; gesture vs. verbal/diagram: p = .156). Correlations were also computed between children's mental folding score and the total consolidation score accuracy. The focus was on determining if children with higher mental folding skills were more accurate in producing gestures in the consolidation questions. There were no significant correlations between participant's mental folding scores and consolidation question scores (concrete model: r = .04, p = .861; gesture: r = .28, p = .190; verbal/diagram: r .19, p = .373). There was therefore no evidence that children with higher mental folding scores were more accurate in producing gestures, in response to the consolidation questions.

5.2.3.4 Near transfer question analysis

To examine the effect of teaching condition, a separate analysis was conducted on near and intermediate transfer scores. A general linear model was constructed using the GLM univariate model build option in SPSS. In the model, teaching condition (gesture, concrete model or verbal/diagram) was included as a between participants factor. In addition, the participant's prior knowledge score was included as a continuous covariate to control for prior knowledge. Finally, to determine if participant's level of spatial ability significantly predicted transfer performance, and if this varied between conditions, mental folding scores were included as a covariate, and added as an interaction term between mental folding and condition.

The analysis first revealed a main effect of condition, F (2,68) = 3.53, p = .035, η_p^2 = .09. Pairwise comparisons revealed that children in the gesture condition performed better on the near transfer questions (M = 10.98, SD = .52), than children in the verbal/diagram condition (M = 9.02, SD = .53), see Figure 51, p = *.033*. There were no other significant differences between conditions (gesture vs magnet: p = *.236*, verbal-diagram vs magnet: $p > .99$). Mental folding was not significant in the model, F (1,68) = 1.49, p = .226. In addition, there were no significant interactions between mental folding and condition, F (2,68) = .206, p = .814.



* $p < .05$ ** $p < .001$

Figure 51: Mean near transfer score by condition

5.2.3.5 Intermediate transfer question analysis

In a parallel model for intermediate transfer outcomes, there was no significant effect of teaching condition, F (2,68) = .034, p = .967 (see Figure 52). Mental folding was significant in the model, F (1,68) = 12.53, p = .001, η_p^2 =. 16. There were no significant interactions between mental folding and teaching condition, F (2, 68) = .726, p = .488.

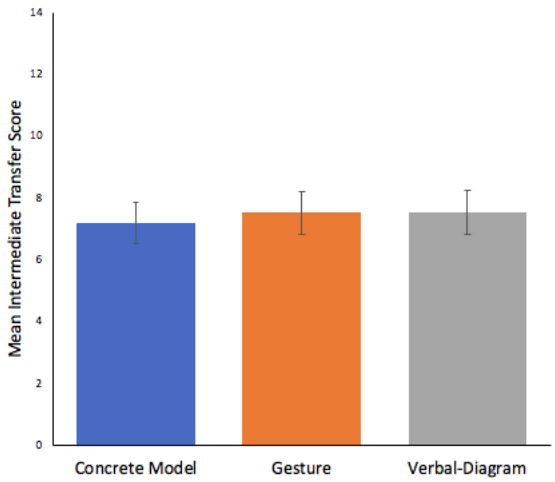


Figure 52: Mean intermediate transfer score by condition

5.2.3.6 Post-test gesture production analysis

The total number of gestures produced by participants during the post-test were then analysed. Videos were reviewed for use of iconic gestures. Specifically, gestures were only recorded if they represented an action, or an object, or both (i.e., a magnet moving). Each gesture was recorded as being from the moment when the hands first engaged to when they came to rest. Gestures such as dietic (pointing) gestures were not recorded, even if they were relevant to the questions (i.e., pointing at an illustration within a question). A second coder viewed 20% of the videos (across study 1 and 2) resulting in a high inter-rater correlation (r = .98, p < .001). Due to technical failure (video equipment failing to record) or data loss, four videos were missing and therefore the analysis was based on 71 participants.

First, the number of gestures produced in the post-test was compared across conditions. This was achieved by running two one-way ANOVAs, with teaching condition as an independent variable in each analysis, and the total number of iconic gestures produced during the post-test, as the dependent variable. A summary of descriptive statistics is provided in Table 46. There were no significant differences in the number of gestures produced between teaching conditions, F < 1. Second, a Pearson's correlation was calculated, between the number of gestures produced during each transfer section of the post-test, and each post-test score (i.e., near and intermediate transfer) score, for each condition. The aim was to investigate if there was any evidence that producing gestures in the post-test, contributed to stronger learning outcomes. Gesture production in the near transfer section of the post-test, for the gesture teaching condition, did not significantly correlate with near transfer scores, ($r = .150$, $p = .504$). This suggests that the benefit seen for the near transfer questions in the gesture condition was not due to participants producing supportive gestures in the post-test. There was, in fact, a negative correlation between gesture production frequency during the near-transfer section of the post-test, and near transfer scores, for the verbal/diagram condition ($r = -.454$, $p = .022$). There were no other significant correlations between either transfer score (near or intermediate), and gesture production accuracy, for any condition ($p > .05$ for all).

Lastly, a correlation was calculated between mental folding scores and the number of gestures produced in each condition. There were no significant correlations, for any condition (concrete model: $r = .237$, $p = .287$; gesture: r -.144, p .502, verbal/diagram: $r = .098$, $p = .643$).

Table 46: Frequency of iconic gestures produced overall during post-test by condition.

Concrete model	Gesture	Verbal/diagram
Mean: 3.18	Mean: 4.46	Mean: 4.76
SD: 3.50	SD: 4.01	SD: 5.27
Range: 0-13	Range: 0-12	Range: 0-20

5.2.4 Discussion

The aim of study 1 was to compare the effectiveness of model-based, gesture-based and diagram/verbal based methods for teaching the science topic of magnetism. In this first study, children in each condition were given one-minute of familiarisation with bar magnets, prior to teaching.

Hypothesis 1 and 2 were, respectively, that model-based teaching and gesture-based teaching would both be more effective than verbal/diagram based teaching, for the near transfer questions. Theories of embodied cognition suggest that the magnet and gesture conditions would both be more effective, because they involve further directed, sensorimotor experiences related to the scientific concept, following the initial familiarisation with magnets. Furthermore, in both of these conditions, the teaching was paired with the sensorimotor experience. In support of hypothesis 2, for the near transfer questions, there was a small but significant benefit in favour of gesture-based teaching, relative to the verbal/diagram teaching condition. However, in contrast to hypothesis 1, the concrete model teaching condition was no more effective than the verbal-diagram teaching condition. This suggests that having further concrete model experience in the concrete model teaching section, did not lead to significantly greater learning, relative to the verbal/diagram condition. That is, there was no significant additional benefit to having *directed* concrete model experience in the concrete model teaching session, compared with the *non-directed* physical pre-experience, gained in the verbal/diagram condition. More generally, it suggests that having additional sensorimotor experiences, per se, was not beneficial.

The video analysis revealed no relationship, for the gesture teaching condition, between the frequency of iconic gesture production in the post-test and transfer scores. Moreover, the frequency of iconic gestures did not differ significantly between conditions. This suggests that the benefit of the gesture condition, over the verbal/diagram condition, for the near transfer questions, was not due to increased gesture production in the post-test. Instead, in addition to reactivation of motor areas in the brain, the gesture teaching may have resulted in participants developing a rich internal representation of the bar magnet, or paying closer attention to the spatial features of the bar magnet itself, during teaching. Given the difference was apparent only for near transfer questions, schematisation and abstraction are less likely mechanisms.

Hypotheses 3 and 4 were that the gesture-teaching condition would lead to greater transfer performance than the concrete model teaching condition in the near and intermediate transfer questions, respectively. However, for near and intermediate transfer questions, there were no significant differences between these two conditions, against prediction.

These findings are partly in line with findings from both studies in Stieff et al. (2016). This previous study found that having students reproduce gestures was more effective than a text/diagram-based teaching condition. They also found that having students reproduce gesture was no more effective, overall, than the use of concrete models. However, students performed worse in the model group if they did not have the model available during the post-test. In contrast the gesture group's performance was not worse, without a concrete model. The availability of the concrete model during the post-test was not manipulated in the current study, and so it is not possible to make direct comparisons.

Hypothesis 5 and 6 were that mental folding would predict performance on the near and intermediate transfer problems, respectively. There was no support for hypothesis 5: mental folding was not a significant predictor of the near transfer problems. However, in line with hypothesis 6, mental folding was a significant predictor of the intermediate transfer problems. In this study, all children had some exposure to the concrete magnets either before or during teaching. It is possible that spatial ability was not related to performance on the near transfer questions, which were more related to the teaching, because it was less essential to use transformation and spatial manipulation to understand the core concepts of the lesson. Because the children could, at some point, actually manipulate the models, spatial transformation skills were less central. For intermediate transfer questions, children encountered more novel magnets, and may have therefore needed to do more active spatial transformation.

5.3 Study 2

5.3.1 Overview

The findings of study 1 indicated that there was a slight benefit to gesture-based teaching, but only relative to the verbal/diagram condition, and only for near transfer scenarios. Furthermore, given that gesture-based teaching was more effective than verbal/diagram teaching, but no more effective than concrete model teaching, this suggests that, practically, there are some advantages in teachers choosing gesture-based teaching.

It is possible that the reason for the lack of significant difference between the gesture and concrete model conditions was that the brief pre-experience with the magnets masked the subsequent benefits of the gesture teaching. For example, participants could partly rely on the affordances of the magnets, because they had the brief action-based experience, first. Moreover, one of the mechanisms through which gesture may be beneficial, over action, is because gesture does not provide tactile feedback (Kita et al., 2017; So, Ching, Lim, Cheng, & Ip, 2014) meaning that the spatial representation is more malleable. Having the initial familiarisation with magnets did provide tactile feedback, even if gesture did not.

In study 2, the pre-experience section was replaced with photographs of magnets, rather than physical experience with magnets. The aim was to control between the experiments so that in both experiments, the child had a magnet-related, non-directed experience, before beginning the teaching session. The predictions were the same as study 1. It was expected that both gesture and concrete model conditions would outperform the verbal/diagram condition, and that the gesture teaching condition would outperform the concrete model condition.

5.3.2 Method

5.3.2.1 Participants

75 children from UK Year 4 participated in this study, (mean = 8.93, SD = .30). Children were recruited from four different schools to study 1 (school 1, n = 9; school 2, n = 20; school 3, n = 23; school 4, n = 23). Ethical approval was given by the UCL ethics committee, and parents gave informed consent to take part in the study. Three participants were excluded due to equipment failure, resulting in data loss of the post-test recordings (audio and video). The final sample therefore consisted of 72 children. The overall proportion of males (44.4%) and females (55.6%) did not differ statistically, χ^2 (1,72) = .899, p = . 346. As with study 1, a power analysis, conducted in G*Power indicated that a sample size of 90 was needed to detect a medium effect (f^2 = .15). A medium-sized effect was chosen as the anticipated effect size, based on previous research (e.g., Stieff et al., 2016). The achieved sample size of 72 was therefore somewhat underpowered to detect a medium sized effect.

5.3.2.2 Materials and procedure

5.3.2.3 Pre-teaching familiarisation activity

Children were first shown six colour photographs. Each photograph filled the laptop screen. Photographs featured the same two magnets used for the pre-experience in study 1, and for the teaching in study 1 and 2. In two of the photographs the magnets were arranged to be attracting, in two they were arranged to be repelling, and in two photographs they were arranged in a neutral position. There were no labels or further information on screen. Each photograph remained on the screen for 10 seconds. Therefore, the pre-experience was matched in time to the pre-experience in study 1. No description or explanation of the photographs was given by the researcher. The child was asked to simply look at the photographs, and they were also told that they did not need to explain or describe the photographs.

After the familiarisation phase, the procedure was the same as study 1.

5.3.3 Results

5.3.3.1 Descriptive statistics

Parametric testing was used throughout. Suitability of parametric testing was determined using the criteria used in study 1.

A summary of descriptive statistics is included in Table 47. To explore the descriptive statistics, as in study 1, separate ANOVAs were carried out, each with condition (gesture, concrete model, verbal/diagram) as an independent variable. Each ANOVA contained one of the following dependent variables: age, mental folding score, prior knowledge score and average engagement score as dependent variables. As expected, the experimental conditions did not differ significantly by age $F < 1$, gender, $\chi^2 = .889$, $p = .346$, mental folding score, $F\ (2,69) = 2.58$, $p = .083$, prior knowledge score, $F < 1$, or mean engagement score, $F < 1$. Preliminary ANOVAs with gender and condition included as an independent variable and near or intermediate transfer as dependent variables indicated no effect of gender (near transfer: $p = .610$; intermediate transfer: $p = .194$), or interaction between gender and condition (near transfer: $p = .152$; intermediate transfer: $p =. 507$), and so gender was not considered further in the subsequent analyses.

Table 47: Descriptive statistics for study 2

	Concrete Model *M* (SD)	Gesture *M* (SD)	Verbal / diagram *M* (SD)
Age (years)	8.95 (.29)	8.96 (.31)	8.87 (.33)
Gender (% female)	48%	47.8%	70.1%
Mental Folding	9.8 (1.84)	8.48 (2.54)	8.46(2.67)
Prior Knowledge	2.56(.870)	2.20 (.90)	2.42(.97)
Mean Engagement	4.11 (.65)	4.13 (.50)	4.10 (.59)
Near Transfer	10.04 (2.73)	10.30 (3.10)	10.54(2.90)
Intermediate Transfer	7.52 (4.03)	6.08(4.33)	7.20 (3.82)

5.3.3.2 Consolidation question analysis

Descriptive statistics for consolidation question accuracy is displayed in Table 48. Due to technical failure (failure to record video) consolidation question data was not available for two children. The analysis is therefore based on 70 children.

Table 48: Descriptive statistics for percentage accuracy on first attempt, by condition.

	Concrete Model	Gesture	Verbal/diagram
% accuracy on first attempt (per participant)	Range: .50-100 Mean: .88 SD: .16	Range: .00-100 Mean: .77 SD: .26	Range: .00-100 Mean: .61 SD: .32

A one-way ANOVA with condition as a between participant factor revealed a main effect of condition, F (2,67) = 6.51, p = .003, η_p^2 = .16. Pairwise tests revealed that accuracy for the verbal/diagram consolidation questions were significantly lower than for the concrete model consolidation questions, p = .004 only (all other comparisons, p >. 05). There were no significant correlations between participant's mental folding score and number of correct consolidation items, in the concrete model (r = .145, p = .498) or gesture (r = .156, p = .478) condition. However, there was a strong positive correlation between correct consolidation items and mental folding, in the verbal/diagram condition (r = .686, p < .001).

5.3.3.3 Near transfer question analysis

A parallel analysis was performed to study 1, to investigate the effect of teaching condition on near transfer performance. Teaching condition was again included as an independent variable, with prior knowledge score, mental folding scores as continuous covariate variables. An interaction term between teaching condition and mental folding was included again in the model. In this model, there was no main effect of condition, F (2,65) = .140, p = .870 (see Figure 53).

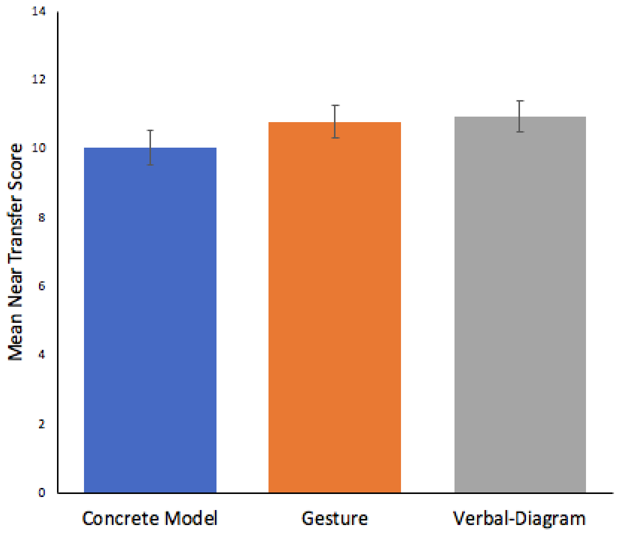


Figure 53: Mean near transfer score by condition

Mental folding was a significant predictor, F (1,65) = 11.27, p = .001, η_p^2 = .15. In addition, there was a significant interaction between teaching condition and mental folding scores, F (2,65) = 5.54, p = .007, η_p^2 = .14.

To investigate this interaction effect, correlations were calculated between mental folding and near transfer scores, per teaching condition. As can be seen in Figure 54, mental folding was not significantly related to near transfer performance for children in the concrete model condition, r = - .12, p = .573. However, there was a strong correlation between mental folding and near transfer performance in the gesture condition, r = .57, p = .005, and verbal/diagram condition, r = .77, p <.001. This explains the interaction.

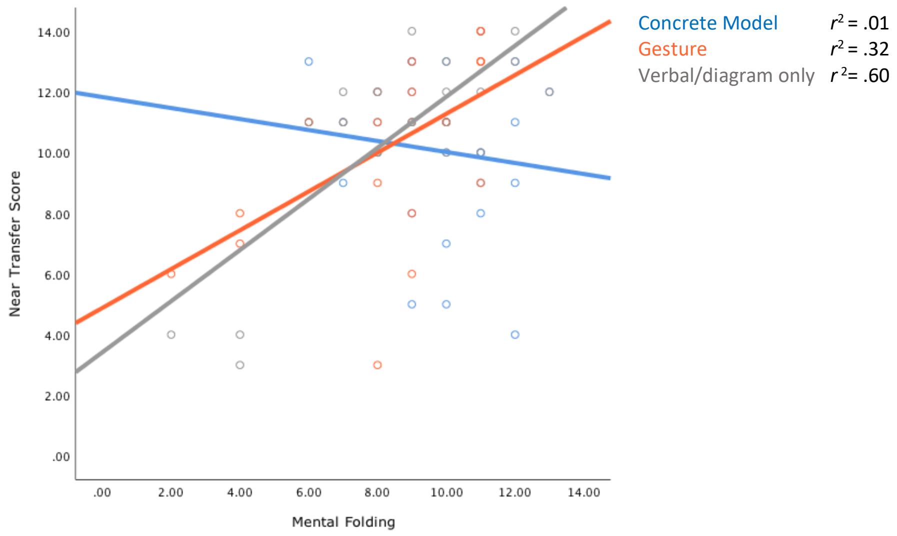


Figure 54: Scatter plot of near transfer performance against mental folding scores, by condition.

5.3.3.4 *Intermediate transfer question analysis*

The model for intermediate transfer outcomes, which included the same variables as the near transfer model, revealed there was no main effect of teaching condition, F (2,65) = .33, p = .723, see Figure 55. Mental folding was significant within the model, F (1,65) = 10.28, p = .002, η_p^2 = .14. The interaction between mental folding scores and teaching condition was not significant, F (2,65) = .50, p = .606, η_p^2 = .015.

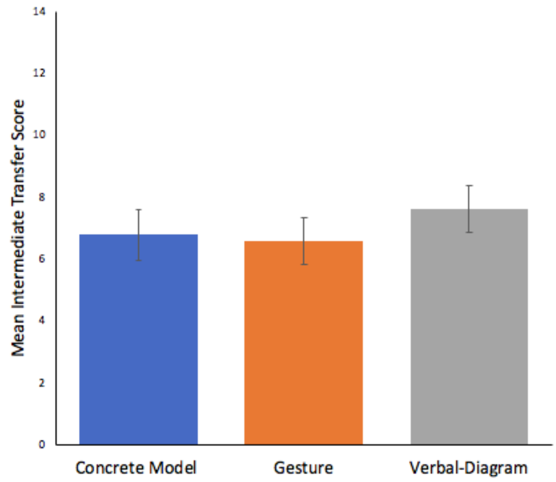


Figure 55: Intermediate transfer performance by condition

5.3.3.5 Post-test gesture production analysis

Due to technical failure (video equipment failing to record) or data loss, four videos were missing and therefore the analysis was based on 71 participants. As with study one, ANOVAs were conducted with condition as the between group factor and transfer score as the dependent variables. There were no significant differences in the number of gestures produced between teaching conditions: $F < 1$. See Table 49 for a summary of descriptive statistics. There were no significant correlations between either near of intermediate transfer score, and gesture production accuracy, for any condition ($p > .05$ for all). Finally, there were also no significant correlations between mental folding scores and total number of gestures produced, in any condition ($p > .05$ for all).

Table 49: Frequency of iconic gestures produced during post-test, by condition.

Concrete Model	Gesture	Verbal/diagram
Mean: 3.71	Mean: 3.56	Mean: 3.08
SD: 3.78	SD: 4.05	SD: 4.05
Range: 0-14	Range: 0-14	Range: 0-16

5.3.4 Discussion

The aim of study 2 was to determine if the presence of concrete models to familiarise children prior to teaching in study 1, masked or reduced the effectiveness of gesture-based teaching. To investigate this, the concrete model familiarisation was replaced with photographs of the concrete models. The hypotheses were the same as study 1.

The most important finding in study 2 was that, for the near transfer questions, the impact of mental folding skills interacted with the teaching condition. Specifically, mental folding significantly correlated with performance in the verbal-diagram and gesture conditions, but, not the concrete model condition. This therefore provides some support for hypothesis 5, which stated that mental folding would be predictive of performance on the near transfer questions. Recall that mental folding did not correlate with near transfer scores in any of the conditions in study 1.

In both the verbal-diagram and the gesture condition, in this study, children could not rely on the affordances of the object at any point. For example, in the verbal-diagram condition, considering the familiarisation phase and the teaching phase together, the only sources of information were photographs, diagrams and verbal descriptions. Mental folding skills were thus more crucial, than in other conditions, because it may have been more necessary to mentally 'animate' the diagrams, or descriptions, of the magnets during teaching (e.g., when actions of the magnets were described). The role of mental folding skills in the verbal-diagram teaching condition in this study was also supported by the significant correlation between mental folding scores and consolidation question scores, for this condition only. In study one, when participants were familiarised with the concrete model before teaching, there were no significant correlations between mental folding scores and scores in the consolidation questions.

Considering the gesture condition, there was no evidence that children with higher mental folding scores were more able to produce the gestures accurately, in the consolidation questions. Thus, it did not appear that individual differences in gesture production ability were the source of the correlation between mental folding and near transfer scores, for this group. When participants in the gesture condition produced gestures during the teaching section, they co-ordinated between their own gestures and the researcher's gestures, whilst noting any additional information in the presentation (e.g. the magnet diagrams). It may have been that children with higher mental folding scores were both able to produce and manipulate the gestures accurately, and, also able to map and co-ordinate the gestures with the simultaneous diagrammatic representation given within the teaching presentation. This integration between the gestures and diagrams was more critical in the gesture condition in this study, due to the absence of the concrete-model, during familiarisation.

In line with hypothesis 6, mental folding skills were also significantly correlated with the intermediate transfer problems, as was the case in study one. As outlined previously, because children did not receive direct instruction about these magnets, it may have been more necessary to generate and perform novel spatial transformation, within these questions.

The finding that strength of the relationship between spatial thinking skills and learning depends on the teaching approach, or the type of visualisations through which learning is delivered, supports prior research. For instance, Sanchez and Wiley (2014) found that multiple object dynamic spatial ability was only related to geoscience learning, when participants received instruction through static diagrams, and not, through animations. The author's interpretation was that spatial ability was important for 'animating' the static illustrations, but less important when the images were animated for the participant. Similarly, a meta-analysis of studies by Höffler (2010) found that spatial ability was more strongly related to learning with static animations than dynamic animations, and also, with 2D rather than 3D illustrations. Participants in the verbal-diagram condition, in the current study, learnt with 2D and static illustrations, only. Therefore, mental folding was perhaps more necessary because of the increased need to visualise and infer 3D, dynamic relationships from 2D, static images. Similarly, the concrete model, provided within the other conditions, perhaps acted in some respects as a 3D, dynamic visualisation; in a physical form, rather than an on-screen animation.

There was no empirical support for the core hypotheses of the study, overall. In particular, there was no evidence that learning in the concrete-model and gesture conditions was greater than learning in the verbal-diagram condition (hypothesis 1 and 2). Moreover, there were also no significant benefits evident for children in the gesture teaching condition, relative to the concrete model condition, for either the near or intermediate transfer questions (hypothesis 3 and 4). Possible reasons for this are included in the general discussion below. There was therefore no support for the hypothesis that the magnet familiarisation in study one had the effect of masking benefit of the gesture condition.

Although there was no evidence that learning overall was greater in any condition, there were associations between mental folding scores and near transfer scores, for the verbal-diagram and gesture conditions. Therefore, children with lower mental folding skills would still at a disadvantage, relative to children with stronger mental folding skills, in a classroom setting with these two specific science teaching approaches.

5.4 Additional analysis for study 1 and 2

Because the design of study 1 and 2 differed only in the type of pre-instruction familiarisation the participants received, it was also possible to combine the data and run an additional overall analysis. An additional linear model was run in which pre-experience (magnets vs photographs) was included as one between-group factor, and instructional condition (concrete model vs gesture vs verbal/diagram) was included as a second between-subjects factor. As before, pre-test scores were included, as well as centred mental folding scores. An interaction term was included between pre-experience and instructional condition. In addition, interaction terms were included, between mental folding and pre-experience, mental folding scores and instructional condition, and a three-way interaction between mental folding scores, instructional condition and pre-experience. The latter interaction was included because of the two-way interaction observed only in study 2.

In the linear model for the near transfer scores, there was a main effect of pre-test scores, F (1,134) = 13.51, $p < .001$, $\eta_p^2 = .092$, and a main effect of mental folding scores, F (1,134) = 9.12, p = 003, $\eta_p^2 = .064$. There were no significant main effects of instructional condition, F (2,134) = 2.07, p = 131, $\eta_p^2 = .030$, or pre-experience, F (1,134) = 3.17, p = .458, $\eta_p^2 = .004$. The interaction between pre-experience and instructional condition was also not significant, F (2,134) = 2.22, p = .113, $\eta_p^2 = .032$. The interaction between mental folding and instructional condition was not significant, F (2,134) = 2.13, p = .123, $\eta_p^2 = .031$, and the interaction between pre-experience and mental folding was not significant, F (2,134) = 3.27, p = .073, $\eta_p^2 = .024$. However, the three-way interaction between mental folding, instructional condition and pre-experience was significant, F (2,134) = 3.90, p = .022, $\eta_p^2 = .055$.

In the linear model for the intermediate transfer scores, there was a main effect of pre-test scores, F (1,134) = 9.85, p = .002, $\eta_p^2 = .068$, and a main effect of mental folding scores, F (1,134) = 20.70, p < 001, $\eta_p^2 = .134$. There were no main effects of instructional condition, F (2,134) = .237, p = 790, $\eta_p^2 = .004$, or pre-experience, F (1,134) = 3.20, p = .076, $\eta_p^2 = .023$. The interaction between pre-experience and instructional condition was also not significant, F(2,134) = .18, p = .834 $\eta_p^2 = .003$. The interaction between mental folding and instructional condition was not significant, F (2,134) = .74, p = .481, $\eta_p^2 = .011$, and the interaction between pre-experience and mental folding was not significant, F (2,134) = .42, p = .520, $\eta_p^2 = .003$. The three-way interaction between mental folding, instructional condition and pre-experience was also not significant, F (2,134) = .35, p = .705, $\eta_p^2 = .005$.

5.5 Overall discussion

The aim of these studies was to investigate the effectiveness of gesture, concrete model and diagram-verbal based teaching on the topic of magnets and magnetism. Although one prior study with adults (e.g., Stieff et al., 2016) compared concrete model versus gesture based science teaching methods, this finding was limited to the domain of biochemistry. Moreover, the only research investigating gesture as a teaching tool, with primary-school aged children, was within the discipline of mathematics (Novack et al, 2014; Congdon et al., 2018).

Hypotheses 1 and 2 stated that the gesture and concrete model teaching conditions would demonstrate greater transfer performance than the verbal-diagram condition, on the near transfer questions. Across studies 1 and 2, the evidence was somewhat mixed. For instance, although the gesture condition outperformed the verbal/diagram condition in study 1, the concrete model condition did not. Moreover, in study 2, there was no evidence that learning outcomes in the verbal/diagram condition were significantly lower than either of the sensorimotor conditions.

Overall, this suggests that the prohibition of sensorimotor movements during the teaching of a topic with a strong spatial-relational basis, does not necessarily always result in significantly reduced learning, compared to more physically embodied experiences. These findings contrast to the findings of Atit et al., (2015) who found that prohibiting participants from gesturing during an explanation regarding a geologic concept, reduced gains on a measure of spatial thinking, relative to a gesture instructed group.

It should be noted that the verbal/diagram condition was designed in a way that it was as well-matched as possible to the other conditions. For example, children were actively involved verbally during the main teaching. Furthermore, children answered diagram-based questions at an equivalent point when children in the other two conditions were actively engaged with sensorimotor actions. The verbal/diagram condition was therefore not a ‘passive’ verbal control. Having additional opportunities to practice using diagrams was most likely beneficial, given the importance of diagrams in the post-test. Therefore, any relative decrease in transfer performance as a result of not having sensorimotor experiences, may have been offset by the benefits of the diagram consolidation questions. Gesture was also prohibited by asking children to hold a clipboard, rather than asking them to sit on their hands (as was the case in Atit et al., 2015). Overall, it is therefore in some ways unsurprising that the verbal/diagram condition resulted in reasonably good learning outcomes.

Hypotheses 3 and 4 stated that the gesture teaching group would demonstrate greater transfer performance than the concrete model group, on near and intermediate transfer questions, respectively. There was no support for Hypothesis 3 or 4, across the two studies. Thus, the hypothesised benefits to gesture teaching, over concrete model teaching, were not evident. These results therefore contrast somewhat to the most closely related findings from Stieff et al. (2016), who found that that teaching a biochemistry topic through gesture was, in some respects, more effective than through concrete models. In particular, students who learnt through gesture, were less dependent on the concrete model at the post-test. However, this specific factor was not directly manipulated in the current study. Comparing only the gesture and concrete model conditions for participants who were not provided with a concrete model at post-test, is a more direct comparison to the current study. Participants in the gesture condition in Stieff et al. (2016), who did not have a concrete model at post-test, performed better in the post-test than the concrete model condition, who also did not have a concrete model at post-test. These findings contrast with the findings of the current studies.

The findings also differ from mathematics research, from Novack et al. (2014) who found that gesture was more effective in promoting transfer of learning about basic algebra (mathematical equivalence). The findings of the current studies, however, align more closely with Congdon et al. (2018). Congdon et al. (2018) found that, for children with higher prior knowledge, learning about measurement through action and gesture were both equally as effective, whereas for children with lower prior knowledge, only action was effective. Although prior knowledge was not considered as a moderating factor in the current study, the finding that gesture was never more effective than action, is more in line with these findings.

Several reasons could account for the finding that gesture was no more effective than concrete models. First, the children in the current study had a moderate level of prior knowledge. It is possible that for children with less prior knowledge, there would be a more definite advantage for gesture-based teaching. This could be addressed in further research. It may have also been the case that the specific gestures in this study were not conducive to learning. In particular, the north/south mapping on different fingers may have been overly complicated and inaccessible for young learners. Therefore, the results of this study do not mean that gesture is not helpful for science learning per se. Rather, more research is needed to determine if the specific gestures were unhelpful. Relatedly, future research might investigate the role of abstract gestures. The magnet gestures were fairly close to the concrete representation of a bar magnet. Novack et al. (2014) found that only abstract gestures, and not concrete gestures, were effective in promoting far transfer. Likewise, in comparing their results to that of Novack et al. (2014), above, Congdon et al. (2018), suggest that a benefit of teaching through gesture, over action, may be restricted to more abstract topics, such as algebra, rather than measurement, which is more concrete. Perhaps, then, learning about magnetism, at this level, is not suited to gesture-based instruction, because the gestures and actions are too similar. Thus, perhaps teaching the topic at a later stage, at a more conceptually advanced level (e.g., learning about fields of force within the magnet), may result in a greater benefit for learning through gesture, compared to action. Likewise, a different topic, within physics, where gestures add a more abstract level of knowledge, beyond action, may be more suited to gesture-based instruction.

From a ‘contrasting cases’ perspective (Schwartz and Bransford, 1998), perhaps the gesture-based instruction would have been more effective in this context if children were given more time and opportunity to compare the gestures to their more concrete referents. Although the gestures were presented alongside the teaching slides which contained diagrams of bar magnets, this occurred whilst instruction was occurring and verbal content was being delivered. In another study, children might be asked to produce the gestures, and shown corresponding images, but also be given specifically allocated time to compare between and align the representations. Finally, learning was only assessed immediately after the teaching intervention. Future research could also include a follow-up element, perhaps after one or two weeks, to determine if gesture results in longer-lasting learning. This has been the case in prior gesture research in mathematics (e.g., Cook, Mitchell, & Goldin-Meadow, 2008).

In summary, considering hypotheses 1-4, there was no strong evidence that any of the instructional approaches were more effective than another. The only difference was evident in study 1 between the verbal-diagram and gesture condition, for the near transfer questions. However, in the overall ANOVA, which included both studies, the interaction between instructional condition and pre-experience was not significant, and the non-significant p-value was associated with a small effect size. This suggests that overall the instruction effect in study one was not a major factor, when taken together with the findings from study 2.

The most interesting and important finding, particularly in the context of the wider thesis, was that the correlation between mental folding skills and learning outcomes, varied by teaching condition, and transfer type. Hypothesis 5 stated that mental folding would predict performance on the near transfer questions. In partial support of this, mental folding only significantly correlated with performance on the near transfer questions in the gesture condition and verbal-diagram condition, in study 2. This was supported in the overall ANOVA, which included both studies, by a three-way interaction between pre-experience, instructional condition and mental folding score. The gesture and verbal-diagram condition in study 2 were the two conditions in which participants had no experience with the concrete model at all. Mental folding scores did not correlate with near transfer scores in any of the conditions in study 1, where participants were familiarised with the concrete-model first, or, in the concrete-model condition, in study 2. Within the verbal-diagram condition in study 2, the only sources of information were photographs, diagrams, and verbal description. Mental folding skills were therefore perhaps more critical to enabling children to animate and generate dynamic mental images of the learning content; both verbal and non-verbal. In the gesture condition, in study 2, although participants did not have experience with the concrete-model, the gestures did provide a sensorimotor experience. Children were able to produce the gestures in the consolidation questions, as well as children with higher mental folding skills. Therefore, the importance of mental folding ability, for the gesture condition, may have been due to the increased need to integrate information from the diagrams, with the produced gestures, and to visualise the relationship between the two, in the absence of a concrete model.

Hypothesis 6 was that mental folding would predict performance on the intermediate transfer questions. In support of this, mental folding was consistently correlated with scores on the intermediate transfer questions, in both study 1 and 2, and this did not interact with teaching condition. It may have been that mental folding was consistently important for these questions, because participants had never been directly taught about these novel magnets, and so more spatial transformation was required. In line with this, as outlined in Chapter 3, Cromley et al., (2016) found that diagram training was most effective for items in which diagrams were central to the questions. That is, for diagrams relaying novel content, which students had not been taught in class, and where the information critical to the question was contained within diagrams. Thus, it could be inferred that, in the current study, stronger mental folding skills were more important in the intermediate transfer questions, particularly because of the inclusion of novel diagrams, which contained information central to the problems. More specifically, in order to answer the problems, it was more essential to be able to ‘read’ and use the diagrams.

A limitation of the study is the relatively small sample size. It is possible that significant differences would have been observed given a larger sample size, and greater power. In addition, because there was no measure of verbal ability, it is possible that the interaction effect with mental folding was at least partly related to general ability, rather than spatial ability, exclusively. However, the findings of the previous chapters, which included varying levels of control for general ability, support a specific role of mental folding in children’s physics reasoning. Another limitation is that there was no control group included which did not include any of the spatial tools, such as a verbal description only group. Practically, a verbal-description only instructional approach is not a realistic representation of approaches used in the classroom. Nevertheless, this may have highlighted the general benefit of these approaches, relative to a condition in which verbal description only was used.

Finally, the instructional approaches were designed in a way that they could be educationally useful, relevant and effective, rather than to be very tightly controlled in terms of variable manipulation. There were therefore several causal mechanisms under which different instructional approaches could operate. For instance, in the gesture instructional condition, the researcher produced the gesture, and then the child was asked to reproduce that gesture. The process of the researcher producing the gesture and the child observing the gesture is a potentially separate learning mechanism to the child then also producing the gesture. This is also the case for the concrete model condition. In both cases, it is therefore unclear if both observing and producing the gestures, or observing and manipulating the concrete model, are important for the learner. This could be addressed in future research. In terms of the verbal-diagram instructional condition, as discussed previously, the diagram consolidation questions may have provided a separate diagram-specific practice and learning mechanism, to the content learning itself.

To conclude, taking the two studies in this chapter together, there was no strong evidence that gesture was a more effective instructional approach or, that any of the instructional methods were more or less effective for learning about magnetism, overall. The only condition in which scores were significantly different to another was the gesture condition, in study one, compared to the verbal-diagram condition, for the near transfer questions. As discussed, the additional analysis which included data from study 1 and 2 also suggested that the instruction effect in study one, which was significant between two groups only, was not a major finding.

Nevertheless, the findings of study two indicated a positive correlation between near transfer learning outcomes and mental folding scores, specifically, for the verbal/diagram and gesture conditions. Within a classroom situation, and these teaching approaches, children with lower spatial thinking skills would therefore still be at a disadvantage, relatively. An implication of this is that, to learn effectively in science, children with lower spatial thinking skills may require support from a combination of interventions, such as: spatial transformation training, diagram training, or support at integrating multiple sources of information, perhaps in combination with guided use of a concrete model, or gesture.

In terms of classroom implementation, the data suggest that the instructional approach in study 1 would be optimal. In study 1, when children had experience with the concrete model in the pre-experience phase, mental transformation skills were not a significant predictor of learning, for the near transfer questions. Therefore, children with weaker mental transformation skills were not at a disadvantage, relative to children with stronger mental transformation skills. In the classroom, children could be provided with magnets to have free experience with prior to whole-class instruction on magnetism from the classroom teacher. Of course, a practical limitation of such an approach is the need for a whole class set of magnets. The whole class instruction could then consist of children continuing to use and manipulate the magnet, as modelled by the teacher. Alternatively, the teacher might introduce the gesture-based instruction by modelling the gestures and then asking children to produce the gestures.

Chapter 6

Discussion

6.1 Introduction

Scientific literacy is important for both personal and economic participation in society. However, concerns over the shortfall in STEM graduate numbers in the UK are reflected in a shortage of qualified scientists within domains such as physics, meteorology, and geology and science-related disciplines, such as engineering ("Shortage Occupation List," 2018). There is also a shortage of specialist science teachers, particularly in disciplines such as physics, within secondary schools (Sibieta, 2018). The focus of this thesis was on spatial cognition as one factor that contributes to learning, achievement and engagement in science.

Spatial thinking plays a central role in science; both for the professional scientist, and the novice learner in the classroom. Empirically, a significant body of research also connects individual differences in spatial thinking ability to STEM learning with adults and adolescents. In particular, large-scale longitudinal studies (Wai et al., 2009; Lubinski & Benbow, 2006) provide particularly compelling evidence for an association between adolescent's spatial ability and later STEM engagement and achievement. In addition, several studies, also with adults and adolescents, provide more specific correlational evidence of a link between spatial thinking skills and aspects of science learning (Stull, Hegarty, Dixon & Stieff, 2012; Delialioglue & Askar, 1999). However, little research has investigated the relationship between spatial thinking skills and science learning with children. Much of the research that has addressed the relationship with children has not been thorough or comprehensive. In particular, past research had not delineated between different spatial skills or science skills. This criticism can also be directed at much of the research with adults and adolescents. This dearth in knowledge served as the primary motivation for this thesis.

Prior research indicates that spatial thinking skills may be more important for learning earlier in development, than later in development (Hambrick et al., 2012; Mix et al., 2016; Uttal & Cohen, 2012). Thus, increasing our knowledge of the role of spatial thinking in science learning for children in the primary school years, at a time when spatial thinking may be particularly crucial, is important for the effective development of interventions and training. Moreover, having a deeper understanding of which spatial skills are predictive of different aspects of children's science learning, means that future spatial interventions can be more targeted, which in turn will positively impact the effectiveness of an intervention. For example, the identification of particular spatial skills which are consistently associated with science learning, suggests that they may be better candidates for spatial training interventions.

One reason for the limited scope of spatial thinking skills within this type of research is the spatial skill typologies upon which they are based (e.g., Carroll, 1993). In particular, prior spatial typologies do not encompass the breadth of skills and processes which, research more broadly indicates, fall under the umbrella of spatial cognition. These shortcomings were addressed by a theoretically and empirically driven typology (Uttal et al., 2013). However, there has been little direct evaluation of this model to date. Therefore, another focus of the thesis was on evaluating this model through several lines of evidence, to learn more about the structure of spatial cognition in childhood.

Spatial tools such as gestures and concrete models are another aspect of spatial cognition. Spatial tools can be used within science instruction to support the learner's understanding of spatially demanding concepts. In addition, spatial tools may moderate the impact of spatial thinking skills, such as mental rotation, on learning. To date, no prior research has compared different spatial tools as instructional methods for children's science learning, and considered the possible moderating role of spatial thinking skills. The final studies in the current thesis addressed this knowledge gap by comparing gesture, concrete models and diagrams as instructional approaches for learning about magnetism, and, by considering the moderating role of mental folding ability.

The discussion that follows is organised around these broad aims, as set out in the introduction. Following this, there is a discussion of the educational implications and directions for future work.

6.2 The structure of spatial thinking skills in childhood

As outlined above, despite the existence of a significant body of psychometric data, based predominantly on adult samples, several researchers central to the field of spatial cognition argued that there had yet to be a satisfactory typology of spatial thinking (e.g., Hegarty & Waller, 2005; Newcombe & Shipley, 2015). Dominant models, for example, from Carroll (1993) and Linn and Peterson (1985), have been criticised for not being sufficiently theory-driven, and, for excluding several types of spatial thinking. Many prior spatial typologies exclude skills such as spatial perspective taking and multiple object dynamic spatial ability, and do not provide the scope for inclusion of abilities such as navigation. A more theoretically driven model by Uttal et al. (2013) thus aimed to account for a wider range of spatial skills. However, the dimensions of the model, particularly the static-dynamic distinction, had received limited attention within the literature. A further issue is that few studies have investigated the structure of spatial cognition in childhood specifically, and taken together, the prior studies that do exist, paint an inconclusive picture. One of the aims of the thesis was, therefore, to investigate the structure of spatial cognition in childhood, particularly, in light of the Uttal et al. (2013) model.

The dimensions of the 2 x 2 model were assessed psychometrically within Chapter 2 and Chapter 4. In both cases, due the number and range of spatial tasks administered, it was only possible to assess the two dimensions separately, i.e., intrinsic-extrinsic, and static-dynamic, and not the full 2x2 (i.e., four factor) model. The cross-sectional data in Chapter 1 also meant that the model could be addressed developmentally. In Chapter 4, spatial cognition was investigated as a predictor of understanding about sound within a science lesson. This provided a larger sample of children of a similar age, and thus higher statistical power. Variation in the task choices within Chapter 4 also meant that subtleties of the Uttal et al. (2013) model could be more closely investigated. For instance, it was unclear whether multiple object-dynamic spatial ability tasks fall within the extrinsic-dynamic category, or, are a distinct type of spatial ability, perhaps falling outside the model.

Only one prior study had addressed both dimensions of the model. Mix et al. (2018) found support for the intrinsic-extrinsic dimension for younger children (age 6 and 9 years) only. There was, however, no support for the full, four-factor model, or the static-dynamic models, for any age group. For the older children (aged 12) neither of the two-factor models fit better than the baseline, one-factor model, and, the one-factor model did not provide a good fit to the data. In addition, the four-factor model could not be tested with the oldest children, due to there being an insufficient number of measures per factor. Age-appropriate versions of tasks were administered to each age group (e.g., 2D rotation for younger children vs 3D mental rotation for older children). Thus, within this study, it is difficult to make strong claims regarding changes in the factor structure with development.

In Chapter 2, in the current thesis, there was no evidence that either of the two-factor models fit significantly better than the one-factor model, for either 6-8 year olds, or the 9-11 year olds. There was therefore no support for the Uttal et al. (2013) model in these findings. These findings differ from those of Vander Heyden et al. (2016), who reported an increased dissociation between intrinsic and extrinsic skills across development. This study reported that object-transformation (intrinsic-dynamic) skills were dissociated from viewer transformation/perspective taking (extrinsic-dynamic) skills at the age of 10.5-12, but not at ages 7.5-10. The authors proposed that the two types of skill were not psychometrically dissociated for younger children because spatial perspective taking skills were yet to develop fully, and were therefore not yet a 'specialised' spatial skill.

These findings also differ to the findings of Mix et al. (2018), described previously. This discrepancy between studies may be due to methodological factors. For example, the study by Mix et al. (2018) study administered a larger number of measures per dimension (four spatial measures for each factor in the two factor model). The studies in the current thesis had two or three measures per factor. The Mix et al. (2018) study was also based on a larger sample size (approximately 250 participants per age group versus approximately 90 in chapter 2). This means that the Mix et al. study had greater statistical power to detect a smaller effect.

Considering intrinsic-dynamic tasks, the 2D monkey mental rotation task loaded poorly onto all models, in contrast to the mental folding task. It is possible that 2D mental rotation tasks involving relatively simple transformations, and other tasks involving more complex object transformations (i.e., mental folding), may need to be considered as spatially separate skills. Such a model configuration would not necessarily align entirely with the Uttal et al. (2013) dimensions, because mental rotation and mental folding would be considered distinct, despite being within the same Uttal et al. (2013) sub-domain.

The CFA in Chapter 4 (spatial cognition as a predictor of children's understanding of sound) provides further insight into the structure of spatial cognition. With a sample of 9-10 year olds, there was also no evidence that the two-factor models were a significantly better fit than the one factor-model. A key difference in spatial tasks within this chapter was within the extrinsic-dynamic domain. In Chapter 4, this was measured through an arrival judgment task. This task loaded poorly onto all model configurations. This finding supports prior research on dynamic spatial ability, the findings of which indicated that these tasks measure an ability distinct from other traditionally measured spatial thinking tasks (Hunt et al., 1988; D'Oliveria, 2004; Larson, 1996).

Prior to this thesis, the spatial tasks with which arrival-judgment type tasks had been compared, had been intrinsic, within the Uttal et al. (2013) model. Therefore, the finding that tasks such as the arrival judgment task were dissociated from other traditionally measured spatial thinking tasks, could have been interpreted to mean that they are dissociated from intrinsic spatial tasks. However, the findings of Chapter 4, suggest that this dissociation extends to other spatial thinking skills, more broadly. It is likely that a two-factor model, containing the four other spatial tasks from Chapter 4, and a separate factor with several of these 'dynamic' spatial tasks, would have been a better fit, than a one-factor model. Indeed, a recent framework of spatial abilities (Buckley, Seery, & Canty, 2018), includes two broad categories: static spatial factors (which includes both mental rotation and perspective taking as sub-factors) and dynamic spatial factors, including speed judgment. The distinguishing characteristic of these two overarching factors is whether the tasks loaded onto them involve externally generated movement of stimuli.

To summarise, there was no significant evidence psychometrically within the thesis that either of the Uttal et al. (2013) dimensions provide a better fit to the data than a one-factor model. This therefore suggests that there was no evidence that the intrinsic-extrinsic or static-dynamic constructs are evident in childhood, within the data in the current thesis.

Other data in the thesis, beyond the direct psychometric evidence, provide further insight into the structure of spatial cognition. First, the cross-sectional nature of the data in Chapter 2, also allowed the model to be applied in terms of relative improvement in accuracy of the skills, throughout development. The developmental data revealed that for mental folding and mental rotation only, the intrinsic-dynamic tasks, development was better described by a quadratic trend. This was driven by rapid development of skills until the age of 8, and then a significantly slower rate of development thereafter. A linear trend described development across age for the other tasks. For all of the spatial tasks, there was significant development in accuracy between the ages of 6-8 years. However, for mental folding and the CEFT only, there was significant development between two consecutive age groups earlier within middle childhood, i.e., between 7-8 years. For the perspective taking task, accuracy of the 11-year-olds was also significantly higher than the 8-year-olds. One possible conclusion from this data was that the intrinsic tasks (mental rotation, mental folding, CEFT) tended to show a greater amount of development earlier within middle childhood. In contrast, the extrinsic tasks (spatial scaling and spatial perspective taking) more so showed continued, steady development, throughout middle childhood. However, this was not quite the case for the CEFT as performance on this task showed a linear pattern of development.

The finding that spatial perspective taking (extrinsic-dynamic skills), in particular, showed a somewhat different developmental pattern to mental rotation (intrinsic-dynamic skills), was in line with predictions, and also findings from Crestectini et al. (2014). Moreover, Xistouri and Pitta-Pantazi (2006), also found that perspective taking continued to develop significantly, beyond 9 years. The findings of Chapter 2 demonstrate that this contrast extends to other intrinsic-dynamic skills, such as mental folding.

In contrast to object-based transformations, mature spatial perspective taking performance may rely more on an allocentric framework (i.e., use of landmarks and external reference points). Prior research also suggests that allocentric strategy use is not fully mature until the age of 10 (Bullens et al., 2010). Thus, the relatively more linear and continued later development of spatial perspective taking may be attributed to the later acquisition of allocentric skills, or an increasing tendency with development to spontaneously use allocentric strategies. It may also reflect the protracted development of executive functions, including working memory and inhibitory control (Gathercole, Pickering, Ambridge & Wering, 2004). This is because effective performance on the perspective taking task required the participant to directly inhibit their own egocentric perspective, because one response option, for non 0^{o} trials, was an image of the array from their viewpoint.

It is also possible that the relatively more linear development of spatial scaling, compared to the intrinsic-dynamic skills, could be linked with the later development of allocentric skills. The process of spatial scaling requires the participant to encode a target location, possibly using the edges of the maps as a landmark, and then align this location with one or more additional maps. Allocentric involves the encoding of space based on the relationship between objects and landmarks. Thus, the need to encode target location on the referent spaces, in relation to the map edges, may have drawn upon allocentric skills.

The relationship between spatial skills and aspects of science learning also provide additional insight into the Uttal et al. (2013) model. First, in Chapter 4 (the relationship between spatial skills and understanding of sound) the three tasks 'dynamic' spatial tasks were the best predictors of overall performance on the problems, whereas neither of the static tasks were, for any science measure. However, given that, as outlined above, the arrival judgment task loaded poorly onto the dynamic factor, this does not necessarily provide support for the Uttal et al. (2013) model.

In Chapter 3 (spatial skills as a predictor of performance on a curriculum-based assessment) after accounting for receptive vocabulary and age, 2D mental rotation did not predict any aspect of science performance, whereas mental folding did. Likewise, in Chapter 4, the analysis of predictions and explanations in the science assessment revealed that mental rotation accounted for unique variance in the predictions aspect of the assessment, whereas mental folding accounted for unique variance in the explanations aspect. This therefore suggests a distinctive role for mental rotation and mental folding. Given they fall into the same intrinsic-dynamic sub-domain, this does not support the 2 x 2 model. Taken together, the findings of both Chapter 3 and 4, appear to be more linked to the dimensions within Carroll's (1993) model (spatial visualisation: folding vs. spatial relations: rotation) than the intrinsic-dynamic Uttal et al. (2013) sub-domain. The mental rotation tasks required more rapid judgments, whereas the mental folding task involved more complex transformations. The distinction between Carroll's (1993) spatial visualisation/spatial relations distinction is also supported by differences in working memory/executive function (Miyake et al., 2001). Alternatively, the apparent distinction between mental folding and mental rotation might provide support for the distinction between rigid (rotation) non-rigid transformations (folding) transformations (Atit et al., 2013).

To summarise, there was no significant evidence of a psychometric dissociation along either of the two Uttal et al. (2013) dimensions. However, there are several reasons why this does not provide definitive evidence that spatial ability is a unitary construct in middle childhood. It may be that methodological factors may have played a role in the lack of a significant dissociation, e.g., sample size, statistical power, number of tasks loaded per factor. Second, across Chapters 2 and 4, a single spatial factor did not account for more than 66% of the variance in any spatial task.

Moreover, a single spatial factor did account for a satisfactory amount of variance in performance on certain spatial thinking tasks (e.g., mental rotation, arrival judgment). This suggests that there are aspects of some of the spatial tasks that are not shared with the other spatial tasks, and, that a two-factor, or more, structure may exist, given different measures or model configurations. This may particularly be the case for measures of multiple object dynamic spatial ability. Third, developmentally, there was also evidence that the intrinsic tasks (particularly mental rotation and mental folding) show more rapid development earlier within middle childhood, compared with the extrinsic tasks (particularly perspective taking). A general theoretical and empirical critique of the Uttal. et al (2013) model is given below.

6.2.1 General empirical and theoretical critique of Uttal et al. (2013) model in light of available evidence.

Empirically, only the findings from the current thesis and the previous study by Mix et al. (2018) have directly addressed the 2 x 2 model. There was no psychometric support for either of the two major dimensions in the current thesis. However, there was some support for the intrinsic vs extrinsic dimension in the Mix et al. (2018) study, at least for 6 and 9 year olds. Given the methodological strengths of the Mix et al. (2018) study (sample size, statistical power), it would be reasonable to conclude that there is at least some support for the intrinsic vs extrinsic spatial distinction in middle childhood, with the tasks used in that particular study. There was no support for the static vs dynamic distinction in either the current thesis or in Mix et al. (2018), which calls into question the validity of this distinction. Finally, there was no support for the full four-factor 2 x 2 model in Mix et al. (2018). Synthesising this evidence together, based on the currently available evidence, the intrinsic-extrinsic dimension of the model appears to have at least some support. In addition, it is worth noting that in descriptions of the model (e.g., Newcombe 2018) also emphasise the intrinsic-extrinsic factor.

However, though there is some support for the intrinsic vs extrinsic distinction, when considering the literature more broadly, several caveats should be mentioned. First, the findings in the thesis also point to the fact that the intrinsic dimension includes, in the same dimension, skills which are partly distinct. Specifically, mental rotation and mental folding, demonstrated consistently distinct differential predictiveness of science learning. Indeed, one of the authors of the model (Newcombe, 2016), in a later description of the model, emphasises the distinctiveness of mental folding vs mental rotation. For instance, mental rotation often shows gender differences, whereas mental folding does not. Newcombe (2016) for this reason states that the 2 x 2 model may be too simplistic. It may be that a further version of the model is needed which formally sub-categorises intrinsic-dynamic skills.

Second, it is unclear if and how multiple object dynamic spatial abilities fit into the 2 x 2 model. Such abilities might be labeled as extrinsic-dynamic. However, the evidence from this thesis suggests, as outlined above, that tasks involving moving stimuli and speed-time judgements may be distinct from static stimuli. The recent model by Buckley et al. (2018) provides an interesting alternative to the 2 x 2 typology, by distinguishing broadly between static and dynamic stimuli. In this model, visualisation and spatial orientation are grouped in the same larger 'static' spatial factor.

Finally, as discussed in section 1.2.2.2, tasks labelled as intrinsic can be solved with extrinsic strategies, and vice-a-versa. Thus, this may introduce noise into psychometric analyses, making a dissociation difficult to detect, which may account for the null result in the current thesis.

To conclude, there appears to be some merit in the intrinsic vs extrinsic aspect of the Uttal et al. (2013) model, but this data is based around one main prior study (Mix et al., 2018); further research with adequately sized samples and suitable task selection is needed. However, the intrinsic vs extrinsic distinction may still be an oversimplification. There is currently no support for the full four-factor model, although this was not tested in the current thesis, and it may not be practical to do so in a rigorous way, given the number of measures needed per sub-domain. Moreover, further models should integrate and take into account the static vs dynamic distinction in terms of onscreen movement of stimuli.

6.3 The relationship between individual differences in children's spatial thinking skills and science learning

Given that there are no firm conclusions regarding the validity of the Uttal et al. (2013) model, and that the spatial tasks across the thesis were chosen based on a priori decision based on the model, the evidence regarding the connections between spatial thinking and science that follow are discussed in relation to the Uttal et al. (2013) sub-domains. However, further discussion and evaluation of the connection between these somewhat contradictory lines of evidence is provided in section 6.6.

6.3.1 Review of findings from Chapters 3 and 4

Prior to the thesis, a small number of studies (e.g., Tracy, 1990; Harris, 2018) provided an initial indication that spatial thinking skills are a predictor of primary-school aged children's science learning outcomes. However, the findings of prior research were limited in several ways. First, prior studies included only intrinsic-dynamic spatial thinking skills, and in some cases, composite measures of several different intrinsic-dynamic spatial skills. Second, the studies only included children of a single age group, and it was therefore not possible to examine possible age-based interactions. Finally, the studies were not designed in a way that allowed comparisons to be made between different types of scientific skills (e.g., factual recall vs. conceptual understanding). In addition, other than the use of a standardised science assessment used Tracy (1990), the studies lacked applicability to the classroom. To address this, in Chapter 2 and 3, the contribution of spatial thinking skills to primary-school aged children's science learning was investigated through distinct but complementary methodologies. In Chapter 3, science was assessed through a more general curriculum-based measure, and in Chapter 4, there was a focus on science learning outcomes, within the classroom. In both studies, spatial thinking tasks were chosen using the Uttal et al. (2013) model. The model had not previously been applied to science learning at any stage of science education.

Chapter 3 presented data on the relationship between 8 to 11 year old's spatial thinking skills and broad science achievement. The science assessment was based upon questions designed to assess the achievement of the UK science curriculum for this age range. Within each question, which focused on a specific conceptual sub-topic, items varied in their requirements: this included factual recall, problem-solving, and aspects of scientific investigation skills. The results of the study showed, overall, that, controlling for receptive vocabulary, spatial thinking skills significantly predicted children's science achievement. This broad finding therefore supported previous results with children in the middle childhood age range described above (Mayer et al., 2014; Tracy, 1990; Harris, 2018) as well as other research with adults (e.g., Kozhevnikov & Thornton, 2006) and adolescents (e.g., Delialioglue & Askar, 1999). More specifically, mental folding and spatial scaling were the strongest overall predictors of total science scores, and, mental folding was a stronger predictor overall than spatial scaling. There was also some variation in this pattern, when considered by area of scientific discipline. For questions based around biology sub-topics, mental folding and scaling were again the best predictors. The strongest predictors for questions based around chemistry sub-topics scores were the CEFT and spatial scaling task. Finally, for physics-related questions, the strongest predictor was mental folding.

To fit the above findings to the Uttal et al. (2013) spatial typology, an intrinsic and an extrinsic spatial skill were the best predictors of overall science scores; this pattern was mirrored for both chemistry and biology. This finding adds to the limited number of previous studies which have compared intrinsic and extrinsic spatial tasks, within the Uttal et al. (2013) model, as predictors of aspects of science learning. In previous research with undergraduates, Sanchez and Wiley (2014) reported that a multiple object dynamic (extrinsic) spatial ability task, uniquely predicted geoscience understanding, whereas mental folding, an intrinsic-dynamic task, did not. However, this finding should be considered in the context of findings of the arrival judgment task, discussed in the previous section. That is, there was evidence that arrival judgment tasks, involving moving stimuli, are somewhat distinct from spatial tasks involving static stimuli. Furthermore, in the domain of biology, Cohen and Hegarty (2007) reported that spatial perspective taking was a stronger predictor of the ability to identify 2D cross-sections of 3D shapes than mental rotation. However, this study did not include both measures within a single regression analysis, so it is unclear whether this association reflected unique variance in perspective taking (i.e., if perspective taking uniquely predicted performance, after accounting for mental rotation). The role of extrinsic-dynamic skills in this study was likely to have reflected the requirement to visualise what an object would look like, when viewed from a particular viewpoint (i.e., inside the object).

The finding in Chapter 3, that spatial scaling predicted science achievement overall, and specifically scores on questions based around biology and chemistry sub-topics, is a novel finding in the literature focused on children's science learning. It is also novel within research investigating the relationship between spatial cognition and science learning in adolescents and adults. The curriculum-based, composite nature of the science assessment means that it is only possible to speculate about the mechanisms involved in this relationship. First, having an understanding of scale and proportion is important in science, for example, when learning about concepts which vary significantly in size (e.g., a cell versus a galaxy). Relatedly, when learning in science, it is necessary to switch between representations at different scales, such as models and diagrams. Alternatively, there may have been online-demands of the problems themselves which utilised extrinsic-static skills. For instance, the use of tables and specific types of diagrams (e.g. classification keys) may draw more heavily on extrinsic-static spatial skills.

The role of mental folding in children's science learning, and in children's physics learning more generally, was aligned with prior research by Harris et al. (2018). Mental folding was discussed within Chapter 3 in relation to mental model construction, of both concepts and individual problems. The role of intrinsic-dynamic spatial thinking skills, and mental folding specifically, in mental model construction will be outlined in depth, in relation to Chapter 4, below.

The role of the CEFT for questions based on chemistry sub-topics, only, is somewhat unclear. On the one hand, it is possible there were features specific to these questions which were particularly supported by static form processing. For instance, specific diagrams or visualisations within the questions which may have benefited from visual discrimination skills. It is also possible that it was a reflection of model construction and manipulation, along with mental folding. Recall from Chapter 1 that disembedding, the skill proposed to be measured by the CEFT, is only sometimes dissociated from tasks loading onto Carroll's (1993) visualisation factor (e.g., mental folding). It has been suggested that this is because they both require the maintenance and representation of an object, whilst performing additional processing. In the case of intrinsic-dynamic tasks, the additional processing required relates to the ability to perform a spatial transformation, whereas in the CEFT, the additional processing requirement relates to the ability to search for the hidden shape. For chemistry scores, the CEFT was the stronger spatial predictor, compared to spatial scaling. Thus, the pattern was similar to that observed for overall scores and biology. That is, in all cases, mental folding and the CEFT were the strongest spatial predictors, and spatial scaling accounted for additional variance in science scores, above and beyond that.

Finally, the results of the study in Chapter 4 showed that age did not interact with the effect of spatial ability on science achievement. This suggested that, in contrast to predictions based on previous research (Mix et al., 2016; Hambrick et al., 2012), the impact of spatial ability was not necessarily greater for younger versus older, or more experienced, learners. However, the age range in the study was relatively small, and, if the age range had been extended upwards or downwards, age-based interactions may have been observed.

To summarise, the main contribution of Chapter 3; through the use of a broader range of theoretically-identified spatial measures than had been administered previously, it was possible to isolate particular spatial skills which best predicted children's science achievement. In addition, the other main contribution of the chapter was the finding that this relationship did not vary developmentally, across the 7-11 age range. These two factors combined are novel within the literature in childhood, and with respect to the first point, the literature as a whole, with adults and adolescents.

The science questions used in Chapter three were of the type that children would typically encounter within school, and so the findings also have ecological validity, from an educational perspective. Conversely, the main disadvantage to the methodology of Chapter 3 was that these questions were designed to assess children's science achievement in a broad sense, and not for research purposes to elucidate specific relationships. Another limitation was that it was unclear how much of the observed relationship may have been accounted for by a more general reasoning ability. Despite this, the findings presented several avenues for future research.

Building on these findings, Chapter 4 focused on a specific physics topic, sound, in-depth, in the context of children's learning in the classroom. In this study, children completed a measure of factual knowledge about sound, and, a measure of conceptual understanding, which related to reasoning about novel problems within the topic of sound. The problem-solving questions were then further divided into predictions and explanations, also related to sound. In contrast to Chapter 2, the study also included non-verbal/fluid reasoning, as well as a measure of vocabulary.

After controlling for vocabulary and non-verbal reasoning, in line with predictions, spatial ability was not a significant predictor of factual knowledge; the only cognitive task that uniquely predicted factual knowledge scores was vocabulary. Spatial ability was, however, a significant predictor of performance, overall, on the problem-solving questions (i.e., predictions combined with explanations). Specifically, the three spatial tasks which had been theoretically defined as being dynamic (mental rotation, arrival judgment and mental folding) were the best predictors of problem-solving scores. However, mental folding was the only unique predictor, when all three spatial tasks were simultaneously included within the regression model. After controlling for the two IQ measures, neither spatial scaling, or the visual discrimination task, were significantly related to any of the science learning outcome measures.

When the problems were analysed in more depth, the results revealed that mental folding was more closely related to explanations, whereas mental rotation was more closely related to predictions. Additionally, spatial ability overall accounted for a larger proportion of variance in conceptual explanations than predictions. Given that spatial ability was most closely related to performance on the problem-solving questions (and therefore to the application of conceptual understanding, and not factual knowledge), it was proposed that spatial ability played a role in the construction and manipulation of a mental model of sound. This finding supports prior research with adults, in which the authors link spatial thinking skills to the construction and use of mental models of physical processes (Hegarty & Just, 1993; Narayanan & Hegarty. 2002; Sanchez & Wiley, 2014). The possible mechanisms involved in the relationship between spatial ability and mental model construction are discussed in detail in the next section.

The finding that children's spatial thinking skills generally predicted aspects of science learning, and more specifically physics learning, mirrored the results of Chapter 3. However, in contrast to Chapter 4, the results of Chapter 4 revealed that this relationship existed only when children were required to apply their conceptual understanding to scenarios and problems, and not to recall facts. Similar findings were reported by Rhodes et al. (2015); however, this previous study focused on visuospatial working memory, and a chemistry topic, with adolescents. In addition, the unique role of mental folding, intrinsic-dynamic spatial thinking skills, for physics learning, also mirrored the findings of Chapter 3. The findings of Chapter 4, also suggested that the relationship remained after controlling for fluid reasoning, which suggested that the relationship was not better accounted for by general reasoning ability. Finally, the overall effect size between Chapter 3 and 4 was similar; spatial ability accounted for approximately 8-10% of the variability in science scores, after controlling for the IQ measure or measures used within the respective chapter.

6.3.1.1 Intrinsic-dynamic thinking and mental models

In the previous section, the correlation between intrinsic-dynamic spatial thinking skills (mental rotation and mental folding) and problem-solving scores, related to sound, were linked with the construction and use of mental models. Mental model construction was also proposed as a general mechanism for the role of mental folding in physics scores and science achievement, overall, in section 6.3.1, in relation to Chapter 3. This section explores the role of intrinsic-dynamic skills in mental models in more depth, drawing on the theoretical approaches outlined in Chapter 1. In particular, the findings of Chapter 4, on sound, are applied to these theoretical approaches. However, this analysis could be modified to apply to other areas of physics, and science, more broadly.

One possibility is that intrinsic-dynamic spatial thinking skills support the construction of transient mental models of problems. This proposal is broadly in line with the theoretical approach of Johnson-Laird (1980). When approaching a problem or scenario, the learner encounters verbal information, and in the case of the questions in the current thesis, is also provided with diagrams and supportive illustrations. In order to answer the problem, the learner must construct a mental representation of the content, which integrates across sources of information, and goes beyond the verbatim text-level information (Sanchez & Wiley, 2014). This may be particularly important in physics problems, where the spatial relations between entities within a problem are crucial. In Chapter 4, for example, children were presented with scenarios involving people and objects, and, in order to answer the problem, it was necessary to process the spatial relations involved between the entities. Therefore, children with stronger intrinsic-dynamic skills may be more able to construct and maintain a mental model from the given information, and, integrate and coordinate between different elements of the model, or models.

However, in order to answer science questions, the learner must also utilise an existing conceptual mental model or models, or at least some form of knowledge, based on the science concept the question relates to (Vosniadou & Brewer, 1992). For instance, reading the sentence ‘Josh threw the ball to Michael’ may both generate a new rudimentary mental model, but, also activate an existing scientific mental model related to gravity. The existing mental model, in this case of gravity, may constrain the new model, relating to the described event.

Several theoretical approaches to the structure of conceptual understanding were outlined in Chapter 1. From a 'knowledge-as-theory' perspective (e.g., Vosniadou, 1994) children's understanding of sound is organised in a coherent and organised manner into a single mental model. This interpretation was supported by prior research by Mazens and Lautrey (2003). The authors specifically argued that children's understanding of sound is initially constrained by a framework theory of objects/matter, until they adopt a theory of process, after which they hold a more vibrational model of sound. In contrast, from a 'knowledge-in-pieces' perspective (e.g. diSessa,1988), knowledge is fragmented, based around p-prims, and loosely organised.

Determining the precise structure of children's concept of sound was not a central aim of the thesis. Thus, because the structure of children's sound concept is unclear, the possible mental model mechanisms, in relation to the findings of Chapter 4, will be outlined in relation to both theoretical approaches. Specifically, a possible role for intrinsic-dynamic spatial thinking skills is discussed in relation to mental model construction of concepts, based on each theoretical approach.

diSessa acknowledges the possibility of mental models as a form of knowledge structure (diSessa 1996, 2002). Importantly, however, diSessa's theoretical approach is that mental models are not constrained by a single overarching ontological framework. Instead, he proposes that mental models consist of a small number of causal elements, some of which may be p-prims (phenomenological primitives). Recall from Chapter 1 that diSessa suggests that p-prims are self-explanatory, simple, shallow interpretations of physical processes, which are constructed from everyday experiences. From this perspective, in the context of sound, a mental model of sound might consist of a collection of loosely associated, fragmented and simplified causal elements, linked to p-prims. For instance, an element of the model might be a vibrating object and perhaps a small number of vibrating particles. In diSessa's view, this element of the model might be linked to the 'Ohm's law' or the 'force as a mover' p-prims. According to diSessa, the Ohm's law p-prim consists of three factors: an energising force or impetus; a result, often in the form of action or motion; some resistance or interference effect between action and result. In the sound example above, the energising force might be the vibrating object, the result might correspond to the vibration of the air particles, and the resistance effect, between the object and the vibrating air particles. The 'force as a mover' p-prim is the belief that an action will continue, for as long as a force continues to be exerted. Thus, for example, in the context of sound, a sound is heard for as long as the object continues to vibrate.

As outlined in Chapter 1, mental simulation and animation is proposed to occur in a piecemeal fashion (Hegarty, 2004). Thus, when providing conceptual explanations, children might mentally simulate the action of particles and objects in a piecemeal fashion, using intrinsic-dynamic skills, based on the type of fragmented knowledge outlined above. These piecemeal simulations could in turn support verbal explanations. Within the science lesson, children were shown videos containing visual models which support this process. For example, one visual model included balls attached to their own ropes, representing molecules. The researcher hit a flat surface, which in turn, struck the first ball, and affected the other balls. Simplified versions of the visual models provided could have formed part of the fragmented knowledge base of the mental model. Thus, through this approach, overall, children with stronger intrinsic-dynamic spatial thinking skills may be more able to run individual aspects of the model.

Vosniadou (1994), in contrast, argues that children hold a single coherent core mental model, which is constructed from a set of closely related beliefs. Some of these beliefs originate from socially-transmitted knowledge (e.g., science instruction). In the case of sound, in order for children to grasp the vibrational model, they must draw more heavily on this socially transmitted knowledge, because action at the molecular level is not observable. In the context of the lesson in Chapter 4, this knowledge was transmitted through verbal descriptions, images, and animations/videos. For example, the verbal description ‘striking an object causes it to vibrate’ could in itself constitute a belief within the specific theory. In order to be integrated, however, the learner must go beyond the surface level meaning. This could theoretically involve the learner generating and maintaining structured mental images, based on the belief. At a later point, when the learner is generating or recalling their mental model of sound, based on these beliefs , intrinsic-dynamic visualisation spatial skills might support the ability to integrate and mentally represent (i.e, simulate) the model.

Vosniadou’s theoretical approach is that mental models of scientific concepts are constrained by a single ontological framework. In the context of understanding the earth, Vosniadou and Brewer (1992) argued that children initially categorise the earth as a physical object, and therefore erroneously apply to the earth the physical constraints of objects. Similarly, Mazens & Lautrey (2003) argue that by the time children have adopted a more advanced conceptual model of sound, they have, gradually, ceased to attribute to sound the properties of solid objects. Perhaps, then, from this theoretical perspective, spatial skills also support the construction and manipulation of mental models of sound, because they play a role in challenging these presuppositions.

For example, consider a scenario in which a learner is asked to compare the rate at which sound travels through a gas and liquid. If a learner attributes to sound the properties of solid objects, it might lead to the belief that sound travels more easily through a gas; a solid object would indeed travel more easily through a gas. Presenting children, in the lesson, with visual models representing the structure of matter, and how this interacts with sound propagation, in part, perhaps played a role in challenging these presuppositions. For example, a visual model shown to children showed the particles within a solid being packed closely together, and how sound travels easily through the solid as a result. Children with stronger intrinsic-dynamic spatial skills may be better able to integrate, manipulate and simulate these representations, which may go some way in challenging object-based presuppositions.

Drawing on the discussion above, the following processes are proposed as one mechanism to account for the role of spatial thinking skills, particularly intrinsic-dynamic/spatial visualisation skills, in reasoning about physical processes, within the context of problems or given scenarios. Children possess mental models of the scientific concept on which the problem is based, constructed from prior experience, observations and socially-transmitted knowledge. When approaching a problem or scenario, children also construct a mental model of the problem, which integrates verbal and non-verbal information presented within the question or problem. In many cases, this includes spatial-relational information which is central to the problem (e.g., entities at different locations, and different objects). Spatial thinking skills are involved in the construction of the problem model and the interpretation and integration of the elements of the model. In order to answer the question, or solve the problem, the learner then uses or applies the conceptual model of the scientific concept. This process might involve accessing a coherent conceptual model (Vosniadou), or, loosely connected fragments of knowledge (diSessa). For example, a scenario involving sound traveling through a wall might reactivate simulations of simplified fragments of knowledge, involving a small number of closely packed vibrating particles. This reactivation might then facilitate the generation of verbal explanations, or, the reactivation of existing verbally-based beliefs.

Although this example is closely aligned to physics, this process could apply to other areas of science in which a process is involved. In addition, the general process of mental model construction of concepts, and models in response to questions, could more broadly apply to areas which are less process-based. For example, in biology, in the construction of a mental model of a plant, which includes representations of the functions and relationship between parts.

6.3.1.2 Summary of findings from Chapters 3 and 4

Intrinsic-dynamic skills (i.e., spatial transformation skills, particularly mental folding) appear to be most important in children's science learning, relative to the other spatial skills administered. This finding was replicated across both Chapters 3 and 4. Mental model construction and use was proposed as a primary mechanism. The role of intrinsic-dynamic spatial thinking skills in relation to mental models was outlined mainly in the context of physical processes, because this was the area on which there was the most data. However, with modifications, this mechanism could apply broadly to many areas of science, process-based or not. In addition, there was no evidence that intrinsic-dynamic spatial thinking, or spatial ability per se, predicted children's ability to recall factual information. The role of intrinsic-dynamic skills appear, therefore, to be related to understanding and comprehension; either of concepts, or problems, or to support the relationship between the two.

Spatial scaling emerged as a predictor within Chapter 3 only; for overall scores, chemistry and biology scores, but not physics scores. In contrast, spatial scaling was not a significant predictor of any learning outcomes about sound, in Chapter 4. Several possible mechanisms were proposed for the role of spatial scaling. Likewise, the CEFT also emerged as a predictor of chemistry only, in Chapter 3. This may have been related to specific aspects of the chemistry concepts or problems. Alternatively, skills assessed by the CEFT could also have been playing a similar role to mental folding, i.e., the association could be related to mental model construction. However, future research is needed to determine the role of the CEFT and spatial scaling in children's science learning and achievement.

6.4 The role of gesture as a spatial tool to support science learning

The final aim of the thesis was to investigate the role of gesture as a spatial tool for supporting children's science learning, alongside the possible moderating role of spatial thinking skills. In Chapter 5, the effectiveness of learning through gesture was considered in relation to two other spatial tools: concrete models and diagrams. Both gesture and concrete models (in this case, magnets) are spatial tools which provide physically embodied experiences. The expectation was that both gesture and concrete models would lead to stronger learning outcomes, relative to the verbal/diagram condition, as a consequence of these physically embodied experiences, related to the concept of magnetism. Based on previous research (Stieff, et al., 2016; Novack, Congdon, Hemani-Lopez & Goldin-Meadow, 2014) the prediction was that the gesture-based teaching condition would provide additional benefits to a purely action-based teaching (concrete model) condition. In particular, because gestures do not provide tactile feedback, and because gestures do not provide direct access to affordances of the object (e.g., repelling poles of the magnet), the use of gestures may in some contexts encourage the learner to construct a rich internal representation. In addition, gestures schematise spatial information, by stripping away unimportant information (e.g. the cuboid shape of the bar magnet) and leaving the core information intact. Stripping away irrelevant information, and focusing on the essential features, may also support the transfer to untaught contexts (in this case, un-taught magnets).

The predictions above, however, were not always borne out. At a more general level, the results of Chapter 5 suggested that physically embodied spatial tools (gesture and concrete models) do not always provide significant additional science learning benefits, relative to learning tools which do not include sensorimotor experiences (verbal description and diagrams). On the one hand, in study one, performance on the gesture teaching condition did outperform performance on the verbal-diagram teaching condition. Yet, also within study one, performance on the concrete model (magnet) condition did not outperform that on the verbal-diagram teaching condition. Moreover, the results of an overall ANOVA which included both studies, did not reveal a significant interaction effect to support the specific instruction effect in study one. Furthermore, in study two, there was no evidence that children performed significantly worse in the conditions in which they learnt with diagrams and verbal description only, relative to the condition in which they learnt with concrete models.

These latter findings, in particular, differ somewhat to previous research from Atit, Gagnier & Shipley (2015), who reported that prohibiting participants from gesturing during a geoscience explanation, reduced the degree of improvement on a measure of spatial thinking. However, there were some differences between the two studies. For example, in the studies reported in Chapter 5, gestures were not prohibited during the explanation phase. Participants were prohibited from producing gestures during the consolidation phase, and not during the post-test, where explanations were generated.

Taken together, the findings above suggest that the combination of diagrams and verbal description can be a reasonably effective spatial learning tool in science. Although it is not possible to determine precisely what features of this teaching method were effective, the verbal/diagram condition did involve the active participation of the learner. For example, children had scaffolded practice at using diagrams of magnets, prior to the assessment questions. Having the opportunity to use and practice with diagrams related to the concept was likely beneficial, given that the post-test assessment required the use of diagrams. Furthermore, although it was not the intention of the verbal-diagram teaching condition, it is possible that through the structured questioning with diagrams prior to the assessment, the diagram/verbal condition indirectly provided general diagram training. Prior research by Cromley et al. (2016) reported that structured instruction in the use of diagrams is effective in promoting learning within science. These findings are also partly in line with the work by Fiorella and Mayer (2017) who found that the use of spatial strategies during learning, such as drawing mind maps and diagrams, significantly predicted learning outcomes, in the domain of biology.

Considering teaching through gesture specifically, contrary to predictions, there was no strong evidence, overall, that gesture as a spatial tool was more effective than the other spatial tools. In study one, gesture-based teaching was more effective than verbal/diagram-based teaching, for the near transfer questions only; but, no more effective than teaching through concrete models. In study two, teaching and learning through gesture was no more effective than any other condition. These findings differ from previous research in the domain of biochemistry with adults (Steiff et al., 2016) and mathematics with children (Novack et al., 2014), who both found a benefit of gesture over concrete models. The findings of the two studies in Chapter 5, were, however, more in line with research by Congdon et al. (2018). This study found that for children with a higher level of prior knowledge, although gesture and action were both effective for learning about measurement, neither was more effective than the other. Though prior knowledge was not considered as a moderating factor in Chapter 5, the key point is that, in Congdon et al. (2018) gesture was only ever as effective as action, for all children. Taking this prior research into account, the findings of Chapter 5, could therefore be attributed to the specific topic. Magnetism, at the conceptual level taught at this stage of education, is perhaps not suited for learning through gesture. It may have been that the gestures, and the topic, was too ‘concrete’, in comparison to algebra, for example.

From a theoretical perspective, both the gesture-as-simulated action theory (Hostetter & Alibali, 2008) and the gesture-for-conceptualisation theory (Kita et al., 2017), see gestures as being produced through a ‘general purpose action generator’. This is also broadly in line with theories of embodied cognition (Gaselle & Lakoff, 2005; Barsalou, 1999; Glenberg, 2010). Predictions generated from these theories, together, would suggest that having sensorimotor experiences during the teaching session, would enable the subsequent reactivation of these areas during the post-test, and thus, support post-test reasoning. However, the finding that the condition with no sensorimotor movements at all, in study 2, was not ineffective, relative to the more physically embodied conditions, contradicts this. One possibility is that children may have previously had physical experiences of using magnets in everyday life. If this were the case, viewing photographs during the familiarisation phase could have induced imagined movement and reactivation of the sensorimotor areas, in the absence of the models themselves. Sensorimotor reactivation could have therefore occurred during the post-test, even if participants had no physical experiences related to the concept, during instruction.

The other main area of findings, linking to section 6.3 of this chapter, concerned the relationship between mental folding skills and learning outcomes, within each condition. The key finding was that the contribution of mental folding scores to learning outcomes depended on the pre-experience, the teaching condition, and the transfer type. Considering first near transfer outcomes, a correlation was evident between mental folding scores and near transfer scores in the verbal-diagram condition, in study 2, only. Mental folding scores were also correlated with consolidation question accuracy for the verbal-diagram condition in study two, but, did not significantly correlate with consolidation question scores in the verbal-diagram condition, in study one. Moreover, the results also revealed that mental folding scores were significantly correlated with near transfer scores, in the gesture conditions, again, only in study two. However, gesture production accuracy, within the consolidation questions, was not related to mental folding scores. This suggested that the correlation between mental folding scores and performance in the gesture condition was not due to children with lower mental folding scores being less able to produce the gestures.

Compared to any other conditions, in both the gesture and the verbal-diagram condition in study 2, spatial transformation skills were more important because the concrete model was never provided to the children at any point. For example, in the verbal-diagram condition in study 2, having stronger mental folding skills was perhaps more critical, because, in this condition, it was necessary to mentally ‘animate’ the static diagrams and photographs. Within this condition, children had no sensorimotor experiences at all. Thus, in the absence of an external model, being able to generate an internal mental model was more crucial. Furthermore, considering the gesture condition, when the children were asked to produce gestures and ‘imagine’ that one part of the gesture represented a particular part of the magnet, their only symbolic referents were the photographs, and the static 2D diagrams. Thus, mental folding skills may have been useful in the gesture condition, in study 2, to enable children to visualise the connections between the gestures and information presented to them in the teaching presentation through diagrams, and verbally by the researcher. Authors of other previously discussed research with adults (Sanchez & Wiley, 2014; Höffler , 2010) also concluded that spatial ability is particularly important for ‘animating’ static diagrams, but less important when visualisations are externally animated for the learner.

Mental folding scores correlated with intermediate transfer scores consistently, across studies 1 and 2. The intermediate transfer questions presented information about novel magnets, in a verbal and diagrammatic form. During the instruction phase within all conditions, children learnt about bar magnets, in the context of the bar magnets. In order to comprehend the requirements of the intermediate transfer questions, it was essential to be able to independently decode the information contained within the diagrams. Therefore, it may have been the case that mental folding skills were more important within the intermediate transfer questions, than the near transfer questions, due to the novelty of the presented spatial information. An implication of this finding is that children with lower mental folding skills are consistently at a disadvantage in more novel transfer contexts, where no prior direct instruction is provided.

To summarise, the main aim of this aspect of the thesis was to determine the role of gesture as a spatial tool in learning science, particularly in relation to other spatial tools commonly used in the classroom. There was no strong evidence that gesture per se was more effective as a spatial learning tool, than the other tools investigated. The findings also indicated that the combination of diagrams with verbal description, without any type of physical or embodied experience, can be a reasonably effective spatial teaching tool. Finally, the findings suggested that having additional embodied / physical experiences do not always provide significant additional significant learning benefits in science. Nevertheless, there were positive correlations between mental folding scores and scores and the verbal/diagram and gesture conditions, in study 2. Moreover, regardless of teaching approach, there were correlations between the intermediate transfer scores and mental folding skills. Children with lower spatial skills still therefore require additional support, in order to learn as effectively as children with stronger spatial skills, particularly in lesson contexts based on these teaching methodologies.

6.5 Overall summary thesis findings

With respect to the structure of spatial cognition in childhood, there was no strong evidence, psychometrically, that, within middle childhood, spatial thinking skills are clearly dissociated along the intrinsic-extrinsic or static-dynamic dimensions proposed, by Uttal et al. (2013). However, a single spatial factor did not always account for a satisfactory amount of variance in spatial scores. These issues considered, combined with the pattern of observed relationships with science throughout the thesis, and the developmental data, suggest that, at the very least, further research is needed to fully determine the structure of spatial cognition in childhood.

The key finding of the thesis was that individual differences in children's spatial thinking skills predict both general science achievement, and, more directly, the outcomes of science instruction. First, intrinsic and extrinsic spatial thinking skills, within Uttal et al.'s (2013) theoretical model, emerged as significant predictors of 8-11 year olds' general science achievement. Furthermore, in a whole-class learning context, with instruction based on verbal description, static images and dynamic animations, spatial thinking skills predicted 9-10 year olds' ability to apply conceptual understanding of sound to novel contexts. Intrinsic and extrinsic dynamic spatial skills were the best spatial predictors of learning outcomes related to the problem-solving questions. However, spatial thinking skills did not relate to the recall of factual scientific information. In a one-to-one learning context on the topic of magnetism, mental folding skills were a significant predictor of learning outcomes closely related to the taught-concept (near transfer), in the two conditions in which a concrete model was not provided, and spatial transformation skills were more important (i.e., the verbal-diagram and gesture conditions, in study 2). Furthermore, problem-solving about magnets at the level of intermediate transfer, i.e., problem-solving about novel magnets, was always related to mental folding ability. In this case, spatial thinking skills may be more important due to the processing of novel spatial representations. These conclusions, however, must be considered with the caveat of the varying levels of statistical control for general ability, between Chapters 3, 4 and 5. Across the thesis, a specific role of non-rigid, intrinsic-dynamic transformations (i.e., mental folding) was evident in a variety of contexts.

From the perspective of spatial tools, in Chapter 5, it was revealed that diagrams, models and gesture can all at least be reasonably effective as spatial tools for learning about magnetism. For questions less closely related to the core learning (i.e. intermediate transfer) all three spatial tools were equally as effective as each other. Given brief familiarisation with concrete models prior to teaching, although diagram-based teaching was the least effective method for the near-transfer problems, it still led to relatively good outcomes. When children were familiarised with photographs first, there were no meaningful differences between conditions. Further research is needed to compare these findings to other conceptual topics.

6.6 Overall evaluation of main lines of evidence in current thesis

Methodological limitations impact the strength of the conclusions which can be drawn in the thesis. First, a central issue concerns the somewhat contradictory findings regarding the structure of spatial cognition, when synthesising evidence across the empirical chapters. In particular, there was no strong evidence that spatial cognition had a two-factor structure, based on psychometric modelling across Chapters 2 and 4. Yet, there were distinctive patterns in how different spatial thinking tasks predicted different aspects of science learning, across Chapters 3 and 4. In addition, developmentally there was some evidence of an intrinsic-extrinsic dissociation in Chapter 2. It is therefore important to evaluate each line of evidence, to determine which conclusions are most convincing.

The psychometric modelling was methodologically limited by the number of tasks loaded onto each of the two factors. It was possible to only load two extrinsic, and three intrinsic tasks within the CFA models; 2 tasks per factor is the absolute minimum number recommended. In addition, the sample size for the CFA analyses was the minimum needed for the analyses for Chapter 4, and slightly under the minimum for Chapter 2. As discussed, the potential for using intrinsic strategies for extrinsic tasks and vice-a-versa potentially adds a significant amount of variability within tasks, making a between factor dissociation more difficult to detect. Therefore, the lack of support for either two-factor spatial model in the current study is not necessarily strong or convincing evidence in favour of a unidimensional structure.

Considering the developmental findings in Chapter 2, the differences in developmental trajectories between intrinsic and extrinsic skills were reasonably convincing and there was no strong evidence that null results linked to power were masking important differences in between-group comparisons. The major inconsistency in the findings of Chapter 2 lay in the trend analyses. Specifically, the developmental trend of the Embedded Figures task across development was described by a linear trend only, whereas for the other intrinsic tasks, it was also described by a quadratic trend. In this case, power issues may have resulted in the lack of significant quadratic trend for the embedded figures task, given there was only one major inflection along the developmental trend. However, in this case, a significant difference here would have strengthened the intrinsic-extrinsic distinction.

The differential predictiveness of spatial skills in relation to science learning across Chapter 3 and 4 also provided quite convincing evidence in terms of the structure of spatial cognition. As above, some of the regression models were not significantly powered to detect small effects, which may have accounted for some null results. For instance, in Chapter 4, the mental folding task emerged as the only unique predictor. The arrival judgment and mental rotation tasks were non-significant. It is possible that with a larger sample, these tasks would have also been unique predictors in the model. However, as with the developmental data, these findings would have provided more, and not less, evidence for a spatial skill dissociation.

Finally, it is also important to note that, it is possible to find a one-factor structure psychometrically through factor analysis, yet also find that some spatial skills make unique predictive contributions to science learning outcomes. Across the thesis, the single factors accounted for between 18% and 66% of variance in spatial thinking tasks. It is possible that non-shared (i.e., unique) variance associated with individual spatial thinking tasks are predictive of science learning.

Overall, weighing up the evidence, the data from the thesis suggest that more research is needed, particularly rigorous psychometric modelling, to more definitively confirm the structure of spatial cognition in childhood. In addition, given the available evidence, it is not currently possible to discount the possibility of a four-factor model of spatial cognition.

A second major line of evidence concerns the importance of spatial cognition in children's science learning, and on specific links between spatial skills and different spatial thinking tasks. The thesis provides convincing evidence that spatial thinking skills predict science learning overall. The finding was particularly strong in Chapter 4 given the inclusion of a measure of general reasoning ability. However, in Chapter 3, the distinctive pattern of significant spatial predictors in relation to the domains of biology, chemistry and physics was limited by the nature of the items and questions used with the assessment. Due to the nature of the questions, which were drawn from existing standardised assessments, there were also different types of items within questions split by science domain. Thus, this adds an additional confounding variable when comparing between science domains. In Chapter 4, the specificity of spatial thinking skills being related to problem-solving questions versus the factual knowledge was possibly limited by the relatively low number of items and reliability of the factual knowledge test. Moreover, though spatial thinking was a slightly stronger predictor of predictions versus explanations, the difference was around 4% variance explained. In addition, explanations had a higher overall number of possible marks compared with predictions, which may have influenced the extent to which spatial thinking was a predictor. Thus, overall, though the overall connection between spatial thinking and science was convincing, further work is needed to confirm the more specific links.

The final line of evidence for evaluation concerns the instructional manipulations within Chapter 5. As outlined, the data do not necessarily provide convincing evidence that gesture as a spatial tool is not useful for science learning in the primary school years. The particular gesture used in this chapter may not have been well-suited to support reasoning about magnetism. Moreover, the lack of 'verbal only' control group means that the data do not necessarily provide convincing evidence that spatial tools per se are the most effective instructional approach for learning about magnetism. That is, it is possible that children would have learned as effectively via a verbal only instructional approach.

6.7 Implications for education

The finding that spatial thinking skills, overall, are a significant predictor of children's science learning, suggests that children who have weaker spatial thinking skills might benefit from extra support in the classroom. At a fundamental level, making teachers aware of this finding could be generally useful in their day-to-day practice. For instance, embedding spatial thinking activities within the daily routine of the classroom could be beneficial for all learners. Moreover, given the findings of Chapter 5, if a teacher was planning a science lesson which was heavily based on diagrammatic representation and verbal description only, they might consider ways of tailoring their teaching for children with weaker spatial thinking skills. Furthermore, the findings of the thesis suggest that spatial skills are particularly important is for developing, using and applying conceptual understanding of scientific concepts. Therefore, similarly, being more aware of this might impact upon teachers' planning of this type of science lesson.

One possible way of supporting children with weaker spatial thinking skills, discussed in the thesis, might be to provide children additional experience with concrete models, for certain topics, particularly in physics. The reasoning behind this suggestion is that if children with weaker spatial skills struggle to visually simulate physical processes, then providing them with more experience with a physical object itself, would provide them with an experience which can be better simulated. The findings of Chapter 5, however, were not straightforward in this regard. On the one hand, children with weaker spatial skills demonstrated poorer near transfer learning outcomes, relative to children with higher spatial skills, in the two conditions in which they did not receive the concrete model, and visualisation skills were more critical. In a parallel fashion, providing a concrete model (either during the familiarisation phase, or during teaching) was successful in eliminating the significant correlation with mental folding scores, for the near transfer questions. However, there was always a correlation between mental folding scores and scores on the intermediate transfer questions; this was not affected by the presence of a concrete model. This suggests that providing a concrete model may support children with lower spatial thinking skills to reason about very related content, on which they received direct instruction. However, providing a concrete model does not overcome their underlying deficit with spatial visualisation skills, which are more critical when reasoning with novel problems.

Another, related, consideration is that, both in and out of the classroom, it is not always the case that a concrete model is available. For instance, in an exam situation, children would not be provided with a model, and in the 'real world' is not always practical, or efficient, to always to have or use a concrete model to problem solve. Taking these factors into account, although there is some benefit to providing additional experience with concrete models, the data suggests that this type of intervention should be combined with other approaches. One approach might be to provide children with practice at integrating information from multiple sources (gesture, diagrams, verbal information). Alternatively, children with lower mental folding skills might benefit from practice or training at visualisation skills. As discussed in Chapter 3, one option might be, in class, to encourage children to visualise, either before or during discussion of a process (see future work).

These implications also link to findings regarding the structure of spatial cognition, described above. On the one hand, the finding that a one-factor model best fit the data might suggest that training any of the spatial skills in this thesis would confer benefits. However, taking a different viewpoint, mental folding was consistently the strongest predictor of science achievement, throughout the thesis. The findings, discussed in detail at the beginning of the chapter, indicate that a large proportion of variance is indeed shared across the spatial thinking tasks. However, there might also be aspects of non-rigid, intrinsic-dynamic transformations that are particularly important to science learning, above and beyond this shared variance. Therefore, mental folding might be a relatively better candidate of focus for future intervention work. The findings of Chapter 2 also suggested that training mental folding skills between the age of 7 and 8 years might lead to the greatest gains, because of the significant development between these consecutive ages.

6.8 Limitations and future work

One remaining question concerns the role of any further cognitive variables, not controlled for in this thesis. The role of executive function and working memory was outlined in Chapter 1, and this is the most obvious confounding variable not directly considered within the thesis. However, when considering the role of mental folding, in Chapter 4, for example, by controlling for vocabulary, fluid reasoning, and both mental rotation and arrival judgment, a significant amount of working-memory, goal-directed type processes would have been, at least partly, controlled for. For example, there is a high correlation between fluid reasoning ability and working memory performance (e.g., Chuderski, 2015). Another related issue is that of over-controlling (see Newcombe, 2003). As outlined in Chapter 1, prior research indicates that visuospatial working memory and executive functioning predict performance on spatial thinking tasks (Miyake et al., 2001). To hold these abilities constant, given that they are important to performing these tasks, therefore, would likely lead to a significant underestimation of the contribution of spatial thinking skills to science learning. Nevertheless, the study outlined by Mayer et al. (2014) found that spatial ability did predict scientific reasoning skills, after controlling for reasoning ability, reading ability, inhibition and problem-solving/executive function. Future work might take a similar approach to Mayer et al., but, utilise structural equation modelling and path analysis to determine the nature of the relationship between these variables in predicting science learning.

Future research might also further investigate the structure of spatial cognition in childhood. As discussed, the findings of the thesis do not provide strong support for the 2 x 2 typology, but, when compared with the literature as a whole, the picture remains unclear. It was not possible within the current thesis to test the four-factor model. The prior study by Mix et al. (2018) also had only the minimum number (two) of spatial measures for each of the four spatial dimensions, and, it did include a test of the four-factor model at all, with the older children. Future research might involve administering three to four measures, for each category in the model, with a sufficiently large sample, and then test the 4-factor model. As outlined, however, this may not be practical with children, due to the number of spatial tasks they would need to complete.

Second, the role of spatial scaling warrants further investigation. Spatial scaling was only a significant spatial predictor in Chapter 3. However, the composite nature of the science assessment in Chapter 3 means its role in science learning was unclear. One possibility would be to compare the contribution of spatial scaling skills to learning about areas that do and do not depend, to the same extent, on an understanding of scale. For instance, learning about adaptation and habitats does not on the surface seem to involve such an understanding. In contrast, learning about the hierarchical organisation of cells, organs and systems in biology, does.

Third, the findings of Chapter 4 (spatial cognition as a predictor of understanding sound) could be extended to investigate additional conceptual areas; either within physics, or other areas. Mental folding was a particularly strong predictor of conceptual understanding, which was partly attributed to the mental simulation of physical processes. Because only one conceptual topic was included, however, it is not possible to conclude that this is topic specific. Indeed, mental folding was a predictor of areas other than physics, in Chapter 3. A future study might include three contrasting topics, using a similarly structured approach to the assessment of science learning to Chapter 4, to more closely determine how the effects of different spatial skills vary by science topic. Relatedly, future work could look more closely into skills aligned with scientific investigation skills. For example, research might investigate spatial thinking skills in relation to the manipulation and control of variables, within practical contexts.

Finally, as discussed in the implications section, one future area of research might be based around interventions to train intrinsic-dynamic spatial thinking skills. On the basis of their meta-analysis of 217 spatial training studies, Uttal et al. (2013) concluded that spatial thinking skills are malleable. More specifically, that the training of spatial skills results in moderately sized, durable effects which also transfer to other, untrained spatial skills (i.e. intermediate transfer). However, there have been only limited research findings to date indicating that training spatial thinking skills also improves performance in wider domains (i.e., far transfer). In the domain of mathematics, the findings of a limited number of studies suggest that training spatial thinking skills can indeed lead to improved performance in aspects of numerical cognition (Cheng & Mix, 2014), whereas others have failed to find an effect (Hawes, Moss, Caswell, & Poliszczuk, 2015).

Two studies with adults, in the domains of geoscience and engineering, are also relevant. First, Sanchez (2012) reported that spatial training using a first-person shooter game resulted in improved geoscience comprehension, compared to an interactive word game. The size of the training effect was moderate (η_p^2 = .07). Interestingly, this study included the same conceptual area of geoscience, and essay topic, as the previously reported study (Sanchez & Wiley, 2014). The results of Sanchez and Wiley (2014) indicated that multiple object dynamic spatial significantly predicted geoscience learning. Therefore, the spatial skill associated with learning about the topic was similar to the effective method of training, which also involved interacting dynamic elements. The spatial training occurred immediately before the learning/assessment of learning, without a follow-up, and therefore the durability of the training effects are unclear.

Second, a study by Sorby, Veurink, and Streiner (2018) reported the impact of a spatial skills intervention delivered to 3766 engineering students across five cohorts. First-year engineering students were required to enrol in the spatial intervention if they scored below 18 on the Purdue Spatial Visualisation Test. The intervention consisted of a term-long spatial skills course, with one class per week. A regression discontinuity analysis revealed that participants in the intervention group performed at higher levels than would have been expected if they had not participated in the intervention, particularly for engineering class grades (d = 0.22, small effect). The lack of an active control group, however, means it is not possible to rule out effects of generally increased engagement or motivation. Nevertheless, both identifying students who might benefit from an intervention, and, embedding an intervention into the existing curriculum, are approaches which could be adopted in future research.

Given the findings of the thesis, an intervention to train conceptually-related science skills in middle childhood, might be based around training mental transformation-related skills, such as mental folding. The training procedure might involve repeated practice, direct instruction in strategies, or a combination of both. In addition, the training could be embedded within the context of science learning. For example, in line with the theoretical proposals, children could be asked to generate piecemeal simulations of individual components, processes or systems linked to scientific concepts. This could be delivered in a structured way, and paired with teaching. In line with the research by Sanchez (2012), and the study in Chapter 3, children could then be asked to provide conceptual explanations. The training might also be integrated into the curriculum, across an extended period, in a manner similar to the design of Sorby et al. (2018).

6.9 Concluding remarks

The main aim of this thesis was to investigate the role of spatial cognition in children's science learning. The thesis incorporated spatial thinking skills which encompassed the broad theoretical typology presented by Uttal et al. (2013). This model had previously received little evaluation to date, and had not been applied to learning at any stage of science education. In addition, the definition of spatial cognition expanded to include the use of spatial tools, including gesture, diagrams and concrete models. Science learning referred to more general science achievement, linked to the outcomes of a curriculum-based assessment, and also to the direct outcomes of individual and whole-class science instruction. The major contribution of this work was, therefore, that it investigated the connection between spatial cognition and children's science learning in a more comprehensive manner than had been carried out previously.

The main finding of this thesis is that individual differences in children's spatial thinking skills account for variance in science learning outcomes in a diverse range of contexts. The role of some of the identified spatial skills (e.g., spatial scaling) requires more detailed investigation in the future. However, the findings across the studies do more conclusively point to a strong role for non-rigid, intrinsic-dynamic mental transformations (e.g., mental folding), especially for using and applying conceptual understanding in science. Further psychometric modelling is needed to connect these findings to theoretical models of spatial cognition. Nevertheless, these findings have the potential to inform the development of targeted interventions and training studies, aimed at children in middle childhood.

References

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, *77*(6), 1698–1716. https://doi.org/10.1111/j.1467-8624.2006.00968.x

Anastakis, D. J., Hamstra, S. J., & Matsumoto, E. D. (2000). Visual-spatial abilities in surgical training. *The American Journal of Surgery*, *179*(6), 469–471. https://doi.org/10.1016/S0002-9610(00)00397-4

Atit, K., Gagnier, K., & Shipley, T. F. (2015). Student gestures aid penetrative thinking. *Journal of Geoscience Education*, *63*(1), 66–72. https://doi.org/10.5408/14-008.1

Atit, K., Shipley, T. F., & Tikoff, B. (2013). Twisting space: Are rigid and non-rigid mental transformations separate spatial skills? *Cognitive Processing*, *14*(2), 163–173. https://doi.org/10.1007/s10339-013-0550-8

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *Psychology of Learning and Motivation* (8th ed., Vol. 8, pp. 47–89). New York, NY: Academic Press.

Baker, S. R., & Talley, L. (1972). The relationship of visualization skills to achievements in freshman chemistry. *Journal of Chemical Education*, *49*(11), 775. https://doi.org/10.1021/ed049p775

Baker, Y. (2018). STEM Skills Indicator. Retrieved November 3, 2018, from https://www.stem.org.uk/sites/default/files/pages/downloads/stem-skills-indicator-findings.pdf

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037//0033-2909.128.4.612

Barratt, E. S. (1953). An analysis of verbal reports of solving spatial problems as an aid in defining spatial factors. *Journal of Psychology: Interdisciplinary and Applied*, *36*, 17–25. https://doi.org/10.1080/00223980.1953.9712874

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(4), 637–660. https://doi.org/10.1017/S0140525X99532147

BBC. (2007). How does sound travel through air? Retrieved September 1, 2017, from https://www.bbc.com/bitesize/clips/ztwkjxs

Bean, T. W., Searles, D., Singer, H., & Cowen, S. (1990). Learning concepts from biology text through pictorial analogies and an analogical study guide. *Journal of Educational Research*, *83*(4), 233–237. https://doi.org/10.1080/00220671.1990.10885961

Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological Science*, *21*(11), 1605–1610. https://doi.org/10.1177/0956797610385353

Bergey, B. W., Cromley, J. G., Kirchgessner, M. L., & Newcombe, N. S. (2015). Using diagrams versus text for spaced restudy: Effects on learning in 10th grade biology classes. *British Journal of Educational Psychology*, *85*(1), 59–74. https://doi.org/10.1111/bjep.12062

Bigelow, G. S. (1971). Field dependence-field independence in 5-to 10-year-old children. *The Journal of Educational Research*, *64*(9), 397–400. https://doi.org/10.1080/00220671.1971.10884202

Bodner, G. M., & McMillen, T. L. B. (1986). Cognitive restructuring as an early stage in problem solving. *Journal of Research in Science Teaching*, *23*(8), 727–737. https://doi.org/10.1002/tea.3660230807

Boyles, S. (2017). Sound Waves: The Symphony of Physics. BBC.

Broadbent, H. J., Farran, E. K., & Tolmie, A. (2014). Object-based mental rotation and visual perspective-taking in typical development and Williams syndrome. *Developmental Neuropsychology*, *39*(3), 205–225. https://doi.org/10.1080/87565641.2013.876027

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Publications.

Buckley, J., Seery, N., & Canty, D. (2018). A heuristic framework of spatial ability: a review and synthesis of spatial factor literature to support its translation into STEM Education. *Educational Psychology Review*, *30*(3), 947–972. https://doi.org/10.1007/s10648-018-9432-z

Bullens, J., Iglói, K., Berthoz, A., Postma, A., & Rondi-Reig, L. (2010). Developmental time course of the acquisition of sequential egocentric and allocentric navigation strategies. *Journal of Experimental Child Psychology*, *107*(3), 337–350. https://doi.org/10.1016/j.jecp.2010.05.010

Burton, L. J., & Fogarty, G. J. (2003). The factor structure of visual imagery and spatial abilities. *Intelligence*, *31*(3), 289–318. https://doi.org/10.1016/S0160-2896(02)00139-3

Busch, J. C., Watson, J. A., Brinkley, V., Howard, J., & Nelson, C. (1993). Preschool embedded figures test performance of young children: age and gender differences. *Perceptual and Motor Skills*, *77*(2), 491–496. https://doi.org/10.2466/pms.1993.77.2.491

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2004). Action observation and acquired motor skills: an FMRI study with expert dancers. *Cerebral Cortex*, *15*(8), 1243–1249.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Castillo, R. D., Waltzer, T., & Kloos, H. (2017). Hands-on experience can lead to systematic mistakes: A study on adults' understanding of sinking objects. *Cognitive Research: Principles and Implications*, *2*(28). https://doi.org/10.1186/s41235-017-0061-8

Chatterjee, A. (2008). The neural organization of spatial thought and language. *Seminars in Speech and Language*, *29*(3), 226–238. https://doi.org/10.1055/s-0028-1082886

Cheng, Y. L., & Mix, K. S. (2014). Spatial Training Improves Children's Mathematics Ability. *Journal of Cognition and Development*, *15*(1), 2–11. https://doi.org/10.1080/15248372.2012.725186

Cherubini, P., & Johnson-Laird, P. N. (2004). Does everyone love everyone? The psychology of iterative reasoning. *Thinking and Reasoning*, *10*(1), 31–53. https://doi.org/10.1080/13546780342000025

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General*, *137*(4), 706. https://doi.org/10.1037/a0013157

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences*, *85*, 98–104. https://doi.org/10.1016/j.paid.2015.04.046

Cohen, C. A., & Hegarty, M. (2007). Individual differences in use of external visualisations to perform an internal visualisation task. *Applied Cognitive Psychology*, *21*(6), 701–711. https://doi.org/10.1002/acp.1344

Congdon, E. L., Kwon, M. K., & Levine, S. C. (2018). Learning to measure through action and gesture: Children's prior knowledge matters. *Cognition*, *180*, 182–190. https://doi.org/10.1016/j.cognition.2018.07.002

Contreras, M. J., Colom, R., Hernández, J. M., & Santacreu, J. (2003). Is static spatial performance distinguishable from dynamic spatial performance? A latent-variable analysis. *Journal of General Psychology*, *130*(3), 277–288. https://doi.org/10.1080/00221300309601159

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010

Craik, K. (1943). *The nature of explanation*. Cambridge University Press.

Crescentini, C., Fabbro, F., & Urgesi, C. (2014). Mental spatial transformations of objects and bodies: Different developmental trajectories in children from 7 to 11 years of age. *Developmental Psychology*, *50*(2), 370–383. https://doi.org/10.1037/a0033627

Cromley, J., Weisberg, S., Dai, T., Newcombe, N., Schunn, C., Massey, C., & Merlino, F. (2016). Improving middle school science learning using diagrammatic reasoning. *Science Education*, *100*(6), 1184–1213. https://doi.org/10.1002/sce.21241

D'Oliveira, T. C. (2004). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *The International Journal of Aviation Psychology*, *14*(1), 19–38. https://doi.org/10.1207/s15327108ijap1401_2

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1–2), 25–62. https://doi.org/10.1016/0010-0277(89)90005-X

Delialioğlu, Ö., & Askar, P. (1999). Contribution of students' mathematical skills and spatial ability of achievement in secondary school physics. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *16*, 34–39.

Department for Education. (2013a). *Science programmes of study: key stages 1 and 2*. London, UK.

Department for Education. (2013b). *The national curriculum in England*. London, UK.

Department for Education. (2017). *Schools, pupils and their characteristics: January 2017.* London, UK.

Diamond, A. (1990). 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*(1), 637–676. https://doi.org/10.1111/j.1749-6632.1990.tb48913.x

DiSessa, A. (1988). Knowledge in pieces. In *Constructivism in the computer age* (pp. 49–70). Columbus, Ohio: Merrill Publishing.

diSessa, A. A. (2014). The construction of causal schemes: Learning mechanisms at the knowledge level. *Cognitive Science*, *38*(5), 795–850. https://doi.org/10.1111/cogs.12131

Disessa, A. A. (1996). What do" just plain folk" know about physics. In D. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching, and schooling* (pp. 709–730). Malden, MA: Blackwell Cambridge, MA.

Disessa, A. A. (2002). Why "conceptual ecology" is a good idea. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 28–60). Dordrecht, Boston, London: Kluwer.

diSessa, A. A., Gillespie, N. M., & Esterly, J. B. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, *28*(6), 843–900. https://doi.org/10.1016/j.cogsci.2004.05.003

Donnelly, C. M., & McDaniel, M. A. (1993). Use of Analogy in Learning Scientific Concepts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 975–987. https://doi.org/10.1037/0278-7393.19.4.975

Dryden, R., Morgan, M. G., Bostrom, A., & Bruine de Bruin, W. (2018). Public Perceptions of How Long Air Pollution and Carbon Dioxide Remain in the Atmosphere. *Risk Analysis*, *38*(3), 525–534. https://doi.org/10.1111/risa.12856

Duffy, G., Sorby, S., & Bowe, B. (2016). Visualizing Electric Circuits: The Role of Spatial Visualization Skills in Electrical Engineering. In *Proceedings of the 70th midyear technical conference: graphical expressions of engineering design* (pp. 113–117).

Dunn, L. M., & Dunn, D. M. (2009). *The British picture vocabulary scale* (3rd ed.). Windsor, UK: NFER.

Ebisch, S. J., Perrucci, M. G., Mercuri, P., Romanelli, R., Mantini, D., Romani, G. L., … Saggino, A. (2012). Common and unique neuro-functional basis of induction, visualization, and spatial relationships as cognitive components of fluid intelligence. *Neuroimage*, *62*(1), 331–342. https://doi.org/10.1016/j.neuroimage.2012.04.053

Efron, D. (1941). *Gesture and environment.* Oxford, UK: King's Crown Press.

Ehrlich, K., Mani, K., & Johnson-Laird, P. N. (1979). *Mental models of spatial relations*.

Ekstrom, R. B., Dermen, D., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests* (Vol. 102). Educational Testing Service Princeton, NJ.

Elliot, J., & Smith, I. M. (1983). *An international dictionary of spatial tests*. The NFER-Nelson Publishing Company, Windsor, United Kingdom.

Estes, D. (1998). Young Children's Awareness of Their Mental Activity: The Case of Mental Rotation. *Child Development*, *69*(5), 1345–1360. https://doi.org/10.2307/1132270

Evangelou, F., & Kotsis, K. (2019). Real vs virtual physics experiments: comparison of learning outcomes among fifth grade primary school students. A case on the concept of frictional force. *International Journal of Science Education*, *41*(3), 330–348. https://doi.org/10.1080/09500693.2018.1549760

Farah, M. J., Levine, D. N., & Calvanio, R. (1988). A case study of mental imagery deficit. *Brain and Cognition*, *8*(2), 147–164. https://doi.org/10.1016/0278-2626(88)90046-2

Fennema, E., & Tartre, L. (1985). The use of spatial visualization in mathematics by girls and boys. *Journal for Research in Mathematics Education*, *16*(3), 183–206. https://doi.org/10.2307/748393

Fiorella, L., & Mayer, R. E. (2017). Spontaneous spatial strategy use in learning from scientific text. *Contemporary Educational Psychology*, *49*, 66–79. https://doi.org/10.1016/j.cedpsych.2017.01.002

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology*, *17*(1), 99–103. https://doi.org/10.1037/0012-1649.17.1.99

French, J. W. (1951). *The description of aptitude and achievement tests in terms of rotated factors*. Chicago, IL: University of Chicago Press.

Frick, A., & Baumeler, D. (2017). The relation between spatial perspective taking and inhibitory control in 6-year-old children. *Psychological Research*, *81*(4), 730–739. https://doi.org/10.1007/s00426-016-0785-y

Frick, A., Hansen, M. A., & Newcombe, N. S. (2013). Development of mental rotation in 3- to 5-year-old children. *Cognitive Development*, *28*(4). https://doi.org/10.1016/j.cogdev.2013.06.002

Frick, A., Möhring, W., & Newcombe, N. S. (2014a). Development of mental transformation abilities. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2014.05.011

Frick, A., Möhring, W., & Newcombe, N. S. (2014b). Picturing perspectives: Development of perspective-taking abilities in 4- to 8-year-olds. *Frontiers in Psychology*, *5*, 1–7. https://doi.org/10.3389/fpsyg.2014.00386

Frick, A., & Newcombe, N. S. (2012). Getting the big picture: Development of spatial scaling abilities. *Cognitive Development*, *27*(3), 270–282. https://doi.org/10.1016/j.cogdev.2012.05.004

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3–4), 455–479. https://doi.org/10.1080/02643290442000310

García-Madruga, J. A., Moreno, S., Carriedo, N., Gutiérrez, F., & Johnson-Laird, P. N. (2001). Are conjunctive inferences easier than disjunctive inferences? A comparison of rules and models. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *54*(2), 613–632. https://doi.org/10.1080/713755974

Gardner, M. F. (1996). *Test of visual-perceptual skills (n-m) - revised*. New York, NY: Psychology and Educational Publications.

Garg, A., Norman, G. R., Spero, L., & Maheshwari, P. (1999). Do virtual computer models hinder anatomy learning? *Academic Medicine*, *74*(10), 87–89. https://doi.org/10.1097/00001888-199910000-00049

Garg, A. X., Norman, G., & Sperotable, L. (2001). How medical students learn spatial anatomy. *Lancet*, *357*(9253), 363–364. https://doi.org/10.1016/S0140-6736(00)03649-7

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The Structure of Working Memory from 4 to 15 Years of Age. *Developmental Psychology*, *40*(2), 177–190. https://doi.org/10.1037/0012-1649.40.2.177

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1016/S0364-0213(83)80009-3

Gentner, D, & Gentner, D. R. (1983). Flowing waters or moving crowd: Mental models of electricity. In *Mental Models*. Hillsdale, NJ: Erlbaum.

Gentner, D, & Stevens, A. L. (1983). *Mental models*. Hillsdale, NJ: Erlbaum.

Gentner, D. (2001). Mental Models, Psychology of. In N. Smelser & P. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences*. Oxford, UK: Elsevier. https://doi.org/10.1016/B0-08-043076-7/01487-X

Gentner, Dr, & Gentner, D. (1983). Flowing waters or teeming crowds: mental models of electricity. In D Gentner & A. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Erlbaum.

Gifford, R., Kormos, C., & McIntyre, A. (2011). Behavioral dimensions of climate change: Drivers, responses, barriers, and interventions. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(6), 801–827. https://doi.org/10.1002/wcc.143

Gilligan, K. A., Hodgkiss, A., Thomas, M. S. C., & Farran, E. K. (2018). The use of discrimination scaling tasks: A novel perspective on the development of spatial scaling in children. *Cognitive Development*, *47*, 133–145. https://doi.org/10.1016/j.cogdev.2018.04.001

Glauert, E. B. (2009). How young children understand electric circuits: Prediction, explanation and exploration. *International Journal of Science Education*, *31*(8), 1025–1047. https://doi.org/10.1080/09500690802101950

Glenberg, A. M. (2010). Embodiment as a unifying perspective for psychology. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(4), 586–596. https://doi.org/https://doi.org/10.1002/wcs.55

Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: the case of gesture. *Perspectives on Psychological Science*, *5*(6), 664–674. https://doi.org/10.1177/1745691610388764

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522. https://doi.org/https://doi.org/10.1111/1467-9280.00395

Guilford, J. P., & Zimmerman, W. S. (1948). The Guilford-Zimmerman Aptitude Survey. *Journal of Applied Psychology*, *32*(1), 24–34. https://doi.org/10.1037/h0063610

Guillot, A., Champely, S., Batier, C., Thiriet, P., & Collet, C. (2007). Relationship between spatial abilities, mental rotation and functional anatomy learning. *Advances in Health Sciences Education*, *12*(4), 491–507. https://doi.org/10.1007/s10459-006-9021-7

Gunderson, E. A., Ramirez, G., Beilock, S. L., & Levine, S. C. (2012). The relation between spatial skill and early number knowledge: The role of the linear number line. *Developmental Psychology*, *48*(5), 1229–1241. https://doi.org/10.1037/a0027433

Habacha, H., Moreau, D., Jarraya, M., Lejeune-Poutrain, L., & Molinaro, C. (2018). Dissociating object-based from egocentric transformations in mental body rotation: effect of stimuli size. *Experimental Brain Research*, *236*(1), 275–284. https://doi.org/10.1007/s00221-017-5125-y

Hambrick, D. Z., Libarkin, J. C., Petcovic, H. L., Baker, K. M., Elkins, J., Callahan, C. N., … LaDue, N. D. (2012). A test of the circumvention-of-limits hypothesis in scientific problem solving: The case of geological bedrock mapping. *Journal of Experimental Psychology: General*, *141*(3), 397. https://doi.org/10.1037/a0025927

Harris, J., George, N. R., Hirsh-Pasek, K., & Newcombe, N. S. (2018). Where will it go? How children and adults reason about force and motion. *Cognitive Development*, *45*, 113–124. https://doi.org/10.1016/J.COGDEV.2018.01.002

Harris, J., Newcombe, N. S., & Hirsh-pasek, K. (2013). A new twist on studying the development of dynamic spatial transformations: mental paper folding in young children. *Mind, Brain, and Education*, *7*(1), 1–7. https://doi.org/10.1111/mbe.12007

Hawes, Z., LeFevre, J.-A., Xu, C., & Bruce, C. D. (2015). Mental rotation with tangible three-dimensional objects: A new measure sensitive to developmental differences in 4-to 8-year-old children. *Mind, Brain, and Education*, *9*(1), 10–18. https://doi.org/10.1111/mbe.12051

Hawes, Z., Moss, J., Caswell, B., Naqvi, S., & MacKinnon, S. (2017). Enhancing Children’s Spatial and Numerical Skills through a Dynamic Spatial Approach to Early Geometry Instruction: Effects of a 32-Week Intervention. *Cognition and Instruction*, *35*(3), 236–264. https://doi.org/10.1080/07370008.2017.1323902

Hawes, Z., Moss, J., Caswell, B., & Poliszczuk, D. (2015). Effects of mental rotation training on children’s spatial and mathematics performance: A randomized controlled study. *Trends in Neuroscience and Education*, *4*(3), 60–68. https://doi.org/10.1016/j.tine.2015.05.001

Haxby, J. V., Horwitz, B., Ungerleider, L. G., Maisog, J. M., Pietrini, P., & Grady, C. L. (1994). The functional organization of human extrastriate cortex: a PET-rCBF study of selective attention to faces and locations. *Journal of Neuroscience*, *14*(11), 6336–6353. https://doi.org/10.1523/JNEUROSCI.14-11-06336.1994

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900–1902. https://doi.org/10.1126/science.1128898

Hegarty, M. (2014). Spatial thinking in undergraduate science education. *Spatial Cognition & Computation*, *14*(2), 142–167. https://doi.org/10.1080/13875868.2014.889696

Hegarty, M, & Just, M. A. (1993). Construting mental models of machines from text and diagrams. *Journal of Memory and Language*, *32*, 717–742. https://doi.org/10.1006/jmla.1993.1036

Hegarty, M, & Kozhevnikov, M. (1999). Types of visual-spatial representations and mathematical problem solving. *Journal of Educational Psychology*, *91*(4), 684. https://doi.org/10.1037/0022-0663.91.4.684

Hegarty, M, Montello, D. R., Richardson, A. E., Ishikawa, T., & Lovelace, K. (2006). Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. *Intelligence*, *34*(2), 151–176. https://doi.org/10.1016/j.intell.2005.09.005

Hegarty, M, Shah, P., & Miyake, A. (2000). Constraints on using the dual-task methodology to specify the degree of central executive involvement in cognitive tasks. *Memory & Cognition*, *28*(3), 376–385. https://doi.org/10.3758/bf03198553

Hegarty, M, & Steinhoff, K. (1997). Individual differences in use of diagrams as external memory in mechanical reasoning. *Learning and Individual Differences*, *9*(1), 19–42. https://doi.org/10.1016/S1041-6080(97)90018-2

Hegarty, M, & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence*, *32*(2), 175–191. https://doi.org/10.1016/j.intell.2003.12.001

Hegarty, M, & Waller, D. (2005). Individual Differences in Spatial Abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 121–169). New York, NY: Cambridge University Press. https://doi.org/10.1017/CBO9780511610448.005

Hegarty, M. (1992). Mental Animation: Inferring Motion From Static Displays of Mechanical Systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/0278-7393.18.5.1084

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285. https://doi.org/10.1016/j.tics.2004.04.001

Hegarty, M, & Kriz, S. (2008). Effects of knowledge and spatial ability on learning from animation. In R. Lowe & W. Shnotz (Eds.), *Learning with animation: Research implications for design* (pp. 3–29). Cambridge, UK: Cambridge University Press.

Heil, C. (2018). A dissociation between two classes of spatial abilities in elementary school children. In *German Conference on Spatial Cognition* (pp. 228–243).

Herrlinger, S., Höffler, T. N., Opfermann, M., & Leutner, D. (2017). When Do Pictures Help Learning from Expository Text? Multimedia and Modality Effects in Primary Schools. *Research in Science Education*, *47*(3), 685–704. https://doi.org/10.1007/s11165-016-9525-y

Höffler, T. N. (2010). Spatial ability: Its influence on learning with visualizations—a meta-analytic review. *Educational Psychology Review*, *22*(3), 245–269. https://doi.org/10.1007/s10648-010-9126-7

Hostetter, A. B., & Alibali, M. W. (2008a). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514. https://doi.org/10.3758/PBR.15.3.495

Hostetter, A. B., & Alibali, M. W. (2008b). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review*, *15*(3), 495–514. https://doi.org/10.3758/PBR.15.3.495

Hostetter, A. B., & Alibali, M. W. (2010). Language, gesture, action! A test of the Gesture as Simulated Action framework. *Journal of Memory and Language*, *63*(2), 245–257. https://doi.org/10.1016/j.jml.2010.04.003

Hunt, E., Pellegrino, J. W., Frick, R. W., Farr, S. A., & Alderton, D. (1988). The ability to reason about movement in the visual field. *Intelligence*, *12*(1), 77–100. https://doi.org/10.1016/0160-2896(88)90024-4

Ioannides, C., & Vosniadou, S. (2002). The changing meanings of force. *Cognitive Science Quarterly*. https://doi.org/10.1017/S0008197302001630

Isaak, M. I., & Just, M. A. (1995). Constraints on the processing of rolling motion: The curtate cycloid illusion. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(6), 1391. https://doi.org/10.1037/0096-1523.21.6.1391

Jaeger, A. J., Taylor, A. R., & Wiley, J. (2016). When, and for whom, analogies help: The role of spatial skills and interleaved presentation. *Journal of Educational Psychology*, *108*(8), 1121–1139. https://doi.org/10.1037/edu0000121

Jansen, P., Schmelter, A., Quaiser-Pohl, C., Neuburger, S., & Heil, M. (2013). Mental rotation performance in primary school age children: Are there gender differences in chronometric tests? *Cognitive Development*, *28*(1), 51–62. https://doi.org/10.1016/j.cogdev.2012.08.005

Jarvis, H. L., & Gathercole, S. E. (2003). Verbal and non-verbal working memory and achievements on national curriculum tests at 11 and 14 years of age. *Educational and Child Psychology*, *20*(3), 123–140.

Joh, A. S., Jaswal, V. K., & Keen, R. (2011). Imagining a way out of the gravity bias: Preschoolers can visualize the solution to a spatial problem. *Child Development*, *82*(3), 744–750. https://doi.org/10.1111/j.1467-8624.2011.01584.x

Johnson-Laird, P.N. (1980). Mental models in cognitive science. *Cognitive Science*, *4*(1), 71–115. https://doi.org/10.1207/s15516709cog0401_4

Johnson-Laird, P N, & Khemlani, S. (2017). Mental models and causation. In M. Waldmann (Ed.), *Oxford Handbook of Causal Reasoning* (pp. 1–42). Oxford, UK: Oxford University Press.

Johnson-Laird, Philip N, Girotto, V., & Legrenzi, P. (1998). Mental models: a gentle guide for outsiders. *Sistemi Intelligenti*, *9*(68), 33.

Kali, Y., & Orion, N. (1996). Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching*, *33*(4), 369–391. https://doi.org/10.1002/(SICI)1098-2736(199604)33:4<369::AID-TEA2>3.0.CO;2-Q

Keehner, M., Guerin, S. A., Miller, M. B., Turk, D. J., & Hegarty, M. (2006). Modulation of neural activity by angle of rotation during imagined spatial transformations. *NeuroImage*, *33*(1), 391–398. https://doi.org/10.1016/j.neuroimage.2006.06.043

Keig, P. F., & Rubba, P. A. (1993). Translation of representations of the structure of matter and its relationship to reasoning, gender, spatial reasoning, and specific prior knowledge. *Journal of Research in Science Teaching*, *30*(8), 883–903. https://doi.org/10.1002/tea.3660300807

Kessler, K., & Thomson, L. A. (2010). The embodied nature of spatial perspective taking: embodied transformation versus sensorimotor interference. *Cognition*, *114*(1), 72–88. https://doi.org/10.1016/j.cognition.2009.08.015

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*(3), 245–266. https://doi.org/10.1037/rev0000059

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freska & C. Habel (Eds.), *Spatial cognition. An interdisciplinary approach to representing and processing spatial knowledge* (pp. 1–17). Springer.

Klein, J., & Koroghlanian, C. (2004). The effect of audio and animation in multimedia instruction. *Journal of Educational Multimedia and Hypermedia*, *13*(1), 23–46.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. New York, NY: Guilford publications.

Kontra, C., Lyons, D. J., Fischer, S. M., & Beilock, S. L. (2015). Physical experience enhances science learning. *Psychological Science*, *26*(6), 737–749. https://doi.org/10.1177/0956797615569355

Kosslyn, S M. (1994). *Image and Brain: The Resolution of the Imagery Debate*. MIT Press, Cambridge (MA). https://doi.org/10.1162/jocn.1995.7.3.415

Kosslyn, Stephen M., & Pomerantz, J. R. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology*, *9*(1), 52–76. https://doi.org/10.1016/0010-0285(77)90004-4

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory and Cognition*, *29*(5), 745–756. https://doi.org/10.3758/BF03200477

Kozhevnikov, M., Hegarty, M., & Mayer, R. (2002). Spatial abilities in problem solving in kinematics. In M. Anderson, B. Meyer, & P. Oliver (Eds.), *Diagrammatic representation and reasoning* (pp. 155–171). London: Springer.

Kozhevnikov, M., Kosslyn, S., & Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory and Cognition*, *33*(4), 710–726. https://doi.org/10.3758/BF03195337

Kozhevnikov, M., & Thornton, R. (2006). Real-time data display, spatial visualization ability, and learning force and motion concepts. *Journal of Science Education and Technology*, *15*(1), 111. https://doi.org/10.1007/s10956-006-0361-0

Lambrey, S., Doeller, C., Berthoz, A., & Burgess, N. (2012). Imagining being somewhere else: Neural basis of changing perspective in space. *Cerebral Cortex*, *22*(1), 166–174. https://doi.org/10.1093/cercor/bhr101

Larson, G. E. (1996). Mental rotation of static and dynamic figures. *Perception and Psychophysics*, *58*(1), 153–159. https://doi.org/10.3758/BF03205484

Leopold, C., & Mayer, R. E. (2015). An imagination effect in learning from scientific text. *Journal of Educational Psychology*, *107*(1), 47–63. https://doi.org/10.1037/a0037142

Levine, S. C., Goldin-Meadow, S., Carlson, M. T., & Hemani-Lopez, N. (2018). Mental Transformation Skill in Young Children: The Role of Concrete and Abstract Motor Training. *Cognitive Science*, *42*(4), 1207–1228. https://doi.org/10.1111/cogs.12603

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child Development*, *56*(6), 1479–1498. https://doi.org/10.1111/j.1467-8624.1985.tb00213.x

Logie, R. H. (1995). *Essays in cognitive psychology. Visuo-spatial working memory.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Lohman, D. F. (1988). Spatial abilities as traits, processes, and knowledge. In R. Sternberg (Ed.) (pp. 181–248). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lohman, D. F. (1993). Spatial Ability and G. In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and assessment* (pp. 97–116). Hillsdale, NJ: Erlbaum.

Lubinski, D., & Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science*, *1*(4), 316–345. https://doi.org/10.1111/j.1745-6916.2006.00019.x

Lufler, R. S., Zumwalt, A. C., Romney, C. A., & Hoagland, T. M. (2012). Effect of visual--spatial ability on medical students' performance in a gross anatomy course. *Anatomical Sciences Education*, *5*(1), 3–9. https://doi.org/10.1002/ase.264

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130. https://doi.org/10.1037/1082-989X.1.2.130

Marmor, G. S. (1975). Development of kinetic images: When does the child first represent movement in mental images? *Cognitive Psychology*, *7*(4), 548–559. https://doi.org/10.1016/0010-0285(75)90022-5

Mayer, D., Sodian, B., Koerber, S., & Schwippert, K. (2014). Scientific reasoning in elementary school children: Assessment and relations with cognitive abilities. *Learning and Instruction*, *29*, 43–55. https://doi.org/10.1016/j.learninstruc.2013.07.005

Mayer, R. E. (2002a). Rote versus meaningful learning. *Theory into Practice*, *41*(4), 226–232. https://doi.org/10.1207/s15430421tip4104_4

Mayer, R. E. (2002b). Understanding conceptual change: a commentary. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 101–111). Dordrecht, Boston, London: Kluwer. https://doi.org/10.1007/0-306-47637-1_5

Mazens, K., & Lautrey, J. (2003). Conceptual change in physics: children's naive representations of sound. *Cognitive Development*, *18*(2), 159–176. https://doi.org/10.1016/S0885-2014(03)00018-2

McDaniel, M. A., & Donnelly, C. M. (1996). Learning with Analogy and Elaborative Interrogation. *Journal of Educational Psychology*, *88*(3), 508–519. https://doi.org/10.1037/0022-0663.88.3.508

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago press.

Messer, D. J., Pine, K. J., & Butler, C. (2008). Children's behaviour and cognitions across different balance tasks. *Learning and Instruction*, *18*(1), 42–53. https://doi.org/10.1016/j.learninstruc.2006.09.008

Meyer, K., & Damasio, A. (2009). Convergence and divergence in a neural architecture for recognition and memory. *Trends in Neurosciences*, *32*(7), 376–382. https://doi.org/10.1016/j.tins.2009.04.002

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception and Psychophysics*, *68*(2), 327–337. https://doi.org/10.3758/BF03193680

Mix, K., Levine, S., Cheng, Y.-L., Young, C., Hambrick, D., Ping, R., & Konstantopoulos, S. (2016). Separate But Correlated: The Latent Structure of Space and Mathematics Across Development. *Journal of Experimental Psychology: General*, *145*(9), 1206–1227. https://doi.org/10.1037/xge0000182

Mix, K. S., Hambrick, D. Z., Satyam, V. R., Burgoyne, A. P., & Levine, S. C. (2018). The latent structure of spatial skill: A test of the 2x2 typology. *Cognition*, *180*, 268–278. https://doi.org/10.1016/j.cognition.2018.07.012

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621–640. https://doi.org/10.1037//0096-3445.130.4.621

Moore, D. S., & Johnson, S. P. (2008). Mental rotation in human infants: A sex difference. *Psychological Science*, *19*(11), 1063–1066. https://doi.org/10.1111/j.1467-9280.2008.02200.x

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*. https://doi.org/10.1207/S15328007SEM0904_8

Narayanan, N. H., & Hegarty, M. (2002). Multimedia design for communication of dynamic information. *International Journal of Human-Computer Studies*, *57*(4), 279–315. https://doi.org/10.1006/ijhc.1019

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

Neuburger, S., Jansen, P., Heil, M., & Quaiser-Pohl, C. (2011). Gender differences in pre-adolescents' mental-rotation performance: Do they depend on grade and stimulus type? *Personality and Individual Differences*, *50*(8), 1238–1242. https://doi.org/10.1016/j.paid.2011.02.017

Newcombe, N. S. (2003). Some controls control too much. *Child Development*, *74*(4), 1050–1052. https://doi.org/doi.org/10.1111/1467-8624.00588

Newcombe, N. S. (2010). Picture this. *American Educator*, *34*(2), 29–43. https://doi.org/10.1037/A0016127.

Newcombe, N. S. (2016). Thinking spatially in the science classroom. *Current Opinion in Behavioral Sciences*, *10*, 1–6. https://doi.org/10.1016/j.cobeha.2016.04.010

Newcombe, N. S. (2018). Three Kinds of Spatial Cognition. In *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (pp. 1–31). https://doi.org/10.1002/9781119170174.epcn315

Newcombe, N. S., & Shipley, T. F. (2015). Thinking About Spatial Thinking: New Typology, New Assessments. In J. Gero (Ed.), *Studying Visual and Spatial Reasoning for Design Creativity* (pp. 179–192). New York, NY: Springer. https://doi.org/10.1007/978-94-017-9297-4_10

Newcombe, N. S., Uttal, D. H., & Sauter, M. (2013). Spatial development. In P. D. Zelazo (Ed.), *Oxford handbook of developmental psychology* (Vol. 1, pp. 564–590). Oxford, UK: Oxford University Press.

Nilsson Jobs, E., Falck-Ytter, T., & Bölte, S. (2018). Local and Global Visual Processing in 3-Year-Olds With and Without Autism. *Journal of Autism and Developmental Disorders*, *48*(6), 2249–2257. https://doi.org/10.1007/s10803-018-3470-8

Novack, M. A., Congdon, E. L., Hemani-Lopez, N., & Goldin-Meadow, S. (2014). From Action to Abstraction: using the hands to learn math. *Psychological Science*, *25*(4), 903–910. https://doi.org/10.1177/0956797613518351

Palmquist, C. M., Keen, R., & Jaswal, V. K. (2017). Visualization instructions enhance preschoolers' spatial problem-solving. *British Journal of Developmental Psychology*, 1–10. https://doi.org/10.1111/bjdp.12205

Parsons, L. M., Fox, P. T., Downs, J. H., Glass, T., Hirsch, T. B., Martin, C. C., … Lancaster, J. L. (1995). Use of implicit motor imagery for visual shape discrimination as revealed by PET. *Nature*, *375*, 54–58. https://doi.org/10.1038/375054a0

Pasta, D. J. (2009). Learning when to be discrete: continuous vs. categorical predictors. In *Proceedings of the SAS Global Forum* (Vol. 248).

Perry, M., Breckinridge Church, R., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development*, *3*(4), 359–400. https://doi.org/10.1016/0885-2014(88)90021-4

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn vandenberg and kuse mental rotations test - different versions and factors that affect performance. *Brain and Cognition*, *28*(1), 39–58. https://doi.org/10.1006/brcg.1995.1032

Piaget, J. (1971). *Les Explications causales*. Paris, France: Presses universitaires de France.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. London: Routledge & Kegan-Paul.

Pillow, B. H., & Flavell, J. H. (1986). Young children's knowledge about visual perception: projective size and shape. *Child Development*, *57*(1), 125–135. https://doi.org/10.1111/j.1467-8624.1986.tb00013.x

Pine, K. J., Lufkin, N., & Messer, D. (2004). More gestures than answers: children learning about balance. *Developmental Psychology*, *40*(6), 1059–1067. https://doi.org/10.1037/0012-1649.40.6.1059

Pribyl, J. R., & Bodner, G. M. (1987). Spatial ability and its role in organic chemistry: A study of four organic courses. *Journal of Research in Science Teaching*, *24*(3), 229–240. https://doi.org/10.1002/tea.3660240304

PricewaterhouseCoopers. (2017). *The economic contribution of the UK Life Sciences industry*. London, UK.

Pruden, S. M., Levine, S. C., & Huttenlocher, J. (2011). Children's spatial thinking: Does talk about the spatial world matter? *Developmental Science*, *14*(6), 1417–1430. https://doi.org/10.1111/j.1467-7687.2011.01088.x

Qualifications and Curriculum Authority (QCA). (1996). *Key Stage 2 Science Test*. Carrickfergus, UK.

Qualifications and Curriculum Authority (QCA). (1999). *Key Stage 2 Science Test*. Carrickfergus, UK.

Qualifications and Curriculum Authority (QCA). (2000). *Key Stage 2 Science Test*. Carrickfergus, UK.

Qualifications and Curriculum Authority (QCA). (2001). *Key Stage 2 Science Test*. Carrickfergus, UK.

Qualifications and Curriculum Authority (QCA). (2002). *Key Stage 2 Science Test*. Carrickfergus, UK.

Qualifications and Curriculum Authority (QCA). (2009). *Key Stage 2 Science Test*. Carrickfergus, UK.

Rapp, D. D. (2005). Mental Models: Theoretical Issues For Visualizations in Science Education. In J. Gilbert (Ed.), *Visualization in Science Education* (pp. 43–60). Dordecht, Netherlands: Springer. https://doi.org/10.1017/CBO9781107415324.004

Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology*, *91*(2), 137–157. https://doi.org/10.1016/j.jecp.2005.01.004

Rhodes, S. M., Booth, J. N., Campbell, L. E., Blythe, R. A., Wheate, N. J., & Delibegovic, M. (2014). Evidence for a role of executive functions in learning biology. *Infant and Child Development*, *23*(1), 67–83. https://doi.org/10.1002/icd.1823

Rhodes, S. M., Booth, J. N., Palmer, L. E., Blythe, R. A., Delibegovic, M., & Wheate, N. J. (2016). Executive functions predict conceptual learning of science. *British Journal of Developmental Psychology*, *34*(2), 261–275. https://doi.org/10.1111/bjdp.12129

Roberts, R. J., & Aman, C. J. (1993). Developmental differences in giving directions: spatial frames of reference and mental rotation. *Child Development*, *64*(4), 1258–1270. https://doi.org/10.1111/j.1467-8624.1993.tb04199.x

Rochford, K. (1985). Spatial learning disabilities and underachievement among university anatomy students. *Medical Education*, *19*(1), 13–26. https://doi.org/10.1111/j.1365-2923.1985.tb01134.x

Sanchez, C. A. (2012). Enhancing visuospatial performance through video game training to increase learning in visuospatial science domains. *Psychonomic Bulletin & Review*, *19*(1), 58–65. https://doi.org/10.3758/s13423-011-0177-7

Sanchez, C. A., & Wiley, J. (2014). The role of dynamic spatial ability in geoscience text comprehension. *Learning and Instruction*, *31*, 33–45. https://doi.org/10.1016/j.learninstruc.2013.12.007

Sattler, J. M. (1992). Assessment of children. In *San Diego, CA: Author*. San Diego, CA: Jerome M. Sattler Publishers.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, *8*(2), 23–74.

Schultz, K. (1991). The contribution of solution strategy to spatial performance. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(4), 474.

Schwartz, D. L., & Black, J. B. (1996). Analog imagery in mental model reasoning: depictive models. *Cognitive Psychology*, *30*, 154–219. https://doi.org/10.1006/cogp.1996.0006

Schwartz, D. L., & Bransford, J. D. (1998). A Time For Telling. *Cognition and Instruction*, *16*(4), 475–5223. https://doi.org/10.1207/s1532690xci1604_4

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

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Shortage Occupation List. (2018). Retrieved November 3, 2018, from https://www.gov.uk/guidance/immigration-rules/immigration-rules-appendix-k-shortage-occupation-list

Sibieta, L. (2018). *The teacher labour market in England*.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520. https://doi.org/10.1016/0010-0285(76)90016-5

So, W. C., Ching, T. H. W., Lim, P. E., Cheng, X., & Ip, K. Y. (2014). Producing gestures facilitates route learning. *PLoS ONE*, *9*(11), 1–21. https://doi.org/10.1371/journal.pone.0112543

Sodian, B., Thoermer, C., & Metz, U. (2007). Now I see it but you don't: 14-month-olds can represent another person's visual perspective. *Developmental Science*. https://doi.org/10.1111/j.1467-7687.2007.00580.x

Sorby, S., Veurink, N., & Streiner, S. (2018). Does spatial skills instruction improve STEM outcomes? The answer is 'yes.' *Learning and Individual Differences*, *67*, 209–222. https://doi.org/10.1016/j.lindif.2018.09.001

Spelke, E. (1991). Physical knowledge in infancy : reflections on Piaget's theory. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition*. Hillsdale, NJ: Erlbaum.

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Stieff, M., Lira, M. E., & Scopelitis, S. A. (2016). Gesture supports spatial thinking in STEM. *Cognition and Instruction*, *34*(2), 80–99. https://doi.org/10.1080/07370008.2016.1145122

Stull, A. T., Hegarty, M., Dixon, B., & Stieff, M. (2012). Representational translation with concrete models in organic chemistry. *Cognition and Instruction*, *30*(4), 404–434. https://doi.org/10.1080/07370008.2012.719956

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., … Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*(2), 273–281. https://doi.org/10.1162/0898929053124965

Thurstone, L. L. (1948). *Primary mental abilities. Science (New York, N.Y.)*. Chicago, IL: University of Chicago Press. https://doi.org/10.1080/00221309.1971.9711309

Titze, C., Jansen, P., & Heil, M. (2010). Mental rotation performance and the effect of gender in fourth graders and adults. *European Journal of Developmental Psychology*, *7*(4), 432–444. https://doi.org/10.1080/17405620802548214

Tolmie, A. K., Ghazali, Z., & Morris, S. (2016). Children's science learning: A core skills approach. *The British Journal of Educational Psychology*, *86*(3), 481–497. https://doi.org/10.1111/bjep.12119

Tracy, D. M. (1990). Toy-playing behavior, sex-role orientation, spatial ability, and science achievement. *Journal of Research in Science Teaching*, *27*(7), 637–649. https://doi.org/10.1002/tea.3660270704

Triona, L. M., & Klahr, D. (2003). Point and Click or Grab and Heft: Comparing the Influence of Physical and Virtual Instructional Materials on Elementary School Students' Ability to Design Experiments. *Cognition and Instruction*, *21*(2), 149–173. https://doi.org/10.1207/S1532690XCI2102_02

Trofatter, C., Kontra, C., Beilock, S., & Goldin-Meadow, S. (2015). Gesturing has a larger impact on problem-solving than action, even when action is accompanied by words. *Language, Cognition and Neuroscience*, *30*(3), 251–260. https://doi.org/10.1080/23273798.2014.905692

Uttal, D. H., & Cohen, C. A. (2012). Spatial thinking and STEM education. When, why, and how? *Psychology of Learning and Motivation - Advances in Research and Theory*, *57*, 147–181. https://doi.org/10.1016/B978-0-12-394293-7.00004-2

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. https://doi.org/10.1037/a0028446

Vander Heyden, K. M., Huizinga, M., Kan, K.-J., & Jolles, J. (2016). A developmental perspective on spatial reasoning: Dissociating object transformation from viewer transformation ability. *Cognitive Development*, *38*, 63–74. https://doi.org/10.1016/J.COGDEV.2016.01.004

Vasilyeva, M., & Huttenlocher, J. (2004). Early development of scaling ability. *Developmental Psychology*, *40*(5), 682–690. https://doi.org/10.1037/0012-1649.40.5.682

Vosniadou, S. (1994). Capturing and modelling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69. https://doi.org/10.1016/0959-4752(94)90018-3

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*(4), 535–585. https://doi.org/10.1016/0010-0285(92)90018-W

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, *101*(4), 817–835. https://doi.org/10.1037/a0016127

Wanzel, K. R., Hamstra, S. J., Anastakis, D. J., Matsumoto, E. D., & Cusimano, M. D. (2002). Effect of visual-spatial ability on learning of spatially-complex surgical skills. *The Lancet*, *359*(9302), 230–231. https://doi.org/10.1016/S0140-6736(02)07441-X

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: The Psychological Corporation.

Wechsler, David. (2003). *Wechsler Intelligence Scale for Children-WISC-IV*. New York, NY: Psychological Corporation.

Wellington, J. J. (1988). The place of process in physics education. *Physics Education*, *23*(3), 150–155. https://doi.org/10.1088/0031-9120/23/3/304

Witkin, H. A., Otman, P. K., Raskin, E., & Karp, S. (1971). *A Manual for the Embedded Figures Test*. California: Consulting Psychologists Press.

Wraga, M., Creem, S. H., & Proffitt, D. R. (2000). Updating displays after imagined object and viewer rotations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 151. https://doi.org/10.1037/0278-7393.26.1.151

Wraga, M., Shephard, J. M., Church, J. A., Inati, S., & Kosslyn, S. M. (2005). Imagined rotations of self versus objects: an fMRI study. *Neuropsychologia*, *43*(9), 1351–1361. https://doi.org/10.1016/j.neuropsychologia.2004.11.028

Xistouri, X., & Pitta-Pantazi, D. (2006). Spatial rotation and perspective taking abilities in relation to performance in reflective symmetry tasks. In *30th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 5, pp. 425–432).

Zacks, J. M., & Michelon, P. (2005). Transformations of visuospatial images. *Behavioral and Cognitive Neuroscience Reviews*, *4*(2), 96–118. https://doi.org/10.1177/1534582305281085

Zacks, J. M., Vettel, J. M., & Michelon, P. (2003). Imagined viewer and object rotations dissociated with event-related fMRI. *Journal of Cognitive Neuroscience*, *15*(7), 1002–1018. https://doi.org/10.1162/089892903770007399

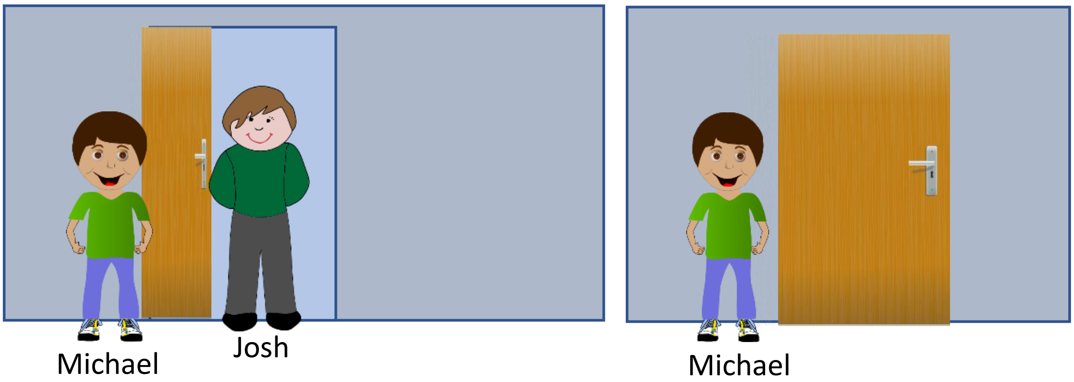
Zelazo, P. D., & Carlson, S. M. (2012). Hot and cool executive function in childhood and adolescence: Development and plasticity. *Child Development Perspectives*, *6*(4), 354–360. https://doi.org/10.1111/j.1750-8606.2012.00246.x

Zimmerman, C. (2000). The development of scientific reasoning skills. *Developmental Review*, *20*(1), 99–149. https://doi.org/10.1006/drev.1999.0497

Appendices

8.1 Example problem-solving question, as presented to children (Chapter 4)

1. Josh stood in his classroom next to the door. Michael then closed the door. Josh then shouted to Michael on the other side of the door.



a. Imagine you were Michael. **Predict** what happened when Josh shouted with the door **closed**.

____

b. **Explain** as fully as you can **how** this happened, starting from the moment Josh shouts.

Josh shouts

8.2 Near transfer questions (magnetism assessment; Chapter 5)

Problem 1) (part [a] source (Qualifications and Curriculum Authority [QCA], 2001)

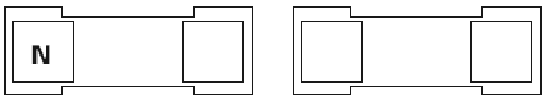
Sam has two magnets. He holds them in different ways. His observations are recorded below.

a)

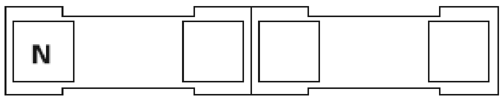
Write **N** (North) or **S** (South) on each end of each magnet below to explain Sam's observations. Some have been done for you.

Sam's observations

The ends of these magnets push away from each other.



The ends of these magnets pull together.

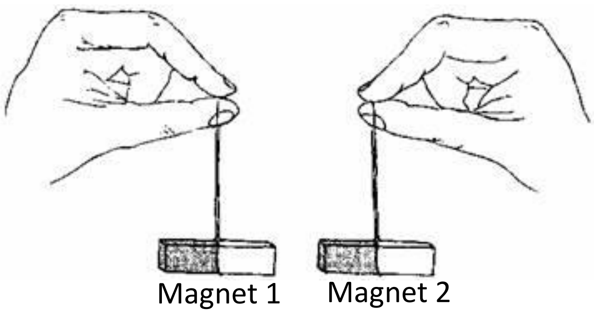


b) Why do the first two magnets push away from each other?

Problem 2) (part [a] source (Qualifications and Curriculum Authority [QCA], 2002)

Exploring magnets

(a) Wayne has two bar magnets. He hangs each bar magnet from a piece of string. He holds them close together like this:



What happens to the magnets when he holds them like this?

....

b) Why does this happen?

c) Draw an arrow on each magnet to show the direction of the magnetic forces

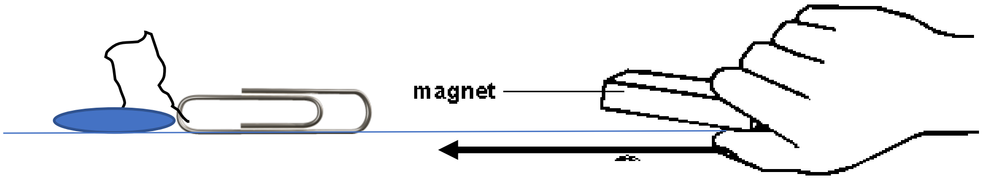
d) Wayne could change the way the magnets react to one another by moving magnet 1 in **two different** ways. What two ways could he move them?

He could…

He could also…

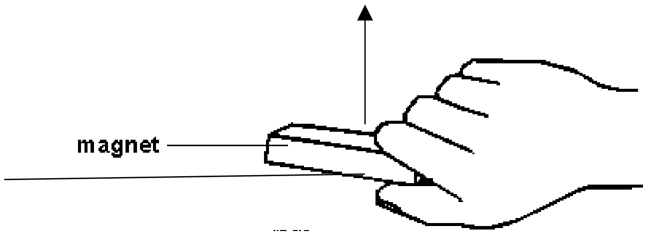
Problem 3) (hand/magnet image and question scenario adapted from (Qualifications and Curriculum Authority [QCA], 2000)

Azeem has an iron paperclip attached to a piece of string, and the piece of string is attached to the table with blu-tac. Azeem holds a magnet to the paperclip on the table, like this:



a) Predict what you would expect to happen when he moves the magnet in the direction of the arrow

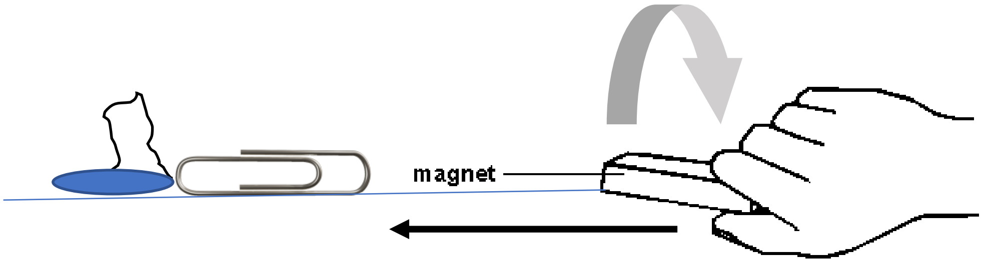
b) Explain why you think this



c) He then lifts the magnet **slightly upwards away** from the table, from its **new position.**

d) Predict what you would expect to happen to the paper clip

e) Azeem then holds the magnet again in the same position as in part (a).



This time, he turns the magnet over so the **other end** is now facing the paper clip and moves the magnet again in the direction of the arrow again.

Do you predict that the paper clip will react in the **same way** as in question (a) or a **different way** to question (a)?

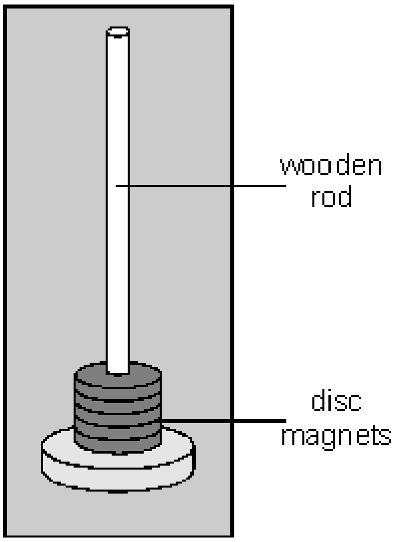
☐ It would react in the same way ☐ It would react in a different way

If you chose the **same way**, explain why you think this. If you think it would react in a **different way**, explain what you think would be different and explain why.

8.3 Intermediate transfer questions (magnetism assessment; Chapter 5)

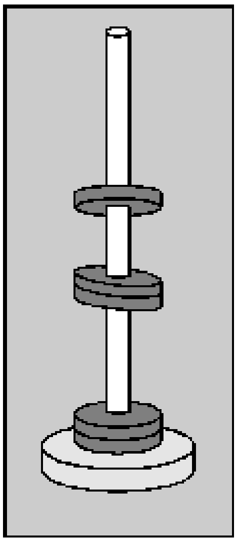
Problem 1) (images on this page, and part [a], and question scenario adapted from (Qualifications and Curriculum Authority [QCA], 1999)

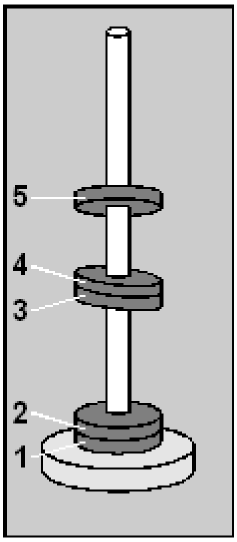
Wayne has some disc magnets on a wooden rod.



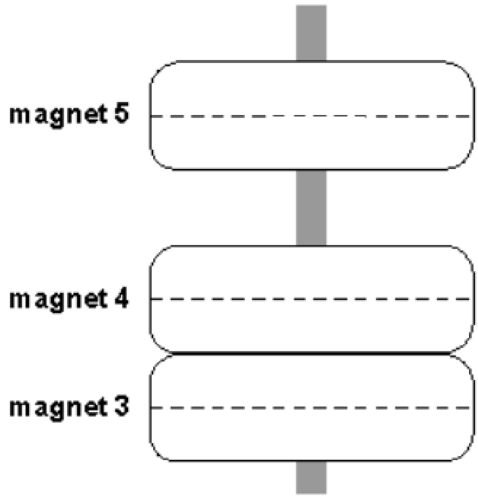
He takes the magnets off the rod. He turns some of them over, then puts all the magnets back on the rod.

Some of the magnets float.





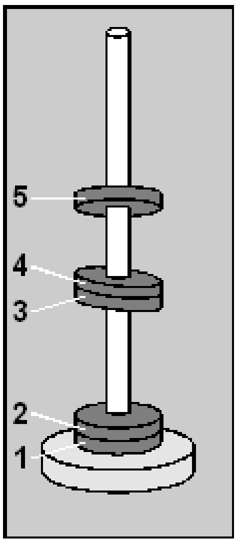
a) On the diagram below, label 'N' and 'S' to show how the arrangement of the poles of the magnets might be, now, for magnets 3, 4 and 5.

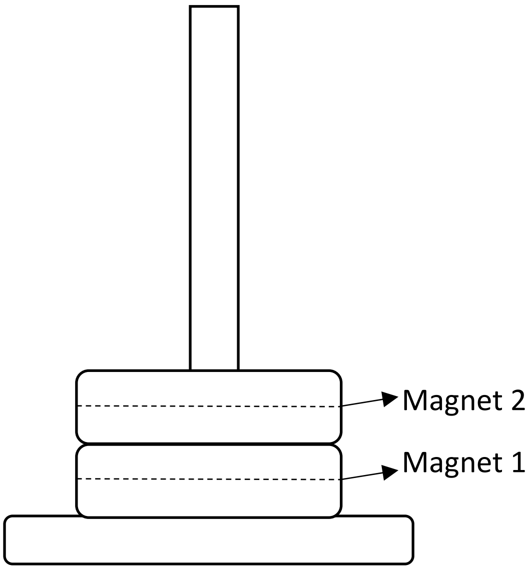


b) If Wayne wanted magnets 3,4 and 5 **to all float together above magnet 1 and 2,** what would he need to do?

c) Draw all the magnets on the diagram of the rod using the shape of the magnets shown (as above) if **magnets 3, 4 and 5 were all floating together above magnet 1 and 2**.

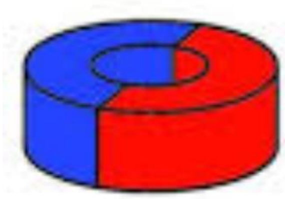
d) Label the poles of the magnets you have drawn.





Problem 2)

Josh also has some disc magnets, but, they look like this:



a) It is possible for Josh to arrange the magnets on the rod, on top of each other, in the same way as Wayne's magnets did in part (d), so that magnets 3,4 and 5 float together above magnets 1 and 2?

Yes No

b) If **yes**, explain how. If no, **explain why not**:

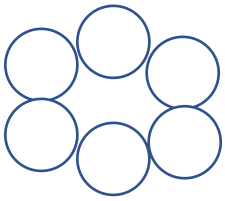
c) How else could this type of magnet be arranged so they join in a **different direction** and **without** the rod, so they were not stacked on top of each other?

You may draw your answer:

Problem 3)

Shelly has some ball-shaped magnets. She makes a bracelet using the ball-shaped magnets as shown in the picture below.

The diagram shows a simplified representation of the bracelet, with fewer balls.



a) Draw lines on each ball to separate each of them into two poles

b) Label N and S on each ball

The balls do not all need to be magnets to make the bracelet.

c) How else could she make the bracelet so that it is still held together by magnetism?

d) Draw and label this diagram to go with your explanation: