The Role of Executive Function, Metacognition, and Support Type in Children's Ability to Solve Physics Tasks

University of Cambridge

Elaine Gray

Darwin College

Supervisor: Dr Sara Baker

This dissertation is submitted for the degree of *Doctor of Philosophy*

June 2018

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee (80,000).

The Role of Executive Function, Metacognition, and Support Type in Children's Ability to Solve Physics Tasks

Abstract

Some research has suggested that guided play (GP) is a better support type for children to learn, but other research has suggested direct instruction (DI) is better for teaching children physics. These research fields formed the basis of this study, in addition to also considering the role of executive function (EF) and metacognition (Mc) due to their potential links to physics task performance.

This research was carried out with 38 3- and 4-year-olds over three time points (TP), six weeks apart. Children completed the same EF, Mc, and physics task at each time point, as well as a transfer ramps physics task at TP3. Children were split into one of two support type two groups to carry out the balance beam tasks: GP or DI.

No significant links between EF and Mc were detected, and no role of EF or Mc in physics task performance was seen. A small association between Mc rate and strategies used was seen at one TP only. A significant difference in Mc behaviours displayed by each group during the balance beam task was found at each TP, due to GP scoring significantly higher than DI, but no significant difference in Mc interview scores was found. No significant link between GP's higher Mc rate scores and other measures was detected.

There was a significant difference in balance beam performance between the groups at TP3, due to DI scoring significantly higher than GP. The results from the balance beam task did not significantly correlate with the transfer ramps task, suggesting support type did not have a strong transferable effect to another physics task. It was found that vocabulary was associated with EF and Mc interview scores, suggesting language was an important individual factor.

The study has highlighted that young children's learning of balance beam concepts is complex, with individuals showing a variety of strategies to solve different balance beam problems. It provides support that DI could be a better support type for teaching children balance beam concepts. The data are discussed with reference to different theories and to the issues surrounding the small sample size and low statistical power, which are potentially impacting the conclusions that can be drawn.

Acknowledgements

I am very grateful to the LEGO Foundation and the Faculty of Education for funding my PhD.

I wish to thank Dr Sara Baker and Professor Christine Howe for their advice, guidance, and feedback during this work.

Thanks to all the children and nurseries who participated.

My biggest thanks go to all of my friends who have supported me over the last few years.

Abbreviations

DI	Direct instruction
EF	Executive function
FIST	Flexible item selection task
GP	Guided play
GR	Graded representations
OW	Overlapping waves
Mc	Metacognition
MK	Metacognitive knowledge
MR	Metacognitive regulation
RR	Representational redescription
SD	Standard deviation
TP	Time point
WM	Working memory

Contents

Declaration iii
Abstract v
Acknowledgements vii
Abbreviations ix
List of Figures xiv
List of Tables xv
1 Chapter 1 Introduction 1
2 Chapter 2 Literature Review 3
2.1 Physics 3
2.1.1 Children's understanding of physics 3
2.1.2 Adults' understanding of physics 20
2.2 Executive function 22
2.2.1 Function and structure of EF 22
2.2.2 EF and reasoning 25
2.2.3 EF and physics tasks 27
2.3 Metacognition 31
2.3.1 Structure and function of Mc 31
2.3.2 Strategy use 33
2.3.3 Mc and physics tasks 36
2.4 How are EF and Mc related 37
2.5 Vocabulary and visual-spatial skills 39
2.6 Support type 42
2.7 Theoretical accounts and predictions 46
2.7.1 Diamond (2013) 47
2.7.2 Halford et al. (2002) and the relational complexity theory 48
2.7.3 Karmiloff-Smith's RR model (1992) 48
2.7.4 Siegler (1976) and the OW theory (Siegler, 1996) 50
2.7.5 Munakata's (2001) GR account 51
2.7.6 Schapiro and McClelland's (2009) connectionist model 51
2.8 Conclusions and aims of the study 52
2.9 Research questions and hypotheses 53
3 Chapter 3 Methodology 58
3.1 Research design 58
3.2 Measures 59
3.2.1 Pilot Study 1 59
3.2.2 Pilot study 2 78
3.2.3 Pilot study 3 89
4 Chapter 4 Main study 91
4.1 Main study participants 91
4.2 Main study participant groups 91
4.3 Main study procedure 93
4.4 Background measures 96
4.5 Balance beam task 96
4.6 Support type 97
4.6.1 Instruction and information before the balance beam trials 97
4.6.2 Instruction given to both GP and DI 98

4.6.3 Instruction given to GP 99
4.6.4 Instruction given to DI 99
4.6.5 Comparison of length of instruction in the two groups 100
4.6.6 Comparison of how much information was provided by the adult 101
4.6.7 Comparison of how much information each group had before starting the trials 101
4.6.8 Feedback after the balance beam trials 102
4.7 Ramps task 102
4.8 EF measures 105
4.9 Mc measures 106
5 Chapter 5 Methods for data analyses 107
5.1 Mc coding scheme and inter-rater reliability 107
5.1.1 Mc coding scheme 107
5.1.2 Inter-rater reliability 109
5.2 Validity and reliability throughout study 112
5.3 Statistical tests employed to answer the research questions 112
5.4 Statistical analyses: power analyses and effect sizes 115
5.5 Statistical analyses: correcting for multiple comparisons 117
5.6 Ethical considerations 119
6 Chapter 6 Results 121
6.1 Results: Exploring the background measures, EF measures, and Mc measures 121
6.1.1 Background measures 121
6.1.2 EF measures 123
6.1.3 Mc measures 132
6.1.4 How are the EF and Mc scores related to each other? 139
6.1.5 Exploratory analyses summary 141
6.2 Results: Balance beam performance and strategy development 142
6.2.1 Balance beam performance per TP and per problem 143
6.2.2 How does balance beam performance relate to the background measures? 147
6.2.3 Strategy development 148
6.2.4 Balance beam data summary 157
6.3 Research question 1: What role do EF and Mc have in children's performance on physics tasks? 158
6.3.1 Research question 1 summary 170
6.4 Results: Research question 2: What impact does support type have? 170
6.4.1 Is there a difference in EF scores between the groups? 171
6.4.2 Is there a difference in Mc between the groups? 172
6.4.3 Is there a difference in balance beam performance between the groups? 174
6.4.4 Is there a difference in strategy development between the groups? 177
6.4.5 Transfer ramps task data 192
6.4.6 Is there a difference in performance on the ramps task between the groups? 200
6.4.7 Research question 2 summary 203
7 Chapter 7 Discussion 204
7.1 Aims and hypotheses of the present study 204
7.2 Exploratory analyses 205
7.3 Balance beam and strategy development 210
7.4 Research question 1: What role do EF and Mc have in children's performance on physics tasks? 213
7.5 Research question 2: What impact does support type have? 217
7.5.1 EF 218

7.5.2 Mc 218
7.5.3 Balance beam task 220
7.5.4 Strategy development 223
7.5.5 Transfer physics task 224
7.6 Limitations of the present study and recommendations for future studies 227
7.7 Contributions of the present study 232
7.7.1 Theoretical contributions 232
7.7.2 Educational contributions 234
7.7.3 Methodological contributions 237
8 Chapter 8 Conclusion 239
9 References 241
10 Appendices 252
10.1 Appendix A C.Ind.Le codes selected for assessing Mc 252
10.2 Appendix B Examples of coding used during the physics tasks 255
10.3 Appendix C Balance beam support type instruction used in the main study 257
10.4 Appendix D Comparison of length of instruction in the two groups 261
10.5 Appendix E Comparison of how much information was provided by the adult 262
10.6 Appendix F Comparison of how much information each group had before starting the trials 264
10.7 Appendix G Feedback after the balance beam trials 266
10.8 Appendix H Ramps task instructions and trials 268
10.9 Appendix I Comparison of the ramps data between the two support groups 272
10.10 Appendix J Mc interview questions used after the balance beam task 273
10.11 Appendix K Information sheet and consent form used in the main study 274
10.12 Appendix L Data screening for age, BPVS, and NEPSY data 278
10.13 Appendix M Data screening for the individual and composite EF measures 279
10.14 Appendix N Data screening for the EF and background measures 280
10.15 Appendix O Data screening for Mc rate 281
10.16 Appendix P Data screening for Mc interview scores 282
10.17 Appendix Q Data screening for the balance beam performance data 283
10.18 Appendix R Data screening for the strategy development data 284
10.19 Means and SDs for the EF and Mc scores for each strategy pattern at each TP 285
10.20 Appendix T Data screening for the EF performance data for each group 290
10.21 Appendix U Data screening for the Mc rate data for each group 291
10.22 Appendix V Data screening for the Mc interview scores data for each group 292
10.23 Appendix W Data screening for the balance beam performance data for each group 293
10.24 Appendix X Is there a difference in the balance beam support type protocol between the groups? 294
10.25 Appendix Y Data screening for complete sample's ramps data 297
10.26 Appendix Z Data screening for each group's ramps task trials data 298
10.27 Appendix AA Data screening for each group's Mc data during the ramps 299

List of Figures

Figure 1. Photo of the balance beam used in pilot study 1. 7
Figure 2. Balance beam used by Siegler (1976). 8
Figure 3. Balance beam task used by Schrauf et al. (2011). 12
Figure 4. Ramps task used in Klahr and Nigam (2004). 17
Figure 5. Door task (Baker et al., 2011) 28
Figure 6. The type of balance beam used by Karmiloff-Smith, taken from Peters, Davey, Messer, and Smith (1999). 49
Figure 7. Photo of the ramps used in pilot studies 1, 2 and 3 and the main study. 69
Figure 8. Photo of the balance beam used in pilot studies 2, 3, and the main study, set up with the dinosaur characters. 79
Figure 9. Grass/snow 105
Figure 10. Modified Corsi blocks 105
Figure 11. FIST 106
Figure 12. Percentage of each strategy used during the 2 balance trials. 151
Figure 13. Percentage of each strategy used during the 4 balance trials. 152
Figure 14. Percentage of each strategy used during the 2 conflict balance trials. 153
Figure 15. Percentage of each strategy used during the 3 conflict balance trials. 154
Figure 16. Percentage of each strategy used during the 3 conflict balance dissimilar trials. 155
Figure 17. Percentage of each strategy used by GP for the 2 balance trials. 179
Figure 18. Percentage of each strategy used by DI for the 2 balance trials. 179
Figure 19. Percentage of each strategy used by GP for the 2 balance trials. 181
Figure 20. Percentage of each strategy used by DI for the 2 balance trials. 181
Figure 21. Percentage of each strategy used by GP for the 4 balance trials. 183
Figure 22. Percentage of each strategy used by DI for the 4 balance trials. 183
Figure 23. Percentage of each strategy used by GP for the 3 conflict dissimilar trials. 185
Figure 24. Percentage of each strategy used by DI for the 3 conflict dissimilar trials. 185
Figure 25. Percentage of each strategy used by GP for the 3 conflict balance trials. 187
Figure 26. Percentage of each strategy used by DI for the 3 conflict balance trials. 187

List of Tables

Table 1 Pilot study 1 participant information: age, sex, location, BPVS scores, and NEPSY scores 61
Table 2 Pilot study 1: conditions used in the prediction trials of the balance beam 63
Table 3 Pilot study 1: conditions used in the production trials of the balance beam 64
Table 4 Pilot study 1: individual performance for each balance beam problem type in the prediction task 65
Table 5 Pilot study 1: individual performance for each balance beam problem type in the production task 66
Table 6 Pilot study 1: number of error strategies used for each solution type in the production task 68
Table 7 Pilot study 1: overview of individuals' scores on each EF task during pilot study 1 74
Table 8 Pilot study 1: Mc rate in the production physics tasks 76
Table 9 The ramps task trials from pilot studies 2 and 3 81
Table 10 Pilot study 2: overview of the six EF tasks including the procedure and scoring used 83
Table 11 Pilot study 2: individual EF scores (percentage correct) from pilot study 2 86
Table 12 Mc interview scores from the balance beam task used in pilot studies 2 and 3 88
Table 13 Performance on the balance beam during pilot study 3 90
Table 14 Main study: descriptive information for the overall sample and the two groups, along with t-test results and effect sizes 92
Table 15 Overview of testing sessions in the main study 94
Table 16 Main study: number of children in each group who completed each task at which TP 95
Table 17 Balance beam trials used in the main study 97
Table 18 Ramps trials used in the main study 104
Table 19 Examples of coding used during the physics tasks 108
Table 20 Inter-rater reliability for the Mc coding 111
Table 21 Mean (SD) and range for age, BPVS scores, and NEPSY scores at TP1 121
Table 22 Kendall Tau correlations between age, BPVS scores, and NEPSY scores 122
Table 23 Mean (SD) and range for each EF score at TPs 1, 2, and 3 123
Table 24 Kendall Tau correlations between the EF measures at TP1 124
Table 25 Kendall Tau correlations between the EF measures at TP2 125
Table 26 Kendall Tau correlations between the EF measures at TP3 125
Table 27 Kendall Tau correlations between the inhibition scores over the three TPs 125
Table 28 Kendall Tau correlations between the WM scores over the three TPs 126
Table 29 Kendall Tau correlations between the shifting scores over the three TPs 126
Table 30 Kendall Tau correlations between EF composite scores at the three TPs 126
Table 31 Kendall Tau correlations between EF, age, BPVS scores, and NEPSY scores 128
Table 32 Spearman partial correlations between EF scores at TP1, controlling for BPVS 129
Table 33 Spearman partial correlations between EF scores at TP2, controlling for BPVS 129
Table 34 Spearman partial correlations between EF scores at TP3, controlling for BPVS 130
Table 35 Spearman partial correlations between inhibition scores over the three TPs, controlling for BPVS 130

Table 36 Spearman partial correlations between WM scores over the three TPs, controlling for BPVS 130
Table 37 Spearman partial correlations between shifting scores over the three TPs, controlling for BPVS 131
Table 38 Spearman partial correlations between composite scores over the three TPs, controlling for BPVS 131
Table 39 Mc rate means, SDs, and range at each TP 133
Table 40 Mean (SD) and range for each Mc rate at TPs 1, 2, and 3 134
Table 41 Kendall Tau correlations between the Mc rates 134
Table 42 Kendall Tau correlations between Mc rate and age, BPVS and NEPSY over the three TPs 135
Table 43 Mc interview scores means, SDs, and range at each TP 136
Table 44 Kendall Tau correlations between the Mc interview scores over TPs 137
Table 45 Kendall Tau correlations between Mc interview scores and age, BPVS scores, and NEPSY scores over the three TPs 137
Table 46 Spearman partial correlations entering Mc interview scores while controlling for BPVS scores 138
Table 47 Spearman partial correlations between Mc rate and Mc interview scores, entering BPVS as a covariate 139
Table 48 Spearman partial correlations between EF scores and Mc rate, controlling for BPVS scores 140
Table 49 Spearman partial correlations between EF scores and Mc interview, controlling for BPVS scores 141
Table 50 Means (SDs) and ranges for the balance beam scores at each TP 143
Table 51 Kendall Tau correlations between balance beam scores at each TP 144
Table 52 Means, SDs, and ranges for the balance beam problem types over all TPs 145
Table 53 Kendall Tau correlations between balance beam problem types 146
Table 54 Kendall Tau correlations between balance beam scores and the background measures 147
Table 55 Consistency scores' means, SDs, and ranges for the different balance beam problems 149
Table 56 First correct scores' means, SDs, ranges for the different balance beam problem, and percentage that solved the problem on the first trial 149
Table 57 Number of children in each classification per balance beam problem 156
Table 58 Spearman partial correlations between balance beam scores and EF scores, controlling for BPVS scores 158
Table 59 Kendall Tau correlations between balance beam scores and Mc rate 159
Table 60 Spearman partial correlations between balance beam scores and Mc interview scores, controlling for BPVS scores 160
Table 61 Spearman partial correlations between first trial correct and consistency scores and EF scores, controlling for BPVS scores 161
Table 62 Kendall Tau between first trial correct and consistency scores and Mc rate 163
Table 63 Spearman partial correlations between first trial correct and consistency scores and Mc interview scores, controlling for BPVS scores 164
Table 64 ANCOVAs examining the difference in EF scores based on strategy use classification for the 2 conflict balance beam problems 165
Table 65 ANOVAs examining the difference in Mc rate based on strategy use classification for the 2 conflict balance beam problems 165
Table 66 ANCOVAs examining the difference in Mc interview scores based on strategy use classification for the 2 conflict balance beam problems 166

Table 67 ANCOVAs examining the difference in EF scores based on strategy use classification for the 2 weight balance beam problems 166
Table 68 ANOVAs examining the difference in Mc rate based on strategy use classification for the 2 weight balance beam problems 167
Table 69 ANCOVAs examining the difference in Mc interview scores based on strategy use classification for the 2 weight balance beam problems 167
Table 70 ANCOVAs examining the difference in EF based on strategy use classification for the 4 weight balance beam problems 168
Table 71 ANOVAs examining the difference in Mc rate based on strategy use classification for the 4 weight balance beam problems 168
Table 72 ANCOVAs examining the difference in Mc interview scores based on strategy use classification for the 4 weight balance beam problems 168
Table 73 EF mean scores and SDs at each TP for each group 171
Table 74 Mc rate means and SDs at each TP for each group 172
Table 75 Mc interview means and SDs at each TP for each group 173
Table 76 Balance beam performance means and SDs at each TP for each group 174
Table 77 Balance beam problem type performance means and SDs for each group 175
Table 78 Strategy classifications for each group per problem type 188
Table 79 Consistency scores means and SDs for each group 190
Table 80 Mean trial when correct strategy was first used and SDs for each group 191
Table 81 Means, SDs and ranges for total percentage correct, percentage of first trials correct, Mc rate, and Mc interview scores during the ramps task 192
Table 82 Kendall Tau correlations between the percentage of total trials correct, percentage of first trial correct, Mc rate, and Mc interview scores 193
Table 83 Kendall Tau correlations between the background measures and the ramps task data 194
Table 84 Kendall Tau correlations between the ramps task data and EF at each TP 195
Table 85 Kendall Tau correlations between the ramps task data and balance Mc rate at each TP 196
Table 86 Kendall Tau correlations between the ramps task data and balance Mc interview scores at each TP 197
Table 87 Kendall Tau correlations between the ramps task data and balance beam performance scores at each TP 198
Table 88 Kendall Tau correlations between the ramps task data and the balance strategy development data 199
Table 89 Means and SDs for the ramps trials total percentage correct and percentage solved on the first try 200
Table 90 Means, SDs and ranges for the ramps task Mc rate and Mc interview scores 202

1 Chapter 1

Introduction

This research examined the role of EF and Mc on children's ability to solve physics tasks with either GP or DI support. GP and DI were selected as the support types due to opposing research concerning which may be better for children's learning. Physics was selected as the area of interest since there is some research with physics tasks with different age groups, although less research with preschool children. Some physics research has also examined how support types may impact physics task performance, which this work builds on.

The age group selected was 3- and 4-year-olds, since this is an important time in the United Kingdom's educational system and the time many children will begin nursery, so the question of support type becomes important. This is an age of vast cognitive changes, with the first five years of life said to be crucial for EF development (Garon, Bryson, & Smith, 2008; Miyake & Friedman, 2012). EF is the cognitive process required for goal-directed behaviour, to control actions, thoughts, and emotions (Vandenbroucke, Verschueren, & Baeyens, 2017). Mc can be defined as one's knowledge about their own cognitions and it is thought to begin to emerge around age 3 (Kuhn, 2000). As children start nursery, the development of EF and Mc becomes important for beginning formal education. This age group will not have received formal teaching in the physics areas to be tested, making it a good concept to examine. (In England, weight is formally taught at age 5-6 and friction and incline taught at age 9-10, Department for Education, 2013.) The focus of this research is on GP (child-led) and DI (adult-led) support types and whether they impact children's learning of physics. The work also considered the role of EF and Mc in how well children perform on the physics tasks. Vocabulary and visual-spatial skills were considered as potential factors but were not the main focus of the work. The links between these various measures were examined to see whether any significant relationships exist and if they help explain physics task performance.

This research is interesting as each child was visited up to eight times over three TPs over 15 weeks in order to track changes in the variable outcomes. This meant changes could be examined over the course of the study to see if and when differences emerged. By taking three measurements, each around six weeks apart, it was expected that measurable changes in performance could be recorded. It was hoped that measuring changes within and between

groups would help identify whether children in one support type performed better. A beneficial difference for one group on any measures could have implications for how children should be taught. A detailed examination of the strategies employed during the balance beam task was utilised to uncover whether children began the task knowing how to solve the problems, having no knowledge of the physics concepts being tested, or perhaps held a misconception concerning the correct strategy. It has been documented that children struggle with some physics misconceptions, which sometimes continue into adulthood, so this work will consider whether evidence of misconceptions can be seen here and if the support types help change children's understanding of physics concepts. It is ultimately hoped this work will help others identify ways to facilitate children's learning and perhaps not just during physics tasks.

The question of whether support type provides benefits within the task was one focus, but an additional question was whether knowledge gained in one physics task transfers to another physics task. This is important when considering domain-general skills and if there is a need to teach such skills to young children.

The outcomes of this work will also add to the field of cognitive development by examining the links between EF and Mc and their role in physics performance. There is still debate as to the structure and function of EF and Mc in this age group, so this data examining how they relate and if they have a role in physics performance adds to the on-going debate.

This thesis contains seven further chapters. Chapter 2 will review the literature in the field of physics; EF; Mc; how EF and Mc are related; vocabulary and visual-spatial skills; support type; theoretical accounts of performance; and the final section will conclude by drawing together links, which form the research questions. Chapter 3 will describe the methodology, rationale, and findings of pilot studies 1, 2, and 3 and how they shaped the main study. Chapter 4 will describe the main study and chapter 5 will detail the analyses methods used. Chapter 6 will present the analyses and findings of the study with reference to each research question. Chapter 7 will discuss the results in relation to the research questions, hypotheses, and literature review to explain the findings. Chapter 7 will also address the limitations of the study, future recommendations, and the contributions of this work. Chapter 8 presents the conclusion.

2 Chapter 2
Literature Review

This chapter will discuss the literature that led to the research questions. The first section will address physics research with children and what can be said about their knowledge, followed by some work with adults and misconceptions. Work on EF and Mc will then be presented to justify the need to examine their relationship to physics performance, followed by how EF and Mc could be related. A section on vocabulary and visual-spatial skills will address the need to include these measures as potentially playing a role in performance. The two support types will then be discussed with reasons for selecting them. Theoretical accounts will be presented to try and draw these areas together to consider how physics performance could be accounted for. The chapter will end by drawing links from each of these sections to highlight where the gaps in research and knowledge are, which are presented in the form of the research questions.

2.1 Physics

This section will introduce research on children's understanding of physics, what children typically know at what age, data from the balance beam and ramps task (the two tasks used here), along with some explanations for differences in performance. The second section will briefly consider some work carried out with adults and why some misconceptions are seen past childhood.

2.1.1 Children's understanding of physics

Research suggests that very young infants have some understanding of physical properties, insofar as they can detect various violations of physical reality. By their first birthday, infants show knowledge of solidity, continuity, and cohesion, all of which are said to develop through experience. Spelke, Breinlinger, Macomber, and Jacobson (1992) claim that some representations concerning the physical world are innate (such as continuity and cohesiveness), but agree that through experience, knowledge is strengthened with the addition of further information. Experience includes viewing objects move behind other objects (occlusion), seeing objects go inside other objects (containment), and seeing objects being covered by other objects (covering) (Baillargeon, 2008).

Using the violation of expectation paradigm – whereby if a child looks longer at an

unexpected outcome they are said to have recognised the violation (Baillargeon, 2008) – infants have been found to recognise violations in physics. By 3 months, infants show recognition of solidity violations and will look longer at a solid object that moves through a solid wall compared to when it stops at the wall (Baillargeon, 2008). Recognition of inertia violations are seen at around 7 months (Kim & Spelke, 1999), for example, when a ball rolls off a surface and falls straight down. Some recognition of gravity violations are seen at 2-years-of-age (Lee & Kuhlmeier, 2013), such as when viewing a ball placed in a curved tube and understanding the ball does not fall straight down, as if through the walls of the tube, but instead travels down inside the curved tube and comes out in a spot not directly below where it started. Thus, different violations are recognised at different ages: solidity by 3 months, inertia at 7 months, and gravity at 2 years. The different ages could be because some physical properties are said to be innate, but others require experience (and some more than others) before being understood.

However, a distinction needs to be made between *recognising* unexpected outcomes in tasks and being able to *predict* the outcomes. Predicting correct outcomes to solidity problems is seen by around 2-and-a-half-years of age (Hood, Carey, & Prasada, 2000), to some inertia problems from 5-years-of-age (Kim & Spelke, 1999), to gravity problems at two-years-old (Lee & Kuhlmeier, 2013), and some say to balance problems at two years of age (Halford, Andrews, Dalton, Boag, & Zielinski, 2002). These ages are often different to the age when they can successfully *recognise* a violation, as stated in the above paragraph. Research has repeatedly shown this discrepancy in performance when the same children are tested on recognition problems compared to prediction problems (such as Howe, Taylor & Devine, 2012), but the reasons for the discrepancy are still debated. There are also tasks that require children to actively *produce* an answer, which can be considered as a slight deviation from the prediction tasks.

The discrepancy in performance between recognition and prediction tasks could be due to having a lack of knowledge, having incorrect knowledge, that prediction problems are more difficult since they involve reasoning to obtain an answer (Howe et al., 2012), the response type required (such as pointing or speaking) (Munakata, 2001), how long they have to respond (Kozhevnikov & Hegarty, 2001), and if EF plays a role (Lee & Kuhlmeier, 2013). These factors are important when considering whether implicit or explicit knowledge is being tested in recognition, prediction, and production tasks. Implicit knowledge is said to be

unconscious, is often used when a fast response time is required, is built up through experience in viewing different actions (and these experiences mean perceptual representations can be formed, which are drawn upon when implicit knowledge is used) (Kozhevnikov & Hegarty, 2001). Explicit knowledge can be learnt, is conscious, and can be verbalised (Kozhevnikov & Hegarty, 2001).

The production physics tasks in the current work are believed to draw on explicit knowledge since the children had to produce answers to different problems, there was no time constraint, and some were asked to verbalise their thoughts. The distinction between prediction and production tasks is important, as the same data cannot be collected in each task. Balance beam prediction tasks require children to indicate whether a beam already set up with weights will balance or tip. The production tasks used here require children to actively balance the weights they are given. It was thought that a better online measure of Mc during the task could be captured with production rather than prediction tasks, and a detailed analysis of strategy use, which is one reason they were selected. However, during the production tasks not all problem types can be tested, for example, production trials require weight *and* distance to be considered when the child places the weights. This means the distance and weight variables cannot be separated, which may impact what conclusions can be drawn about children's knowledge, as well as impact how the data can be explained by different theories, as will be explained through this work.

The production physics tasks used here are believed to access children's explicit knowledge, which could potentially be modified through the support type (as the knowledge could be updated with instruction), verbal and visual feedback, and various trials and sessions (experience). Thus, the focus of the physics literature review will be on tasks that are thought to examine children's explicit knowledge.

When considering performance, a distinction needs to be made between children failing a physics problem because of a lack of knowledge and because they used incorrect knowledge, believing it to be correct (misconception). The children in this study are unlikely to have received any formal education concerning the physics tasks involved, but they may have some experience with balance and ramps, so it is possible that they already hold knowledge, whether correct or incorrect knowledge. It was hoped by examining the strategies used that this question could be unravelled. If a child uses trial and error, it could indicate a lack of

knowledge and if they continuously use the incorrect strategy it could indicate a misconception. However, this may be complicated if a child has the correct knowledge, but is unable to apply it, perhaps due to EF demands. By examining strategy use in relation to EF scores, it is hoped these ideas will be unravelled and conclusions drawn on what knowledge children have concerning balance beam concepts.

The physics tasks used in this study were the balance beam task and the ramps tasks, but the tubes task was also considered. The teaching of forces (in England) does not occur during nursery according to the statutory framework for the early years' foundation stage (Department for Education, 2013). That means the children in this study should not have received formal teaching in the areas of physics being examined here and thus not be proficient in such tasks. The framework states that forces – including weight, friction, and gravity – should be formally taught from Year 5 (age 9). The balance beam task tackles weight, the ramps task tackles incline and friction, and the tubes task tackles gravity. The National Curriculum for Science's aims states that children should "develop scientific knowledge and conceptual understanding" and "understand the processes and methods of science" (Department for Education, 2013, p3). It also states that children should learn how to work scientifically from Year 1 (5 years old), such as through using science equipment, observation, asking questions, answering questions, and testing (Department for Education, 2013). The knowledge, understanding and methods described in the aims, as well as what constitutes working scientifically, overlap with the support types to be implemented here, which will be discussed later. Each support type incorporates different aspects of working scientifically, with GP involving more questions and DI involving more observation. Since this is covered from Year 1, it suggests that by 5 years of age children are capable of learning how to work scientifically, but it also means children should not have already received formal instruction on working scientifically. An aim of this work is to see how children's understanding of scientific concepts changes over time, depending on the support type provided, so using forces as the topic to be tested should be interesting since the children should not have formally-taught knowledge and could show a noticeable degree of learning.

Quite a lot of work has been carried out examining children's knowledge of balance, but less so with the ramps task. There is disagreement over when children start to understand balance concepts, what the task is testing in terms of knowledge, and the reasons for differing findings. These will be discussed next, with the focus being on prediction and production

tasks.

2.1.1.1 Balance beam task

The balance task was first developed by Inhelder and Piaget (1958) and has been used to investigate children's physics knowledge, cognitive processes, and to test different cognitive theories. The balance beam has pegs (or another means to place weights) on each side of the beam for weights to be placed. An example of a balance beam can be seen in Figure ***1***. There are different methods to use the balance beam: one is *prediction*, which is when the adult places the weights on the beam when it is propped up so as not to tip, and the child has to predict what will happen (tip or balance). Another method is to let the children actively place the weights themselves, termed *production* here.

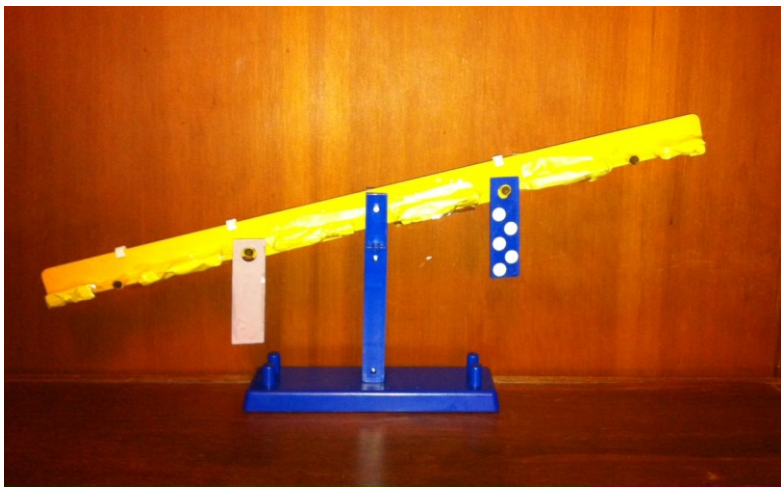

Figure 1. *Photo of the balance beam used in pilot study 1.*

Much of the work carried out with the balance beam has built on the work of Inhelder and Piaget (1958). Siegler (1976) furthered this work and was able to categorise children's ability based on their predictions to different balance beam problems. 4- and 5-year-olds' predictions were mostly based on whether there was the same weight on each side of the beam (the greater weight would tip) (termed rule I). 8- and 9-year-olds' considered distance if the weight on each side of the beam was the same (the greater distance would tip) (rule II), and considered distance if the weights on each side were not the same (either the greater weight would tip or the greater weight and greater distance would tip), but struggled if both the weight and distance differed (rule III), otherwise they reverted to rule I (i.e. if weight and distance differed on each side the greater weight would tip). 12- and 13-year-olds used rule III – being able to consider different distances and different weights. Rule IV was deemed

quite advanced and requires calculating the force of each weight based on its distance, which adults cannot always do (Siegler, 1976). Siegler (1976) used the percentage correct for each problem and children's explanations of the outcome to aid categorising children, which resulted in 107 children being categorised as using one of the four rules.

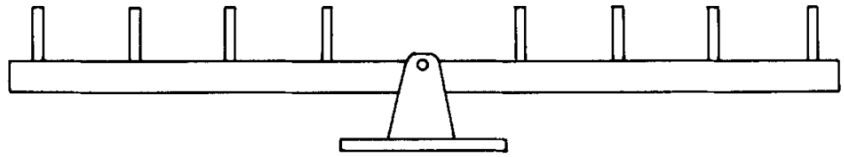

Figure 2. *Balance beam used by Siegler (1976).*

Siegler (1976) noticed that the younger children did not appear to take notice of the distance aspect during the task and so carried out a follow-up task with 5- and 8-year-olds. Children had to replicate how an adult set up a beam, and it was seen that the 5-year-olds were less likely to incorporate distance when placing the weights, indicating they did not register distance, but the older children did. In a further study, children received instruction which drew their attention to both the number of weights on the beam and the distance they were placed at, which resulted in children being more successful on replicating the beam set-ups (around 50% accuracy) (Siegler, 1976). Siegler (1976) concluded that the findings could be explained by children's ability to encode both weight and distance in the trials, termed the encoding hypothesis. This hypothesis suggests younger children are less able to incorporate various pieces of information and draw them together, unlike older children who are better abled to encode new information (Siegler, 1976). Although the last experiment saw an improvement in accuracy, and Siegler acknowledged there could be an attention aspect in being able to replicate the beam, it may have been beneficial to also have a measure of attention or working memory from the children, to see if this played a role in performance.

Overall this work highlights two important points to consider in the current work: there appears to be performance shift with age based on the problems children can and cannot solve, and drawing children's attention to the different variables of the balance beam improves their encoding of information, which improved performance. The instruction provided could be an important aspect of support type to be examined here. One downfall in this work is that not all children can be categorised based on their rule use, suggesting the knowledge children hold and/or can apply is not as simple as these four rules. This work is

important as it has driven much more research in this field, but this staircase approach to describe what knowledge children hold may not be the best method, since all children could not be accounted for.

Siegler and Chen (1998) followed on from Siegler's work, and also found younger children tended to make predictions based on weight alone, so they sought to examine whether children could learn more complex rules. They used a similar balance beam as Siegler (1976), although with five pegs on each side. Siegler and Chen (1998) worked with 70 4- and 70 5-year-olds and manipulated whether children were given the opportunity to learn rule I or II in order to see the pattern of problem types that could be solved. In the pre-test, children had to predict and explain what would happen when different combinations of weights and distances were set up on the beam. They also completed an encoding task, involving replicating balance beam set-ups, similar to Siegler (1976). During the feedback phase (on another day), children either saw trials that would aid learning rule I *or* rule II and had to predict and explain what would happen to the balance beam the adult had set up. The adult provided feedback during these trials by telling them whether they were correct but still asked the child to explain why the balance beam tipped or did not tip over. The post-test was carried out on another day and repeated the pre-test problems and encoding task. The pre-test results showed over half of the 4-year-olds showed no rule use (compared to a quarter of the 5-year-olds), around a third of the 4-year-olds appeared to understand the role of weight (compared to around half of the 5-year-olds) and none appeared to understand the role of distance (compared to 7% of the 5-year-olds). Of the 4-year-olds who used no rule or the wrong rule in the pre-test and received the weight trials in the feedback phase, 56% (of 25 children) correctly used rule I in the weight trials, but 0% used rule II in the distance trials. Of the 4-year-olds who used no rule or the wrong rule in the pre-test and received the distance trials in the feedback phase, 26% (of 35 children) correctly used rule I in the weight trials and 6% correctly used rule II in the distance trials. The encoding task revealed 4-year-olds were not as successful at encoding the variables of weight and distance as the 5-year-olds.

These results show that over a short time with trials, feedback, and being asked to explain the outcomes, some 4-year-olds can learn how both rules work, although learning distance is more difficult. It showed that children who only received distance trials in the feedback phase still managed to learn how to solve the weight trials and more so than the distance trials. The

same was not seen for the children who received the weight trials – none went on to solve any distance trials. Siegler and Chen (1998) discuss the findings in terms of knowledge held before starting the task and cognitive processes that occur during the task, such as the ability to encode the variables. Their regression model entering age, pre-test rule category, and the ability to encode predicted half of the variance in being able to learn rule II (incorporating distance). They found that the 5-year-olds knew more about the balance concepts before the task, which likely aided their ability to learn more about the balance beam during the task. This study supports Siegler (1976) insofar as finding that weight is an easier variable than distance (despite the same feedback phase protocol) and there is an age trend. Individual knowledge before starting the task is therefore not the only factor that could influence performance, and a mixture of other factors could be involved, potentially including feedback and instruction.

Halford et al. (2002) carried out a similar study to Siegler and Chen (1998), but with 2-year-olds. They used the same type of beam as Figure 2, but on the last peg of each side was a stuffed animal, and children had to predict which animals would go up/down or if they would stay the same. The task was similar because children learnt weight *or* distance problems, but it was dissimilar as the instruction phase differed to Siegler and Chen's (1998) feedback phase. In Halford et al.'s (2002) study 22 children received a 25-minute familiarisation session, during which the adult showed and explained how to solve either weight or distance problems, were provided feedback, and then the child was asked to have some goes to make a particular side of the beam tip over. The nature of this instruction could be compared to DI in terms of children being shown and told how to solve a particular problem and the adult providing feedback as to why something happens, but being allowed to try themselves is more like GP. Children completed six trials of the different problems and correctly solving two or more trials was considered above chance. The pre-test results showed the groups performed above chance on items requiring them to only consider weight on each side of the beam (when distance was the same), but not above chance on trials testing distance. This suggests children as young a 2-years-old already have some knowledge of the concept of weight, although it should be noted the sample size here is quite small. The post-test results for the distance trials showed that children who were taught the distance problems performed above chance, but the children who were taught the weight problems did not. Both groups continued to perform above chance on weight problems.

Halford et al. (2002) then used the same familiarisation procedure to examine children's knowledge of conflict trials (when the side of the beam with either a greater weight or a greater distance does not tip, or conflict balance trials where weight and distance differ and the beam balances). Halford et al. (2002) found performance for the 50 3- and 4-year-olds on the conflict balance trials was not above chance, but performance on the conflict-weight and conflict-distance trials were. The 54 5- and 6-year-olds scored above chance on all the problems. The group means were calculated as the proportion of correct answers from the trials they were given (which sometimes differed between children). These findings support the idea that younger can solve balance beam problems after a familiarisation. The inclusion of the stuffed animals on the beam to make it more game-like perhaps positively contributed, as performance is higher for these young children than other studies would predict.

Halford et al. (2002) discuss the results in terms of how many variables must be considered when solving the problems – termed the relational complexity theory. They suggest that trials that involve only one variable (weight *or* distance) can be coded by 2-year-olds, and trials that involve two variables can be solved by 5-years-old. This somewhat corresponds with the idea of Siegler's (1976) encoding hypothesis: younger children are less able to incorporate as many pieces of information as older children, which results in the performance scores being lower for younger children.

Another similarity between Halford et al. (2002) and Siegler (1976) is the idea of classifying children based on how they solved the trials. Halford et al. (2002) also suggested four categories: if a child scored less than two-thirds on both the weight and distance trials (strategy 0); if they scored more than two-thirds on either weight or distance trials but less than half on conflict trials where weight and distance had to be considered (strategy 1); if they scored more than two-thirds on either weight or distance trials and more than half on the conflict trials where weight and distance had to be considered (and the beam would tip), but less than two-thirds on the conflict balance trials (when the beam would balance) (strategy 2); and if they scored more than two-thirds on weight and distance trials and more than half on the conflict trials where weight and distance had to be considered (and the beam would tip), and more than two-thirds on the conflict balance trials (strategy 3). Both suggest children's knowledge can be categorised based on the problems they can solve, but their different findings suggest elements in the task may be important, such as the apparatus, instruction, feedback, and practice.

Opposing Halford et al.'s (2002) claim that 2-year-olds can solve balance beam problems is Schrauf, Call, and Pauen (2011), who argue that the finding could be due to visual feedback available to the children (such as seeing two versus one weight), and not because children were actually displaying knowledge about balance. Schrauf et al. (2011) tested this claim with 60 3- and 4-year-olds to try to unravel the issue. They used a variation of the balance beam task (Figure 3) and utilised a plausible heavy weight and implausible light weight to tip the beam – the latter to challenge children's thinking about the role of weight.

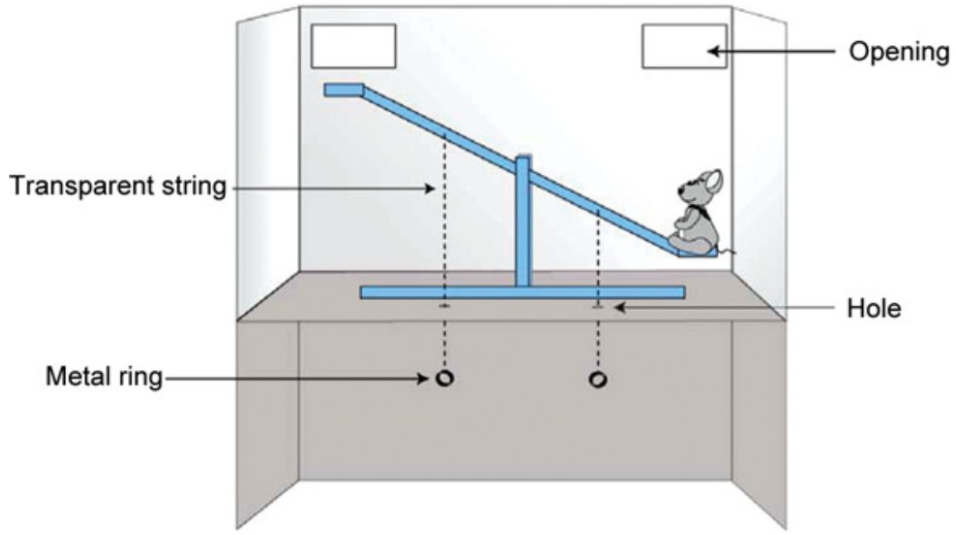

Figure 3. *Balance beam task used by Schrauf et al. (2011).*

The task was set up as a game where the child had to place a weight on one side of the beam to get the other side of the beam to lift up, so they could reach the toy. 3-, 3-and-a-half, and 4-year-olds were split into either the plausible or implausible condition to complete 12 trials. In the implausible group, the beam was secretly manipulated by the experimenter to tip the wrong way, so the light weight would tip down. The aim was to see whether children considered the physical feeling of the weights before putting them on the beam or used the visual feedback, which challenged expectation. The results showed that the 3- and 3-and-a-half-year-olds performed similarly to each other and similarly in the two conditions, however the 4-year-olds performed better in the plausible condition. The 3-year-olds performed better than the four-year-olds in the implausible condition, suggesting the four-year-olds were struggling with the implausible trials when the beam tipped in an unexpected way.

Schrauf et al. (2011) therefore suggested the 3-year-olds' similar performance between

conditions provided evidence of them using visual feedback from the beam rather than knowledge of weight to solve the trials. The 3-year-olds saw the heavy weight did not tip the beam as it should, so they used a different weight, but the 4-year-olds seemed less able to accept that the heavy weight did not tip the beam as expected and made more incorrect tries. Although this finding is important to consider, it does not correspond to findings that show children often begin the balance beam task showing knowledge of weight, such as Halford et al.'s (2002) pre-test findings. Schrauf et al. (2011) found no difference in whether children used a heavy or light weight on the first trial to tip the beam, suggesting some children may not have had the knowledge that heavy weights are more likely to tip the beam. There appears to be two notable differences between Schrauf et al.'s (2011) study and Halford et al.'s (2002) study: using one location on each side of the balance beam task versus several pegs, and producing a solution versus predicting an outcome. Perhaps in Schrauf et al.'s (2011) study the negative feedback changed how the 3-year-olds consider the role of weight or maybe selecting the correct weight and producing solutions was more challenging. The present work included 3-year-olds and a more typical balance beam apparatus with two pegs on each side of the beam for children to place weights. Children were provided with weights and asked to make them balance, rather than select which weights to use. It was hoped by examining performance and strategy development that light could be shone on whether 3-year-olds begin the task already showing knowledge of balance beam concepts and if they show improvement over time.

The studies discussed so far differ on several major variables: whether children predict outcomes or produce solutions, the apparatus used (how many pegs and weights and if it is game-like), and the type of support received. In Schrauf et al.'s (2011) study children had to actively select and place the weight to solve the problems and they concluded 4-year-olds show an understanding of weight, but 3-year-olds do not. In Siegler's (1976), Siegler and Chen's (1998), and Halford et al.'s (2002) work, children had to predict the outcome, each study found that weight was easier than distance, and the two studies that tested whether distance could be taught found it could, although it was more difficult than weight. Halford et al. (2002) found 2-year-olds could solve balance problems, and Siegler (1976) and Siegler and Chen (1998) found 4-year-olds could solve balance problems. Halford et al.'s (2002) and Siegler's (1976) work states that children perform much better when only one variable must be considered – weight or distance – although weight is seen to be the dominant variable and learnt first. Schrauf et al. (2011) only tested weight and found 4-year-olds displayed

knowledge of this concept.

The apparatus used in these studies also differed from one another, perhaps contributing the differences seen. The beam Halford et al. (2002) used had three pegs on each side of the fulcrum, Siegler's (1976) work used four pegs on each side, Siegler and Chen's (1998) work used five pegs on each side, and Schrauf et al.'s (2011) only had one seat on each side of the beam. It may have been expected that more locations would make it more difficult for the children, particularly on conflict trials when both weight and distance must be considered, but Schrauf et al. (2011) found 3-year-olds could not solve weight trials. The other issue to consider is whether children relied on visual information to solve the task, such as counting the weights or distance, without necessarily understanding the role of either (as suggested by Schrauf et al., 2011). However, it is difficult to tease apart whether using visual aids is necessarily not showing knowledge of balance concepts, as complex conflict problems (differing in both weight and distance) are often solved with the torque calculation – explained by Siegler's (1976) rule IV – which states multiplying the weight and distance is often the way to find the correct answer.

The instruction given in each study also differed. Halford et al. (2002) provided 25 minutes of familiarisation where they showed and explained how to solve different problems, provided feedback, and allowed children to have some goes. Siegler and Chen (1998) asked children to predict and explain the outcomes and the adult gave feedback, and Schrauf et al. (2011) provided no instruction or feedback. Siegler (1976) found if children were taught to take notice of both the number of weights on a beam and the pegs they were on then children were better able to replicate the beam set up when the beam was removed, as if they had not fully understand the role of these two concepts before. The three studies that provided instruction of some sort all found the youngest children succeeded or improved on various problems, but only the 3-year-olds in Schrauf et al.'s (2011) study (where they were not provided instruction) performed poorly. During Schrauf et al.'s (2011) study children were also required to select which weight to use, which may have been more challenging. The question of whether instruction is important in learning balance beam concepts therefore seems key, especially since young children appear to benefit. This is why support type is being examined in the present study and these findings will be referred back to, when support type is discussed.

The studies classified children on whether they passed or failed a problem and often examined group differences based on performance means or the frequency of how many children were in which category. Some studies had higher participant numbers than others and different trials numbers, which may account for some variation in findings. The present study will take account of overall performance and performance on the different problem types, as well as how many children can solve which types of problems over time.

In sum, there is still some debate as to when children show understanding of balance concepts. Most would agree that weight is an easier concept than distance, but some say children show knowledge of weight at 2-years-old (Halford et al. 2002), some say 4-years-old (Schrauf et al. 2011), some say it is at 5 years (Case 1985) or even as old as 7 years (Karmiloff-Smith & Inhelder, 1974). It has been noted here that distance can be learnt by 2-year-olds (Halford et al. 2002) and by 4-year-olds (Siegler & Chen, 1998), but conflict balance problems are more difficult and cannot be solved by 2-year-olds (Halford et al. 2002). Overall, the evidence suggests that at 4-years-old children can attempt and solve some of the problems presented to them. How much is due to knowledge and how much is due to differing variables of prediction versus production, the apparatus (weights and pegs), and the instruction is unclear and these will all be considered in the present study. The production trials required children to actively place the weights, while considering both weight and distance, rather than seeing a beam already set-up and indicate whether it will balance or not. Prediction trials would not allow for strategies to be coded in such a detailed manner, which is a key aspect of this work. Although children could be classified based on the knowledge they displayed regarding which beams will balance, it does not provide the same rich data as allowing children to try different strategies. The possible rules or strategies children use will be considered further later, but as the production problems used here do not test the range of problems required to classify children based on rules (weight alone, distance alone, conflict trials) it is not a central focus of this work. The balance beam task was selected as the main task to be completed at each TP since the expected outcomes at different ages have been documented (although there is disagreement), the different problems have different difficulties, and the strategies used are relatively easy to classify. The evidence concerning the elements of support type during the task is less well documented, so this work will examine these aspects. The ramps task was selected as the transfer task as it also encompasses problems of different difficulties and a range of strategies can be used, although they are less easy to classify, as will be discussed in the next section.

2.1.1.2 Ramps task

A transfer task was used in the present study to examine whether performance on one physics task carries over to another. During the balance beam task, children should ideally have learnt that considering both weight and distance is important for successfully solving problems. During the ramps task, children had to take account of the height and surface of the ramps. During the balance beam task, children had multiple attempts of the same trial over several sessions, but in the ramps task they had multiple attempts within one session. The transfer task allowed for the impact of support type in another task to be examined. There is little previous research looking at transfer tasks in physics with young children, and little research examining how support type might impact transfer, so this is a focus here.

Much less research has been carried out with the ramps task, unlike the balance beam task, and far fewer conclusions have been drawn on the development of knowledge, knowledge of variables, and strategy use. Studies using the ramps task tend to focus on control of variables strategy (CVS – understanding the need to test one variable at a time), which is different to what the focus will be here (knowledge of how the surface and incline of a ramp changes how far a ball rolls). The next sections will discuss research with the ramps task examining CVS, the influence of incline and friction, how a different number of variables impacts performance, and how children perform at different ages.

Chen and Klahr (1999) investigated 7- to 10-year-olds' ability to reason and implement CVS in a ramps task, with differing types of support. The CVS strategy requires all but the variables in question to be kept constant to examine the effect of these variables. Children first received an exploration session where they could use and change the ramps apparatus to compare one of the variables (such as the weight of the ball) (see Figure 4 for an illustration of the ramps task). The children were asked probe questions and had to explain why they had set the variables as they had. There were three experimental conditions: explicit instruction with probe questions, no instruction with probe questions, and no instruction and no questions. The children had to compare two variables (like in the exploration session, but with one new variable they had not tested), and how well they employed CVS was scored. Chen and Klahr (1999) found that only the children who received explicit instruction with probe questions showed an improvement on CVS use, suggesting CVS at this age without instruction is difficult and children are not capable of thinking in this way without explanation. The addition of probe questions could potentially have tapped Mc, as the

children had to think about what they had learnt and explain it, which may have strengthened their knowledge and performance. However, probe questions without explicit instruction was not enough to improve CVS use, indicating it was the instruction that was the important element. The findings show that the children were not able to easily compare two variables and it may be at this age within this type of task it is too difficult without explicit instruction.

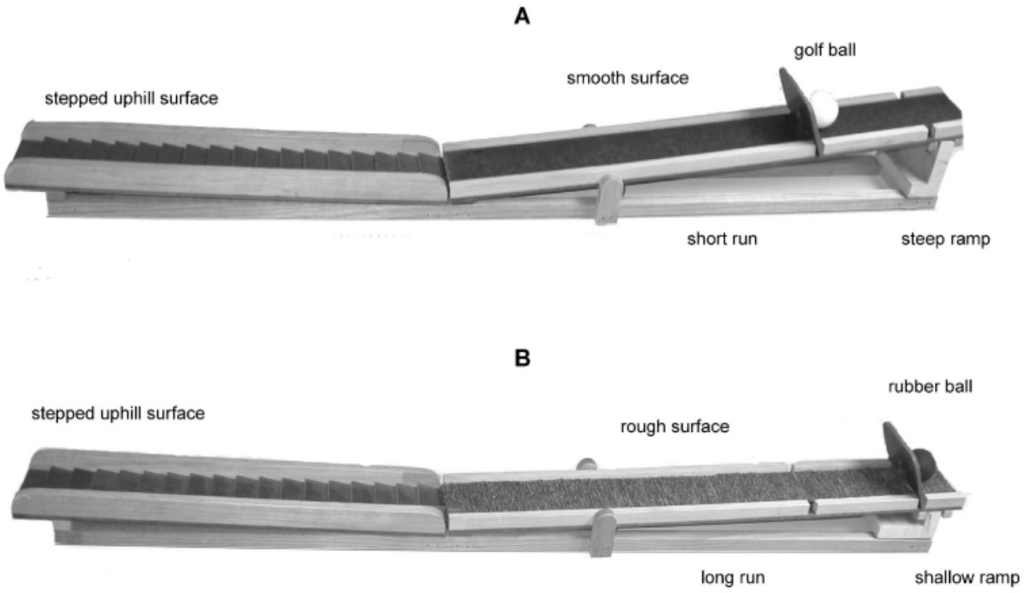

Figure 4. *Ramps task used in Klahr and Nigam (2004).*

Klahr and Nigam (2004) examined how 112 8- to 10-year-olds in either DI or discovery learning support (which is similar to GP in that children had to design their own experiments and received no feedback, but dissimilarly, they received no prompts) performed on a transfer science task. The ramps CVS task (Figure 4) was the main task and evaluating science posters via a structured interview (including questions on whether it was a good study, how to improve it, does the conclusion fit the data) was the transfer task. Children first received an exploration session where they could try the ramps to see how changing the set-up changed how far the ball rolled. Children's baseline knowledge was assessed with four experiments where they had to assess incline and ramp length; support type was then implemented and children completed four more trials. The DI group saw examples and heard explanations of good and bad CVS designs while the discovery learning group continued to use the apparatus to test different variables. Children then completed four more experiments examining ramp length and surface, but the support type differed to before, as all children received no

feedback after the trials. The transfer task took place a week later and their answers blindly coded.

Klahr and Nigam (2004) found the DI group performed significantly better than the discovery learning group from pre- to post-test, as seen by the number of unconfounded tests they designed. They found that those who performed well on the ramps task tended to perform well on the poster task, regardless of support type. Klahr and Nigam (2004) conclude that it is how much children learnt in the ramps task that is important, not which support they received. So the few discovery learning children who performed well on the ramps task performed as well on the poster task as the DI children who performed well on the ramps task. Although there were few test trials to compare performance, the addition of the baseline strengthens these results and adds to research that suggests DI is a better way to teach certain science concepts. Although Klahr and Nigam (2004) state that it is not support type that predicts transfer scores, it was support type that resulted in DI performing better than the discovery learning group, so it could be argued that there is some influence of support type in transfer scores. In the present study performance between the two physics tasks was examined both within the whole sample and between groups to see if there was a relationship between the two tasks' performance scores and a difference between the groups.

Van der Graaf, Segers, and Verhoeven (2015) also used the ramps task to examine children's use of CVS. Van der Graaf et al. (2015) worked with 45 children aged 4 years and 6 months to 6 years and 3 months. The children were first shown how the apparatus worked and then joined in by helping change variables and measuring how far the ball went. The variables used were: the weight of the ball, the starting point on the ramp, the surface of the ramp, and the incline height of the ramp. The children were told they could test how changing the variables affected how far the ball went. Children were asked to set a particular variable to show how it would change how far the ball went (examining one variable), and if they got this correct they were then asked to set two variables, and so on. On each trial, the experimenter set the other variables, which children could not change. Children had to explain each time why they set the variable as they did, they received feedback on why it was correct (if it was correct), and asked to try again if it was incorrect. Children were allowed two tries on each trial and had to pass one of the trials at each variable level to move on to trials with an additional variable.

Van der Graaf et al. (2015) scored how many trials children got correct (maximum 16) and how many variables they correctly changed (maximum 60). All the children managed to pass at least one trial with one variable and 40/45 children passed at least one trial with two variables. The younger children got on average 5.14/16 trials correct and 14.68/60 for correctly setting up variables. The scores may have been influenced by children being allowed a second attempt to pass the trial, but passing trials with two variables indicates it is less likely to be due to chance. The results indicate that 4-and-a-half-year-olds can show CVS use in this kind of physics task. However, it should be noted that they found age and non-verbal reasoning to be related to children's ability to implement CVS. Age may not be unexpected since previous work has shown children's ability to consider more variables increases with age (Siegler, 1976; Halford et al., 2002). The relationship with non-verbal reasoning could be due to some overlap in the potential processes used, but this will be discussed further later. These findings provide reasons to consider them as possible factors to control in the present study.

Van der Graaf et al.'s (2015) finding challenges that of Chen and Klahr (1999) and Klahr and Nigam (2004), as the latter two found instruction was an important aspect for children performing well. The children in Van der Graaf et al.'s (2015) study were younger than in the other two studies, so it could be the design and scoring method that resulted in the performance rates. Van der Graaf et al.'s (2015) study allowed children a second attempt to correct a variable if they got it incorrect on the first try, which may account for the performance rates. As Van der Graaf et al.'s (2015) methodology was quite different to the other two studies there is some doubt over whether children could really solve the ramps task trials or only corrected the variable on their second try, and as the results of whether children were correct on their first or second trial were not reported it remains unclear.

Past studies using the ramps task have tended to work with children older than 3- and 4-years-old, and it is only in the past few years that research has focused on this younger age group. Some research has suggested the ramps task is more difficult than the balance beam task, but there is also less work with the ramps. It could be that trying to assess CVS use in young children is more challenging, or finding a format that young children understand is the difficult aspect. Assessing the variables and strategies used in the ramps task is not as straightforward as in the balance beam task, which could be one reason for the lack of research in this area. It is also difficult to tease apart the question of lack of knowledge versus

incorrect knowledge and to examine strategy development. Multiple variables can differ on each trial (incline, surface, and which ramp each ball is rolled down). Chen and Klahr (1999) recorded the CVS strategies the children used and they found that the children who received explicit instruction showed abrupt changes in the strategies used, unlike those who did not receive explicit instruction. They also found that the children showed a range of strategies over the sessions, but as a group, the 7-year-olds did not improve on CVS use. Due to a lack of work examining strategy development in the ramps task with young children, it was unclear how children in the present study would use different strategies during the ramps task (this will be discussed further later).

To conclude this section on the physics tasks to be used in the present study, it has been seen that a lot of work has been carried out examining children's knowledge of balance, but less work has been carried out examining young children's knowledge of ramps, specifically, distance resulting from motion down an incline and friction, which the present study assesses. Both physics tasks offer a range of problem types of differing difficulties and a range of possible strategies that children can use. It has also been seen that the majority of balance tasks have used prediction tasks, but the ramps tasks have used production tasks. The present study used production tasks for both physics tasks to allow for a more direct comparison of the two tasks' data. It was thought production trials would be more engaging for the children and the strategy development data could be richer. The task protocols used in the present study were selected based on the children's engagement, performance and strategy scores that could be obtained, and how well children understood the aims of tasks.

The next section will briefly outline some work with adults to illustrate why physics learning in childhood could be important for teaching difficult concepts.

2.1.2 Adults' understanding of physics

Work with adults has identified that some adults struggle with physics misconceptions. For example, Dunbar, Fugelsang, and Stein (2007) found many adults believe the different seasons are caused by a change in Earth's distance from the sun and McCloskey, Washburn, and Felch (1983) found that adults often incorrectly predicted how an object would fall from a moving carrier (that it would fall straight down or even backwards). In a study by Kaiser, Jonides, and Alexander (1986) adults were found to have difficulty when asked to show the trajectory of a ball leaving a curved tube, with many indicating the ball would continue on a

curved path, however, when the same adults were asked to indicate the direction of water leaving a hose they found this easier, which the authors suggest could be due to experience and the adults' familiarity with each task. The question here was whether misconceptions can be identified in children and if they could be corrected. The other question is whether misconceptions always exist and if it is that they can be overcome through the use of EF, so these links will be examined. For example, Masson, Potvin, Riopel, and Foisy (2014) suggest a person can hold a misconception as well as accurate knowledge concerning a concept, and through EF (inhibition) the misconception can be suppressed in order for the accurate knowledge to be used. This idea was also put forward by Brookman (2015; cited in Tolmie, Ghazali, & Morris, 2016), as she found adolescents' inhibition scores predicted their scores on a task examining common misconceptions in science and maths, suggesting that EF is used to overcome misconceptions.

These everyday misconceptions are likely based on intuitive physics and may be difficult to change without evidence to challenge them (Karmiloff-Smith, 1992). However, Vosniadou and Brewer (1992) would argue that misconceptions can be changed – they found 6- to 11-year-olds held misconceptions about the shape of the Earth, but through formal teaching, the children gathered more information on the topic and their conceptions changed. By challenging misconceptions during childhood incorrect knowledge could be corrected and this was examined here.

This question is important for several reasons: one is whether evidence for children using incorrect information can be seen in the present study (through strategy development and consistently using an incorrect strategy), whether links between strategy use and EF are seen (supporting the idea that EF can suppress a misconception, aiding performance), and if support type impacts strategy development, potentially correcting a misconception, and possibly whether EF plays a role in this relationship. The theories to be discussed later would agree that experience can change concepts; some may suggest there is a role for instruction (connectionist model, graded representations (GR) account, the RR model), and maybe EF playing a role (Diamond, 2013, connectionist model, GR account). These research findings give reason to investigate the role of EF in strategy development and physics performance. EF will be discussed in the next section and strategy development will be discussed later.

2.2 Executive function

This section will discuss the debate on the function and structure of EF in children and adults, How EF and reasoning are related, and how EF may contribute to physics task performance.

2.2.1 Function and structure of EF

EF will be examined to see if it can explain why some children perform better than others on physics tasks. EF may play a role in problem solving (Diamond, 2013), so it could be that stronger EF skills aid solving physics tasks. There is also some evidence to link EF to performance on physics tasks, but the research is limited, so the present study will add to this field. The children in this study are 3- and 4-years-old and it is at this age that EF is thought to undergo a lot of developmental changes (Best & Miller, 2010), making it an interesting age to study.

There is debate over what EFs are actually part of EF; some argue that there any many components, including attentional control (including self-regulation, self-monitoring, and inhibition), goal setting (including planning and strategic organisation), information processing, and cognitive flexibility (including utilising feedback and WM) (Anderson, 2002). Despite the debate, most acknowledge three "core" components of EF: inhibition, WM (sometimes referred to as updating if remembering information and updating it is required), and shifting (also known as cognitive flexibility). Inhibition involves suppressing a prepotent response, WM is responsible for updating information held during a task, such as remembering a rule or additional information and updating these as the task goes on, and shifting involves the skill of moving from one rule or task to another (Miyake et al., 2000). Since these are the EFs most discussed in the literature (see Miyake et al., 2000; Brydges, Reid, Fox, & Anderson, 2012; Willoughby & Blair, 2011), and typically seen as the "core" EFs, these are the focus of the work here.

There is also debate as to the structure of EF and whether the structure of EF during childhood is the same as in adults. Miyake et al. (2000) examined EF constructs in adults by collecting data from 137 adults who each completed nine tests (three examining each component). Through confirmatory factor analysis (CFA), to examine how data from each test/component related to one another, they found the three EF components to form one unitary construct, but to also be somewhat distinguishable from each other. Miyake et al. (2000) suggest the linkage between components could be the role of the central executive.

Garon et al. (2008) reviewed a large number of studies examining the different components of EF in young children (aged 3 to 5 years) and concluded, as Miyake et al. (2000) found with adults, that the three components are separate, but they are linked. Garon et al. (2008) worked through the research to develop a framework to account for the findings, which resulted in the data indicating that different components may develop at different times and may rely on one another to strengthen and develop. Garon et al. (2008) suggest inhibition and WM may develop earlier than shifting and acknowledge that shifting tasks must incorporate both inhibition and WM – thus inhibition and WM components must be developed before shifting skills can develop. This idea is perhaps reflected in tasks typically used with 3-year-olds: there are more tasks of inhibition and WM available than there are for shifting, as shifting is more difficult for 3-year-olds. The work of Miyake et al. (2000) and Garon et al. (2008) support one another and provides strong support for the proposal that EF components are separate, but linked.

An established theoretical model that supports the above theories is that of Diamond (2013), whose model illustrates that each EF component is separate, but that they interact with one another. Like Garon et al. (2008), Diamond (2013) suggests that inhibition may feed into shifting, and so WM and inhibition must be in place before shifting can develop. The model also illustrates that these three EF components play a role in "higher-level" EFs, such as reasoning, problem solving and planning.

However, others have found that EF components are not so distinguishable in childhood. Wiebe et al. (2011) also completed a CFA with data from 228 3-years-olds (with a mean age of 3 years and 1 month). Children completed nine EF tests, although these only examined two EF components – WM and inhibition. Wiebe et al. (2011) found the model best fitted one construct rather than two separate components. Although others suggest there are three distinct EFs, it may be that one construct emerges in some work due to WM feeding into both inhibition and shifting, thus being the main contributor of both components, which may be reflected in the tasks used. Most studies use different tests (see Diamond (2013) and Garon et al., (2008) for examples of the wide variety of tasks available), which can make comparisons difficult. Garon et al. (2008) suggest the EF components could emerge at different times, so it could be possible that results are due to the different ages in the samples (ages 3 to 5 versus 3) and that the components become more distinguishable with age.

A similar study was carried out by Monette, Bigras, and Lafrenière (2015), who worked with 272 5- and 6-year-olds who each completed nine EF tests targeting the three core EFs. Also using CFA, Monette et al. (2015) found the model best accounted for two factors: inhibition and the second factor comprised of WM and flexibility. They found inhibition and WM to be distinct (unlike Wiebe et al., 2011), but WM shared variance with flexibility, meaning it was not seen as a separate factor. They also found inhibition and WM correlated with one another, suggesting overlap or links between the two, just as Miyake et al. (2000) and Garon et al. (2008) found. As before, the varying results when examining the different EF components could be due to age or the tasks used.

There is another theory that does not suggests EF is not a set of separate or combined components, but instead is an interactive system of graded representations – this is Munakata's (2001) GR account. This interactive account suggests that different tasks tap different representations, which vary in strength and can be influenced by the situation in which the task is presented (Munakata, 2001). The situation and factors influencing the task can change the strength of the representation, which can impact the result or outcome that the task requires. This means if an infant is presented with a task that relies only on weak representations (such as an implicit recognition task) then they are likely to do well, but if they are presented with a task that relies on stronger representations (such as involving reaching or pointing to an object) then they may do less well (Munakata, 2001). Support for this theory comes from Lee and Kuhlmeier (2013) who had two-year-olds complete a computerised version of the tubes task (a test of gravity and solidity). In this task, a ball is dropped down one of three opaque intertwined tubes and the child must find it in the correct cup at the bottom of the tube. Lee and Kuhlmeier (2013) found that although the children looked to the correct location of the ball's endpoint (as measured by eye-gaze), many pointed to the wrong location. They suggest their findings could be accounted for by the GR account due to motor responses being required to solve the prediction task and processing load and EF (WM and inhibition) being required to remember information and not to point to the same location as the previous trial. In the current work it is difficult to test whether the representations held are weak or strong or whether there is a disassociation between knowledge and action. This theory would not allow for a claim of unitary or dissociable EF components to be made, as it is said to be an interactive system

It could be that the mixed findings and lack of consensus in work with children and adults are due to the overlap in different EF components being tested in the different tasks. All tasks will require some element of WM to keep the aim in mind, and maybe some form of attention and planning or thinking through problems – all of which are different elements of EF. Wiebe et al. (2011) acknowledge this issue and suggest EF could also involve other potential variables (language, motor, and visual-spatial skills, further complicating how EF skill may not be reflected in EF task performance. It may be that cognitive processes are all very overlapped, making it difficult to know for sure what is responsible for each process.

Despite more research into EF in recent years, there is still no consensus on the development of EF throughout childhood and into adulthood, so as it stands the debate between the structure of EF continues. The present study considers the three core EFs separately and measures of each were taken with the aim to examine them to see whether they appear as a unitary construct or dissociable. Thus, the results chapter will consider whether to merge them into one or two components or to keep them as three, depending on how they relate to one another here.

EF's connection to physics performance is examined here, and the next section briefly considers how EF relates to reasoning and whether physics tasks can be seen as measuring reasoning.

2.2.2 EF and reasoning

This section will look at some work that has identified links between EF and reasoning. This is being considered here, as the physics tasks likely involve an element of reasoning, due to the task requirements. Reasoning is said to involve an element of uncertainty and through making inferences a conclusion can be drawn based on the information available (Barbey & Barsalou, 2010). Barbey and Barsalou (2010) state that problem solving is the stage whereby inferences are made, and thus part of reasoning. They also state problem solving involves a stage of planning to construct a way to approach the task and that tasks requiring 'prediction and explanation' require causal reasoning and tasks that involve generating 'new predictions and explanations' require analogical reasoning (Barbey & Barsalou, 2010, p.35). Some say reasoning and problem solving are two separate components (Diamond, 2013) and some say problem solving is part of reasoning (Barbey & Barsalou, 2010). The physics tasks to be used here will require children to come to an answer, use a strategy or method, observe

information, and consider the outcome of the strategy or method. When it comes to making predictions and verbalising reasons for new problems without the aid of knowledge to guide the answer, analogical reasoning can be employed, which is using knowledge in one domain and applying it to another to come to an answer (Barbey & Barsalou, 2010). Thus, causal reasoning, problem solving, and analogical reasoning are all types of reasoning, as the physics tasks here will employ all of these skills. The physics tasks will tap the conceptual knowledge children hold concerning the physics concepts, but there is an element of tapping cognitive skills, such as reasoning, because of the task requirements.

Evidence that EF relates to problem-solving skills comes from Senn et al. (2004) who worked with 2- to 6-year-olds and found WM and inhibition were related to one another, and that together they accounted for problem-solving ability in their study (shifting did not account for additional variance in the model). Van der Sluis et al. (2007) worked with 9- to 12-year-olds and found updating and shifting were related to non-verbal reasoning, but inhibition did not emerge as a separate factor accounting for any additional variance. There are also theoretical accounts that incorporate EF as linked to reasoning and problem-solving skills – see Munakata, (2001) and Diamond (2013). The complexity and overlap between EF and reasoning and problem-solving skills may be difficult to tease apart, but due to EF links in previous work should be considered a potential factor in physics task performance.

When children develop the skill to reason has been debated, and the debate may be due to the definition and methods each researcher has used. Further, there is also research on scientific reasoning, which is reasoning in the context of science, i.e. using information, testing ideas, learning from results, and drawing conclusions (Gopnik, 2012). Gropen, Clark-Chiarelli, Hoisington, and Ehrlich (2011) would argue that children can begin to reason scientifically from as young as 3 years of age, insofar as being able to test hypotheses, which involves being able to predict an outcome to a problem, test the prediction, and make the necessary adjustments to correct the solution. Gropen et al. (2011) state that children's ability to test hypotheses is largely down to EF ability due to the need to use WM (for example, to remember the aims of the task) and inhibition (perhaps to reduce an impulsive incorrect response) to succeed on a problem. They also state that at 3- to 4-years-old children can only consider one variable at a time (Gropen et al., 2011), possibly due to EF demands (perhaps WM in particular).

Cook, Goodman, and Schulz (2011) found evidence that when 3- to 5-year-olds are provided with unambiguous evidence about variables' causal relationship they tend not to explore as much as children who are given ambiguous information about a causal relationship. These children were either shown how four beads could make a machine work or how two of the four made it work. When children were then presented with two pairs of these beads, two that could be separated and two that could not, the children who were originally shown that all four beads made the machine work were far less likely to try and split up the beads or test the different beads to see which worked, unlike those who were told that two of the beads made the machine work. The children who were shown that two of the beads made the machine work were more likely to try each bead against the machine to investigate which made it work and to split apart the pair that could be split to isolate the beads to test. Cook et al. (2011) suggest children are able to recognise what knowledge they believe to be true (accepting the information that all four beads make the machine go) and what they do not to be certain (which of the two beads makes the machine go), and will attempt to unravel the uncertain information. The behavioural differences between the two groups suggest it was not simply children playing that brought about the difference, but children acted differently through their exploratory play because of the information they were presented with and used scientific reasoning to work through problems.

In sum, reasoning likely involves some or all combination of causal reasoning, problem solving, and analogical reasoning. The focus here is not on what processes are involved when solving physics tasks, but it is acknowledged that there is an element of reasoning involved and other factors have been highlighted. The physics tasks will tap both physics knowledge, but also cognitive skills involved with reasoning, which is why EF and Mc's role will be considered when examining physics task performance.

2.2.3 EF and physics tasks

Beyond EF and reasoning, there is some research to indicate EF plays a role in physics task performance, but also research that finds no statistically significant link. Overall there is little research in this area, with some focusing on general science or scientific reasoning and not specifically physics. The mixed findings and little research are reasons to investigate EF – to see whether a link exists between EF and physics tasks, whether a link exists at all TPs, or whether it perhaps develops over time with experience.

Some evidence that supports the role of EF in children's performance on physics tasks comes from Baker, Gjersoe, Sibielska-Woch, Leslie, and Hood (2011). This work examined inhibition (delay and the reverse-categorisation task) in relation to young children's (29 to 36 months old) knowledge of solidity. The physics task used was the door task (Figure ***5***), which required the child to find a toy hidden behind a door. A wall was visibly inserted in front of one of the doors, which stopped the toy when rolled down the ramp. After controlling for age and raw receptive vocabulary scores, Baker et al. (2011) found the children's performance on the delay inhibition task showed a significant amount of overlap with their performance on the wall task, but the reverse categorisation inhibition task did not. The finding suggests that stronger delay inhibition skills relate to a decreased chance of selecting the incorrect door, perhaps being able to suppress a prepotent response allows time to select the correct door.

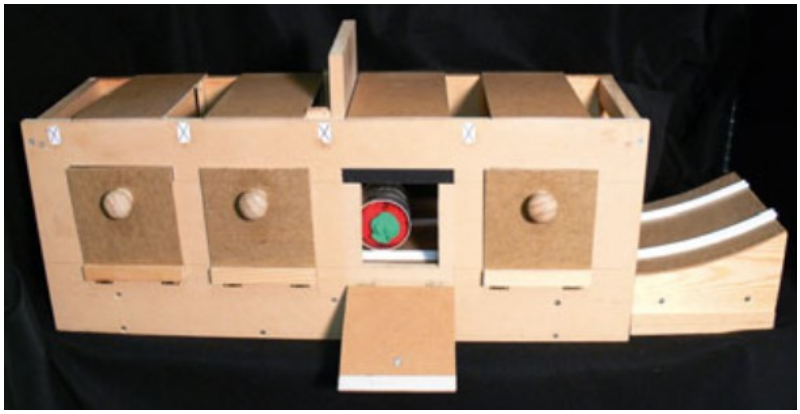

Figure 5. *Door task (Baker et al., 2011)*

In another study, working with 36- to 48-month-olds, Baker et al. (2011) used the tubes task (dropping balls down opaque tubes) along with tasks of delay inhibition and conflict inhibition. Baker et al. (2011) found significant correlations between the children's first responses on the tubes task and their performance on both the delay inhibition task and the conflict inhibition task. This again suggests that there is a link between inhibitory control and performance on these physics tasks – perhaps related to being able to inhibit an unsuccessful but frequently used strategy after a new strategy is discovered.

These studies provide some correlational evidence to support inhibition playing a role in physics tasks, but the lack of a link involving the reverse categorisation task suggests it is not all EF components. It may be due to the task requiring the use of switching in addition to

inhibition, and this task is often used as a switching task and not as only testing inhibition. It may be that particular tasks are better for measuring EF or that only particular aspects of EF are linked with physics performance, which makes finding strong evidence for such links more challenging. The pilot studies will assess different EF tasks before selecting ones for use in the main study.

Nayfeld, Fuccillo, and Greenfield (2013) investigated EF's link to science learning in 4-year-olds. Their test of science was not specifically physics, but the Preschool Science Assessment test, a test of science knowledge and content skills, assessed using a picture book requiring the children to answer/respond to the items. This was taken at two time points in the year, along with measures of maths, vocabulary, and listening comprehension. Five tests of EF were taken: spatial conflict (inhibition), pig game (inhibition), something's the same (cognitive flexibility), operation span (WM), and pick the picture (WM). They carried out factor analysis to confirm that the five EF tests loaded onto one variable, which was then used during structural equation modelling (SEM). SEM revealed that EF predicted gains in scores for science readiness, maths, vocabulary, and listening comprehension, with the strongest relationship between EF and science readiness. Nayfeld et al. (2013) suggest this could be due to the similar skills required for each, such as reasoning, and comparing, reflecting and acting on information, and therefore, science skills may also strengthen EF skills.

Evidence for a link between EF and science (again, not specifically physics) with older children has also been found. Latzman, Elkovitch, Young, and Clark (2010) worked with 11- to 16-years-olds and found various links between EF and academic performance. The EF tests used were taken from the standardised Delis-Kaplan EF System and included eight measures, targeting inhibition, monitoring (updating and evaluating of WM), and conceptual flexibility. As with many EF tests available, the inhibition tasks selected included a component of shifting, making the result less clear-cut. The measures of academic achievement were taken from the Iowa Tests of Basic Skills and Iowa Tests of Educational Development and examined reading, maths, social studies, and science. The Kaufman Brief Intelligence Test was also administered and the verbal and non-verbal results included in the hierarchical regressions. Latzman et al. (2010) found that different EF components predicted different academic strengths. They found inhibition and conceptual flexibility predicted science, conceptual flexibility and monitoring predicted reading, monitoring predicted social

studies, and inhibition predicted maths. They conclude that different EFs are responsible for different academic skills due to what the subjects involve. This finding is likely impacted by it being carried out with older children, who may have more developed EF. The tests of academic performance in science are perhaps not the best comparison to physics here, but overall the research does highlight links between EF and academic performance, including science.

Following on from this work is the findings of a link between EF and physics performance due to EF playing a role in inhibiting misconceptions, which then impacts physics performance (Brookman, 2015, cited in Tolmie et al., 2016; Masson et al., 2014). Other research has examined the relationship between EF and performance on physics tasks and found no significant link. Tolmie (2014, cited in Tolmie et al., 2016) found that 5- to 10-year-olds' understanding of freezing, melting, evaporation, and condensation showed no link to their EF. However, some evidence of a link between semantic inhibitory control and children's descriptions of two of the concepts (freezing and melting) was found, which was attributed to children inhibiting certain information in order to produce accurate descriptions of the processes. It instead suggests a possible link between language and EF, which is then reflected in physics performance.

These mixed findings regarding EF and physics/science could be due to several factors, including the age of the children, the tests of EF used, the tests of physics/science used, and the overall lack of consistent research available to draw conclusions. It could also be influenced by how many variables (and maybe, therefore, the cognitive resources required) children must consider during the physics tasks. The research in this section used a variety of methods and analyses, which may also contribute to the mixed results. The current work considers EF's relation to physics performance, whether children can reason with different variables, and what progress is made over several sessions. The analyses focus on means, group differences, and correlations, so it includes several of the analyses described. It is expected that greater EF skills will be related to better performance on the physics tasks, although if this is found it may have implications for training EF skills in young children or teaching physics concepts or science skills.

The next section will discuss Mc, which in the present study was considered in a similar way to EF to see whether links to physics performance exist.

2.3 Metacognition

This section will outline what the structure and function of Mc are said to be, some measurement methods, the importance of examining strategy development, and the possible link Mc has to physics task performance.

2.3.1 Structure and function of Mc

Mc is thought to comprise of different components, with most agreeing to the components, but some disagreeing over the labels. The definitions stated by Flavell et al. (2002), Zohar and Barzilai (2013), and Whitebread et al. (2009b) are used here since they encompass the parts of Mc that are often cited in the literature. Flavell et al. (2002) state that Mc comprises of *metacognitive knowledge* (*MK* – the knowing and understanding of strategies, tasks and people), *metacognitive monitoring and self-regulation* (*MS* – the planning, monitoring, control and evaluation of thinking and learning), and *metacognitive experiences* (*ME*). Zohar and Barzilai (2013) state that many label Flavell's definition of MS as *metacognitive skills* (also *MS*). Whitebread et al.'s (2009b) Cambridgeshire Independent Learning (C.Ind.Le) measure refers to *MK* (knowledge of persons, tasks, and strategies, in agreement with Flavell et al., 2002), and *metacognitive regulation* (*MR*), which includes planning, monitoring, control, and evaluation. MR is the equivalent of Flavell et al.'s (2002) definition of MS and Zohar and Barzilai's (2013) definition of MS. Since the C.Ind.Le coding system is used in the present study, the labels MK and MR are used. The C.Ind.Le also includes *emotional and motivational regulation*, similar to ME, which are the links to experiences and affect (Zohar & Barzilai, 2013). Since emotional, motivational, and affect are not a focus in the current work, ME will not be discussed or assessed – only MK (metacognitive knowledge) and MR (metacognitive regulation) were measured (see Appendix A for the components, as measured by the C.Ind.Le). The C.Ind.Le provides Mc rate – the rate of Mc per minute, calculated from the total Mc behaviours recorded divided by the length of time of the observation.

As can be seen in the definitions, self-regulation is encompassed within MR, although they have often been referred to separately in past research. In more recent literature they are usually considered together, such as in Whitebread et al.'s (2009b) work in developing the C.Ind.Le to measure Mc in young children. This is due to the considerable overlap and interaction between metacognition and self-regulation, which comprise of components focused on the individual's thoughts, knowledge and learning, the reflection of such

processes, how changes can be made and applied, and considering the outcomes made. The focus is on MK and MR since the work concentrates on children's knowledge of physics, the strategies used to solve physics problems, and children's ability to monitor their behaviours during the tasks and adjust accordingly. It is said that MK and MR are constantly updating one another through experience (Roebers, 2017), so it is expected links will be seen between these and to physics tasks, with the potential for change in the relationships over time. However, as there is little research with young children it was unclear whether a distinction would be seen between MK and MR and their subcomponents. The data were analysed for a decision to be made on whether the data can support the components or subcomponents being analysed separately or whether Mc should be analysed as a whole via total scores.

When Mc develops has been debated, but some suggest it has already started to emerge by age three (Kuhn, 2000). Roderer and Roebers (2014) suggest monitoring and control do not develop until primary school age and that monitoring skills may precede control skills, and Roebers (2017) states that monitoring and control skills are very separate, even in older children. Most would agree that Mc continues to develop with age (Veenman, Van Hout-Wolters, & Afflerbach, 2006) and experience (Flavell et al., 2002). Some suggest the apparent growth in Mc during childhood is due to the methods used to assess Mc (Paulus, Proust, & Sodian, 2013), as assessing Mc in young children is difficult. Some believe that Mc in young children may not be accurate due to children often overestimating their abilities and knowledge (Roderer & Roebers, 2014), thus links between Mc and performance would not be expected. Due to the lack of consensus and the difficulties in measuring Mc in young children the research is lacking, particularly in relation to performance on physics tasks.

Assessing Mc in older children is easier as verbal reports and questionnaires are available, so more research has been done with older children. Measures of Mc with 3-year-olds are problematic due to the methods used – if verbal reports are used it could result in a poor Mc score, due to it relying on the child having a well-formed expressive vocabulary. Whitebread, et al. (2009b) have shown that observation is a possible method for assessing Mc behaviours in young children and have developed the C.Ind.Le based on observations of children within the 3- to 5-year age range. Whitebread et al. (2009b) coded children's displays of Mc (speech and non-verbal behaviour) during problem-solving tasks and concluded observation is a viable Mc measurement method. Others have since used the C.Ind.Le, such as Robson (2016) with 4- and 5-year-olds, and stated its usefulness. Robson (2016) also used a reflective

dialogue task, using the videos coded during observation and found it to be a worthwhile measure as well. They also found that over a 10-month period, children's Mc scores during the reflective task increased, suggesting that practising talking about their cognitions became easier (Robson, 2016), although it could be the children developed the language to talk more effectively about their cognitions. They did find differences between the two measures though: the observation resulted in higher scores of MR for the children, but the reflective task resulted in higher MK and ME scores. Robson (2016) suggests two tasks together could provide a more accurate measure of Mc, rather than relying on only one, apparently acknowledging the difficulties.

There is debate as to whether Mc is responsible for strategy use and development, although those working in the Mc field, and as seen in the definitions above, say that Mc controls strategy use. The next section will consider strategy use and the debate on what may be responsible before considering how Mc may play a role in physics tasks.

2.3.2 Strategy use

A strategy can be defined as a method employed to achieve a goal (Roberts, Taylor, & Newton, 2007). Strategies can be taught and learned, but also learnt implicitly through interactions with the world (Jansen & van der Maas, 2002). Kuhn (2000) states that over time strategy use changes with feedback from each strategy selection and through this process metacognitive skills become more conscious and leads to better strategy selection. This suggests that children should be able to understand and explain why they chose to solve a problem in a particular way, as well as be able to use feedback from errors to improve future strategies, thus improving performance. Siegler and Stern (1998) agree that strategy selection may start out as an unconscious process but develop into a conscious process, so it was hoped here that through experience on the balance beam strategy development could be tracked.

One question here is: what is it that is responsible for selecting and implementing a strategy? Some say problem-solving skills include the ability to select a strategy (Barbey & Barsalou, 2010), but some say that Mc is responsible for strategy use (Zohar & Barzilai, 2013). Some would say that problem-solving skills can be defined as a higher-level EF (Diamond, 2013), therefore implying EF could be responsible. Some have even found links between Mc and EF (Bryce, Whitebread, & Szucs, 2015), so the connections between Mc, EF, and strategy use should be explored further.

One methodological problem is the issue of examining strategies used by young children, rather than just observing the solutions they have used. Devising methods purely to measure what strategy has been employed to solve a physics task is beyond the scope of this work, but examining the solutions could give some insight. Jansen and van der Maas (2002) say that children's answers and performance on the balance beam task can be assumed to be a reflection of the different strategies employed. The solutions used by children during the physics tasks in this work were therefore considered a reflection of the strategies employed and will be referred to as such throughout this work. For example, during the balance beam task, this could be placing the same weights at the same distance or placing all the weights on one side of the beam; for the ramps task this could be setting one high incline carpet ramp and one high incline wood ramp and rolling one ball down each ramp or setting one high incline carpet ramp and one low incline wood ramp and rolling both balls down the same ramp. By observing the different children they solve the trials over time, when they first correctly solved a trial, and when they consistently solved a trial will aid the examination of strategy development over time. Microgenetic designs do just this by obtaining multiple data points over longer periods of time in order to track changes. The current work used three distinct TPs, so it was not possible to carry out a microgenetic analysis, but some of the microgenetic analysis methods are used here, especially when considering strategy development, which will be discussed next.

Siegler and Svetina (2002) used a microgenetic approach to examine non-verbal reasoning to demonstrate that even when children discovered the correct strategy for a problem they sometimes reverted to an old strategy in the next session. Strategy development is not stable or predictable, and by examining change over time a more in-depth analysis can be carried out by measuring the strategies used in each trial at each TP. Siegler and Stern (1998) also used a microgenetic design to investigate children's strategy use to solve maths problems and found children tended to try multiple strategies at each TP, but there was a pattern of strategy advancement in the groups. The idea of multiple strategies being available for use supports Siegler's overlapping waves (OW) theory (Siegler 1996), whereby numerous strategies (some better than others) are available and can be used to try and solve the same problem, as opposed to a staircase model where strategy use generally improves and only certain problem types are seen to be solved at a time. Siegler (2016) suggests it is through experience that children select better strategies and the OW theory also states that new strategies can be

developed through such things as instruction and experience in solving different problems (Siegler, 2016).

There are several theories as to what is responsible for strategy development and use, with much debate still surrounding whether it is cognitive or metacognitive, explicit or implicit, conscious or unconscious, or whether it is a mixture of these depending on the situation. Some theories include the idea that cognitive styles (individual differences) determine strategy use. Support for this comes from studies with adults that showed verbal skills and visual-spatial skills may have a role in which strategy is selected during a task, based on individual strengths, as measured by solution times and error rates (Roberts et al., 2007). However, Bacon, Handley, Dennis, and Newstead's (2008) work with adults using tasks that tapped verbal and visual reasoning concluded that the strategy used during the task did not correlate with reasoning performance. This work also indicated that EF and Mc were not contributors to the strategy selection process and Bacon et al. (2008) suggest that conscious Mc is not a contributing factor in strategy selection. The opposing leading theory for what is responsible for strategy use is that Mc is responsible, as stated earlier, with much support coming from the Mc field. What is responsible for selecting and implementing strategies is explored more in the sections discussing EF and Mc and physics.

Devising a method to determine what is actually responsible for strategy development and use is beyond the scope of this work, but measures of vocabulary, visual-spatial skills, EF, and Mc were taken from the children in order to see whether links to strategy use can be found. If links from physics performance or strategy development to vocabulary or visual-spatial skills are found it might support the idea of individual differences playing a role in strategy selection, if links to EF are found it might support the idea that EF is responsible, and if links to Mc are found, it might support the idea that Mc is responsible for strategy use.

If certain measures are found to be linked to children's strategy use it could have implications for promoting and training the responsible components if a benefit is seen. The next section will focus on Mc's possible role in physics tasks and the section after that will consider the relationship between Mc and EF.

2.3.3 Mc and physics tasks

There is little research examining Mc alongside performance scores in young children and no research examining physics performance was identified. This could be due to the methodological issues of measuring Mc or that research tends to focus on other subjects, such as maths or reading, or focuses on older age groups. At this time there are no studies to report examining Mc in 3- or 4-years' old ability to solve physics tasks, but one other study with older children will be reported.

Rozencwajg (2003) worked with adolescents to examine how Mc relates to physics. 12- to 13-year-olds completed a crystallised intelligence task, as measured by a maths test and a sentence completion task, and a fluid intelligence test, as measured by a matrix task. The children were observed completing a 42-item paper and pen physics task examining electricity concepts (to examine Mc). MK was measured using a five-item questionnaire examining strategy use and monitoring (part of MR) was examined via response latency times in a reflection-impulsivity computer task. Rozencwajg (2003) found that MK was related to crystallised intelligence and monitoring was related to fluid intelligence. She also found the strategy types used by the students during the physics problem-solving task showed differing relationships with MK, for example, one strategy was linked to high MK, but another strategy used by some students (which also resulted in the correct answer) was linked to lower MK scores. Rozencwajg (2003) suggest students' MK could determine which strategy is used and it could therefore be worth either trying to strengthen MK or teach children strategies to help improve performance. Examining strategy use in the current study is therefore worthwhile, as it could be linked to Mc or other measures.

An important reason Mc is being considered in the current work is because it is thought to be responsible for strategy selection and development, as discussed earlier. Mc was measured so it could be examined against physics strategy use to see whether any links exist. Mc is thought to have a link with EF, so examining EF was important here. The lack of work in the field is a third reason to consider Mc. Taking an accurate measure of Mc could be challenging, but this is addressed in the methodology chapter. It is expected that stronger Mc will be associated with better strategy use and therefore performance. The potential link between Mc and EF will be examined next.

2.4 How are EF and Mc related

As touched on already, there is some suggestion that EF and Mc are linked. One of the main reasons to consider EF and Mc here is due to the behaviours and strategy use that children should implement during the physics tasks. The overlap between EF and Mc is of particular interest, with some describing each as part of self-regulation (Roebers & Feurer, 2016), complicating teasing apart what might be responsible for strategy use, as well as obtaining accurate measurements. Anderson (2002) believes planning and strategic organisation are part of goal setting, one of four EFs he identified. Jurado and Rosselli (2007) list several authors who claim strategy control / monitoring / generation / and implementation are part of EF. Roebers and Feurer (2016) however seem to suggest that strategy selection and development is through Mc, but implementing a strategy is through EF. However, those in the field of Mc would argue these are Mc skills. Overall, there is no consensus on how EF and Mc might be connected, and if they are, whether it is due to the overlap in what EF and Mc are each responsible for.

Diamond's (2013) theoretical model states inhibition and self-regulation are intertwined, with self-regulation responsible for response and attention inhibition, and maintaining ME (motivational and emotional arousal), as well as cognitive arousal. Roberts and Erdos (1993) state Mc is responsible for strategy selection, but EF could be responsible for implementing it. When selecting a strategy to use to solve a problem there may be several available, so they must be considered and one must be selected and applied – this is perhaps when EF plays a role in strategy use (Roberts & Erdos, 1993). There could be a link between EF and Mc through EF playing a role in strategy selection by implementing inhibitory control to select from two competing strategies (Kuhn, 2000). For example, it could be that better EF skills aid Mc and strategy selection since WM could aid being able to consider different strategies, inhibition could aid inhibiting less useful strategies, and shifting could aid choosing different strategies in a task with numerous trials which each require different strategy responses. It may be that greater EF skills aid Mc and strategy use, perhaps through WM providing the resources to consider different strategies, shifting being able to switch between strategies, and inhibition inhibiting incorrect strategies. However, high EF skills do not guarantee a high Mc score, since there could be a failure in either EF or Mc, resulting in poor performance.

Bryce, Whitebread, and Szucs (2015) investigated the link between monitoring and control behaviours (coded during a task), EF (inhibition and WM), and academic achievement in 5-

and 7-year-olds. They found some significant correlations between inhibition and monitoring, although more so for the younger children, which they suggest is due to developmental changes occurring with age (Bryce et al., 2015). However, it would perhaps have been expected that older children show a stronger link, based on developmental changes and previous research findings. This finding suggests that not all of EF relates to all of Mc and instead it could be subcomponents of each that are related or it is task-specific, which does not allow for claims of strong links between EF and Mc to be made. García, Rodríguez, González, Álvarez, and González (2016) also found some links between EF and Mc in 10-12-year olds. However, their measure of EF was a behaviour rating scale completed by teachers and families and the measure of Mc was a self-report by the students. Students who were rated as having better EF scored higher on the Mc measures, suggesting a link exists, but the measures used do not allow for this to be seen as strong evidence. Roebers, Cimeli, Röthlisberger, and Neuenschwander (2012) worked with 7-year-olds over the course of a year who completed EF measures at the start of the study and EF and Mc measures at the end. Using SEM they found that EF at the start was related to Mc control at the end of the study, and EF and control measures at the end of the study were related, which again highlights that only some components of EF and Mc show significant links.

Spiess, Meier, and Roebers (2016) also investigated the link longitudinally in 8-year-olds over an eight-month period, in which the same tasks were completed at the start and end. Their measure of Mc was a spelling task in which children were asked to rate how confident they were in their answers (monitoring) and to be given the chance to change their answers if they so wished (control). Spiess et al. (2016) found performance on the measures improved over time, but the EF and Mc measures did not significantly correlate either within or over TPs.

Due to previous research finding no strong evidence on whether EF and Mc are linked, it seems necessary for the current work to take multiple measures at multiple TPs in order to examine the relationship as best as possible. Even considering differences in age, strength of the link, or the tasks used, it is difficult to draw any conclusions. Based on some supporting literature it could be expected that a higher EF score will relate to a higher Mc score, and also to a higher physics performance score. Higher Mc and EF scores could provide an advantage with task and strategy knowledge, the ability to work with feedback to adjust future responses, and the ability to apply the behaviours. It may additionally be found that EF plays

a role between Mc and performance, as even if a child has a high Mc score, they may not be able to implement Mc because of an EF failure. It would logically seem plausible that these two cognitive functions, which both control and apply cognitions and behaviours, would be linked in some way. The analyses to be used will try to account for the various possible directional relationships between these factors.

The literature appears quite mixed, with some finding no significant link and others finding a link between EF and Mc, sometimes between particular components only. The mixed findings could be due to the link between EF and Mc changing with age, that the link is just not very strong, or it is very dependent on the tasks used.

The next section will consider the potential role of vocabulary and visual-spatial skills in this work.

2.5 Vocabulary and visual-spatial skills

As referred to throughout this chapter, vocabulary and/or visual-spatial skills appear to play a role in some of the research. Vocabulary and language are likely important for understanding the tasks, for scoring well on verbal Mc measures, and for some of the balance beam classifications that use verbalisations. Visual-spatial skills, a measure of non-verbal skills, have been found to predict scientific reasoning skills in children (Mayer et al., 2014). There is an element of reasoning involved in visual-spatial tasks, so it might be this skill influences physics performance, thus it should be accounted for. It is therefore worth examining whether vocabulary and visual-spatial skills contribute to physics task performance in young children, to perhaps include as covariates if they are found to be influential.

Van der Graaf et al. (2016) found non-verbal reasoning mediated the link between attentional control and the number of attempts needed before solving the trial on a computerised physics task, and vocabulary mediated the link between attentional control and the number of actions carried out during the task. Van der Graaf et al. (2016) acknowledged that vocabulary could be important for thinking through the concepts involved in the task and non-verbal reasoning could be influential since the physics tasks themselves require reasoning. Roberts et al. (2007) found that strategy use was influenced by individuals' differing strengths in verbal and visual-spatial skills, which impacted performance at times. Wiebe et al. (2011) state in their work that EF has a role in language and non-verbal skills and Mayer et al. (2014) state

that it is important to consider verbal and non-verbal abilities whenever examining scientific reasoning skills. All of this work supports the need to measure vocabulary/language and visual-spatial/non-verbal skills in the current study to assess whether either has a role in physics task performance.

Weiland, Barata, and Yoshikawa (2014) investigated the link between (receptive) vocabulary and EF (inhibition, WM, and shifting) in 4-year-olds. They followed 400 children longitudinally overly nearly six months as children completed vocabulary and EF tasks at the start (time 1) and end (time 2). Weiland et al. (2014) used cross-lagged SEM and found EF at the first TP predicted vocabulary at TP2 (controlling for vocabulary at TP1), but the reverse was not found – vocabulary at TP1 did not predict EF at TP2. They suggest this could be due to the increased vocabulary during this age group, as also reflected by vocabulary at TP1 being a weaker predictor of vocabulary at TP2, compared to EF at TP1 being a strong predictor of EF at TP2. Weiland et al. (2014) broach the idea that EF plays a role in vocabulary development – a slightly different reason why the two might be linked. However, they do acknowledge that their finding could be due to using receptive vocabulary, rather than expressive, and that their sample included children who had experience of other languages. Either way, the finding made it worth investigating to see whether links can be found, so vocabulary was measured in the present study.

There is also evidence to suggest language ability is related to inner speech (see Cragg and Nation, 2010, for a short review). Inner speech is said to go through a lot of development from around age 3 and plays a role in thinking through actions/requests, behaviours, and attention (Cragg & Nation, 2010). This is around the same time that vocabulary starts to steadily increase, so it is logical to think that this impacts inner speech. It is suggested that inner speech can help keep track of and guide rules, actions, and strategies, although it is not the only factor involved in such events (Cragg & Nation, 2010). These kinds of processes could be viewed as Mc in nature and relevant to the tasks to be used in the present study. Therefore, vocabulary was measured here and against other variables.

Wagensveld et al. (2015) carried out a study looking at the role of instruction versus discovery in physics CVS use in 9-10 year olds and 11-12 year olds. They found each group (instruction and discovery) had different predictors for gains in CVS scores: vocabulary, verbal reasoning, and reading comprehension predicted gains in the instruction group, and

verbal reasoning and reading comprehension scores predicted gains in the discovery group. It is therefore worth considering vocabulary and visual-spatial skills as possible influential factors in the current study and as having potentially different roles in each support group. DI involves more verbal and visual explanations than GP, so the different skills many impact performance scores.

There is also evidence that suggests prior language concerning a subject could have a role in how well or quickly children learn about the subject. Ghazali (2014, as cited in Tolmie, Ghazali, & Morris, 2016) found expressive and receptive language in 4- to 11-year-olds related to their understanding of science concepts on the topic of biology. Language was not seen to influence performance on the science tasks but instead, performance was mediated by language, with evidence that language related to the task facilitated performance. Philips and Tolmie (2007) found in their study with 8-year-old children who completed the balance beam task with the parents' support that children only benefitted from the parents' descriptions of balance concepts if they (the children) already had some understanding. This idea was also presented in Siegler and Chen's (1998) work, as they found that the knowledge the children in their study held at the start of the physics task likely contributed to their learning during the task. It could, therefore, be important to consider children's vocabulary in the present study, as perhaps those with a wider vocabulary are more prepared for such tasks, instructions, and know the physics terms. Tolmie et al.'s (2016) work supports the idea that language is key is driving implicit knowledge to become explicit, supporting the RR model, to be discussed later. Language can be seen to provide a way to organise concepts, so different concepts can be brought together by the appropriate and necessary language, which vocabulary perhaps provides to children.

Due to various areas of research indicating a possible role of vocabulary and/or visual-spatial skills, both were measured in the current study to see whether links between the variables exist and to see whether they need to be accounted for during the analyses.

The next section will consider the two support types used in the study and what role each may have on the measures to be examined here.

2.6 Support type

Two support types were implemented to see if they impact performance on the physics tasks or any of the other measures. The two types of support to be used here are GP and DI – both frequently cited in the literature with some opposing findings. Support type is important since the method by which children are taught will likely impact their learning, and of interest here, if support type interacts with other cognitive factors (such as EF and Mc). The Department for Education (2017) encourages that children learn through play and are given the freedom to explore tasks themselves to solve problems, but also states there should be a mixture of adult-led and child-led activities. The current work acknowledges that there is a large body of research showing play can be beneficial for children's learning, but there is also research showing that when it comes to physics tasks, instruction-based and adult-led learning is best, which is the focus of the work here. This work will examine whether DI or GP is the best support type for young preschool children to learn about physics and whether links to other measures can be seen which may support promoting one support type over the other. Research for each side of this argument is discussed next.

GP is defined as the child deciding the direction the task takes, although the adult initially sets the structure of the task through setting goals to keep the child on-task but the child is free to achieve these goals through their own exploration (Weisberg et al., 2013). The adult encourages the child through "commenting on their discoveries, co-playing along with the children, asking open-ended questions about what children are finding, or exploring the materials in ways that children might not have thought to do" (Weisberg et al., 2013, p. 105). This is more than just encouragement – the adult's role is to ensure the child stays focused on the goals, but in such a way that the child still decides the direction this takes. DI is defined as the adult controlling the content and structure of information presented to the child while the child listens and does what they are told, rather than directing their own behaviour during the task (Weisberg, Hirsh-Pasek, & Golinkoff, 2013). Weisberg, Kittredge, Hirsh-Pasek, Golinkoff, and Klahr (2015) state the main difference between DI and GP is who directs the task (the child or the adult), but in both support types the adult initiates the task. In the current study, these ideas were followed, alongside each group receiving different feedback. GP were asked if a solution worked or not and why, and DI were told whether their solution worked or not and why.

Some say GP is a more effective way than DI for children to learn because it is more engaging for the child (Weisberg et al., 2013). A study by Fisher, Hirsh-Pasek, Newcombe, and Golinkoff (2013) showed GP to be a more effective teaching method than DI. Fisher et al. (2013) investigated 4- and 5-year-olds' learning of shapes when in one of three support conditions: GP, didactic instruction (the equivalent of DI), and free play. The children who received GP support were asked to help the adult investigate shapes' secrets and the adult and child each wore detective hats before learning about (typical and atypical) shapes. The adult first explained a little about what they were going to do and named the shapes on the cards in the process. The child was asked if they could find out what made some shapes the same and the adult encouraged them through prompting (asking about how many sides they had) and asking the child to remind them why some shapes were the same. After this, the child was required to make some shapes from construction sticks provided and explain why they were the same as the ones they had just discovered. In the didactic instruction group, only the adult wore the detective hat, as they were the one in charge of exploring. The adult explained the shapes to the child (including such information as how many sides the shapes had) and then went on to make the shapes from the construction sticks while explaining why they were the same as the shapes on the cards. The free play group was allowed to play with the cards and construction sticks however they wanted. After 15 minutes of working with the shapes and sticks, all children completed a shape-sorting task to test how much they had learnt.

The results showed that children in GP performed best, as they recognised more typical and atypical shapes and knew when one was not a real shape (i.e., there was a part missing). The didactic instruction group performed well on the typical and incomplete shapes but tended to say the atypical shapes were not real, thus the adult instruction had not been enough for them to understand what made them shapes (Fisher et al., 2013). The children in the free play condition performed poorly, even on the typical shapes, and Fisher et al. (2013) suggest it could be due to the children not knowing the aim of the task – although the children played with the materials, their focus was not on the shapes. Fisher et al. (2013) have shown that even if children receive the same information (as in the didactic and GP groups), the way it is done has an impact on the child's learning. GP elicits more interest from the child, can result in solo discoveries, requires thinking in order for the child to answer the questions and prompts, is more fun, and it is more likely to keep the child engaged and thus more learning will occur. These aspects of GP give more opportunities for children to implement EF and Mc, as the child must direct their own behaviour and thinking, remember what they have

learnt, and repeat, explain and show what they have learnt. This distinction between the support types demonstrates how support type could impact on children's learning. Fisher et al.'s (2013) study has been widely cited as providing evidence for GP over DI. Fisher et al.'s methods were used as the basis for the support types in the current study and will be discussed in more detail in the methodology chapter.

Others, however, have found support type that includes instruction to be better when teaching physics concepts. Chen and Klahr's (1999) study into CVS use found that only the children who received explicit instruction showed an improvement in CVS use, and only the oldest children showed a benefit from the probe questions. Perhaps the older children benefitted from the probe questions because they tapped Mc and the older children had stronger Mc and so were able to make use of it. This finding makes it difficult to tease apart whether explicit instruction alone or only when teamed with probe questions is most effective, but it was seen probe questions alone were not enough. DI does not use probe questions and so questions were not used during DI support in the current study, but questions were used during GP support. This helps to examine the impact of DI (with no questions) and GP (with questions).

Chen and Klahr's (1999) finding corresponds with Wagensveld et al.'s (2015) results, as they found children (aged 9- to 12-years-old) who were taught CVS via instruction showed more improvement in scores than the children who were allowed to try and learn the CVS concepts by themselves (the discovery group, which is similar to GP since children had to design their own experiments and received no feedback, but dissimilar to GP since they also received no prompts). The instruction group received explicit explanations concerning CVS while also viewing how the ramps were set up to see CVS being properly implemented. The discovery group were given a worksheet to try and work through some CVS set-ups by themselves but were not told the aim of the task or their attention brought to the idea of CVS use. This is somewhat similar to the DI and GP support types implemented here – DI heard explanations and saw the problems being worked through, but GP were asked to work out some solutions for themselves, however, each group knew the aims of the task. These findings are also supported by Klahr and Nigam (2004) who investigated DI and discovery learning in 8- to 10-year-olds and found the DI group performed better on CVS use from pre- to post-test.

An important consideration in the support types is how tasks are explained, so the language used needs carefully planned. This could possibly be a contributing factor in some of the

opposing findings, as simple language is crucial with young children, especially if introducing new words. Muentener and Schulz (2012) considered how the language used during children's causal reasoning tasks might influence how children interpret events. They carried out a series of experiments with 3- and 4-year-olds and found that the language used when explaining that a particular action causes an event influences children's performance. The children were filtered to ensure only children who first understood that a moving block caused a toy to move were included in the study. Muentener and Schulz (2012) then tested the impact of language when showing and explaining causal events (i.e., x causes y). They found children who saw the event and heard an explanation utilising causal language were more likely to go on and move the block themselves, compared to children who saw the event and were only told to watch what happens. Muentener and Schulz's (2012) results show that the language used may play a role in children's understanding of causation, despite children seeing the same event. This is an important aspect of the DI and GP support here – DI children heard the reason for something balancing, but the GP children did not. The DI children saw problems along with an explanation, but the GP children only saw the problem outcomes when they moved the weights onto the beam themselves. This could add a further factor between the two groups' performance and is another reason to consider the children's language ability, as it could impact their understanding. The language and explanations used in each support type are explained in detail in the methodology chapter.

Although GP has been given more attention in recent years, there seems to be advantages and disadvantages to both GP and DI. In this work, during GP the children played with the materials themselves to learn things on their own, so they may have been more likely to stay interested and attend to the task. However, learning through GP is only be effective if the child discovers particular aspects of the materials or follows the prompts given. The questioning and prompts should solidify their learning (if responded to), as the children should think about what the adult is saying and they should verbalise their thoughts and explanations about the task. The questioning and prompts (such as why events happen and why trials ended as they did) will tap into Mc and the children could strengthen their MK and MR and subsequently modify their strategy use. On the other hand, DI were provided with all the information needed in order to succeed on the task, but it was perhaps less engaging for the child. However, Chen and Klahr (1999) found that explicit instruction was needed for children to improve on such tasks – and here children had all the information they needed to solve the problems. The lack of questioning and prompts to think of other ways to explore the

task may have resulted in less learning for the child through fewer opportunities to engage with Mc. However, children were given feedback on incorrect trials and this could have provided opportunities to think about what they could change in order to find the correct solution. A consequence of GP support was that children could score higher on the observational Mc measure due to being asked about the task – this issue will be addressed later.

Previous research does not discuss the impact of support type on Mc or EF, but as suggested, GP could aid Mc through questions, although only if the children respond to them, but DI could also aid Mc if children reflect on feedback. With reference to EF, Barker et al. (2014) found that children who spent more time in less-structured activities tended to have better self-directed EF (identifying their goal and directing their behaviour to achieve it), likely due to experience in managing their own behaviour in this way. However, the GP support here may be too structured for it to impact EF, as the adult should try to keep the child focused on the goal. Barker et al. (2014) calculated all the activities children took part in and the structure the activities took for these measurements. In comparison to the tasks used in the current work, it will be a relatively short time and therefore less likely to impact EF, but if there are benefits to EF from the support it is expected to be for children in the GP condition.

Overall, it seems that some research indicates GP could be a better support type for young children, but others say when it comes to physics and science tasks that DI is better for learning. The study will consider both arguments and examine the links from each support type to EF, Mc, and physics performance.

The next section will attempt to draw together some theoretical accounts and predictions to solidify what each theory can and cannot account for in the areas examined.

2.7 Theoretical accounts and predictions

The focus of the current study is whether children can solve physics problems, whether any conclusions can be drawn about whether they understand balance concepts, whether links to EF or Mc exist and if they help account for balance performance scores, and whether links to a transfer physics task exist. This section will discuss theories that could help account for work here and the predictions of each. Some theories have used prediction data and/or

verbalisations and/or other variables to explain their theories. This means there may not be sufficient data here for the theories to be able to predict and account for the data in this study.

Four theories have been presented already, two of which focus on using rules and strategies to categorise children's performance: Halford et al.'s (2002) work and the relational complexity theory, and Siegler's (1976) work and his OW theory (1996). The third theory was Munakata's (2001) GR account, and the fourth was Diamond's (2013) EF account, which incorporates problem-solving and reasoning. Two more are presented here: Karmiloff-Smith's representational redescription (RR) model (1992), which classifies children based on their displayed level of knowledge (strategies and verbalisations), and Schapiro and McClelland's (2009) connectionist model. Of these six theories, they are somewhat divided into staircase or stage models and continuous or connectionist/interactive models. Halford et al.'s (2002) work is a staircase model, Karmiloff-Smith's (1992) is somewhat staircase (but allows for regressions), Munakata's (2001) GR account and Schapiro and McClelland's (2009) are seen as continuous (connectionist/interactive), the OW theory sits between staircase and continuous, and Diamond's (2013) account does not fit either. Staircase models suggest strategy use (and thus knowledge) tends to improve over time (with experience), with little to no regression to earlier strategy use (unless between stages), and that one main strategy/rule is mostly relied on at a given time. The RR model is slightly different since some regression can be accounted for. Continuous models also suggest strategy use tends to improve over time (with experience), but there are a variety of strategies available to use at a given time, and which strategy is used depends on various factors. Next, each theory will be considered in turn and predictions from each will be presented.

2.7.1 Diamond (2013)

Diamond's (2013) theory concerns EF and how it links with Mc and problem-solving skills. Diamond suggests the EF components are separate, but Mc (self-regulation) has an interactive relationship with inhibition. She suggests all of these feed into the higher-level EF, which are reasoning, problem-solving and planning. This model would therefore predict that Mc and inhibition, and EF and physics performance would show positive connections. Diamond (2013) does not discuss knowledge systems, what knowledge children might display during the balance beam task, language, or support type, so no claims regarding improvement can be made based on her theory.

2.7.2 Halford et al. (2002) and the relational complexity theory

As already described, Halford et al. (2002) suggest four strategies that children can be classified into, based on what problems they can and cannot solve. The relational complexity theory includes the claim that by two years of age children can solve problems involving one variable. Their work suggests a staircase model of learning, whereby children can only solve particular problems at one time, because they rely on particular strategies in their repertoire at that time, based on how many variables they can solve at once. This means a child's performance typically shows consistency in the problems they can and cannot solve, with some improvement seen over time through experience. This experience could include instruction and feedback, so it could be predicted that DI shows faster improvement, however, GP have time to work with the materials themselves, so this experience may be just as beneficial. Due to the classification system used, Halford et al.'s (2002) strategies are best suited to prediction tasks, since it includes reference to predicting trials with one or two variables (weight and/or distance) and conflict trials (that will not balance). Halford et al. (2002) make no reference to knowledge systems, how children might display implicit and explicit knowledge, language, or visual-spatial skills, so no predictions are made about these.

2.7.3 Karmiloff-Smith's RR model (1992)

This model suggests learning occurs in a staircase fashion, whereby children use particular strategies at a given time, showing mostly consistency or improvement, with little regression to previous strategies once the correct strategy has been found, unless a child is testing a strategy to see if it is an improvement. Karmiloff-Smith's RR model (1992) also classifies children into four categories based on what they can and cannot solve, in conjunction with verbalisations made. Karmiloff-Smith (1992) suggests that knowledge is first acquired through interactions with the environment and this is what builds up the representations. Changes occur through representations of individual concepts undergoing redescription – a process involving feedback and experience with the concepts, which compares already held information with newly gathered information. This helps account for children being able to show knowledge of different properties at different ages: as redescription takes place, children are likely to ignore outcomes that challenge the representations they have, but after enough opposing evidence their representations undergo redescription to fit with the evidence they have accumulated.

Karmiloff-Smith (1992) states that knowledge starts as implicit (level I), and then becomes

conscious (level E1), then explicit (level E2), then can be accessed and verbalised (level E3). Karmiloff-Smith (1992) states knowledge must be held at level E2 – explicit – in order to pass prediction tasks, and not necessarily at the point of being able to access and verbalise the knowledge (level E3). This means that children who hold representations of information at different levels may show the same pattern of behaviour, but for different reasons, which complicates classifying children somewhat. For example, in a balance beam production task with 4- to 8-year-olds, the youngest and oldest children both performed well, but the 6-year-olds did not (Karmiloff-Smith & Inhelder, 1974-5, cited in Karmiloff-Smith, 1992). This task involved a wood block that could be moved along another wooden block until it balanced (Figure 6). The top wood block could differ in weight, so it was sometimes at the end of the block, sometimes in the middle, and of varying distances in-between. She explains this finding as being due to the 4-year-olds using representations only available at level I which relies on visual feedback from the task to solve the problem, but success is through trial and error and children cannot carry information from one trial to the next resulting in again watching the beam as they try to make it balance and adjusting accordingly. However, the 6-year-olds were using representations available at level E1, which includes information about how a balance beam works – that centring the weight is the best option. This resulted in failure when uneven weights were given and furthermore the children were not able to correct themselves using the feedback or any kind of explicit knowledge as they continued to rely on the knowledge they had. This type of negative feedback (challenging what they knew about the balance beam) is what would drive redescription to level E2. The RR model states that 4-year-olds do not have a conscious/explicit understanding of balance concepts and are at level I, where their knowledge and actions are implicit. However, children at this age and even younger have been found to solve balance beam problems, so it may be Karmiloff-Smith and Inhelder are underestimating children's abilities.

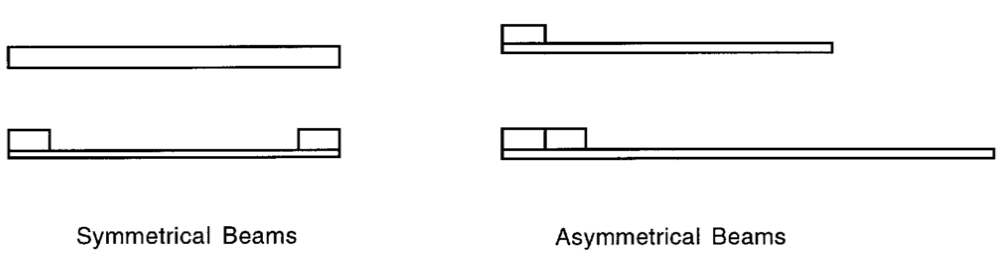

Figure 6. *The type of balance beam used by Karmiloff-Smith, taken from Peters, Davey, Messer, and Smith (1999).*

The RR model is best suited to tasks that can take a detailed analysis of verbalisations from children after every trial to be able to appropriately classify them into a level based on their knowledge. This would be problematic here since only the GP group are asked questions after each trial; if the DI group were asked questions as well it would then not be DI support.

The RR model is popular in the literature and does seem to be able to account for the discrepancies in knowledge as being due to the different levels of representation. The aspect of verbalisations is important as it means children need to rely upon their language skills. This is one reason that language skill should be considered when examining children's performance on the balance beam task – it appears that language is key in driving knowledge from implicit to explicit, through redescription. It could therefore be predicted that DI may perform better than GP, since they received language input via instruction and feedback, which could help drive redescription and potentially aid in strengthening children's knowledge. This theory would predict that children should be categorised into one of the levels for the most part and show some consistency in their ability to solve some problems, although regressions could be seen if a child is attempting to test different ways to solve problems. Overall, an improvement over time would be expected due to experience. No information regarding visual-spatial skills, EF or Mc is offered, so no predictions are made.

2.7.4 Siegler (1976) and the OW theory (Siegler, 1996)

Siegler (1976) also outlined four rules that children use when solving the balance beam, and the rules they use reflect what knowledge they have concerning the role of weight and distance. It is similar to Halford et al.'s (2002) suggestion and is again best suited to prediction trials since it requires knowing whether children can solve conflict trials and distance trials. This indicates it is a staircase model, as the rules suggest only certain problems can be solved at a certain time, with progression over time, but the OW theory suggests children have multiple strategies available to them, so learning is not staircase. Siegler developed the OW theory in response to data that indicated children used multiple strategies at a given time, which indicates he acknowledges learning is more continuous. The OW theory suggests strategies can change through experience and with instruction (Siegler, 2016), so it could be found that DI shows more improvement than GP due to the instruction. The OW theory would also predict that multiple different strategies could be used by children – both within a session and between sessions. It would suggest that improvement in which various strategies are used may be seen, but not large changes in which strategies are used

(Chetland & Fluck, 2007). No information regarding language, visual-spatial skills, EF or Mc are offered, so no predictions are made.

2.7.5 Munakata's (2001) GR account

Munakata's (2001) GR account suggests knowledge is graded rather than there being different knowledge systems for implicit or explicit knowledge. The GR account states that internal representations of the world are graded as a result of experience, resulting in some being stronger than others. This theory emphasises reliance on other components, such as memory and actions to explain why there are performance differences in tasks (Munakata, 2001). This is seen as an interactive continuous model, rather than a staircase model, and representations (knowledge) strengthen with experience. Which strategy is used in a particular situation depends on different factors, such as response type, support type, and EF.

The GR account provides an adequate explanation for differences in performance, not only between tasks, but between children too. This model also acknowledges that performance can change depending on the support (Munakata, 2001) and so a difference between GP and DI support could be seen, although no prediction about which could benefit more is made. The suggestion that EF could play a role in performance could help explain findings too, as the physics tasks used here will likely involve some WM and possibly inhibition if several trials are given. It could, therefore, be predicted that EF will link to physics task performance. The GR account does not mention Mc, visual-spatial skills, or language (although language could perhaps be considered an aspect of support type).

2.7.6 Schapiro and McClelland's (2009) connectionist model

Schapiro and McClelland's (2009) connectionist model is similar to the GR account in that they state knowledge is not split into implicit and explicit, but knowledge is on a continuum, rather than learning being discrete, as seen with the rules and strategies staircase approaches. Schapiro and McClelland (2009) suggest that learning the balance beam task is a graded, continuous process whereby newly acquired knowledge changes the connections between units, resulting in some rules being used, and at times the connections change ever so slightly, to the point of qualitatively different answers being given by the children, which could be interpreted by some as rule changes. These connections also explain the difference in implicit and explicit knowledge, as a weak connection could possibly be outputted as

implicit knowledge. It could be suggested that newly acquired knowledge can be obtained during support types and experience, although no claim on who benefits is made.

Schapiro and McClelland (2009) tested this theory by using a large sample of data from a balance beam task conducted by researchers who classified the children based on rules. Schapiro and McClelland (2009) found that their connectionist model could account for all of the data and the rules, but it could also account for the children that did not fit a particular rule, which is a big issue for those advocating for the rule classification approach. Schapiro and McClelland (2009) suggest that when a balance beam problem is presented, rather than explicitly selecting a rule to use to solve the problem, the graded connections between potential solutions determine which is selected based on what information is presented to the child. This model is useful for considering that learning is continuous and it would predict improvement with experience. It could account for various strategies being used within and between sessions based on information presented and how it could change connections. No predictions on language, visual-spatial skills, EF or Mc, are presented for this model.
In sum, the different theories would predict different outcomes in relation to language, visual-spatial skills, EF, Mc, balance beam performance, and strategy use, although some make no reference to some of these variables. The remainder of this chapter will conclude with what has been discussed so far and present the research questions, theory predictions, and hypotheses.

2.8 Conclusions and aims of the study

To conclude this chapter, the current study aims to address some of the unanswered questions in the field, examine areas lacking in research, and add to the existing literature. There is little research examining 3- and 4-year-olds' knowledge of balance concepts and even less in the area of motion down an incline, and transfer tasks. Although a lot of work has been carried out into EF, there is still debate surrounding the structure of EF – if it is one or multiple components, so the current study will consider this. Work has been carried out with older children to examine Mc, but very little with 3- and 4-year-olds, so this study will dive into a relatively young research field utilising the C.Ind.Le to code observations and through using a puppet interview. Comparing different support types has previously been investigated, but the mix of physics tasks and strategy use, while measuring EF and Mc, and with a young age group over several TPs, is novel.

Based on the work from the literature review, the role of EF and Mc in physics is still debated, so this work will examine not only this question, but also the link between EF and Mc. The link between EF and Mc is considered background analyses. Strategy development data will aid in answering the above aims concerning physics knowledge (held at different points in the study) and learning, and their potential links to the variables measured. The impact of support type on EF, Mc and physics performance, as well as to a transfer physics task was tracked to measure any impact changes.

The findings from the physics work could have implications for if and how it might be best to teach physics to young children and if individual abilities should be considered in the decision of which support type to provide. Pre-schoolers are not required to learn about balance beam or ramps concepts, so if benefits are seen here (such as in the transfer task) it should be seen as supportive evidence for pre-schoolers to cover topics like these. As was noted earlier, scientific reasoning encompasses a lot of skills and so they may translate into other areas of learning, and so it could be valuable to teach these early in life to make later teaching of other topics easier. Detailed individual analyses should aid in making conclusions regarding how to approach teaching individual children. If links are found between EF, Mc, physics, or strategy development it would be worth analysing the data to see whether the direction of the relationship can be uncovered, as this would aid in making conclusions regarding which factors are driving performance.

The conclusions drawn here have formed the basis for the research questions, which are outlined next, along with the hypotheses.

2.9 Research questions and hypotheses

Two research questions emerged from the above conclusions:

1. What role do EF and Mc have in children's performance on physics tasks?
2. What impact does support type have on:
    - EF?
    - Mc?
    - Physics task performance?
    - Strategy development?
    - A physics transfer task?

The hypotheses and theories' predictions for each research question will be discussed next.

Research question 1. What role do EF and Mc have in children's performance on physics tasks?

EF was expected to relate to and predict performance on the physics tasks, as it was thought that EF has a role in controlling behaviour and possibly links to Mc. Mc is expected to show a positive link to physics performance if it has a role in developing and implementing strategies. The relationship between EF, Mc, and physics performance was examined since it was thought that EF could have a role in applying Mc, which could show through the physics task performance. It was hoped that by examining the development of EF, Mc and strategy development over three TPs that it could be possible to tease apart the relationships and investigate how these variables relate. If a link between EF and Mc exists this would support Diamond's (2013) theory, as it suggests an interactive relationship between inhibition and self-regulation. If a link between EF and physics performance exists then this would lend support to Diamond's (2013) theory and the GR account (Munakata, 2001). Diamond's (2013) theory and the GR account (Munakata, 2001) both implicate EF as having a role in problem solving. If visual-spatial skills link to EF or physics performance this may support Diamond (2013) who suggests non-verbal reasoning is a higher-order EF. If language links to balance beam performance this could support the RR model and the GR account.

Considering strategy use during the balance beam, Halford et al.'s (2002) model predicts that children will consistently use the same strategy or show little deviation except when going through periods of improvement. Karmiloff-Smith's (1992) RR model, although staircase-like, does support regressions in that children may attempt other strategies even when they have found the correct solution, as a way to test and check they cannot improve on what they know. The OW theory (Siegler, 1996) predicts that children will use multiple strategies for solving different problems, both within a session and between different sessions. The connectionist model (Schapiro & McClelland, 2009) and the GR account (Munakata, 2001) could also account for multiple strategies within and between sessions, due to changes in weighted connected and the graded representations, dependent on each individual trial.

It will not be possible to compare the data here to the different rules and strategies suggested by Siegler (1976) and Halford et al. (2002) due to using production tasks here (and so not being able to examine all the different problem types required). Comparing the data to the RR

model is also difficult because not all children were prompted for verbalisations after the trials (thus cannot be classified based on explicit knowledge being verbalised). It was hoped the results concerning children's knowledge and strategies could be explained through the use of one the theories previously outlined, but the challenge may be that the data cannot distinguish between the different theories.

The strategy development data was examined to see whether any patterns emerge to support the above accounts and to also help identify whether children might hold a misconception, have a lack of knowledge, or show a change that learning is taking place. If a child consistently uses the same strategy for the same problem it might indicate the child holds a misconception and so continues to use the same wrong strategy, unable to correct it. If they show this pattern to begin with, but by the end of the trials they are consistently using the correct strategy (or a mixture of strategies) it could show they have learnt their misconception is wrong and they are correcting, or as Karmiloff-Smith (1992) would suggest, undergoing redescription. If a child uses trial and error it could suggest they have a lack of knowledge and so are trying different strategies to see what works. Again, if they show a change and begin using the correct strategy it could indicate they have learnt the correct strategy. The question of misconception or lack of knowledge could be unravelled if links between strategy performance and EF are seen. If a link between low EF and a pattern of strategies indicative of misconceptions are seen, and high EF and better performance, it may support the idea of EF having a role in suppressing misconceptions. If this is seen it could provide support for the connectionist model (Schapiro & McClelland, 2009) and the GR account (Munakata, 2001), but it may be difficult to tease apart further in relation to other theories. The staircase models can only support a drastic change in strategy use if children have changed the rule/level they are at. Thus, the pattern of strategies used will be important to try and distinguish which of these models support the data.

In sum, EF and Mc will be examined alongside physics performance and strategy development to see if either plays a role in performance.

Research question 2. What impact does support type have on:

EF?

Mc?

Physics task performance?

Strategy development?

A physics transfer task?

It was thought that both groups might show benefits from the support types, but for different reasons. Considering EF, it was thought GP might show more benefits since the children would need to exercise control over their behaviour to stay on-task. However, it was acknowledged that the length of time given for this task might not be long enough to provide a measurable difference. Considering Mc, it was thought GP might show benefits to Mc through the interaction the support provides, such as questions, prompts, and the opportunity for self-discoveries. The support style required GP to respond to questions about whether something worked or not and why, which would aid their verbal Mc ability and potentially impact on their ability to solve the problems. However, it was thought DI might also show benefits to Mc, since clear feedback regarding why a trial was correct or incorrect would be provided to the children and they could use the information to strengthen their Mc, which would potentially impact their ability to solve the problems.

Since both groups could show Mc gains, considering differences in physics performance and strategy development is also important since the Mc benefits could impact strategy development, which in turn impacts performance. GP had the chance to try different strategies during the play section of the task, unlike DI. Therefore, it was expected that the children in GP would show faster strategy development, which could be reflected in the physics task performance scores. However, the benefit of Mc in either support type may depend on how much the individual makes use of the questions and prompts or the instruction and feedback. All the theories support the idea that experience encourages learning, so it may be GP show higher performance due to the additional time they have exploring the materials themselves. However, some theories also advocate that instruction and feedback could aid performance, so it could be found that the instruction provided to the DI group may aid their performance scores from the start. If experience in the form of instruction and feedback is more beneficial, DI should perform better than GP, and the data would support the same claims as with EF: that it can be accounted for by the RR model

(Karmiloff-Smith, 1992), the OW theory (Siegler, 1996), the GR account (Munakata, 2001), and the connectionist model (Schapiro & McClelland, 2009). If experience in the form of playing with the balance beam materials themselves is more important than GP is expected to perform better than DI. Thus, it is not certain which support type is expected to emerge as more beneficial for learning here. It is hoped by also examining the strategy development data that a pattern will emerge to show learning, such as children starting with a higher performance score or showing they are using new knowledge over the sessions.

It is unclear how the EF, Mc, and balance beam data will relate to the transfer physics task. It could be that if one support type does significantly better during the balance beam task that the children in that support group will perform better on the ramps task. Or it could be found that individual children's performance on the balance beam relates to performance on the ramps task, with no influence of support type as Karmiloff-Smith (1992) found. The theories discussed so far do not necessarily reference how knowledge in one task would relate to another, but as all suggest experience is important it could perhaps be predicted that no significant association will be found between the physics tasks, as all children will receive some support style and time using the ramps.

The next chapter will outline the methodology employed to address these research questions.

3 Chapter 3
Methodology

This chapter will outline the methodology employed in order to address the research questions. The chapter outlines the research design, the tasks trialled in the three pilot studies, and the measures employed for the main study.

3.1 Research design

The study employed a between-subjects design over three TPs, allowing for an assessment of change over time and by group. It is believed that EF and Mc develop rapidly between 3 to 5 years old (Garon et al. 2008; Kuhn, 2000), so the element of time and multiple testing points was important for tracking change, which was believed to be measurable at this age. The design also allowed for strategy development during the balance beam task to be recorded, both over the different sessions and by each balance beam problem type. By observing children at three TPs, individual changes in EF, Mc, physics performance, and strategy use can be carefully examined, which are important aspects of this work.

Taking multiple EF measures mean they can be examined to see how they relate to one another in each session and over sessions, which could help identify whether the components are stable or changing in the timeframe assessed and if they are unitary or distinguishable. The Mc measures will allow for changes through support type to be measured to see if there is a benefit of support type over time. The EF and Mc measures also allow for each individual's EF and Mc relationship to be tracked and for it to be examined against support type and other potential factors. Examining individual progress is important, as group-level data can sometimes conceal changes. A detailed examination of strategy development will allow for each strategy per problem type per session to be examined, which can help tease apart different theories that predict use of one or several strategies and within or between sessions. Again, group level analyses would not reveal individuals' strategy development, which was important here. The strategy development data may reveal what knowledge children hold at different points in the study and if they hold a lack of knowledge or possibly a misconception. If the data support these ideas then some conclusions can be drawn concerning children's learning when presented with information on how to solve the trials.

In order to answer the research questions, the children were split into one of the two support types for the duration of the study, making support type between-subjects. The piloting data defined the differences between the two support types through the use of scripts and protocols. The repeated measures design of the other variables meant measures of EF, Mc, and physics had to be robust enough to use at three TPs to both measure the variable of interest and withstand children losing interest. Three measures of EF were taken: inhibition, WM, and shifting, all of which were quantitative in the form of percentage correct. Two measures of Mc were taken: Mc rate during the physics task (measured using the C.Ind.Le) and an Mc interview after the physics task. Both of the Mc measures were qualitative and scored based on the Mc behaviours seen or verbalised. These scores were then converted into quantitative data – a rate of Mc per minute and a percentage score for the interview questions. Video and audio recording of the sessions were taken to allow me (and second-coders) to look back through them multiple times to score and code behaviours.

All of the tasks were based on measures from the literature (although some were modified) and extensively piloted before the main study to ensure they were appropriate. The main study's tasks were selected following the findings from pilot studies 1, 2, and 3, which are detailed in the next sections. Please note, I carried out all of the data collection, but some aspects may be referenced using third-person language.

3.2 Measures

This section outlines which tasks were piloted and which were selected for the main study and why.

3.2.1 Pilot Study 1

The aims of pilot study 1 were to select the background measures (vocabulary and visual-spatial skills), physics tasks, EF measures, and Mc measures. Participants are detailed first, followed by the measures piloted and why they were selected. Many task changes took place during the course of pilot study 1 and will be reported here. Results will be detailed, where possible, along with conclusions drawn and the changes required for pilot study 2.

3.2.1.1 Participants

10 children aged between 36 months and 57 months took part (Table 1). Parents reported no medical or educational needs and all children had English as their first language. There were

eight females and two males. Six children were seen in the Faculty of Education's Observation Lab and four were visited in a local nursery over three days.

3.2.1.2 Background measures

As determined by findings in the literature review, measures of vocabulary and visual-spatial skills would be taken, to act as covariates, if appropriate.

The vocabulary test selected was the British Picture Vocabulary Scales II (BPVS II) (Dunn, Dunn, Whetton, & Burley, 1997). This is a standardised assessment of receptive vocabulary and is suitable for children aged 36 months and over. Children saw a picture book with four pictures and when they heard a word they had to point to the corresponding picture. Two children refused to finish the task and the other eight received raw scores between 23 and 63 (Table 1).

The visual-spatial ability task selected was the block construction subtest from the NEPSY-II (Korkman, Kirk, & Kemp, 2007). Children watched the adult make a construction with the blocks and then had to copy the design, with later trials requiring them to copy constructions from a stimulus book. This has been standardised for children aged 36 months and older. Of the 10 children, two refused to finish the task and two were maladministered. The other six children received raw scores between 4 and 9 (Table 1).

Raw BPVS and NEPSY scores will be reported throughout all pilot studies and the main study, as it is more meaningful to know whether children's raw ability relates to any measures, rather than whether a child is at the expected level for their age.

Table 1

Pilot study 1 participant information: age, sex, location, BPVS scores, and NEPSY scores

Child	Age (months)	Sex	Location	BPVS raw score	NEPSY raw score
507	36	F	Nursery	41	5
508	37	F	Nursery	45	4
505	38	F	Lab	39	4
504	45	F	Lab	23	Refused
509	45	F	Nursery	30	6
506	48	M	Lab	Refused	6
502	49	F	Lab	63	Maladministered
510	50	M	Nursery	Refused	Refused
503	56	F	Lab	55	9
501	57	F	Lab	61	Maladministered
Total mean	46.10			44.63	5.67
Total SD*	7.43			14.34	1.86

Notes. *Standard deviation.

The conclusions drawn from the two background measures were that they would be suitable for the age group in the main study, fit the time-constraints, and manage to keep most children's attention long enough to complete the task. Therefore these measures were used again in pilot study 2 and the main study.

3.2.1.3 Assessing physics knowledge and strategy use

The aim was to select one or two physics tasks for the main study, although at the time the tubes task, balance beam and ramps task were the only considerations, as they fitted the theme of forces and would have overlapping scientific enquiry concepts. The decision on which to use was based on how long the tasks took, children's understanding of them, performance ranges, and the range of strategies that children produce when trying to solve the problems. Some changes were made throughout pilot study 1 when it became apparent what would and would not work.

Of the changes that were made, one was that the first three children completed the tubes task

and balance beam task as the physics tasks, but the other seven children completed the balance beam task and ramps task. The tubes task will not be discussed further since it was not used in the main study due to the lack of strategies that were used by the children, the repetitiveness of trials, and the likelihood it would be too easy for many children. The following sections will discuss each physics task, performance and strategy data from pilot study 1, and the conclusions made.

3.2.1.3.1 Balance beam: knowledge and strategies

The balance beam task tests children's knowledge of balance with the concepts of weight and distance from the fulcrum. This is a well-established task and many have stated there are different stages / rules / strategies / levels children use, indicating what level of knowledge they have (Inhelder & Piaget, 1958; Siegler, 1976; Halford et al., 2002, Karmiloff-Smith, 1992). The balance beam task seemed appropriate since the problems are likely to elicit a range of performance in 3- and 4-year-olds, the problems are of varying difficulty, and different variables can be set (distance and weight). It was decided this would be a suitable task to use here to track progress over a longer period of time, as it would also allow for strategy development to be examined.

The balance beam used in pilot study 1 had four pegs and two types of weights: heavy (purple) and light (blue with white spots, which was half the weight of the heavy weight) (Figure ***1***). Trials were split into a prediction task where the child saw the beam set-up with weights and had to predict (by pointing to one of three picture outcomes) what would happen when the adult let it go (i.e., would it balance or tip to the left or right) and a production task where the child was given weights and asked to make the beam balance. The conditions chosen were based on the work of Siegler and Chen (1998) and Messer et al. (2008). The conditions outlined in Table 2 are best measured in the prediction task (conflict weight and conflict distance can only be measured in this task) and the production task trials can be seen in Table 3.

Table 2

Pilot study 1: conditions used in the prediction trials of the balance beam

Condition	Weight on each side of the beam		Distance on each side of the beam		Does it balance?
	Same	Different	Same	Different	
Balance	✓		✓		Yes
Weight		✓	✓		No – side with more weight tips
Distance	✓			✓	No – side with more distance tips
Conflict weight		✓		✓	No – side with more weight and less distance tips
Conflict distance		✓		✓	No – side with less weight and more distance tips
Conflict balance		✓		✓	Yes

Table 3

Pilot study 1: conditions used in the production trials of the balance beam

Condition	Weights	Where must weights go
2 balance	Two the same	Same pegs on each side
4 balance	Four the same	Same pegs on each side
2+2 balance	Two light, two heavy	Same pegs on each side
2 conflict balance	One heavy, one light	Heavy near fulcrum and light far from fulcrum
3 conflict balance	Three light	Two light near fulcrum and one light far from fulcrum
2+3 conflict balance	Two heavy, three light	Heavy on same pegs on each side, two light near fulcrum and one light far from fulcrum

Changes were made to the conditions and trials between testing sessions, which resulted in children completing a different number of trials and of each condition. The last four children received similar trials and received a second try on the trial if they got it incorrect, to allow them the chance to use the feedback to improve their strategy. A discontinuation rule was added so if a child got a certain number of consecutive trials incorrect then the task was stopped since it would be assumed that it was too difficult to proceed.

Tables 4 and 5 give an overview of how each individual performed by the number of trials correct in each condition of the prediction and production tasks.

Table 4

Pilot study 1: individual performance for each balance beam problem type in the prediction task

Child	Age (months)	Weight	Distance	Conflict balance
507	36	0/1	1/1	
508	37			
505	38	0/1	2/2	0/1
504	45			
509	45			
506	48			
502	49			
510	50			
503	56	1/4	1/2	
501	57	1/1	1/3	2/3

Table 5

Pilot study 1: individual performance for each balance beam problem type in the production task

Child	Age (months)	2 balance	4 balance	2+2 balance	2 conflict balance	3 conflict balance	2+3 conflict balance
507	36	6/6	2/2		0/2	0/2	0/1
508	37	6/6	2/2		1/4	0/5	
505	38	5/7	2/2		0/2	0/2	0/3
504	45	2/2				0/2	
509	45	0/1					
506	48	2/3	1/1			0/4	1/1
502	49	2/6		1/3	0/3		
510	50						
503	56	6/6				2/3	
501	57	5/6	0/2				0/2

Table 4 shows few children completed the prediction task. This is partly due to administering the production task first, as it was thought to be more engaging for the child, but it meant children were no longer engaged by the time they got on to the prediction task, perhaps due to the introduction to the task taking too long (approximately eight minutes). Table 5 shows that in the production task most children performed well with the various balance trials, but less well with the conflict trials, which included considering both weight and distance. Unfortunately, due to the lack of data, statistical analyses could not be carried out on either set of data.

Of the errors made during the production task, a brief summary of the number of each strategy used per child can be seen in Table 6. The strategies are based on the work outlined earlier by Siegler and Chen (1998). Only two strategies can be used to correctly solve the problems: the same weight at the same distance or different weights at different distances

(depending on the weights given). These represent each side of Table 6, with the error type used displayed.

Table 6

Pilot study 1: number of error strategies used for each solution type in the production task

		When the correct strategy is: same weight at same distance				When the correct strategy is: different weight at different distance				
Child	Age (months)	Same weights at different distances	Different weights at same distances	All weights on one side	Different weights at different distances	Same weights at same distances	Same weights at different distances	Different weights at same distances	All weights on one side	Different weights at different distances (incorrect)
507	36				2			1		2
508	37				5				1	2
505	38			2	2			1	1	3
504	45	1	1							
509	45			1						
506	48	2			3					
502	49	4	2					2	1	
510	50									
503	56				1					
501	57	4			1					

Table 6 shows a range of different strategies were used by the children: some children used only one error strategy, but others used several. The error strategies for production were interesting and were used again. However, completing both the production and prediction tasks took too long, so a mixed task incorporating both was used in pilot study 2 and is explained later.

3.2.1.3.2 Ramps task: knowledge and strategies

The ramps task tests children's knowledge of motion down an incline. The difficulty can be manipulated by including several variables, such as the height of the incline (high, low), weight of the ball (heavy, light), and friction from the surface type (wood, carpet). Until recently, research in this area has tended to work with older children. Using other techniques, such as interviews, it is suggested that children understand the role of the incline by 5 years of age (Hast & Howe, 2013). Van der Graaf et al. (2015) found that children aged 4-and-a-half-years-old could design experiments to test a specified variable in the ramps task. Van der Graaf et al. (2015) do not detail if children succeeded with some variables more than others, so it is unclear if more understanding is present for some variables compared to others. Van der Graaf et al.'s (2015) study showed that young children can understand the demands of the task, but they used a different apparatus (Figure 4) to that used here (Figure 7). At the time of designing this study, the task had not been completed with 3-year-olds and it was thought that some adaptation would be required.

Figure 7. *Photo of the ramps used in pilot studies 1, 2 and 3 and the main study.*

The apparatus used in this study was different to that used by van der Graaf et al. (2015): their apparatus could fix the variables, so the children could not change them. The apparatus used here could not be fixed, which resulted in some issues when the bases were moved or when children tried to change a variable that was not meant to be tested.

Seven children completed this task in pilot study 1. The first three children completed a version with large coloured squares on the floor and they had to set up the ramps so that a ball landed in a particular square (production task) or to predict which square the ball would land in when the adult had set-up the ramps (prediction task). There were two variables: incline and surface. This was carefully designed so that the ramps had to be set-up in a particular way for a ball to land in a particular square, but as mentioned, an issue with this was that the children sometimes moved the ramps' bases, which resulted in the balls not landing where they were meant to. It was decided this was not the best format to assess knowledge of these variables, so the task was changed for the next four children.

The next four children completed a version using three variables: incline, surface, and weight of the ball. There was again a production and prediction task, although no child managed to get on to the prediction task, due to the time taken. As with the balance beam, the introduction to the task took approximately eight minutes and many children did not want to continue with it for much longer and only done a limited number of trials. This version of the ramps task was somewhat based on van der Graaf et al.'s (2015) procedure – two variables were set up and children were asked to set-up one variable to test how it changed how far the ball went. Children sometimes tried to change a variable that they were not meant to, for example, if they were asked to just test the surface of the ramps, they could still change the incline, which meant the adult then had to adjust the incline if they got it wrong, which interfered with the trial. It seemed many children did not grasp what they were supposed to do, so it was determined that a more simplified version of the task with more instruction was required. Another version was piloted and this will be discussed in pilot study 2. Due to the problems with the ramps task in pilot study 1, the data are not reported here.

3.2.1.4 Support type procedure

The DI and GP procedures were based on Fisher et al. (2013), although various modifications were made to the procedure during piloting. It changed between the first six children, based on how each testing session went. The last four children received similar instruction,

although some slight changes were made in the sessions based on how well the session was going, and all of this was fed into a more definite procedure for pilot study 2. An overview of each support type (for the balance beam and then the ramps task) is outlined next.

3.2.1.4.1 Support type for the balance beam task

During DI support, the adult told the child that they were going to play a science game and the adult would be the scientist, so the adult had to put on the science coat (a white laboratory coat) before starting. The adult explained that they were going to test things and gave an example of what testing meant (an example, as used by van der Graaf et al. (2015) was used). Before starting the tasks the adult identified and named the different variables to be used and used the names during the instruction. During the balance beam task, the adult said they would look at what makes it balance and proceeded through a pre-written set of instruction that included demonstrations of what happens when each variable changed (i.e., the different balance beam problems). The children received two practice trials in their support type before starting the trials.

During GP support, the child was told they were going to play a science game and they would be scientists, so they had to put their science coats (white laboratory coats) on before starting. The adult explained that they were going to test things and gave the same example as DI. Before letting them play with the tasks the adult identified and named the different variables to be used. During the balance beam tasks, the adult said they wanted them to test the balance beam and see if they can find out what makes it balance and during the ramps task they asked them to test the ramps to see if they could find out what makes the ball go far. As stated by Weisberg et al.'s (2013) definition of GP, the adult commented on any discoveries the child made, asked open-ended questions about what they were doing and why, and directed them to any variables they appeared to have missed. The adult asked the child what they had found out, why they were doing what they were doing, and commented in any way appropriate for what they were doing. Each child received two practice trials before starting the trials and feedback was given on whether it was correct and why.

The time for both GP and DI was roughly the same in order to make sure the length of time introducing the task each time did not influence performance. In Whitebread et al.'s (2009a) study comparing play and taught conditions, children were allowed five minutes to work with a puzzle and in Cheyne and Rubin's (1983) study, children were given eight minutes to play

with a construction puzzle. Using the support type procedure outlined above, it took approximately eight minutes to complete the introductions (the DI involved the adult explaining the task and GP involved children exploring the task with the adult's input) before the test trials started. Although this time was needed, it seemed that some children's interest waned in the task by the time the trials started and this will be addressed in pilot study 2.

3.2.1.4.2 Support type for the ramps task

At the time of piloting, a decision was still to be made on whether the balance beam or ramps task would be the main study task and which would be the transfer task. The ramps task took a similar format to the balance beam task, with the aim explained as looking at how changing the ramps changes how far the ball went. Each child received two practice trials before starting the test trials and feedback appropriate to the support type, as in the balance beam.

3.2.1.5 EF measures

Three tasks were required for the main study: inhibition, WM, and shifting. Ten tasks were piloted in pilot study 1. No standardised measures were selected since a suitable measure for each EF component could not be found. The tasks were instead selected from the literature, based on previous research carried out with this age group. The Flexible Item Selection Task (FIST) materials were obtained from personal communication with the author (Sophie Jacques), however as there were only 15 test trials and multiple practice trials, one of the additional practice items was included as a test trial, so that all EF tasks piloted had 16 test items and two practice trials.

Pilot study 1 included tests with practice trials, which were later identified as problematic since the practice trials would filter out the children who perhaps understood the task, but failed to employ the relevant EF skills to succeed on the practice trials. Children who failed the practice trials did not receive the test trials, but whether they understood the task, but failed the practice trials, was not taken account of. For this reason, the instructions changed for pilot study 2 (Table 10), so pilot study 1's instructions are not detailed. The mean scores and SDs for the EF measures can be seen in Table 7. ("X" indicates the child failed the practice items and the test trials were not given, "Ref." indicates the child refused to complete the test, "Malad" indicates the test was not administered correctly, and blank cells indicate the test was not given.) The mean percentage correct can be seen, although it should be used with caution, as some tasks have a low *n*. Scores are reported as percentage correct from the

total number of possible trials that could have been solved. One change of EF tests occurred during piloting – less is more was swapped for dog/dragon, due to difficulties administering the test and the time taken to complete it, therefore dog/dragon is not reported.

Validity data for these EF tasks are often sparse and many tasks do not have any validity data associated with them. Despite this, the literature accepts these tasks as examining EF. The tasks are therefore used based on previous work having used them and the research community accepting them as EF tasks. Some of the validity data (for tasks used in pilot study 2), where available, can be seen in Table 10.

Table 7

Pilot study 1: overview of individuals' scores on each EF task during pilot study 1

Child	Age (months)	Sex	Gift delay	Grass/ snow	Spinning pots	Corsi blocks	Forward digit	FIST	Reverse categorisation	DCCS
507	36	F		0	44	X	44	50	100	88
508	37	F		X	44	25	69	X	100	88
505	38	F	0	X	47	X	56	X	63	63
504	45	F	100	X	44	6	Ref.	100	100	63
509	45	F		44	38	13	38	63	100	Ref.
506	48	M	88	Ref.	57	38	69	Malad.	75	75
502	49	F	Ref.	69	67	13	94	31	100	100
510	50	M		Ref.	Ref.	Ref.	Ref.	Ref.	100	100
503	56	F	100	81	53	88	88	Malad.	100	100
501	57	F	100	100	Malad.	81	100		100	100
Number who completed task			5	5	8	7	8	4	10	9
Mean	46		77.60	58.80	49.25	37.71	69.75	61.00	93.80	86.33
SD			(43.69)	(38.64)	(9.29)	(33.64)	(22.98)	(29.13)	(13.37)	(15.69)

Table 7 shows the results for each test are very varied. Most children performed at or near ceiling for gift delay, but few completed it, so it was piloted again. Grass/snow received a range in scores and it was piloted again, but with the typical green and white squares as the stimuli instead. Spinning pots received scores all within a small range, so the number of boxes and stickers was adjusted and piloted again, to increase the difficulty. The Corsi blocks range was quite large, so it was piloted again, but modified to include more blocks to increase the difficulty level. The forward digit was not used again as it was acknowledged that performance might depend on how well the children know their numbers, and it seemed this was an issue here. FIST performance showed a range and was piloted again. The reverse categorisation task was too easy and was not piloted again. The DCCS was a simplified version, as the original is often said to be too difficult for young children, but this resulted in a high performance rate for most children. The DCCS was piloted again, but with the typical pictures of rabbits and boats. All practice trials were removed and replaced with a clarifying question to ensure the child understood before proceeding with the trials.

3.2.1.6 Assessing Mc

As mentioned earlier, assessing Mc in young children is challenging due to the methods available. The best method available for this age range is observation and coding verbal and non-verbal Mc behaviours. Whitebread et al. (2009b) designed the C.Ind.Le coding framework, which is an observational coding scheme to code for Mc in children aged 3- to 5-years-old. There are three sections to C.Ind.Le, but only two have been used here: Mc knowledge and Mc regulation (Appendix A). MK refers to knowledge of persons, tasks, and strategies, and MR refers to planning, monitoring, control, and evaluation. When MK and MR were analysed, a decision whether to examine them separately or together was made. No validity data are available for the C.Ind.Le, but Whitebread et al., 2009b found their inter-rater agreement for 10% of the events to be 74.8%. This indicates the coders agreed on the same code nearly three-quarters of the time, suggesting the coding scheme is detailed enough for this level of reliability.

The C.Ind.Le coding scheme was the best available and so it was used. The codes were somewhat adapted for the task, as the examples in the C.Ind.Le framework are quite task-specific and include codes relevant to working with and referring to peers, which will not be relevant here. The use of the framework was to give the best possible overview of Mc behaviours displayed by each child. Children were videoed completing the physics tasks and

their Mc behaviours later coded. Examples of when a code was applied in the main study can be seen in Appendix B.

Changes to the physics tasks meant that children completed different tasks and trials and received different support types, so a comparison between children was not possible, but some videos were coded to ensure that Mc behaviours could be coded during the physics tasks. Only the production physics tasks were coded and the length of time coding took place was from the start of the trials to the end, so there are different lengths of time for each child. Whitebread et al. (2009b) calculated how many times each behaviour was displayed within the time-period and then calculated each behaviour's occurrence per minute and the same approach was adopted here. Eight children who completed the balance beam and one child who completed the ramps task were coded. The participants, task, length of time, number of behaviours displayed, and the rate per minute can be seen in Table 8.

Table 8

Pilot study 1: Mc rate in the production physics tasks

Participant	Age	Sex	Task	Length of task	No. behaviours	Rate per min
507	36	F	Balance	06:07	16	2.7
508	37	F	Balance	07:20	2	0.3
505	38	F	Balance	09:01	8	0.9
509	45	F	Balance	00:43	3	6
509	45	F	Ramps	04:43	3	0.7
506	48	M	Balance	05:19	5	1
502	49	F	Balance	05:38	13	2.4
503	56	F	Balance	03:33	10	3.3
501	57	F	Balance	10:06	0	0

Table 8 shows that some children displayed none or very few Mc behaviours, whereas some displayed several per minute. The coding used verbalisations and non-verbal behaviours during the physics tasks, but there is the obvious issue of children needing to verbalise their thoughts during the task, otherwise, it could result in a poor score. The analyses may take the child's vocabulary level into account when analysing the data to ensure this is controlled for.

However, some children barely verbalised during the tasks, despite scoring well. This was particularly noticeable for 501 and 508 and these children received poor Mc scores. A further Mc measure was trialled in pilot study 2 and is discussed later.

3.2.1.7 Conclusions from pilot study 1

Pilot study 1 was useful in determining what tasks may and may not be feasible for the main study. The BPVS and NEPSY were found to be appropriate for the main study but were piloted again for practice.

The balance beam task seemed like a good option for the physics task, although more piloting was required. The inclusion of the ramps task depended on the results of pilot study 2. Changes were made to the trials in both tasks to try and examine production and prediction trials in one task rather than two separate tasks because it was seen to take too long. The decision to include both in the main study depended on the time taken and the results of pilot study 2.

The EF tasks were mostly successful, although the practice trials that had to be passed before the test trials were administered needed to be omitted and replaced with simple questions to ensure the child understood what they had to do in the task.

Despite the challenges faced with measuring Mc in this age group, the observational coding did provide some interesting results, and within more structured physics tasks with defined support type, it was thought to be a viable method for the main study. It may be best accompanied by another method, so another measure was taken. All of these issues are discussed in more depth next, in relation to the plans for pilot study 2.

The support type underwent many changes but improved over the course of pilot study 1. Pilot study 2 used a refined version and the aim was to not change support between children to ensure it can be assessed properly.

Pilot study 2 will be explained next, including the conclusions made, and the need for pilot study 3.

3.2.2 Pilot study 2

Pilot study 2 addressed the issues highlighted in pilot study 1. The sections below will discuss the participants, the background measures, EF measures, physics tasks, the way Mc was assessed, and the procedure and support type instructions used.

3.2.2.1 Participants

Pilot study 2 included 11 children, aged between 36 months and 56 months, comprising nine females and two males. Three children visited the Faculty of Education's Observation Lab and eight were seen in a local nursery. (See Table 11.)

3.2.2.2 Background measures

The BPVS and NEPSY were administered to the children who visited the lab and since the results were again positive these will be used in the main study (although an updated edition of the BPVS will be used instead - BPVS – III (Dunn, Dunn, Styles, & Sewell, 2009)).

3.2.2.3 Assessing physics knowledge and strategy use

Pilot study 1 revealed the physics tasks to be too long, so pilot study 2 trialled shorter versions of the balance beam and ramps task, incorporating production and prediction trials into one task. The changes to each physics task will be discussed next.

3.2.2.3.1 Balance beam: knowledge and strategies

The primary changes to the balance beam task were to shorten the task, examine production and prediction within the same task, and make it more engaging. To make it more engaging the task was presented as a dinosaur seesaw game, where children had to find out which dinosaurs had to sit where on the seesaw in order for it to balance so they could seesaw (Figure 8).

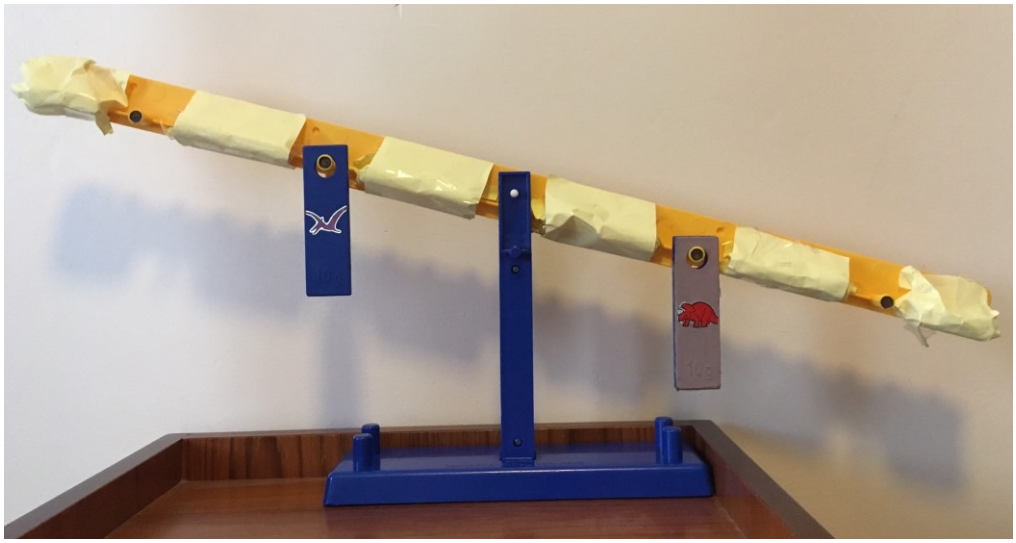

Figure 8. *Photo of the balance beam used in pilot studies 2, 3, and the main study, set up with the dinosaur characters.*

Each trial begun with prediction trials – the child was asked if the weights on the beam would balance and if they said no they had to change something (by removing or moving a weight) to make it balance. There were a few issues with this task: one significant issue was that the strategies could no longer be measured in the same way, as some children opted for the easiest strategy to solve the task, i.e., they always solved the trial using weight and never considered distance. Since the task no longer seemed to be able to measure what was hoped, and without the detailed strategy use data, it was determined that the mixed prediction and production task was not suitable. Due to time constraints, the realisation that production was more engaging for the children, and that a production task could test their knowledge of particular balance concepts, it was decided that the main study would use the production task only. For this reason, the trial data from the balance beam in pilot study 2 will not be presented. The children in pilot study 3 will complete another version of the balance beam task to refine the trials before the main study.

3.2.2.3.2 Ramps task: knowledge and strategies

One challenge with the ramps task in pilot study 1 was to get the children to understand what the task required. Pilot study 2 tried to set the task up as a game to clarify the aim as much as possible. Children were told the game would involve a hungry caterpillar and a bad alien and that they had to make sure hungry caterpillar always went down the "best ramp" so he would end closer to the leaves (printed on paper, propped up) at the end of the ramps. The children were told that the hungry caterpillar had to eat a lot before he could turn into a butterfly, so

they had to help him get the leaves before the bad alien took them away. The terms of closer, nearer, and far were explained with visual examples. The two ramps' variables were explained and the child was asked to set them up for a practice go. The task was set up as prediction and production: for the first four trials the adult set up the ramps using just one variable (incline) and held caterpillar and alien at the top of the ramps and asked the child if the ramps were set up to get the caterpillar to win/closer to the leaves. If the child said no then the adult asked them to change the ramps (by changing the incline of one or both ramps) to make the caterpillar win. Children completed four incline trials and if they got three or four correct they completed four surface trials, and if they score three or four they received four trials testing both variables. If the child incorrectly solved a trial they received a second try on the same trial. It was hoped that this version of the task would be easier for children to understand since it was a bit more structured, with a clear task goal.

Four children completed this version of the ramps task – three from pilot study 2 and one from pilot study 3 (code 536), but all will be considered together here. During the coding of the video, other elements were also coded/noted: the length of time for instruction, play, and trials; how much instruction the children received (a count of the pieces of information); the Mc displayed during the GP section (using the C.Ind.Le coding scheme); the Mc displayed during the trials (using the C.Ind.Le); and the feedback children received after the trials (a count of the pieces of information). These elements were deemed to be important for the main study, and piloting them showed it was possible to measure them.

The trials were much more successful than in pilot study 1, although there was still the issue of changing the variable that was not supposed to be changed, i.e., if the child was supposed to be comparing surfaces there were times they turned one of the ramps over to make it two surfaces the same. Despite reminders, this was difficult to control for and resulted in some trials straying from the protocol – this is reflected in the "other - straying trials" in the trials data overview in Table 9.

Table 9

The ramps task trials from pilot studies 2 and 3

Child	Age (months)	Sex	Condition	Incline trials	Surface trials	Other trials (straying)	Total trials
536	42	M	GP	3/3	2/3	2/2	7/8
530	47	F	GP	4/4	1/1	1/1	6/6
521	54	F	DI	3/3	3/3	1/1	7/7
531	56	F	GP	2/2	3/3	1/2	6/7

As can be seen, the children performed well on the various trials, which resulted in few strategies being displayed. As also seen in the balance beam task, children often wanted to take control of the task rather than predict what would happen and then change something. It was decided a single production task would work better instead. The piloting indicated that this was not the best task to measure physics performance and so would not be used as the main physics task in the main study. However, it was refined further and used as a transfer physics task instead. It was thought that the strategies would be difficult to categorise, since there are several variables that need to be considered when deciding what strategy was used: the incline of each ramp, the surface of each ramp, whether the children roll the ball down only one ramp or two, and whether they roll the ball or bounce/throw it. This makes it better to use as a transfer task than as the main task, where tracking strategy development is very important.

3.2.2.3.3 Physics tasks: conclusions

Pilot study 2 was not successful in finalising the physics tasks to use, but it was decided that the balance beam task would be used for the main study and the ramps task used for the transfer task. The balance beam task was piloted once more with the aim to use it in the main study and the ramps task was therefore refined for use as a transfer task at the end of the main study. The balance beam task will be discussed again in pilot study 3 and the ramps task discussed further in the main study.

3.2.2.4 Support type procedure

The support type procedure to be used changed slightly from that used in pilot study 1 to reflect changes to the aims of the game and such. Overall, the support type principles were

the same as in pilot study 1. The support type procedure went well and the two supports are distinct from each other and follow the principles of GP and DI.

3.2.2.5 Assessing EF

Pilot study 2 was modified to remove the need to pass the test trials and instead children were asked a simple clarifying question to ensure they understood what they needed to do. An additional inhibition task was piloted – head shoulder knees toes (HTKS) (Ponitz, McClelland, Matthews, and Morrison, 2009). During HTKS children learnt they must touch their head when the adult said, "toes" and touch their toes when the adult said, "head". If they performed well it was repeated with knees and shoulders, and if they performed well on these items then the final trials included all four items. As decided after pilot study 1, the grass/snow pictures were changed to a white square and green square. The spinning pots task was changed to have 12 boxes and nine stickers. The number of Corsi blocks increased to eight, the materials were changed to eight brown LEGO blocks fixed to a board. The DCCS task used different pictures – blue and red boats and rabbits. The demonstration was restricted to one card at the start and one question to check the child understands what they need to do. An overview of the EF tasks used in pilot study 2, including the new procedures and scoring can be seen in Table 10 and individuals' scores can be seen in Table 11.

Table 10

Pilot study 2: overview of the six EF tasks including the procedure and scoring used

EF and task	**Procedure and number of trials**	**Scoring**	**Reference and validity if reported**
Inhibition: Gift delay	•1 adult explanation •1 test trial	Seconds before child turned around to peek.	N/A
Inhibition: Grass/snow	•1 adult explanation and demonstration •2 questions to check child understands the task (adult will correct if child unsure and repeat questions once more) •16 test trials •Instructions repeated after 3 consecutive wrong trials and discontinued after 4	Number of times child points to correct picture.	N/A
Inhibition: HTKS	•1 adult explanation and demonstration •2 repetitions if child does not understand •2 practice tries after explanation •Second attempt of practice trials if incorrect •8 trials, if 5 or more correct do 8 more trials, if 5 or more correct do 8 more trials	Scored 0, 1, or 2 depending on action. Number of correct actions.	McClelland et al. (2014) found high test-retest reliability over several TPs ($\alpha \geq .92$).

EF and task	Procedure and number of trials	Scoring	Reference and validity if reported
	•Up to 48 correct actions.		
WM: Spinning pots	•1 adult explanation and demonstration •1 question to check child understands the task (adult will correct if child unsure and repeat question once more) •Up to 16 trials	Performance score of how many stickers found in the number of trials.	N/A
WM: Corsi blocks	•1 adult explanation and demonstration •1 question to check child understands the task (adult will correct if child unsure and repeat question once more) •16 test trials •Instructions repeated after 3 consecutive wrong trials and discontinued after 4	Number of correctly copied sequences.	Cornu, Schiltz, Martin, and Hornung, (2018) found internal consistency to be high (α = .74).
Shifting: FIST	•1 adult explanation and demonstration •1 question to check child understands the task (adult will correct if child unsure and repeat question once more) •16 test trials	Number of correct selections (must get both selections for one point).	N/A

EF and task	Procedure and number of trials	Scoring	Reference and validity if reported
	•Instructions repeated after 3 consecutive wrong trials and discontinued after 4		
Shifting: DCCS	•1 adult pre-switch explanation and demonstration (using one card) •1 question to check child understands the task (adult will correct if child unsure and repeat question once more) •8 pre-switch test trials •8 post-switch test trials •Instructions repeated after 3 consecutive wrong trials and discontinued after 4	Number of correctly sorted cards after the rule switch.	Beck et al. (2011) found the intraclass correlation to be .94.

Table 11

Pilot study 2: individual EF scores (percentage correct) from pilot study 2

Child	Age (months)	Sex	Location	Gift delay	Grass/snow	HTKS	Spinning pots	Corsi blocks	FIST	DCCS
524	36	M	Nursery		0	Refused	67	6		88
526	40	F	Nursery		0	0	89	38	6	94
529	41	F	Nursery		63	Recording error	100	6	0	69
523	42	M	Nursery				78			
525	43	F	Nursery		75	4	89	31	69	75
527	45	F	Nursery		69	27	89	19	38	88
522	46	F	Nursery		75	58	89	25	100	100
528	46	F	Nursery		100	67	78	25	0	82
530	47	F	Lab	100	69	69	89	13	50	100
521	54	F	Lab	1.7	63	67	89	25		100
531	56	F	Lab	100	88	92	89	31	38	100
No. children	11			3	10	8	11	10	8	10
Mean (SDs)	45.09 (5.86)			67.23 (56.75)	60.20 (33.68)	48.00 (33.55)	86.00 (8.65)	21.90 (10.79)	37.63 (35.58)	89.60 (11.32)

It can be seen in Table 11 some children did not complete all the tasks (a blank cell indicates the test was not given). It was not possible to do the gift delay task in the nursery environment, but of the three children who did complete it, two scored 100%, indicating they did not peek. The grass/snow task resulted in a range of scores from the children, from 0 to 100. The HTKS task was difficult to carry out in the nursery environment although those who completed it displayed a range of scores. From the inhibition tasks, it was decided that grass/snow would be used in the main study.

The spinning pots task was time-consuming and scores show a number of children performed well, with little range in scores. The Corsi blocks task was not as successful as hoped and resulted in a range of low scores. Corsi blocks was piloted again in a more engaging format (pilot study 3) and it was decided if it was not successful the spinning pots would be used with an increased difficulty.

The FIST and DCCS took about the same length of time to administer, but the FIST resulted in a wider range of scores, so it was used in the main study.

3.2.2.6 Assessing Mc

Despite the challenges faced with measuring Mc in this age group, the observational coding in pilot study 1 did provide some interesting results. One issue that arose from pilot study 1 was that not all children received an Mc score due to not displaying any Mc behaviours. In order to try and obtain as true a measure of Mc as possible, two other methods were also piloted: an Mc interview and the train track task (Bryce & Whitebread, 2012).

The Mc interview was based on work by Berhenke, Marulis, and Neidlinger (2012), who used an Mc knowledge interview with 32- to 70-month-olds after they completed a puzzle. The children were asked questions about how well they thought they had done on the task, what could they have done better, and if talking while doing the task made it easier (Berhenke et al., 2012). The children were asked 11 questions and their answers were rated from 0-2 depending on how Mc it was deemed. Berhenke et al. (2012) also used a puppet interview with the children to assess motivation after they completed a puzzle. The adult asked the child eight questions such as: did they like the tasks and how they thought they performed on the task. Two puppets were used to retrieve answers from the child – one

puppet responded negatively (i.e., the puzzle was hard) and one positively (the puzzle was easy), and the child had to point to the puppet they agreed with. Pilot study 2 used an interview combining the two aspects Berhenke et al. (2012) used: an Mc interview with puppets. The children here were asked four questions after each physics task: if they thought it was easy, what was easy, what was hard, and what could help with the hard parts. Validity statistics are not available due to this being a novel task. Not all of the children in pilot study 2 completed the task, but some from pilot study 3 did, so they will all be considered together here (Table 12). Children scored 0 if it was deemed not at all Mc, 1 if slightly, and 2 if definitely. Note, the balance beam task used in pilot studies 2 and 3 were different, but it does not affect the data below.

Table 12

Mc interview scores from the balance beam task used in pilot studies 2 and 3

Child	Age (months)	Sex	Q1	Q2	Q3	Q4	Total /8	% score
532	42	M	1	0	0	0	1	12.50
525	43	F	2	0	0	0	2	25.00
533	43	F	2	1	2	0	5	62.50
530	47	F	2	0	0	0	2	25.00
521	54	F	2	0	1	1	4	50.00
527	56	F	2	0	0	0	2	25.00
531	56	F	2	1	0	0	3	37.50

The Mc interview scores show a range in scores, which gives scope for improvement over the TPs. This method does not take long to administer and it would give further information than would be obtained from observation alone, so it was included in the main study. For the main study, the Mc interview score and Mc rate can be compared to see how well they correspond to each other.

The other Mc method that was piloted was the train track task (Bryce & Whitebread, 2012), with the aim of finding a measure of Mc that was independent of the observations during the balance beam task. As with the coding of Mc behaviours during the balance beam task, some children did not display many Mc behaviours, regardless of whether they performed well or

not. It also took a long time to administer (up to 20 minutes). The task was deemed not to be a useful enough tool to keep in the main study due to the time taken to complete the task and the query whether it added to the physics task behavioural coding, as the train track task is coded in the same way. It was therefore decided that the Mc coding during the balance beam task would be more appropriate and the scores from it could be compared to physics task performance. The inclusion of the Mc interview would aid with a secondary, back-up measure, to accompany the Mc coding during the physics task.

3.2.2.7 Pilot study 2 conclusions

Pilot study 2 finalised the inhibition (grass/snow) and shifting (FIST) tasks, and the two Mc measures (Mc rate during the physics task and the Mc interview after the physics task). The Corsi blocks were piloted once more, as was the balance beam task - these two tasks are detailed below. It was decided the balance beam task would be the main physics task and the ramps task would be the transfer task.

3.2.3 Pilot study 3

Pilot study 3 addressed changes to the Corsi blocks and to the balance beam.

The Corsi blocks were piloted a third time, this time set up as a tiger jumping game, where the child had a small tiger (a ball with a small tiger sticker on it) and the adult had a large tiger (a ball with a large tiger sticker on it) and the child (baby tiger) had to copy the "rocks" (brown LEGO blocks) the adult (mummy tiger) jumped on. Four children completed the task and it appeared much more engaging than previous piloting. The percentage correct scores were 0, 31, 31, and 38, so not terribly different from pilot study 2. Since the spinning pots had quite a high range of scores and took much longer to administer, it was decided this version of the Corsi blocks would be used in the main study.

Pilot study 3's version of the balance beam involved giving the child some weights (with dinosaurs on them) and asking them to make it balance (so they dinosaurs could seesaw). The scores for each condition can be seen in Table 13.

Table 13

Performance on the balance beam during pilot study 3

Child	Age (months)	2 balance	4 balance	2+2 and 2+3 balance	3 conflict balance	Total	% correct
532	42	1/2	1/1		0/1	2/4	50.00
534	42	4/4	1/1	0/1	0/2	5/8	62.50
533	43	1/2	1/1		0/1	2/4	50.00
535	47	4/4	1/1	0/1	0/2	5/8	62.50

The 2 and 4 balance trials involved weights the same, the 2+2 and 2+3 trials involved two heavy and two light weights or two heavy weights and three light weights, and the 3 conflict balance trials involved 3 light weights. This version of the task was much easier to administer, the scores were much clearer, and the strategies were easier to code. The children performed well on the 2 and 4 balance trials, as was also seen in pilot study 1, but the other trials, which involved solutions that were not placing the same weights on the same pegs, were more difficult. The production task was used in the main study with a mixture of problems that require children to use different types or numbers of weights.

3.2.3.1 Pilot study 3 conclusions

Pilot study 3 confirmed the final two tasks for the main study: the Corsi blocks as the WM task and the production balance beam task with a mixture of problem types as the main physics task. The next chapter will discuss the main study in detail, bringing together what was outlined in the three pilot studies.

4 Chapter 4

Main study

This chapter will discuss the main study's participants and groups, the procedure employed, the measures used, and some analyses to show how the groups and support types were matched.

4.1 Main study participants

The main study included 38 3- and 4-year old children from nine different nurseries in Cambridgeshire and Suffolk. Nurseries were contacted and informed of the project aims and requirements and shown the participant information sheet and consent form. In participating nurseries, opt-in consent forms were given to all parents of children who had English as a first/main language and all who returned the form giving consent took part in the project. Parents reported no diagnosed medical conditions or educational needs. The number of participants is lower than planned and hoped, but unfortunately no more could be recruited.

4.2 Main study participant groups

Within TP1 all children completed BPVS-III, the block construction subset of the NEPSY-II, and three EF tasks. Based on the scores from the BPVS, NEPSY, EF tasks, age (at first visit), and sex, the children were split into two equally matched groups. Table 14 shows the means, SDs, and ranges for the overall sample and for each group on all the matched measures, as well as the results comparing the support groups, measured using 2-tailed independent *t*-tests and effect sizes. The three EF tasks are presented as percentages correct, so scores could range from 0 – 100 at each TP.

Table 14

Main study: descriptive information for the overall sample and the two groups, along with t-test results and effect sizes

	Complete sample means (SDs)	Complete sample range	GP means (SDs)	DI means (SDs)	Independent 2-tailed *t*-tests (*df* = 36) *t* value	*p* value	Cohen's *d*
Number of children	38		19	19			
Age (complete months)	44.53 (4.81)	36 – 55	44.21 (4.92)	44.84 (4.82)	0.40	.69	.13
Sex (M:F)	19:19		10:9	9:10			
BPVS raw score	55.11 (15.20)	29 - 87	54.26 (13.98)	55.95 (16.66)	0.34	.74	.11
NEPSY raw score	5.55 (1.90)	3 - 12	5.47 (2.14)	5.63 (1.67)	0.25	.80	.08
Inhibition 1	54.61 (36.26)	0.00 – 100	55.59 (35.90)	53.62 (37.57)	0.17	.87	.05
Working memory 1	23.03 (17.20)	0.00 – 50.00	21.05 (16.82)	25.00 (17.80)	0.70	.49	.22
Shifting 1	36.02 (26.17)	0.00 – 87.50	37.17 (24.87)	34.87 (28.05)	0.27	.79	.09

Levene's test for equality of variance was not significant for any of the measures in Table 14 ($p > .05$), therefore variance between the two groups is assumed. No statistically significant differences were detected between two groups on any of the measures and the effect sizes are very small. (Cohen (1988) defines a small Cohen's *d* effect size as 0.2 - see Chapter 5 for more information on power and effect sizes.) Using G*Power (Mayr et al., 2007 and Faul, Erdfelder, Buchner, & Lang, 2009) and entering alpha (*p)* as .05 and power as .8, it states a sample size of 52 would be required to detect a large effect size (with 80% chance). It cannot be said that no difference exists between the two groups, but based on the *p* values and effect sizes the data does not show a difference. However, as the tests are underpowered, it may be that the sample size is too small to detect any significant differences.

4.3 Main study procedure

Data collection for the main study was conducted between February 2016 and July 2016. Nurseries were split into three blocks, depending on their availability, and testing happened over three TPs. Nurseries in block 1 were visited February – May, nurseries in block 2 were visited February – June, and nurseries in block 3 were visited April – July. The TPs were (roughly) a three-week period and each TP was roughly six weeks apart. The testing sessions at each TP and the tasks completed can be seen in Table 15. The number of children who completed each task at each TP can be seen in Table 16.

Table 15

Overview of testing sessions in the main study

Time point	Visit	Measures carried out
TP1	1	BPVS III and NEPSY II (20-35 minutes)
	2	EF1 (10 minutes)
	3	Balance beam 1 and Mc interview 1 (15-25 minutes)
~Roughly six weeks later		
TP2	4	EF2 (10 minutes)
	5	Balance beam 2 and Mc interview 2 (15-25 minutes)
~Roughly six weeks later		
TP3	6	EF3 (10 minutes)
	7	Balance beam 3 and Mc interview 3 (15-25 minutes)
	8	Ramps task and Mc interview (20-30 minutes)

Table 16

Main study: number of children in each group who completed each task at which TP

	TP1			**TP2**			**TP3**		
	Total	GP	DI	Total	GP	DI	Total	GP	DI
BPVS III and NEPSY II	38	19	19						
EF	38	19	19	36	17	19	37	19	18
Balance beam	36	17	19	31[1]	16	15	33[2]	16	17
Mc rate	36	17	19	31[3]	16	15	29[4]	13	16
Mc interview	35	17	18	31[5]	16	15	29[6]	13	16
Ramps and Mc interview							18[7]	8	10

Notes.

[1]32, but 1 DI did not record.

[2]37, but 3 GP and 1 DI without sound.

[3]32 including 1 DI that did not record.

[4]33 including 3 GP and 1 DI without sound.

[5]32 including 1 DI that did not record.

[6]33 including 3 GP and 1 DI without sound.

[7]20 including 2 GP with recording errors.

If a child was absent during session 1 or 2 they completed the missing task(s) in session 3. If they were absent during session 4 they completed the missing task in session 5. If they were absent during sessions 6 or 7 they completed the missing task(s) in session 8. Tasks were not carried over to the next TP. 35 children completed all three EF sessions. 27 children completed all three balance beam sessions and 26 of these completed all three balance beam Mc interviews. Only 18 children completed the ramps task and the ramps Mc interview. The drop in the number of children who completed the ramps task is due to absence and being unable to see them at a later visit, since the ramps task was carried out during the last visit to each nursery. The available space to work with the children also contributed to the drop in participant numbers, as some nurseries did not have space to be able to carry out the ramps task, due to the task requiring a lot of floor space to set up the ramps and to let the balls roll.

Children were seen individually, usually in a space out-with their classroom or away from the class, where other children could not interrupt, however, this changed from nursery to nursery and sometimes week to week. Interruptions and distractions were generally not a problem.

4.4 Background measures

The BPVS-III and the block construction subset of the NEPSY-II, as detailed in the pilot studies.

4.5 Balance beam task

The balance beam task as described in pilot study 3 was used. Children received a practice trial before beginning. Children received four trials and if they got three or four correct they completed another four trials. No child got on to trials 9 – 12, so they will be omitted from reference. The trials and weights given, along with what they tested, can be seen in Table 17.

Table 17

Balance beam trials used in the main study

Trial	Condition	Weights given
Practice	2 balance	Two light
1	2 balance	Two heavy
2	4 balance	Four light
3	2 conflict balance	One light, one heavy
4	2 balance	Two light
Continued if 3 or 4 correct		
5	3 conflict balance (dissimilar)	Two light, one heavy
6	2 conflict balance	One light, one heavy
7	2 balance	Two light
8	3 conflict balance	Three light

The condition number denotes how many weights were given to the child and the 'dissimilar' label distinguishes between the conflict balance trials that used 3 weights the same, as the dissimilar trials involved weights that were not all the same. The balance trials and the conflict balance trials were not merged since they each included different weights and as will be reported later, resulted in different performance rates, so they are reported separately. As previously stated, using production trials meant all the prediction conditions could not be assessed.

Each time the child was given a set of weights (dinosaurs) and asked to make them balance (so they could seesaw). After each trial, the child received feedback appropriate to the support type. The instruction received by each group can be seen in Appendix C and is explained in more detail next.

4.6 Support type

4.6.1 Instruction and information before the balance beam trials

The GP and DI conditions are based on a modified version of Fisher et al. (2013). Both groups received the same instruction at the start concerning the aim of the game, but GP then

got time to play with the balance beam and weights while DI heard the adult explain and show them how to solve the different problems.

4.6.2 Instruction given to both GP and DI

During GP support, the child was told that they were going to play a science game and they would both be scientists, so they had to put on their science coats (white laboratory coats) before starting. During DI support, the adult explained they were going to play a science game and the adult would be the scientist (and the adult wore a white laboratory coat) and show them how the game works.

During both support types, the adult explained that the game involved a seesaw and asked whether they had been on a seesaw before, to get them engaged with the game. Using cartoon pictures, children were shown a seesaw and were told that the best people to play together on a seesaw make it balance because they can make the seesaw go up and down. Children then saw a picture with two characters on a seesaw making it balance. The adult showed them another picture of two characters on a seesaw, but this time one character was making it tip over, so not balancing. The adult again explained that the best people to play together on the seesaw make it balance because if the seesaw tips over then they cannot play together. The adult reiterated that the game involved making the dinosaurs play together on the seesaw, by balancing it so they could seesaw.

The adult explained in simple terms what balance means – that the seesaw was to be straight (demonstrating with pictures and visual gestures) and not tipping over (demonstrating with pictures and visual gestures). The adult showed the child the two different types of dinosaurs – the light dinosaurs, which were on the blue weights, and the heavy dinosaurs, which were on the purple weights. The adult asked the child to put one in each hand to feel the difference to try and emphasise which was heavy and which was light. The adult explained in simple terms what weight is – that light things are easier to pick up and heavier things can be “more tricky” to pick up.

The adult explained and showed that there were different seats the dinosaurs could sit on – two seats on each side of the seesaw, with seats near the middle and seats far from the middle.

4.6.3 Instruction given to GP

Following the above instructions, the adult told the child it was their turn to have some goes at putting the different dinosaurs in the different seats to see what happens and to see who can balance the seesaw and play together. The adult gave them a few minutes to play with the dinosaurs, but how long they played for was dependent on how engaged they were and if they wanted to try a few more things before stopping. However, overall, the play times were kept roughly the same between the children. Based on Weisberg et al.'s (2013) definition of GP, the adult asked them about anything they had found out, about what they were doing, if something worked and why/why not. The adult also directed them to things they had missed, such as not using all the dinosaurs or quite commonly, not realising two dinosaurs could share a seat.

4.6.4 Instruction given to DI

Following the above given to both groups, the child was told that the adult would have a go at weighing the different dinosaurs on the seesaw to see who needs to sit where to make it balance, and then it would be their go to have some tries. All children were shown the same variables in the same order. The first was two light dinosaurs sitting in the same seats on each side of the beam (which made it balance) and the adult explained it was because they weighed the same/they were the same kind of dinosaur and were sat in the same seats. The adult then moved one of the two light dinosaurs to a different seat to show it did not balance. The adult explained that even though they weighed the same and there was one on each side of the beam, they needed to sit in the same seats for it to balance. The adult explained that the dinosaur sitting far from the middle makes the seesaw tip over. So, when the dinosaurs weigh the same they have to be in the same seats on each side to make it balance.

The adult then showed the children what happened when there were two dinosaurs who do not weigh the same (using a heavy and a light dinosaur). The adult showed what happened if they sat in the same seats - they do not balance – because they do not weigh the same and so the heavy dinosaur makes it tip over. The adult then showed what happened if the heavy dinosaur sat in the far seat and the light dinosaur in the seat near the middle – it still does not balance, because the heavy dinosaur in the far seat makes it tip over. They were to remember that heavy dinosaurs try to make the seesaw tip over and dinosaurs who sat in the seat far

from the middle they try to make it tip over. So, when a heavy dinosaur sits in the far seat, they will really try to make it tip over.

The adult then showed what happened when they swapped seats and the heavy dinosaur sat near the middle and light dinosaur far from the middle – it balanced. The adult explained it was because the heavy dinosaur was trying to make it tip over (because that is what heavy dinosaurs do) and the lights dinosaur in the far seat was trying to make it tip over (because that is what dinosaurs in the far seat do), so they were both trying to make it tip over, and so it balanced out.

The adult then showed them that two light dinosaurs weighed the same as one heavy dinosaur. The adult then summarised the information: they need to think about which dinosaur sits where. To make the dinosaurs balance, the dinosaurs must weigh the same and sit in the same seats on each side or they put heavy dinosaurs near the middle and light dinosaurs far from the middle. They need to remember that heavy dinosaurs try to make the seesaw tip over and dinosaurs who sit in the far seat try to make it tip over. They also need to remember that two light dinosaurs weigh the same as one heavy dinosaur. They had to think about both the weight of the dinosaurs and which seats they were in to try and make it balance.

4.6.4.1 Balance beam trials instruction given to both groups

When the trials were to start the adult gave the child the weights they were to balance for that trial and asked them to make the dinosaurs balance, so they could seesaw. The full procedure used for each support type can be seen in Appendix C.

4.6.5 Comparison of length of instruction in the two groups

The time of the GP and DI support was to be equal in order to make sure the length of time introducing the task before the trials did not influence performance. In GP, the time allowed for play is included in the instruction length, while for DI the time taken for the adult to explain and show the problems was included. The means and SDs for each TP can be seen in Appendix D. A significant difference in length of time at TP1 was found (along with a large effect size), due to GP receiving a longer instruction than DI. This is likely due to refining the GP procedure over time. Despite piloting, there was still some learning over the first TP and

as indicated by the means and SDs at TPs 2 and 3, the instruction shortened over time. No statistical differences were seen at TPs 2 or 3, but the effect sizes are medium (defined as between 0.5 and 0.79 by Cohen, 1988). It could be that there was not enough statistical power to detect a significant difference, which is why the effect sizes are seen as medium, but *p* is above significance. G*Power indicates a sample size of 54 is required to detect a large effect size here and 134 to detect a medium effect size. It is therefore not possible to rule out a difference existing between the groups at TPs 2 and 3. (See Chapter 5, Section 5.4 for more information on statistical power.)

4.6.6 Comparison of how much information was provided by the adult

To ensure the amount of information provided to each group matched and differed accordingly (therefore making the groups qualitatively different), how many pieces of information each child received during the instruction was calculated. The groups were to receive the same amount of information during the "generic" instruction at the start, but DI was to then receive the "DI" information on how to solve the different trials depending on the weights provided – this information is in Appendix E. The two groups appeared to have received a similar amount of "generic" information at each TPs 1 and 2, due to the non-significant Mann-Whitney tests and small or very small effect sizes, but there was a difference at TP3, as seen by the significant *p* value and large effect size. This difference was due to GP receiving more pieces of "generic" information. According to G*Power, the tests were underpowered, so there was likely not enough power to detect a significant result, if one exists.

4.6.7 Comparison of how much information each group had before starting the trials

The previous analyses were carried out to compare the amount of information provided to each group by the adult, but how many pieces of information the GP children discovered during the play time is also important. The comparison is to see whether the groups differed in how much information they had before starting the trials. The data showed a significant difference between the total information each group obtained at each TP due to DI having more information (Appendix F). The results show that there is a significant difference between how much information each group obtained during the introduction, even when this included GP's discoveries made during the play times. The *p* values are statistically significant and the effect sizes are large, indicating there is a difference between the groups.

However, as with the other Mann-Whitney tests, there may not be enough power, due to the small sample size, to confidently say a significant difference exists.

4.6.8 Feedback after the balance beam trials

Besides the instruction, the feedback after each trial was the other central difference between the two support types. The GP children were asked if their strategy worked/if it seesawed and why/why not. Depending on the response given (if there was one) there were sometimes other questions. In DI children were told whether it worked/seesawed and why it did/did not work. Essentially, the difference was whether they were asked questions (GP) or told something (DI). The data showed there to be a significant difference between how much feedback was given to each group at TPs 1 and 2, but not TP3 (Appendix G). The significant differences are due to the DI receiving more feedback than GP. The significant differences at TPs 1 and 2 also have large effect sizes, indicating a difference likely exists between the groups, but the tests may be underpowered due to the small sample size, so the small effect size at TP3 may be a true significant difference, but cannot be detected because of low power.

These differences in the protocol will be examined again later see if they link to balance beam performance.

4.7 Ramps task

The ramps task was used as a test of physics transfer and completed in the last session. The protocol was refined from the last piloting to try and ensure the best way of testing their knowledge was used. Children were first told about the ramps inclines and surfaces and asked to set them up, to make them comfortable changing the ramps. They were then given around three to four minutes to use the balls and ramps and were reminded to change the ramps to see how it changes how far the balls go. The caterpillar and alien game, as used during piloting, was then explained (the full protocol and instruction are in Appendix H).

The variables were enforced and pictures of the variables the children were to use were displayed during each trial, i.e., if they were asked to test a high and a low ramp and two wooden ones they would see the corresponding photos showing a high and a low ramp, and a picture of the wooden ramp (4 pictures). The incline used in the first four trials was secured

by removing two of the props the ramps sat on, so the incline could not be changed. It was adjusted between trials to correspond to what was required for that trial. For trials 5 to 10 the children were told which surfaces and inclines to use, but the incline was not set for them. If the child tried to change the variables from those they were told to use they were reminded (at around 45 and 90 seconds) which variables they were supposed to be testing, so if they were asked to use two wood ramps and tried to change one to carpet the adult pointed out the photo reminders.

Children completed one practice trial, followed by four test trials, and if they got three or four correct they moved on to do four more, and if they got three or four correct they moved on to do two more. They were given around two minutes to complete each trial and they could have multiple goes. The child had to say when they got it correct and if they did not succeed in the time limit they were moved on to the next trial. The trials and what they aimed to test can be seen in Table 18.

Table 18

Ramps trials used in the main study

Trial	Condition	Incline props adult set up	Ramps child asked to use
Practice	Surface	Two low	One carpet and one wood
1	Incline	One high, one low	Two wood
2	Surface	Two high	One carpet and one wood
3	Incline	One high, one low	Two carpet
4	Surface	Two low	One carpet and one wood
Continued if 3 or 4 correct			
5	Incline		Two carpet, one low and one high
6	Surface		One carpet and one wood, two high
7	Incline		Two wood, one low and one high
8	Surface		One carpet and one wood, two low
Continued if 3 or 4 correct			
9	Incline and surface		One carpet and one wood and one low and one high
10	Incline and surface		One carpet and one wood and one low and one high

The data were checked to ensure the two groups did not differ on any of the task variables: the time given per child to use the ramps before the trials, the information provided before the free time to use the ramps, the information given when explaining the aim of the game, and the feedback given (calculated as a rate per try). (See Appendix I for the corresponding means, SDs and Mann-Whitney tests comparing the groups.) No significant differences were detected and most effect sizes were very small. The result of the tests comparing feedback given to each group showed a medium effect size. According to the power calculations it may be there is not enough power to detect a difference, should one exist, so it may be that there is a difference between the groups but it was not large enough to be found here.

4.8 EF measures

Three EF tasks were used: grass/snow (inhibition), modified Corsi blocks (WM), and the FIST (shifting). Photos of each can be seen in Figures 9, 10, and 11. The tasks were counterbalanced, so at each TP children completed them in a different order. This was roughly randomised based on sex, group, and nursery, to ensure an equal counterbalanced sample overall. Raw scores were turned into a percentage correct, so each task's scores could be more easily compared. (Had the raw scores been used the results would not have changed since scores were always calculated from the same number of trials at each TP, for example, a raw score of 8/16 would be 50%.)

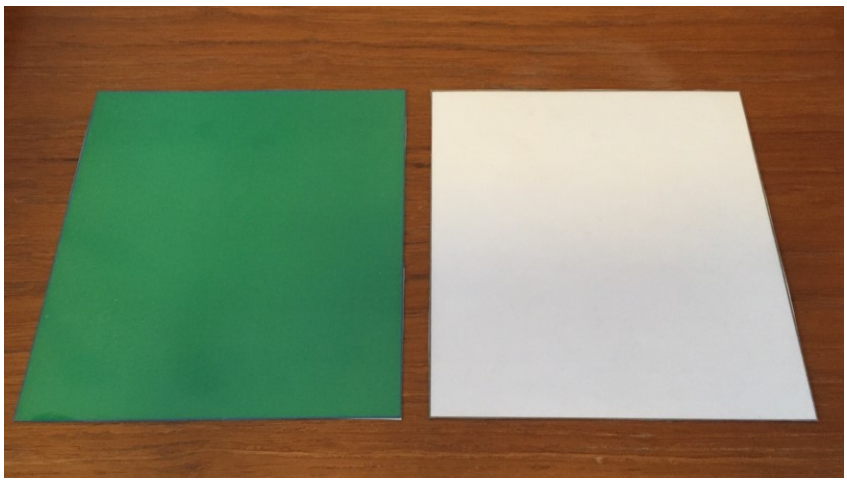

Figure 9. *Grass/snow*

Figure 10. *Modified Corsi blocks*

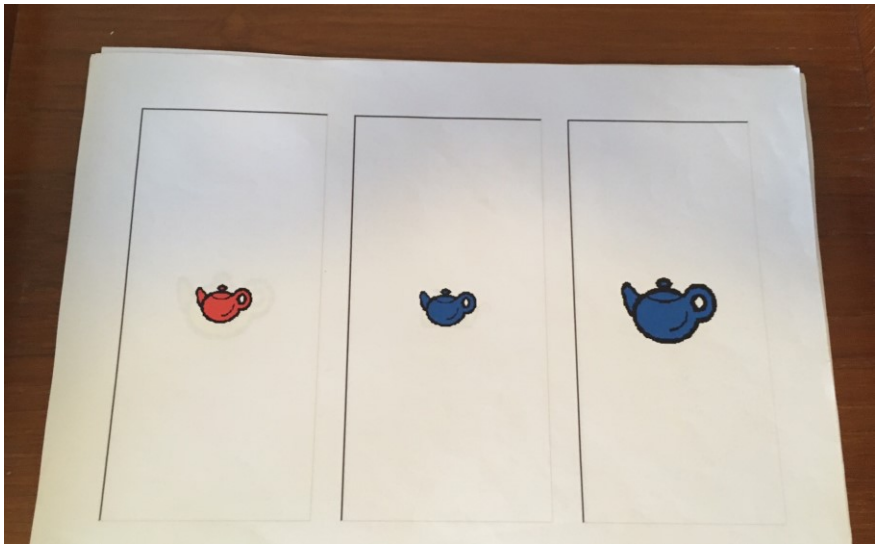

Figure 11. *FIST*

4.9 Mc measures

The two measures of Mc selected: the C.Ind.Le coding scheme was used for coding Mc rate during the physics task, and the Mc puppet interview was carried out after the balance beam task, and another Mc interview was administered after the ramps task (which included the same four questions, but with three additional ones). (See Appendix J for copies of each interview.)

This section has presented the tasks and necessary checks between support types. The next chapter will begin by discussing the coding and reliability of the Mc measures and the validity and reliability of the main study. The analyses used to address each research question will be described, information on aspects of the statistical analysis (including corrections and power analyses) will be discussed, and an overview of the ethics procedure will be provided.

5 Chapter 5
Methods for data analyses

The data were collected and analysed in order to address the research questions. Some data were scored as it happened, but much of it was scored and coded via video afterwards. The BPVS and NEPSY were scored live. The EF tasks were scored live and the videos were watched if I was not sure I had correctly noted something. The Mc coding during the physics tasks and the Mc interviews were coded when the videos were watched back. The strategies used by children during the balance beam task and ramps task were noted down at the time and always watched back afterwards to code and note strategies. Statistical analyses were carried out using IBM SPSS 23.

5.1 Mc coding scheme and inter-rater reliability

As the Mc measures relied on coding of Mc actions and verbalisations, the inter-rater reliability for the measures was checked and is presented next.

5.1.1 Mc coding scheme

The C.Ind.Le coding scheme was used to code the Mc behaviours during the physics tasks. As previously discussed, the codes to measure Mc knowledge and Mc skills were used and piloted three times for practice. The coding took place in several blocks, over the course of 18 months. The sections of the physics tasks that were coded for Mc were: GPs' play during the balance beam, the balance beam trials (and feedback), the Mc interviews after the balance beam and ramps tasks, and the time to use the ramps before the trials. The length of time for each section was also noted.

Coding happened while the data were collected over the six-month period, but some codes developed over this time and my willingness to categorise particular behaviours into certain codes changed, which meant I re-watched and recoded the videos at least three times to correct codes and ensure all videos were coded the same, regardless of when the data were collected.

Due to the structure of the task, it was found that some codes were used more frequently than others. For example, the MK persons code was never used, as no verbalisations related to that

code were identified. Some leniencies were made, such as in the verbalisations relating to MK tasks and MK strategies. If the child made a reference to knowing something about how the task works or how to solve it was coded as Mc, with some leeway in the language they used due to their age. The MR evaluation code was applied in the same way for all children, although again, leniently. This code was for children reviewing task performance or evaluating the quality of others' performance, so if a child commented on the outcome (i.e. if it was correct or incorrect) this counted as an MR evaluation code. For example, when the children in GP were asked if something worked their answers would likely fall into this code. Some examples of the coding used during the physics task can be seen in Table 19; refer to Appendix A for a definition of each code. No MK knowledge of persons codes were applied.

Table 19

Examples of coding used during the physics tasks

Mc component	Example from the current data
MK: knowledge of tasks	Child 600: "and if you put a light one over here it tips over" Child 602: "that one's too heavy it tips over"
MK: knowledge of strategies	Child 625: "because they're both in the same seats" Child 636: "if you put this one over there and this one will be the same seat"
MR: planning	Changing where they decided to put the weights was a frequent occurrence Commenting where the weights were going to go was also a frequent occurrence.
MR: monitoring	Looking back and forth between the pictures and the weights either on the beam/in hand/on table was often seen Commenting on progress was also seen, child 617: "that's not right"

Mc component	Example from the current data
MR: control	Changing strategy and repeating strategies between trials was often seen After seeing the beam tipping some children tried to fix their errors
MR: evaluation	Commenting on the outcome was often seen Child 636: "it looks like that one" comparing it to the picture on the table Child 629: "oh, it's not working"

5.1.2 Inter-rater reliability

To check the reliability of the coding, second-coders were used. Around 25% of the data were coded – see Table 20. Three coders were used: one coder coded the balance beam Mc during the trials and balance beam Mc during the play, another coder coded the ramps trials and the three to four minutes of free time to use the ramps before the trials, and another coder coded the balance beam and ramps Mc interviews. Videos of different children were selected, a mix of males and females, and from a mix of TPs. Coders were given the coding scheme, some of my own examples of the coding (not of the videos they had), and then we practised coding two videos together – one with me leading and another with them leading.

During the time the second coders worked on the data, some amendments were made to some of the codes. This resulted in coders agreeing to change four codes for all children: 1) what would be accepted as a child explaining a strategy, 2) adding in when a child repeated or copied a similar strategy they had already used (during that session), 3) what defined a self-correction, and 4) to take out “incorrect” answers from the children (for example, if children were asked if they beam balanced and they said yes, but it was not balancing). The practice trials were not included in the coding. Strategy changes, even if incorrect were coded. “Incorrect” verbalisations, such as saying the beam was balancing when it was not, were not coded.

Landis and Koch (1977) defined a Cohen’s Kappa (k) value of .61-.80 as showing a substantial level of agreement. Table 20 indicates the coding was reliable. The Kappa values for the Mc coding during the balance beam trials and GP may be slightly lower than hoped,

but likely due to the fact there were 30 codes in the coding scheme, and a higher number of codes gives a higher chance of lower inter-rater reliability. The exact agreement percentage is all above 67%, which is good. Close agreement was calculated for the Mc interviews as well, which resulted in very high Kappa values. Close agreement can be calculated by examining agreement if it is only one score off (and uses ordinal or interval data), for example, coder 1 giving a score of 1 and coder 2 giving a score of 2. Close agreement cannot be calculated for the Mc codes since the data are nominal.

As a comparison, when the C.Ind.Le was established, only 20 events (of 196, so 10.20% of events) were coded and their exact agreement was 74.80% (Whitebread, et al., 2009b). So, these data fair well since many more events were coded and the exact agreement did not drop more than 7.80% from Whitebread, et al.'s (2009b) results.

Table 20

Inter-rater reliability for the Mc coding

	% videos coded	Mc cases coded	% exact agreement	*k* for exact agreement	*p* and level of exact agreement	% close agreement	*k* for close agreement	*p* and level of close agreement
Balance beam Mc during trials	25.00	294	67.00%	.61	$p < .01$, on the border of substantial			
Balance beam Mc during GP	23.80	134	70.00%	.63	$p < .01$, substantial			
Balance beam Mc interviews	25.00	92	74.00%	.62	$p < .01$, substantial	100%	1.00	$p < .01$, substantial
Ramps Mc during free time	27.70	97	71.13%	.62	$p < .01$, substantial			
Ramps Mc during trials	27.70	91	74.73%	.66	$p < .01$, substantial			
Ramps Mc interviews	27.70	35	74.29%	.73	$p < .01$, substantial	97.14%	.94	$p < .01$, substantial

5.2 Validity and reliability throughout study

Various measures were in place to try and ensure the validity and reliability of the data. The piloting studies were carried out to ensure the tasks were reliable for what they were intended for. All of the children were assessed by one individual (me) and tasks were carried out in a familiar environment – the child's nursery – to make it less stressful. The tasks were very structured and due to the children completing them at three TPs and getting to know the adult better they were successfully completed. Background measures of vocabulary and visual-spatial skills were taken to act as control measures. All of the sessions were video-recorded to allow me to watch the tasks multiple times and ensure the data coded and scored for use in the analysis was as accurate as can be. The use of second-coders and high inter-rater reliability of coding allowed me to analyse data with confidence. It is believed the data are as reflective of the children's abilities for the targeted component as could be ensured.

5.3 Statistical tests employed to answer the research questions

The statistical tests used to address each research question are outlined next. This includes the analyses conducted before the research questions were addressed, including exploring the background measures, EF measures, and Mc measures, and how the balance beam and strategy development data are related, all of which were carried out before the research questions were addressed.

All data were plotted in histograms and examined visually for skewness, kurtosis, and outliers. Histograms were used to examine skewness and distribution. Skewness and kurtosis values were checked and if they were not close to zero they were converted to Z scores to check if they were distributed within the acceptable range of +/-1.96 (as suggested by Field, 2013 and Ghasemi & Zahediasl, 2012). Shapiro-Wilk tests were performed for all sets of data to check the normality of the data, as it is suitable for smaller sample sizes (Ghasemi & Zahediasl, 2012). The Shapiro-Wilk test was selected rather than the Kolmogorov-Smirnoff test since it is said to be more powerful and a better measure for testing normality, and not so affected by extreme scores (Ghasemi & Zahediasl, 2012). Due to the small sample size, some caution should be taken with the Shapiro-Wilk tests, but they will be considered alongside the skewness, kurtosis, and histograms, as suggested by Field (2013) and Ghasemi and Zahediasl (2012). Outliers were identified using boxplots and are reported in the results chapter when identified in the data. Scatterplots with fit lines were inspected for linearity. All of this aided

in determining whether the statistical tests selected could be used based on if the assumptions of the data were met.

Correlations were carried out to examine the relationships between the background measures, EF measures, and Mc measures, and between the balance beam and strategy development data. Correlations were selected as they give an indication of the strength of association between two variables (Rasch, Kubinger, Klaus, & Yanagida, 2011). Kendall Tau correlations were selected due to the small sample size, the differing strength of linearity between variables, and because this test is not sensitive to normality in the way the parametric Pearson correlation is (Field, 2013). Some of the data did not meet the assumptions required for Pearson's correlations (including normality), so using the Kendall Tau for non-normally distributed data was more appropriate, especially since the sample size was small. All Kendall Tau correlations reported in this thesis are two-tailed.

It is acknowledged that non-parametric tests (the Kendall Tau and others) may not be as powerful as parametric tests, but non-normally distributed data and small sample sizes are more suited to non-parametric tests to avoid violating the parametric test assumptions (Marusteri & Bacarea, 2009; McCrum-Gardner, 2008). Visual inspection of the data alongside considering normality testing will be used to decide whether to use a parametric or non-parametric test and parametric tests will be used where possible.

Some partial Spearman correlations were also carried out when factoring in covariates. Partial Spearman tests were selected since there is not a covariate Kendall Tau option in SPSS. The Spearman test is not quite as sensitive as the Kendall Tau, but it is more appropriate than parametric Pearson partial correlations due to the small sample size, linearity between variables, and normality of the data (Field, 2013). The outcome of the correlations determined how some of the variables were treated in later analyses. All partial Spearman correlations reported in this thesis are two-tailed.

Research question 1 aimed to examine the role of EF and Mc in children's physics task performance. To answer this, correlations were used to see the strength of the relationship between physics performance (overall and on problem types) and the EF and Mc variables. Again, for the reasons listed above, Kendall Tau or partial Spearman tests were used,

depending on whether a covariate was to be entered. One-way analysis of variance (ANOVA) (with group as the level) and analysis of covariance (ANCOVA) (with group as the level and entering a covariate into the model) were also used to examine group differences in EF or Mc based on children's strategy use classification. (It is noted that independent t-tests are the equivalent to one-way ANOVAs and give the same result.) ANOVAs test for a statistical difference between groups by analysing group means (Rasch et al., 2011). ANOVAs are said to be relatively robust to the assumptions placed on the data before running the test (including heterogeneity and normality) (Field, 2013). Regardless, before each analysis it is stated whether the assumptions of the data were met. The results will indicate if there is a difference in EF or Mc mean scores in the different groups (strategy use classification groups) and therefore whether EF or Mc might have a role in children's strategy use during the balance beam task.

To answer research question 2, what impact does support type have, one-way ANOVAs and ANCOVAs were used. These tests allow for the two support groups' scores to be compared to see if a significant difference in groups' mean EF, MC or physics scores exists at any TP. Separate univariate ANOVAs were used rather than carrying out a mixed ANOVA over the three TPs, due to a drop in *n* if a mixed ANOVA had been used. The separate ANOVAs will not highlight any interactions over the TPs and may incur a higher chance of a Type I error, but the use of one-way ANOVAs results in a slightly larger sample size. As with research question 1, the assumptions of the data are outlined before each analysis and when the assumption of variance was violated then a Mann-Whitney test was used instead, which is a non-parametric test, equivalent to ANOVA, to examine if two groups' data differ (Marusteri & Bacarea, 2009).

To compare the strategy classifications used by the groups for each balance beam problem, Fisher's exact tests were used. This is a version of the Chi-square test to compare the proportion of data in nominal data groups. It is appropriate for small sample sizes (<20) (McCrum-Gardener, 2008) and unequal and low numbers (<5) of observations in each category (Field, 2013). This test is suitable for unequal nominal data and allows for the two groups' classification for each problem to be examined to see if there is an association between them (Field, 2013; McCrum-Gardener, 2008).

5.4 Statistical analyses: power analyses and effect sizes

Statistical power indicates how likely it is to detect a significant effect from chance (Field, 2013). Field (2013) suggests that a power value of .8 is suitable, meaning there is an 80% chance of finding a significant result, should one exist. As corrections were applied it means the chance of type II errors increased, as previously discussed, but due to the number of tests performed it was required to reduce type I errors. Power is calculated using sample size and effect size and based on Field's (2013) suggestion, a power value of .8 will be used, so if the power is .8 or higher the result can be said to have enough power to detect an effect.

Effect sizes indicate the observed effect between what is being tested – the null and alternative hypothesis – so the bigger the difference between the two, the easier it is to find a large effect size (Field, 2013). Larger effect sizes indicate less overlap, meaning more difference between the null and the alternative hypothesis, and smaller effect sizes indicate more overlap, making small effect sizes more difficult to detect unless the tests have sufficient power. When there is not enough power to detect a significant *p* value it means the null hypothesis cannot be rejected, as it could be that there is a significant difference, but it has not been detected (Mayr, Erdfelder, Buchner, & Faul, 2007).

Some suggest effect size is more important than *p* values since they are independent of factors that influence the chance of finding a significant result, such as sample size (Sullivan & Feinn, 2012). Effect size, *p* values, power, and sample size are all linked, but *p* values are reliant on sample size, and power is reliant on sample size and effect size (Sullivan & Feinn, 2012). Effect size indicates the difference between the null and alternative hypotheses' data and *p* indicates whether it has reached statistical significance, but it could be that low power prevents a significant result (in terms of *p*) being detected, even if one exists. This is why effect sizes will be reported alongside *p* values in this thesis, to aid in determining whether to reject the null hypothesis.

Effect size guidelines are taken from Cohen (1988). Field (2013) suggests using Cohen's *d* as it is a standardised measure of effect size, as it is based on the groups' means, calculated in standard deviations. For ANOVAs, Eta squared is sometimes used, but Field (2013) suggests this is not a good measure of effect size as it is based on the results of the ANOVA itself, and is not objective in the same way Cohen's *d* calculates the difference between means, so Eta

will not be used for the majority of ANOVAs here. The strategy classification ANOVAs and ANCOVAs conducted to answer part of research question 1 will report partial Eta squared (η^2_p), as an effect size for that test, as there are multiple groups and Cohen's *d* is used for comparing two groups. Field (2013) states partial Eta squared is the proportion of variance explained that the other variables do not explain. This is not a standardised measure of effect size in the way Cohen's *d* is, but it provides a proportion in relation to the test run with that data (Gray & Kinnear, 2012). The effect sizes for partial Eta squared can be interpreted as .01, .09, and .25 for small, medium and large (Watson, 2018), but it should be acknowledged that it is displayed as a percentage proportion effect size.

Cohen's *d* will be reported for t-tests, ANOVAs and ANCOVAs. The correlation coefficient (*r*) will be reported for the effect size of the Kendall Tau and partial Spearman correlations. For Cohen's *d* a small effect size is 0.2, medium 0.5, and large 0.8 (Cohen, 1988). For *r*, a small effect size is 0.1-0.3, medium 0.3-0.5, and large 0.5 (Cohen, 1988). When calculating effect sizes in G*Power, *F* will be entered during calculations, with 0.1 entered for a small effect size, 0.25 for medium, and 0.4 for large (Cohen, 1988).

The sample size required to detect different effect sizes for each statistical test employed will be outlined next. A priori power analyses were calculated using G*Power version 3.1 (see Mayr et al., 2007 and Faul et al., 2009) entering alpha ($p = .05$), power (.8), and effect size (small, medium, and large – each size determined by the statistical test used) to find the sample size required to detect the different effect sizes. Two-tailed was selected when available for the given test.

For 2-tailed independent t-tests, 52 participants would be required to detect a large effect size, 126 for a medium effect size, and 394 for a small effect size. For 2-tailed (bivariate normal) correlations, 29 participants would be required to detect a large effect size, 84 for a medium effect size, and 782 for a small effect size. This would apply to both the Kendall Tau and partial Spearman correlations used. For one-way ANOVAs, a sample size of 52 would be required to detect a large effect size, 128 for a medium effect size, and 788 for a small effect size. For the one-way ANCOVAs here (with one covariate) a sample size of 52 would be required to detect a large effect size, 128 for a medium effect size, and 787 for a small effect

size. Mann-Whitney tests require a sample size of 54 to detect a large effect size, 134 for a medium effect size, and 824 for a small effect size.

For Fisher's exact test, the contingency tables tab of G*Power will be used, due to the analyses consisting of 2x3, 2x4, and 2x5 contingency tables. Cohen's (1988) *w* is required for this and is defined as .1 for small, .3 for medium, and .5 for large. For the 2x3 tables, 39 participants are required to detect a large effect size, 108 for medium, and 964 for small. For 2x4 contingency tables it is 44, 122, and 1091 participants, and for 2x5 contingency tables it is 48, 133, and 1194 participants, each to detect large, medium, and small effect sizes.

These power analyses will be referred back to when reporting the results of each test. Sullivan and Feinn (2012) recommend reporting *p* and effect sizes in order to provide a better understanding of whether a difference exists, so these will be reported throughout. It is likely that some of the tests will not have sufficient power to detect significant differences if they exist, due to the small sample size. (Note: references to "very small" effect sizes in this thesis are those that fall below the cut-off for small effect sizes.)

5.5 Statistical analyses: correcting for multiple comparisons

One issue that arose during the statistical analyses was the use of multiple corrections when carrying out multiple tests. Not correcting for multiple comparisons increases the likelihood of declaring a result statistically significant when it is instead due to chance through the high number of tests carried out. This is known as a type I error – incorrectly rejecting the null hypothesis when a statistical effect does not actually exist in the data (Field, 2013). By over-correcting for multiple comparisons, a type II error may occur: failing to reject the null hypothesis when there is actually a statistical effect in the data (Field, 2013). A balance between these had to be found and an appropriate correction applied to reduce type I errors, to not increase type II errors. The analyses were set around two research questions, with particular statistical tests to address them. However, there were also some exploratory analyses to investigate how certain data were related to one another, and not set around a research question and hypothesis. There is much debate in the literature about if and when multiple comparisons should be corrected to control for inflated alpha levels.

Streiner and Norman (2011) suggested that corrections do not need to be applied if hypotheses are formulated before data analyses are carried out and correcting for multiple analyses instead increases the chance of type II errors. Streiner and Norman (2011) advise not correcting for such analyses and that instead any significant findings should be highlighted as potential future areas to study, which is why reducing type II errors is important. They also note that multiple corrections should be applied if the analyses are examining the same outcome, but with different data (for example, if vocabulary had been measured with two different tests). Armstrong (2014) also suggested applying corrections if hypotheses and tests are not first formulated.

I decided to follow some of the key papers' recommendations and only corrected for multiple comparisons if the analysis was exploratory and not in response to trying to answer a hypothesis. Thus, the first set of analyses in the results chapter have been corrected, as they were exploratory. I also decided not to correct for every test that was carried out to answer the research questions to reduce the chance of type II errors and declaring results not significant when they actually are. Correcting for every test would be too restrictive and so this balance between correcting exploratory analyses and not correcting when testing hypotheses were chosen. This work is still fairly new to the field, so highlighting potential significant findings was deemed more informative than correcting for very comparison and reducing the chance of finding any significant results. All statistical tests reported in the results chapter are two-tailed due to the unknown direction of data and the unknown existence of links between data.

The correction used was the sequential Holm-Bonferroni (a full explanation of the formula can be found in Eichstaedt, Kovatch, & Maroof, 2013). This method requires that the *p* values are ordered sequentially from lowest to highest (once the statistical tests are carried out) and then each is tested with the Holm-Bonferroni correction applied, starting from the lowest. Alpha is divided by the number of tests, minus the rank of the test, plus one until the tests turn non-significant. This method is said to be less conservative than the Bonferroni method, but as effective due to the sequential methods, which results in more statistical power and fewer type II errors (Eichstaedt et al., 2013).

The Holm-Bonferroni correction should reduce incorrectly rejecting the null hypothesis, but not be too conservative that it increases accepting the null hypothesis when there is, in fact, an effect. The potential issue of a lack of power due to a small sample size may increase type I errors, as the null hypothesis cannot be rejected if there is not enough statistical power to run a test. It is believed that incorporating corrections alongside reporting *p* values, effect sizes, and any power issues will give a balanced conclusion when reporting results.

Other post-hoc corrections that could have been applied include the Bonferroni correction, and Holm's sequential method, as neither assumes the tests are independent and are not restricted to pairwise comparisons only. The Bonferroni is often seen as quite conservative due to the method used, likely increasing type II errors (Field, 2013), and reducing the statistical power of the tests (Eichstaedt et al., 2013). The Holm's sequential method is similarly effective to the Bonferroni method, while maintaining statistical by using a sequential method to order alphas by size (and dividing by the number of tests) to find which are significant, rather than divide by the number of tests used, resulting in a less restrictive type I error rate (Eichstaedt et al., 2013). Thus, using the Holm-Bonferroni method appears to be the most appropriate and justified here.

5.6 Ethical considerations

The appropriate ethics form, risk assessment, and information and consent forms were been submitted to the Faculty of Education and were approved. I obtained a CRB check and the university received a copy.

No known risks were associated with any of the tests or tasks that were used. Complete information was provided to participating nurseries and informed written consent was obtained from parents/guardians with the option to withdraw their child from the study at any point without explanation. Written consent was also obtained from parents/guardians to video their child during each session, with the assurance that the recordings, along with all other information obtained from the family (electronic and hardcopy) will be stored securely. Copies of the information sheet and consent form can be seen in Appendix K.

The tasks were explained to the child in the most simple way and the child was reassured that they did not need to complete anything they did not want to. The children received stickers at

the end of every session, and at the end of the last session they received a book gift set. Nurseries were also provided with a gift and a thank you card.

The video data (labelled by participant number) are stored on encrypted external hard drives and kept in my personal possession. The consent forms with names and accompanying participant codes are also kept in my personal possession and only I have access. All of the electronic data are all anonymised, encrypted, and stored against participant code only. All data will be reviewed for destruction five years after collection and securely destroyed within ten years.

6 Chapter 6

Results

The results will be presented in five sections. The first section will present information on the background measures, EF measures, and Mc measures, including how they relate to one another. The second section will examine the balance beam and strategy development data and how they are related. The third section will answer research question 1: what role do EF and Mc have in children's performance on physics tasks? The fourth section will answer research question 2: what impact does support type have on EF, Mc, balance beam performance, strategy development, and the transfer ramps task (as well as detail information on the ramps task data).

6.1 Results: Exploring the background measures, EF measures, and Mc measures

This section reports on the three background measures: age (in complete months at the first time visited), BPVS (receptive vocabulary) scores (raw scores are reported throughout), and NEPSY (visual-spatial ability) scores (raw scores are reported throughout), as well as the EF measures, and the Mc measures (rate and interview scores). This analysis was important in order to see whether the background measures related to EF and Mc, as a decision had to be made whether to consider any variables as covariates. It was also important to see how the EF and Mc scores presented, as a decision had to be made on whether to combine the different components of each measure.

6.1.1 Background measures

During the first visit, all children completed the BPVS and NEPSY tasks, and age is calculated from this date. The means, SDs, and ranges for age (in months), BPVS scores, and NEPSY scores can be seen in Table 21.

Table 21

Mean (SD) and range for age, BPVS scores, and NEPSY scores at TP1

	Mean	Range
Age	44.53 (4.81)	36.00-55.00
BPVS	55.11 (15.20)	29.00-87.00
NEPSY	5.55 (1.90)	3.00-12.00

Note. $N = 38$

Table 21 shows the mean age to be 44 months. The BPVS mean raw score was 55.11 and shows quite a large range, from 29 to 87. The NEPSY mean raw score was 5.55 and shows a range from 3 to 12. Overall, it appears the participating children show a wide range of abilities on the measures.

Data screening was carried out and can be seen in Appendix L. No issues were seen with BPVS or age, but the NEPSY scores were slightly skewed and the Shapiro-Wilk test of normality was significant ($W = .90$, $p < .01$). Skewness and kurtosis values were converted to Z scores and the data were found to be normally distributed, as seen within the +/-1.96 range (Field, 2013). Therefore the NEPSY data were not transformed and the remaining analyses will be carried out using the original data.

Kendall Tau correlations were carried out between the background measures to see how they related to each other. As this was an exploratory analysis and not in answer to a research question, the Holm-Bonferroni correction was applied – see Table 22.

Note: statistically significant correlations in this results chapter will be **highlighted in bold** throughout the document for clarity, especially since adjustments are often applied, meaning the alpha level will change per analysis. Numbers will be displayed to two decimal places unless more decimal places are required for clarity. The Holm-Bonferroni sequential method was used for all the corrected analyses.

Table 22

Kendall Tau correlations between age, BPVS scores, and NEPSY scores

	Age	BPVS
Age		
BPVS	**.31, *p* < .01**	
NEPSY	**.41, *p* < .01**	.18, *p* = .13

Note. $N = 38$.

Table 22 shows significant correlations were found between age and BPVS scores, and age and NEPSY scores, suggesting age relates to both, which may be expected since raw scores

should increase with age. The correlation between BPVS and NEPSY scores showed no statistically significant relationship and only a small effect size, but the low power could mean a significant relationship was not detected.

6.1.2 EF measures

The performance scores (percentage correct) for the EF measures are presented next. In addition to the EF scores for inhibition, WM, and shifting, a Z score was also computed for each child at each TP. Z scores were calculated by first subtracting each individual's score from the mean for that task and then dividing by the group's SD for the task, and then averaging the three EF Z scores at that TP. Positive Z scores indicate individual scores higher than the mean, displayed in terms of SD. The means, SDs, and the range for each of the EF measures and the composite Z score information (referred to as composite from hereon) at each TP can be seen in Table 23. (Z score calculations result in a group mean of 0.)

Table 23

Mean (SD) and range for each EF score at TPs 1, 2, and 3

EF and TP	*N*	Mean	Range
Inhibition 1	38	54.61 (36.26)	0 – 100
Inhibition 2	36	57.64 (36.23)	0 – 100
Inhibition 3	37	67.06 (32.99)	0 – 100
WM 1	38	23.03 (17.20)	0 – 50.00
WM 2	36	29.51 (17.46)	0 – 56.25
WM 3	37	30.57 (17.23)	0 – 62.50
Shifting 1	38	36.02 (26.17)	0 – 87.50
Shifting 2	36	54.69 (29.29)	0 – 93.75
Shifting 3	37	48.82 (33.03)	0 – 100
Composite 1	38	0 (2.26)	-4.22 – 4.27
Composite 2	36	0 (2.26)	-4.98 – 4.03
Composite 3	37	0 (2.36)	-5.10 – 3.09

Table 23 shows an increasing trend in mean scores over time, with the exception of shifting at TP3. The SDs are roughly the same over the TPs, suggesting there is still a fair amount of variation in the scores.

The assumptions of the EF data were checked and some issues found (Appendix M). Histograms showed some of the EF measures were skewed. The Shapiro-Wilk test of normality showed the data not to be normally distributed for all of the individual EF variables at each TP (all nine tests ***p* < .05**). EF composite scores at TPs 1 and 2 did not violate the assumption of normality, but they did at TP3 (W = .93, p = .03), likely because the scores were positively skewed (i.e., performance was higher at TP3). Skewness and kurtosis Z scores indicated the data were within a normal distribution, except for inhibition at TP1, thus were not transformed. Due to the small sample size, the varying strengths of linearity between variables, and the potential issue of non-normality for some variables, two-tailed Kendall Tau correlations were used to investigate the relationship between the EF variables – these are detailed next.

The EF measures were examined to see how they related to one another. Results are presented at each TP (Tables 24 – 26) and for each measure (Tables 27 – 30). All performance data are percentage correct. The Holm-Bonferroni correction was applied.

Table 24

Kendall Tau correlations between the EF measures at TP1

	Inhibition 1	WM 1	Shifting 1
Inhibition 1			
WM 1	.32, p = .01		
Shifting 1	.23, p = .06	.31, p = .01	
Composite 1	**.55, *p* < .001**	**.64, *p* < .001**	**.58, *p* < .001**

Note. N = 38.

Table 25

Kendall Tau correlations between the EF measures at TP2

	Inhibition 2	WM 2	Shifting 2
Inhibition 2			
WM 2	**.36, *p* = .006**		
Shifting 2	.27, *p* = .03	.23, *p* = .07	
Composite 2	**.66, *p* < .001**	**.61, *p* < .001**	**.50, *p* < .001**

Note. $N = 36$.

Table 26

Kendall Tau correlations between the EF measures at TP3

	Inhibition 3	WM 3	Shifting 3
Inhibition 3			
WM 3	.30, *p* = .02		
Shifting 3	**.44, *p* < .006**	.28, *p* = .03	
Composite 3	**.62, *p* < .001**	**.60, *p* < .001**	**.68, *p* < .001**

Note. $N = 37$.

Table 27

Kendall Tau correlations between the inhibition scores over the three TPs

	Inhibition 1	Inhibition 2
Inhibition 1		
Inhibition 2	**.41, *p* = .006** *n* = 36	
Inhibition 3	**.40, *p* = .006** *n* = 37	**.55, *p* < .006** *n* = 35

Table 28

Kendall Tau correlations between the WM scores over the three TPs

	WM 1	WM 2
WM 1		
WM 2	**.55, *p* < .006**	
	n = 36	
WM 3	**.49, *p* < .006**	**.48, *p* < .006**
	n = 37	*n* = 35

Table 29

Kendall Tau correlations between the shifting scores over the three TPs

	Shifting 1	Shifting 2
Shifting 1		
Shifting 2	**.39, *p* < .006**	
	n = 36	
Shifting 3	**.35, *p* < .006**	**.41, *p* < .006**
	n = 37	*n* = 35

Table 30

Kendall Tau correlations between EF composite scores at the three TPs

	Composite 1	Composite 2
Composite 1		
Composite 2	**.54, *p* < .001**	
	n = 36	
Composite 3	**.48, *p* < .001**	**.62, *p* < .001**
	n = 37	*n* = 35

Tables 24-26 show the three individual EF measures do not always significantly correlate with one another at each TP, but all three significantly positively correlate with the composite scores. At TP1 no significant correlations were detected between the three individual

measures, at TP2 inhibition and WM significantly correlated, and at TP3 inhibition and shifting significantly correlated. The significant correlations show medium or large effect sizes. The non-significant correlations are all small or medium effect sizes, indicating there could be a relationship there, but may not reach statistical significance due to low power. Tables 27-30 show each individual EF and the composite score to be significantly correlated at each TP, and with medium or large effect sizes, indicating that the measures were related and stable over time.

The EF measures' relationships with the background measures were examined to check whether they were related to the EF scores and if they should be considered covariates. Data screening was carried out and no issues found (Appendix N).

Next, two-tailed Kendall Tau correlations were carried out. The Holm-Bonferroni method was used for the exploratory multiple comparisons. The results of the correlations can be seen in Table 31.

Table 31

Kendall Tau correlations between EF, age, BPVS scores, and NEPSY scores

EF and TP	Age	BPVS	NEPSY
Age			
BPVS			
NEPSY			
Inhibition 1	.25, $p = .04$	.16, $p = .18$	.21, $p = .10$
Inhibition 2	.07, $p = .60$	.22, $p = .07$	.14, $p = .29$
Inhibition 3	.13, $p = .30$	.34, $p < .01$	.02, $p = .88$
WM 1	.23, $p = .06$	.28, $p = .02$	.39, $p < .01$
WM 2	.10, $p = .44$	.25, $p < .05$	.26, $p < .05$
WM 3	.18, $p = .14$	.32, $p < .01$	.10, $p = .46$
Shifting 1	.07, $p = .57$	.35, $p < .01$	.01, $p = .91$
Shifting 2	.27, $p = .03$	.34, $p < .01$	.08, $p = .56$
Shifting 3	.08, $p = .51$	.29, $p = .01$	-.08, $p = .55$
Composite 1	.20, $p = .08$	.31, $p < .01$	.25, $p = .04$
Composite 2	.14, $p = .24$	.36, $p < .01$	.17, $p = .17$
Composite 3	.15, $p = .19$	**.43, $p < .001$**	.05, $p = .68$

Notes. $N = 38$ at TP1. $N = 36$ at TP2. $N = 37$ at TP3.

The only statistically significant correlation was between BPVS and EF composite at TP3. Most correlations showed a small or medium effect size, but due to the post-hoc corrections being applied some were not deemed statistically significant. Age and visual-spatial ability were not seen to be significantly related to EF scores. However, the correlations here only had enough power to detect large effect sizes, so it could be that significant differences exist in the data, but were not detected. To check whether BPVS scores are related to the EF composite scores, partial correlations were carried out and can be seen next.

The EF correlational analysis was repeated controlling for BPVS scores to see what effect it had on the EF measures' relationships. As explained in Section 5.3, it is not possible to carry

out partial Kendall Tau correlations in SPSS version 23, so syntax was obtained to carry out Spearman partial correlations (Watson, 2016). The Holm-Bonferroni correction was applied and the results can be seen in Tables 32 – 38.

Table 32

Spearman partial correlations between EF scores at TP1, controlling for BPVS

	Inhibition 1	WM 1	Shifting 1
Inhibition 1			
WM 1	.36, $p = .03$		
Shifting 1	.26, $p = .13$	.27, $p = .11$	
Composite 1	**.71, $p < .004$**	**.74, $p < .004$**	**.65, $p < .004$**

Note. df = 35.

Table 33

Spearman partial correlations between EF scores at TP2, controlling for BPVS

	Inhibition 2	WM 2	Shifting 2
Inhibition 2			
WM 2	.37, $p = .03$		
Shifting 2	.26, $p = .13$	.20, $p = .25$	
Composite 2	**.80, $p < .004$**	**.72, $p < .004$**	**.60, $p < .004$**

Note. df = 33.

Table 34

Spearman partial correlations between EF scores at TP3, controlling for BPVS

	Inhibition 3	WM 3	Shifting 3
Inhibition 3			
WM 3	.24, $p = .17$		
Shifting 3	.46, $p < .01$	.23, $p = .18$	
Composite 3	**.72, $p < .004$**	**.60, $p < .004$**	**.80, $p < .004$**

Note. $df = 34$.

Table 35

Spearman partial correlations between inhibition scores over the three TPs, controlling for BPVS

	Inhibition 1	Inhibition 2
Inhibition 1		
Inhibition 2	**.51, $p < .004$** ($df = 33$)	
Inhibition 3	**.49, $p < .004$** ($df = 34$)	**.66, $p < .004$** ($df = 32$)

Table 36

Spearman partial correlations between WM scores over the three TPs, controlling for BPVS

	WM 1	WM 2
WM 1		
WM 2	**.63, $p < .004$** ($df = 33$)	
WM 3	**.51, $p < .004$** ($df = 34$)	**.53, $p < .004$** ($df = 32$)

Table 37

Spearman partial correlations between shifting scores over the three TPs, controlling for BPVS

	Shifting 1	Shifting 2
Shifting 1		
Shifting 2	.38, $p = .03$ ($df = 33$)	
Shifting 3	.33, $p = .05$ ($df = 34$)	.45, $p = .01$ ($df = 32$)

Table 38

Spearman partial correlations between composite scores over the three TPs, controlling for BPVS

	Composite 1	Composite 2
Composite 1		
Composite 2	**.65, $p < .004$** ($df = 33$)	
Composite 3	**.52, $p < .004$** ($df = 34$)	**.69, $p < .004$** ($df = 32$)

When these Spearman partial correlations controlling for BPVS scores are compared with the earlier Kendall Tau correlations it can be seen that five correlations turned non-significant, suggesting BPVS scores are related to EF scores. The correlation between inhibition and WM at TP2 turned not significant, the correlation between inhibition and shifting at TP3 turned not significant, as did all three correlations between the shifting scores over the three TPs. The shifting scores suggest there may be something different happening with this component. These correlations had the power to detect large effect sizes, so it could be that medium and small effect sizes exist but are not statistically significant due to low power, and this idea is supported by the medium and small effect sizes seen.

Overall, these analyses have found that age and NEPSY scores are not statistically related to the EF scores, and so will not be used as covariates in future EF analysis. However, BPVS scores were found to have a positive relationship with the EF scores, and so will be considered as a covariate in the EF analyses. It was acknowledged throughout that the tests had enough statistical power to identify large effect sizes, but it could be that the tests showing medium effect sizes lacked the power to detect a statistical difference, thus the null hypothesis is not supported but it also cannot be rejected.

The results within each TP suggest that the three EF tasks were likely measuring a related component, as seen by the detection of some significant correlations, the medium to large effect sizes between components, and all the individual components significantly correlating with the composites. The correlations over the TPs suggest reliability in the EF measures and consistency in scores. This finding gives support to the idea that the composite consists of three separate EF components, but all of which are related. Since the EF composite scores significantly correlated with all the individual EF measures and with each other, the composite scores will be used as the main measure of EF throughout the remaining analysis, although the individual EF measures will be reported for comparative purposes where appropriate. (Note the EF composite will be referred to as EF scores hereon.) In the next analyses, the individual Mc measures' relationships with the background measures will be examined.

6.1.3 Mc measures

Next, the Mc measures will be presented in a similar manner to the EF measures, starting with Mc rate then the interview scores.

6.1.3.1 Mc rate

The Mc rate was calculated as the number of Mc behaviours displayed per minute during the balance beam trials. The means, SDs, and the rate range for each Mc code, each Mc regulation (comprising of planning, monitoring, control, and evaluation) and Mc total at each TP, can be seen in Table 39.

Table 39

Mc rate means, SDs, and range at each TP

Mc code and TP	Mean (SD)	Range
Knowledge 1	0.17 (0.34)	0 – 1.36
Knowledge 2	0.28 (0.45)	0 – 1.29
Knowledge 3	0.34 (0.55)	0 – 2.54
Planning 1	0.34 (0.40)	0 – 1.24
Planning 2	0.30 (0.45)	0 – 1.77
Planning 3	0.29 (0.30)	0 – 1.27
Monitoring 1	0.52 (0.52)	0 – 1.74
Monitoring 2	0.56 (0.54)	0 – 1.96
Monitoring 3	0.67 (0.48)	0 – 2.01
Control 1	0.67 (0.39)	0 – 1.36
Control 2	0.78 (0.44)	0 – 1.92
Control 3	0.75 (0.42)	0 – 1.71
Evaluation 1	1.25 (1.03)	0 – 4.59
Evaluation 2	1.35 (1.14)	0 – 3.45
Evaluation 3	1.13 (1.11)	0 – 3.44
Total Mc regulation 1	2.78 (1.22)	.76 – 6.11
Total Mc regulation 2	2.99 (1.30)	.87 – 5.61
Total Mc regulation 3	2.85 (1.35)	.71 – 5.66
Total Mc rate 1	2.95 (1.38)	0.76 – 6.11
Total Mc rate 2	3.27 (1.60)	0.87 – 6.42
Total Mc rate 3	3.19 (1.76)	0.71 – 7.61

Notes. $N = 36$ for TP1. $N = 31$ for TP2. $N = 29$ for TP3.

Table 39 shows the mean Mc rate is quite low for the individual codes and there is not a large increase between TPs. Mean Mc knowledge and monitoring show an increase over time, but the other codes tend to be similar at each TP or show fluctuations. The mean total Mc regulation and Mc total rate are similar at each TP, with some fluctuation, although TP3 is higher than TP1. The maximum rate shows an upward trend over time for the total MC rate. Overall, there does not appear to be a strong change in Mc rate over the three TPs. It was decided to combine the individual codes into a total for each child due to the small means,

SDs, and ranges, and because using a large number of codes will require many corrections to be made, which will reduce the statistical power further, potentially hiding any meaningful findings. These Mc scores for each TP can be seen in Table 40.

Table 40

Mean (SD) and range for each Mc rate at TPs 1, 2, and 3

	N	Mean	Range
Mc rate 1	36	2.95 (1.38)	0.76-6.11
Mc rate 2	31	3.27 (1.60)	0.87-6.42
Mc rate 3	29	3.19 (1.76)	0.71-7.61

As seen in Table 40, the means are relatively small and the ranges show a little more variation in scores. The upper range shows an upward trend over time, but the means and SDs are quite similar over time, with TP2 showing the highest mean.

The assumptions of the data were checked (Appendix O) and no issues were seen for TPs 1 or 2, but issues seen at TP3. Total Mc rate at TP3 showed a positive skew and the Shapiro-Wilk test was significant (W = .93, p = .04). However, Z scores of the skewness and kurtosis fell within the +/-1.96 range, so it was decided to use the data in their current format, rather than transform the Mc rate data. Two-tailed Kendall Tau correlations (with the Bonferroni-Holm correction applied) were carried out to examine how the Mc rate relate (Table 41).

Table 41

Kendall Tau correlations between the Mc rates

	Mc rate 1	Mc rate 2
Mc rate 1		
Mc rate 2	**.34, *p* = .01** (n = 29)	
Mc rate 3	**.32, *p* = .02** (n = 27)	.26, p = .07 (n = 26)

Table 41 shows Mc rate significantly positively correlated at TPs 1 and 2, and 1 and 3, but not between 2 and 3. Two of these tests lack sufficient power to detect large effect sizes (a

sample size of 29 is required), so it could be the non-significant correlation has not got sufficient power to detect a difference, if it exists in the data.
In order to see how age, BPVS, and NEPSY scores relate to these measures Holm-Bonferroni-corrected Kendall Tau correlations were carried out (Table 42).

Table 42
Kendall Tau correlations between Mc rate and age, BPVS and NEPSY over the three TPs

	Age	BPVS	NEPSY
Age			
BPVS			
NEPSY			
Mc rate 1	.14, p = .26	.29, p = .01	-.05, p = .72
Mc rate 2	.07, p = .61	.18, p = .16	-.02, p = .86
Mc rate 3	-.07, p = .59	.24, p = .07	-.04, p = .77

Notes. N = 36 at TP1. N = 31 at TP2. N = 29 at TP3.

No statistically significant correlations were detected between the background measures and the Mc rate. Most of the effect sizes are very small, but there are some small effect sizes, suggesting there could be some small relationships between the data. The low power may have contributed to whether statistically significant relationships could have been detected. The background measures will not be considered as covariates in the Mc rate analyses based on the test results and effect sizes seen. The Mc interview scores will be presented now, followed by how the Mc measures relate to one another.

6.1.3.2 Mc interview scores
The Mc interviews were carried out after the balance beam task and calculated as a percentage correct. The means, SDs, and ranges can be seen in Table 43.

Table 43

Mc interview scores means, SDs, and range at each TP

	N	Mean (SD)	Range
Mc interview 1	35	42.86 (18.26)	12.50 – 75.00
Mc interview 2	31	48.79 (19.19)	25.00 – 100
Mc interview 3	29	48.71 (19.29)	25.00 – 87.50

The Mc interview scores increase from TP1 to 2 but have near identical means and SDs at TPs 2 and 3. The range at TP1 is lower than TPs 2 and 3. Overall the Mc interview scores do not appear to have changed substantively over the three TPs.

The data were screened and non-normality issues were found (Appendix P). The data were positively skewed at TP2 and the Shapiro-Wilk tests were significant at each TP (***p* < .01**, ***p* = .02**, ***p* = .01**). The data were log10 transformed and root squared transformed to investigate whether this would help with the distribution of data. The histograms did not improve and the spread of data and all the Shapiro-Wilk tests that were originally significant were still significant. Due to non-normality still existing after trialling two different transformations, it was decided not to use transformed data and instead use the original data. Two-tailed Kendall Tau correlations were thus used for the correlational analyses, due to the small sample size and because normality is not an assumption of this statistical test. Table 44 shows how the Mc interview scores relate to one another at each TP and Table 45 show how they relate to the background measures.

Table 44

Kendall Tau correlations between the Mc interview scores over TPs

	Mc interview 1	Mc interview 2
Mc interview 1		
Mc interview 2	**.51, p < .01** (n = 28)	
Mc interview 3	**.38, p = .02** (n = 26)	**.32, p = .05** (n = 26)

Table 45 indicates the Mc interview scores significantly positively correlated at each TP, although the tests were underpowered, but the effect sizes are medium and large, supporting the significant association,. This finding indicates that it was perhaps measuring the same construct each time and that it is stable.

Table 45

Kendall Tau correlations between Mc interview scores and age, BPVS scores, and NEPSY scores over the three TPs

	Age	BPVS	NEPSY
Age			
BPVS			
NEPSY			
Mc interview 1	.34, p < .01	**.40, p < .0031**	.28, p = .04
Mc interview 2	.36, p = .01	**.52, p < .0031**	.31, p = .03
Mc interview 3	.01, p = .95	**.44, p < .0031**	-.02, p = .92

Notes. N = 35 at TP1. N = 31 at TP2. N = 29 at TP3.

Table 45 shows no significant correlations were detected between Mc interview scores and either age or NEPSY, but since these tests only had enough power to detect large effect sizes it could be that smaller relationships exist in the data but have not been detected. Mc

interview scores significantly correlated with BPVS at each TP. To investigate this further, two-tailed Holm-Bonferroni adjusted Spearman partial correlations entering Mc interview scores and controlling for BPVS were carried out – see Table 46.

Table 46

Spearman partial correlations entering Mc interview scores while controlling for BPVS scores

	Mc interview 1	Mc interview 2
Mc interview 1		
Mc interview 2	.40, p = .04 (df = 25)	
Mc interview 3	.21, p = .33 (df = 23)	.03, p = .87 (df = 23)

Comparing Table 46 with the earlier Kendall Tau correlations shows that BPVS scores did appear to be significantly related to Mc interview scores since all three correlations turned non-significant. However, the small sample size results in low power, making it more difficult to detect a significant difference, if one exists. The results here indicate BPVS scores to be significantly correlated with the Mc interview scores, so BPVS will be considered a covariate in the Mc interview score analysis.

6.1.3.3 How do the Mc measures relate to each other?

The relationship between the Mc measures was investigated to see how they related to one another. Two-tailed Kendall Tau correlations were carried out between Mc rate and Mc interview scores at each TP, with BPVS added as a covariate due to its link with Mc interview scores. The Holm-Bonferroni method was used to adjust for these exploratory tests and the results can be seen in Table 47.

Table 47

Spearman partial correlations between Mc rate and Mc interview scores, entering BPVS as a covariate

	Mc rate 1	Mc rate 2	Mc rate 3
Mc rate 1			
Mc rate 2			
Mc rate 3			
Mc interview 1	-.02, p = .91 (df = 32)	.23, p = .25 (df = 25)	.29, p = .16 (df = 23)
Mc interview 2	-.06, p = .76 (df = 26)	.14, p = .45 (df = 28)	-.04, p = .86 (df = 23)
Mc interview 3	-.07, p = .74 (df = 24)	-.22, p = .30 (df = 23)	.30, p = .13 (df = 26)

Table 47 shows that no statistically significant relationships between the Mc measures were detected. It was expected that there would be significant relationships between the Mc rate over the three TPs, between the Mc interview scores over the three TPs, and between the Mc rate and Mc interview score at each TP. The findings indicate that the two Mc measures were not strongly related to each other and therefore the decision was made to keep them separate, rather than combine them to make one Mc measure. The sample size for some of the correlations fell below the 29 required to detect a large effect size, which potentially contributed to these findings. Therefore the null hypothesis that the measures are not related cannot be rejected nor fail to be rejected. The effect sizes here range from very small to the cut-off for medium, indicating a range of differences between the scores. It may be that with more power that some significant results may have been detected.

6.1.4 How are the EF and Mc scores related to each other?

The relationship between EF and Mc scores is presented next. Two-tailed Spearman partial correlations were carried out, entering BPVS as a covariate. The correlations between EF scores and the Mc rate can be seen in Table 48 and correlations between EF scores and Mc

interview scores can be seen in Table 49. The Holm-Bonferroni method was applied to each set of analyses.

Table 48

Spearman partial correlations between EF scores and Mc rate, controlling for BPVS scores

	EF 1	EF 2	EF 3
EF 1			
EF 2			
EF 3			
Mc rate 1	.04, p = .84 (df = 33)	-.09, p = .62 (df = 31)	.03, p = .86 (df = 32)
Mc rate 2	-.07, p = .73 (df = 28)	.08, p = .69 (df = 28)	.18, p = .34 (df = 28)
Mc rate 3	.25, p = .21 (df = 26)	.22, p = .27 (df = 25)	0.01, p = .99 (df = 26)

Table 49

Spearman partial correlations between EF scores and Mc interview, controlling for BPVS scores

	EF 1	EF 2	EF 3
EF 1			
EF 2			
EF 3			
Mc interview 1	.25, p = .16 (df = 32)	.32, p = .07 (df = 30)	.15, p = .40 (df = 31)
Mc interview 2	.19, p = .31 (df = 28)	.18, p = .33 (df = 28)	.25, p = .18 (df = 28)
Mc interview 3	.42, p = .03 (df = 26)	.49, p = .01 (df = 25)	.41, p = .03 (df = 26)

Tables 48 and 49 show that no significant correlations were detected between any of the EF scores and any of the Mc rate or Mc interview scores. The effect sizes between EF scores and Mc rate are very small or small, indicating there is likely only a weak relationship between these data, but the effect sizes between EF scores and Mc interview scores range from small to medium, indicating there could be a stronger link between these data. The small sample size for some correlations resulted in power lower than 80% to detect large effect sizes, contributing to whether significant correlations could be detected. This means that the null hypothesis cannot be rejected or fail to be rejected, as the tests do not have enough power to confidently support this.

6.1.5 Exploratory analyses summary

The analyses from section 1 of this chapter have shown that performance on each individual EF measure is somewhat significantly related over the TPs, but each individual measure did significantly relate to the EF composite scores. The low power in the correlations may have resulted in medium and small effect sizes not being detected. This resulted in the decision to use the EF composite scores for each child at each TP for the EF analyses. BPVS scores were found to significantly relate to the EF measures, with some significant correlations

disappearing once it was included as a covariate, so BPVS will be included as a covariate in future EF analysis.

The Mc rates showed small means and SD for each Mc code and it was not evident that keeping the MK, MR, or the sub-codes separate would benefit the analysis and conclusions made, thus it was decided the total Mc rate would be used. Total Mc rate did significantly correlate at TPs 1 and 2, and 2 and 3, indicating it was likely measuring the same construct at each time. The Mc interview scores were found to be significantly related to BPVS scores, and the significant correlations disappeared when BPVS was included as a covariate, so it was decided that BPVS would be included as a covariate in the MC interview analysis.

No significant relationships were detected between the Mc measures, which may indicate they were not measuring the same construct or it could be the low power made it difficult to detect differences. The effect sizes ranged from very small to the cut-off for medium, so based on this the two Mc measures will not be combined and instead will be considered separately in the remaining analysis.

The final results to report are the relationships between EF and Mc: EF did not significantly relate to either Mc measure, resulting in being unable to conclude that EF and Mc are significantly related in this study. However, as stated throughout these analyses, the low power potentially contributed to the findings and may have masked significant associations, as the effect sizes sometimes indicated there might be a relationship in the data, but the tests did not reach statistical significance.

6.2 Results: Balance beam performance and strategy development

This section will present the balance beam performance scores in terms of performance at each TP and performance for each problem type. They will be examined alongside the background measures to see how they relate. The strategy development data (consistency and first correct) will then be discussed and presented alongside a detailed analysis of the strategies used per problem and per session, as well as the coding scheme used to classify children based on their strategy use per problem. The findings from this section will be used in parts of the balance beam analysis in research question 1, when examining what role EF and Mc may have balance beam performance.

6.2.1 Balance beam performance per TP and per problem

First, performance on the balance beam task at each TP will be presented, followed by performance on each problem type as a total over all three TPs. (Scores are presented as percentage correct throughout; had raw scores been used the results would not have changed since scores were always calculated from the total number of trials that could be given and solved.)

The means, SDs, and range for balance beam performance can be seen in Table 50.

Table 50

Means (SDs) and ranges for the balance beam scores at each TP

	N	Mean (SD)	Range
Balance beam 1	36	21.99 (11.29)	0 – 41.67
Balance beam 2	31	23.92 (10.03)	0 – 41.67
Balance beam 3	33	23.99 (12.28)	0 – 50.00

Table 50 shows that the mean percentage correct only slightly increases over the TPs, and TPs 2 and 3 have a near identical mean percentage correct. The range at TPs 1 and 2 are the same, and only slightly increases at TP3.

The data were first screened (Appendix Q) and the Shapiro-Wilk test of normality showed the data not to be normally distributed at each TP (***p* < .05**), but the Z scores for skewness and kurtosis were within the acceptable level (+/- 1.96), thus the data were not transformed. Two-tailed Kendall Tau correlations were carried out, which were selected due to the small sample size and the differing strengths of linearity between the variables. The Holm-Bonferroni correction was applied. The results of correlations between the balance beam scores over the three TPs can be seen in Table 51.

Table 51

Kendall Tau correlations between balance beam scores at each TP

	Balance beam 1	Balance beam 2
Balance beam 1		
Balance beam 2	.29, p =.06 (n = 29)	
Balance beam 3	**.36, p = .016** (n = 31)	**.41, p = .010** (n = 29)

Table 51 shows significant positive correlations between the balance beam scores at TPs 1 and 3, and 2 and 3, but not between 1 and 2, although there is a medium effect size. It could be there was not enough power to detect a medium or small effect size, should it exist in the data. This gives some support to the idea the task is measuring the same variable over time. The next analyses will look at performance on each of the balance beam trial problems.

The balance beam data were broken down to examine the scores for the different problem types to see whether differences existed in the data. The five problem types were expected to be of varying levels of difficulty for the children. No child completed the last four trials of the balance beam task, so these will not be considered in the analyses. The "2 conflict balance" trials involved two unequal weights, which had to be placed on different pegs, i.e., the heavy weight on the inside peg and the light weight on the far peg. The "2 balance" trials involved two equal weights, which had to be placed on the same pegs on each side of the beam for it to be correct. The "4 balance" trials involved four light weights and the same weight had to be placed on the same pegs on each side of the beam to make it balance, whether it was one weight on each peg or two weights on the same pegs. The "conflict balance" trials were split into "3 conflict balance dissimilar" and "3 conflict balance". The 3 conflict balance dissimilar trials involved two light weights and one heavy weight, which meant children had to know two light weights equalled one heavy weight and to place the same weight on the same pegs on each side of the beam. The 3 conflict balance trials involved three light weights and required children to know that two light weights equalled one heavy weight and that heavy weights were best on the inside peg and light weights best on the far peg (as in the 2 conflict balance trials).

The means, SDs, and ranges can be seen in Table 52, and are presented as percentage correct, based on the number of each problem each child completed. All participants are included in the data as it is calculated for all trials completed, but the 3 conflict trials have a lower *n* since some children did not succeed in getting on to those trials. The trial numbers differ due to the number of sessions and trials each child completed.

Table 52

Means, SDs, and ranges for the balance beam problem types over all TPs

	N children	*N* trials	Mean (SD)	Range
2 balance	38	246	65.21 (26.34)	17.00 – 100
4 balance	38	98	91.26 (24.08)	0 – 100
2 conflict balance	38	150	12.32 (19.88)	0 – 67.00
3 conflict balance dissimilar	28	48	2.96 (11.13)	0 – 50.00
3 conflict balance	28	50	0.00 (0)	0 – 0

Table 52 shows that the 4 balance trials had the highest performance rate, followed by the 2 balance trials, although there were many more 2 balance trials. The 2 conflict balance trials were not easily solved, as seen by the low mean despite the high number of trials. The 3 conflict balance dissimilar trials were rarely solved. No child correctly solved a 3 conflict balance trial, so this problem type will be dropped from further analysis. This suggests the non-conflict trials were easiest, followed by the 2 conflict balance trials, and then the 3 conflict balance trials.

The balance beam problem type data were screened and some issues with the distribution of scores were seen. Two outliers were found in the 4 balance data due to low scores, and two outliers were found in the 3 conflict balance data, due to high scores. Histograms showed the data not to be normal due to the skewing of results depending on the difficulty of the problem. The Shapiro-Wilk test of normality showed the data were normally distributed for the 2 balance problems ($p = .70$), but not for the other problem types (***p* < .05**). Z-scores identified issues with the skewness of scores for the 2 conflict problems, the 4 balance problems, and the 3 conflict balance dissimilar problems. Issues with kurtosis for the 4

balance problems and the 3 conflict balance dissimilar problems were also seen. It is likely these issues are due to some problems being too easy and some being too difficult.

Two-tailed Kendall Tau correlations were used, as the assumption of normality does not need to be met before using this test. Two-tailed Kendall Tau correlations (with Holm-Bonferroni corrections applied) examining how performance on the different problem types was related can be seen in Table 53. (The 3 conflict balance and 3 conflict balance dissimilar trials are not included in the analyses due to low or no scoring.)

Table 53

Kendall Tau correlations between balance beam problem types

	2 conflict balance trials	2 balance trials	4 balance trials
2 conflict balance trials			
2 balance trials	-.01, p = .98 (n = 38)		
4 balance trials	.07, p = .65 (n = 38)	.15, p = .28 (n = 38)	
3 conflict balance dissimilar trials	.26, p = .14 (n = 28)	.13, p = .42 (n = 28)	.08, p = .69 (n = 28)

Table 53 shows that no significant correlations were detected between performance on the different balance beam problem types, suggesting that performance on one problem type is not indicative of performance on another problem type. The effect sizes were very small or small, indicating the strength of association. However, the 3 conflict balance dissimilar trials did not have enough power to detect large effect sizes with 80% power and the other tests did not have enough power to detect medium or small effect sizes with 80% power, so differences may have been missed.

The next analyses will look at how the balance beam scores relate to the background measures.

6.2.2 How does balance beam performance relate to the background measures?

The balance beam scores were checked to see whether any correlations existed with age, BPVS scores, or NEPSY scores. Two-tailed Kendall Tau correlations were carried out with the Holm-Bonferroni correction applied – see Table 54.

Table 54

Kendall Tau correlations between balance beam scores and the background measures

	Age	BPVS	NEPSY
Age			
BPVS			
NEPSY			
Balance beam 1	.20, p = .12	.19, p = .13	.13, p = .33
Balance beam 2	-.03, p = .81	.24, p = .09	.00, p = .99
Balance beam 3	-.01, p = .95	.16, p = .24	.14, p = .30
2 balance trials performance	.14, p = .23	.27, p = .02	.08, p = .51
4 balance trials performance	-.01, p = .97	.11, p = .44	-.12, p = .38
2 conflict balance trials performance	-.07, p = .57	-.03, p = .79	.05, p = .69
3 conflict balance dissimilar trials performance	-.01, p = .96	-.14, p = .40	-.10, p = .56

Note. N = 36 for balance beam 1, 31 for balance beam 2, and 33 for balance beam 3. N = 38 for the 2 balance, 4 balance trials, and 2 conflict balance, and 28 for the 3 conflict balance dissimilar trials.

Table 54 shows no statistically significant correlations were found between balance beam scores or the problem type performance scores and the background measures. It likely means the two significant correlations found between the balance beam scores (Table 51) are not influenced by the background measures, as supported by the very small or small effect sizes. However, as already stated, the tests have low power, so it cannot be concluded that there is no link between the background measures and trials data. Based on the correlations and effect

sizes the background measures will not be considered as covariates when examining the balance beam data. Next, the strategy development data are presented.

6.2.3 Strategy development

The strategy development data were analysed by extracting the strategy used in every trial by every child for the different problem types over all the balance beam sessions. The analysis was carried out as a whole over the various sessions since children completed a different number of trials depending on when their trials were discontinued and how many sessions they completed.

Elements of backwards trial graphing (see Siegler and Svetina, 2002) were used to address this. This involved plotting all the strategies used by each child to identify when each child started to *consistently* use the correct strategy to solve the type of problem being assessed. Consistency was calculated as the highest number of trials in a row a child used the correct strategy for that problem type. (This could also be reported as how many children managed to solve the trials x number of times in a row, but due to the low number of trials and often poor performance, this is not possible here.) This number of correct times in a row resulted in four consistency scores for each child (one for each problem type). The *first correct* score was obtained by identifying the trial when the child first used the correct strategy. The first correct data were reverse-scored for the analyses for simplicity in understanding. That means that a higher score indicates the problem was solved earlier in the trials, and a score closer to 0 indicates it was never solved or took more time to solve. Only children who completed all three sessions were included in the consistency analysis and only the children who completed all three balance beam tasks *or* solved the problem were included in the first correct analysis. This was to try and prevent a misleading conclusion based on lower scores due to completing fewer balance beam sessions and through a lack of opportunity to score as high as a child who completed all three balance beam tasks.

The means, SDs, and ranges for the consistency scores (how many consecutive trials the problem was correctly solved) can be seen in table 55. The means and SDs for the first correct scores can be seen in table 56, along with the percentage of children who solved the problem on the first trial of that problem.

Table 55

Consistency scores' means, SDs, and ranges for the different balance beam problems

	N children	Mean (SD)	Range
2 balance trials	27	3.48 (2.31)	1 – 9
4 balance trials	27	2.67 (0.68)	0 – 3
2 conflict balance trials	27	0.56 (0.85)	0 – 3
3 conflict balance dissimilar trials	19	0.11 (0.32)	0 – 1

Table 55 shows the 2 balance and 4 balance trials received the highest consistency scores, indicating they were perhaps the easiest trials, as they were mostly solved correctly 2-3 times in a row. The 2 conflict balance and 3 conflict balance dissimilar trials received low scores for consistency, indicating that these trials were more difficult.

Table 56

First correct scores' means, SDs, ranges for the different balance beam problem, and percentage that solved the problem on the first trial

	N children	Mean (SD)	Range	Solved on first try
2 balance trials	38	2.53 (0.76)	1 – 3	68.42%
4 balance trials	37	1.89 (0.40)	0 – 2	91.89%
2 conflict balance trials	28	1.11 (1.55)	0 – 4	17.86%
3 conflict balance dissimilar trials	19	0.11 (0.32)	0 –1	10.00%

The means in Table 56 show how quickly the problems were solved. The 4 balance problems have the highest percentage of children correctly solving the trial on their first try, followed by the 2 balance problems. The 2 conflict balance and 3 conflict balance dissimilar problems had much lower rates of success on the first trial – only 18% solved the 2 conflict balance trial on the first try (5 children) and only 10% solved the 3 conflict balance dissimilar trial on the first go (2 children – this is lower since fewer children attempted to solve this problem due to the discontinuation rule).

The consistency and first correct data were checked and the results of the screening can be seen in Appendix R. The consistency and first correct data were checked for normality and it was found that all the Shapiro-Wilk assumptions were violated ($p < .01$), with the exception of 2 balance consistency. However, this is not unexpected due to the limited range of trials and the skewed results (such as many children getting trial 1 correct), therefore the data are analysed acknowledging this. Holm-Bonferroni-corrected Kendall Tau correlations between the balance beam consistency and the first trial correct scores were carried out to determine whether the scores could be combined (Appendix R). Only two significant correlations were found: between first correct scores and consistency scores for 2 conflict balance and 4 balance. No significant correlations were detected between the different problem types, indicating performance on one problem type was not strongly related to performance on another problem type, although low power may be a factor and some differences may not have been detected. The decision was made to keep each problem and the consistency and first correct data separate.

Since the strategy development data derives from the problem type data and balance beam data, a detailed analysis of the relationships between these variables will not be carried out. A detailed analysis of the strategy development data are presented next.

Next are five figures that show the strategies used for each of the five problem types at each TP for the 35 children who completed at least two balance beam sessions. The 3 conflict balance data are presented, but will not be analysed further. Over the three sessions (discounting the last 4 trials, which no child completed) children could complete up to nine 2 balance trials, three 4 balance trials, six 2 conflict balance trials, three 3 conflict balance trials, and three 3 conflict balance dissimilar trials.

There were seven different possible strategies that could be used, but only certain strategies could be used for each problem due to the weights the child was given. Each possible strategy is shown in the legends and the black line in each figure is the correct strategy for that problem, for which there can only be one correct solution – these are labelled “Correct” on the legend. “W” stands for weight and “D” stands for distance. Strategy 1 (“Correct - same W same D”) was placing the same weights at the same distance (correct). Strategy 2 (“Incorrect - same W different D”) was placing the same weights at different distances (incorrect).

Strategy 3 (“Incorrect - different W same D”) was placing different weights at the same distance (incorrect). Strategy 4 (“Incorrect - one side”) was placing the weights on one side only or only placing one weight (incorrect). Strategy 5 (“Correct - different W different D”) was placing different weights at different distances, but it was the correct solution for that problem. Strategy 6 (“Incorrect - incomplete”) was placing some of the weights, so it would balance, but having a weight left over (defined as incorrect, as not all of the allocated weights were used). Strategy 7 is the same as strategy 5, placing different weights at different distances, but it is different in that strategy 7 (“Incorrect - different W different D”) solutions were incorrect for that problem. The figures show the 2 balance problems (Figure 12), 4 balance problems (Figure 13), 2 conflict balance problems (Figure 14), 3 conflict balance problems (Figure 15), and 3 conflict balance dissimilar problems (Figure 16).

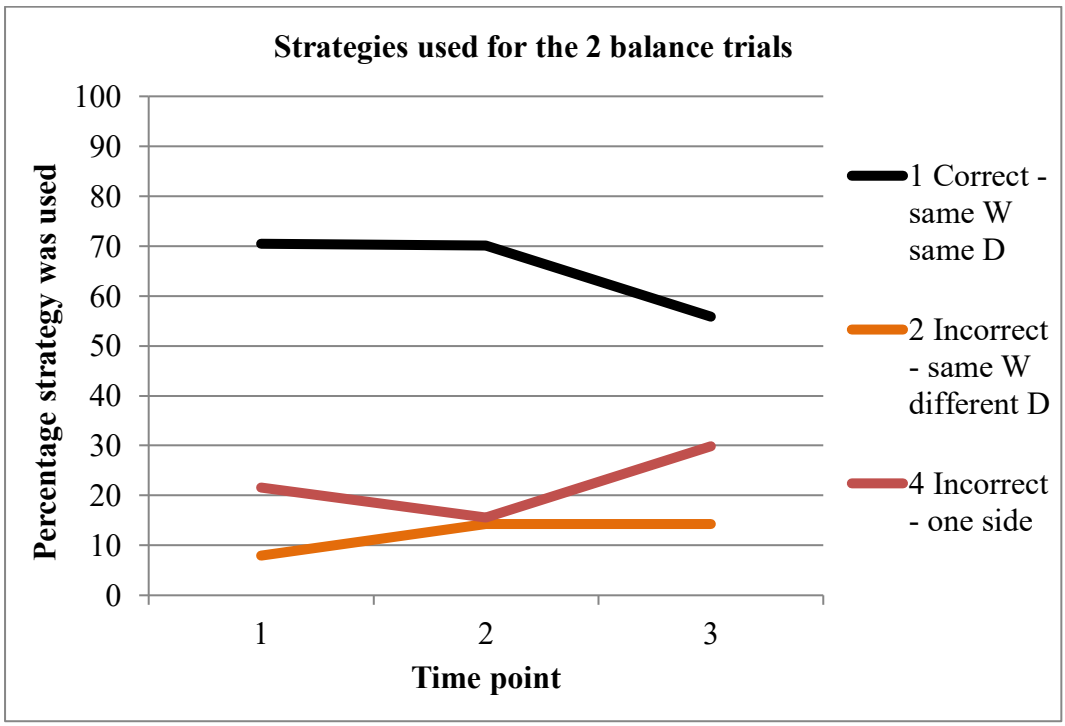

Figure 12. *Percentage of each strategy used during the 2 balance trials.*

Note. *N* trials = 88, 77, and 77, at TPs 1, 2, and 3.

Figure 12 shows a high percentage of children got the correct answers at TPs 1 and 2 but fewer correctly solved the problems at TP3. TP3 shows an increase of the use of strategy 4 – only placing one weight on the beam or placing all weights on one side, indicating no knowledge of how the beam works. Strategy 2 may indicate children have an understanding

of weight, but not distance, as they place the same weight on each side of the beam, but do not take distance into account.

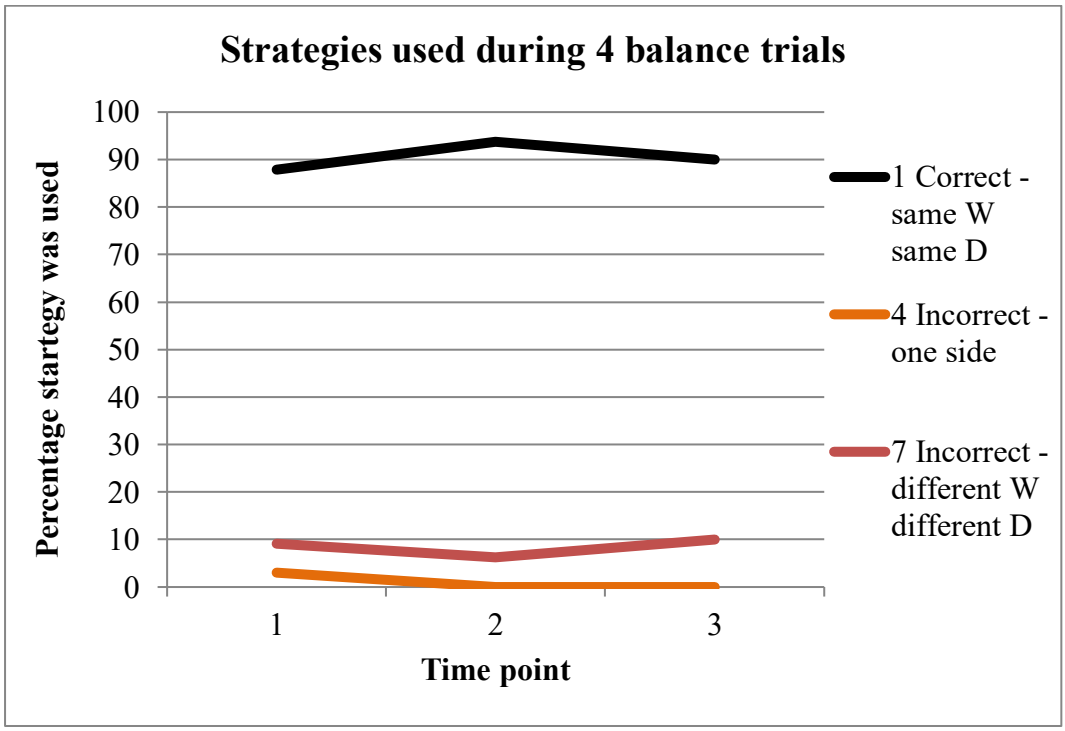

Figure 13. *Percentage of each strategy used during the 4 balance trials.*

Note. N trials = 33, 32, and 30, at TPs 1, 2, and 3.

Figure 13 shows a near ceiling effect, as most children got these problems correct. This indicates this was an easy problem, despite there being four weights (of the same weight), and surprisingly easier compared to when there are only two weights (of the same weight).

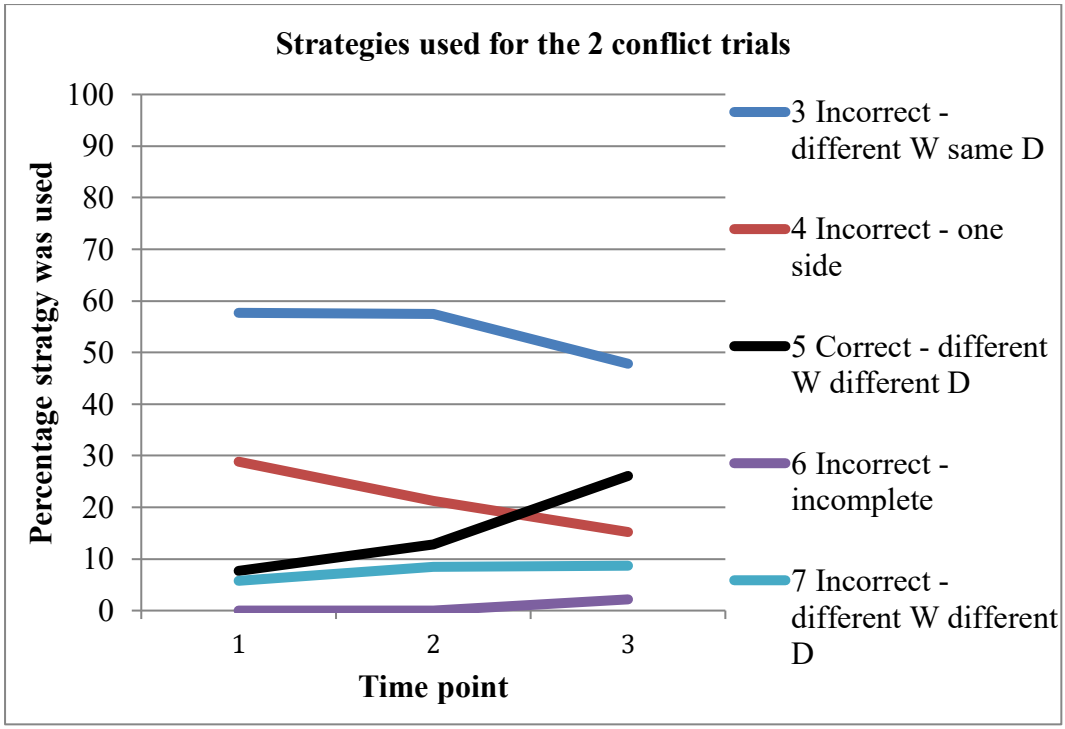

Figure 14. *Percentage of each strategy used during the 2 conflict balance trials.*

Note. N trials = 52, 47, and 46, at TPs 1, 2, and 3.

Figure 14 shows that only a small percentage of children used the correct strategy in the first session (black line), which was to place the two weights (of different weight) at different distances in order for it to balance. More than half the children used strategy 3 – to place the different weights at the same distance – in the first session. This indicates that children do not have an understanding of weight. At TP2 there appears to be a change in strategy use and children begin to use the correct strategy more and strategies 3 and 4 less, with this pattern continuing into TP3. This pattern of the correct strategy being used more over time appears to indicate learning. Strangely, there are a number of children using strategy 4 – to only place one weight or place both weights on the same side of the beam, which shows no knowledge of how the beam works, although this pattern does decrease over time.

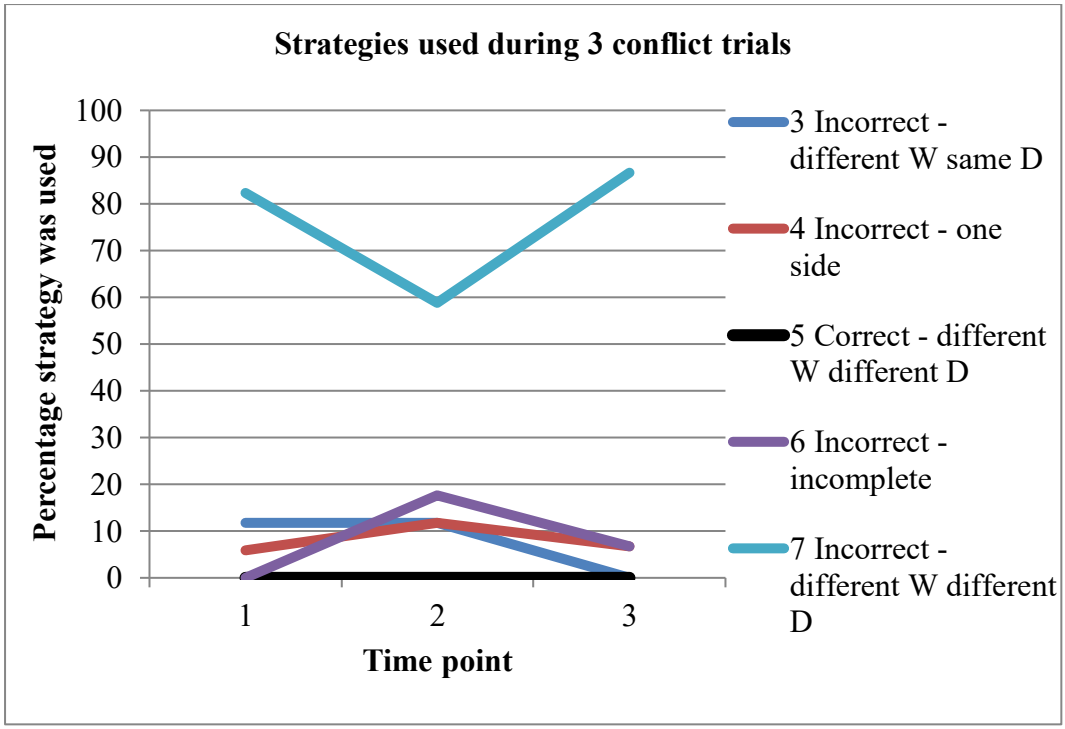

Figure 15. *Percentage of each strategy used during the 3 conflict balance trials.*

Note. N trials = 17, 17, and 15, at TPs 1, 2, and 3.

Figure 15 shows that no child correctly solved the 3 conflict balance problems, which involved three light weights. The most common strategy was to place different weights at different distances – perhaps showing some knowledge that they could not all be placed at the same distance on each side. Overall, it appears this problem was difficult for children.

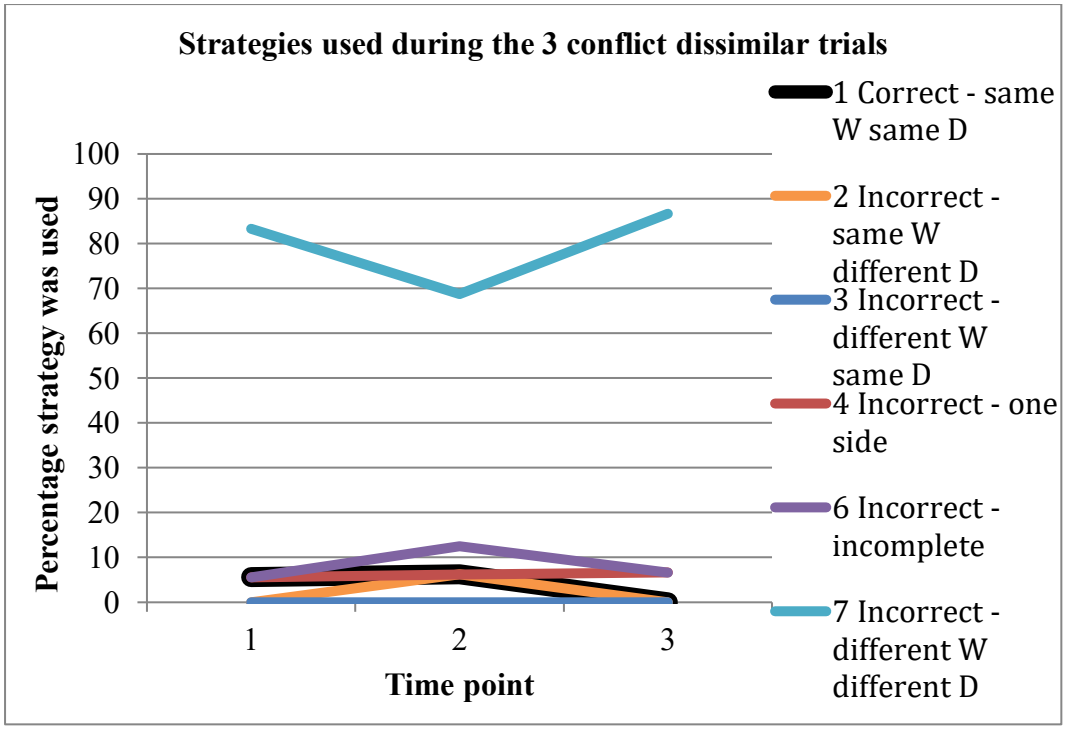

Figure 16. *Percentage of each strategy used during the 3 conflict balance dissimilar trials.*

Note. N trials = 17, 17, and 15, at TPs 1, 2, and 3.

Figure 16 shows the 3 conflict balance dissimilar trials to be very difficult for the children and only a few children correctly solved the problem at TP1. The correct strategy was to place the same amount of weight at the same distances while remembering that two light weights weigh the same as one heavy weight. Instead, children predominately used strategy 7 – to place different weights at different distances. These findings suggest this is a difficult problem for children to solve.

These figures illustrate the variety and frequency of strategies used per session for each problem type. Next, the pattern of strategies individual children used was examined. The responses children made over the various trials for each problem were examined to see whether a pattern of consistently correct or incorrect or a pattern of using trial and error could be seen. The 35 children who completed at least two balance beam tasks were included. Children were classified as showing a particular pattern based on the strategies they used and given a separate classification for the three different problems types reported (2 balance, 4 balance, and 2 conflict). These classifications were somewhat based on Chetland and Fluck (2007) who coded strategy use during an arithmetic task. They classified kids as being

consistent, showing improvement, showing regression, and showing a mixed pattern. Here six classifications were used, which are detailed next. There are six in order to incorporate the different patterns seen in this data.

Classifications are per problem. If over every trial the child used the correct strategy they were coded as "correct". If over every trial they were always wrong and used one incorrect strategy they were coded as "1 wrong strategy". If they were always wrong and used 2 incorrect strategies, but one of the strategies was not used more than a third of the time they were coded as "2 wrong strategies". The reason for coding if the second strategy was not used more than a third of the time was to filter out those who mostly used the same one wrong strategy and those using trial and error. If they began the trials with at least a third of the trials being incorrect but the last third of the trials being correct they were coded as "wrong then correct". They were coded as "trial and error (wrong)" if they used 3 or more strategies and a pattern of solving a third of the trials correctly in a row was not seen. They were coded as "trial and error (correct)" if they used 3 or more strategies and a pattern of solving a third of the trials correctly in a row was seen. The data for each problem type can be seen in Table 57.

Table 57

Number of children in each classification per balance beam problem

Strategy	2 balance trials	4 balance trials	2 conflict trials
Correct	6	29	0
1 Wrong strategy	0	2	14
2 Wrong strategies	0	0	7
Wrong then correct	1	0	3
Trial and error (wrong)	11	0	9
Trial and error (correct)	17	4	2

Note. $N = 35$.

The coding of strategy patterns adds to the strategy development data and can be used in answering research question 1 to see whether EF or Mc relates to balance beam performance. It can be seen that different problem types resulted in different strategy patterns. The 2 balance trials were mostly answered with trial and error, sometimes by children getting the

trials mostly wrong and sometimes by children managing to consistently solve a third of the trials correctly. The 4 balance problems saw the majority of children correctly solving all their trials. The 2 conflict trials saw a spread of patterns: consistently using the one wrong strategy, consistently using two wrong strategies, and using trial and error. In sum, it adds to the idea that the 4 balance problems were easiest, that children show different strategy development, and children often use multiple strategies when attempting to solve a problem.

6.2.4 Balance beam data summary

This section of the chapter first explored the balance beam and strategy development data. The balance beam data showed significant correlations between TPs 1 and 3 and 2 and 3, suggesting some consistency in what was being measured. No significant correlation was seen between TPs 1 and 2, but it could be due to low power. No statistically significant links between the different problems were detected, suggesting performance on one problem type was not strongly related to another problem type, which was support by the effect sizes, but there was only sufficient power to detect large effect sizes. The 4 balance trials were seen to be easiest, as indicated by performance scores and the strategy development data, followed by the 2 balance problems. The other three problem types were seen to be quite difficult for children. No significant relationships were detected between the balance beam data and the background measures. This was supported by the effect sizes, but there was only power to detect large effect sizes, so medium and small effect sizes may not have reached statistical significance. Based on the correlational results and effect sizes it was decided that the background measures would not be considered covariates in any balance beam analyses.

The strategy development data revealed the breadth of strategies used by children for each problem and at each TP. No balance beam or strategy data will be combined for this reason. The classifications used for the different strategy patterns will be used when addressing research question 1 to examine whether any differences here relate to balance beam performance and whether any conclusions can be drawn about the classifications' links to EF or Mc.

6.3 Research question 1: What role do EF and Mc have in children's performance on physics tasks?

Research question 1 aimed to examine the role of EF and Mc in children's performance on the physics tasks. Using the results from the previous sections, the EF (composite) scores, Mc rate and Mc interview scores will be used to examine the relationships to physics task performance in terms of percentage correct or strategy use.

Holm-Bonferroni adjustments were not applied to the analyses in this section since they have been conducted in order to answer the research question.

First, EF was examined to see how it relates to balance beam performance. The EF scores and balance beam performance data were entered into Spearman partial correlations, entering BPVS scores as a covariate.

Table 58

Spearman partial correlations between balance beam scores and EF scores, controlling for BPVS scores

	EF 1	EF 2	EF 3
EF 1			
EF 2			
EF 3			
Balance beam 1	.06, p = .72 (df = 33)	.01, p = .97 (df = 31)	.02, p = .91 (df = 32)
Balance beam 2	.04, p = .82 (df = 28)	.07, p = .70 (df = 28)	.11, p = .58 (df = 28)
Balance beam 3	-.06, p = .74 (df = 30)	-.09, p = .63 (df = 28)	-.12, p = .51 (df = 30)

Table 58 shows no significant correlations were detected between EF scores and balance beam performance. There was only power to detect large effect sizes at 80% power and large

effect sizes just below 80% power, due to the sample size, but the very small effect sizes suggest there is not a strong relationship between EF ability and the ability to solve balance beam problems. As with all low-powered tests, perhaps with more power more statistically significant relationships could have been detected. (Note: for comparative purposes, the individual EF measures were examined and showed the same result.)

The results of the Kendall Tau correlations using the balance beam performance data and Mc rate can be seen in Table 59.

Table 59

Kendall Tau correlations between balance beam scores and Mc rate

	Mc rate 1	Mc rate 2	Mc rate 3
Mc rate 1			
Mc rate 2			
Mc rate 3			
Balance beam 1	.05, p = .68 (n = 36)	-.01, p = .94 (n = 29)	-.10, p = .49 (n = 27)
Balance beam 2	.12, p = .39 (n = 29)	.23, p = .10 (n =31)	.28, p = .07 (n = 26)
Balance beam 3	-.08, p = .59 (n = 31)	.01, p = .92 (n = 29)	-.05, p = .71 (n = 29)

Table 59 shows no significant correlations were detected between the Mc rate and the balance beam performance scores. The very small or small effect sizes suggest there is not a strong relationship between the Mc rate obtained during the balance beam task and the performance score on the balance beam task. Again, there was only power to detect large effect sizes with 80% power and large effect sizes with just below 80% power, so significant associations may not have been detected.

The balance beam performance data and Mc interview scores were entered into Spearman partial correlations, entering BPVS scores as a covariate.

Table 60

Spearman partial correlations between balance beam scores and Mc interview scores, controlling for BPVS scores

	Mc interview 1	Mc interview 2	Mc interview 3
Mc interview 1			
Mc interview 2			
Mc interview 3			
Balance beam 1	.11, p = .54 (df = 32)	.30, p = .13 (df = 26)	-.05, p = .81 (df = 24)
Balance beam 2	.10, p = .62 (df = 25)	.05, p = .81 (df = 28)	.20, p = .34 (df = 23)
Balance beam 3	.21, p = .27 (df = 27)	.09, p = .65 (df = 26)	.14, p = .49 (df = 26)

Table 60 shows no significant correlations between Mc interview scores and balance beam performance scores were detected. The power is only sufficient to detect large effect sizes with 80% power and large effect sizes just below 80% power, so it is possible that differences were not detected. The effect sizes are very small or small, so it is unlikely that a strong link exists in the data, but cannot be ruled out.

(For reference, the same results were seen between EF and Mc and performance totals on the different balance beam problem types, so they are not reported.)

Overall, the results in this section have not detected statistically significant associations between EF or Mc and balance beam performance, but as noted throughout, low power was a limitation. The effect sizes were mostly very small or small, so it is less likely that the sample size and power have resulted in these non-significant results, but it is acknowledged that more power is required for identifying medium and small effect sizes. The DI and GP groups will be explored separately later, so it may be there are differences between the groups, but not overall. Next, the strategy development data will be examined for links to EF or Mc.

In order to further address research question 1 and whether EF or Mc scores relate to performance on the balance beam task, EF and Mc scores were entered into two-tailed Spearman partial correlations with the strategy development data to see whether any significant relationships existed. BPVS scores were controlled for in the EF and Mc interview score analyses.

The partial correlations between the strategy development data and EF scores can be seen in Table 61.

Table 61

Spearman partial correlations between first trial correct and consistency scores and EF scores, controlling for BPVS scores

	EF 1	EF 2	EF 3
EF 1			
EF 2			
EF 3			
2 conflict balance consistency	.05, p = .81 (df = 24)	-.22, p = .29 (df = 24)	-.18, p = .39 (df = 24)
2 balance consistency	.05, p = .81 (df = 24)	.04, p = .86 (df = 24)	.07, p = .74 (df = 24)
4 balance consistency	-.19, p = .36 (df = 24)	-.14, p = .50 (df = 24)	-.09, p = .66 (df = 24)
2 conflict balance first correct	-.16, p = .43 (df = 25)	-.25, p = .21 (df = 25)	-.23, p = .25 (df = 25)
2 balance first correct	.09, p = .60 (df = 35)	.03, p = .86 (df = 33)	.15, p = .39 (df = 34)
4 balance first correct	-.03, p = .87 (df = 34)	.-04, p = .82 (df = 32)	.02, p = .90 (df = 33)

Table 61 shows that no significant correlations were detected between the strategy development scores and EF scores. The power is only sufficient to detect large effect sizes with 80% power and large effect sizes with just below 80% power (depending on the sample size). The very small and small effect sizes suggest there is not a strong link in the data, although it cannot be ruled out. This result is perhaps not surprising since EF scores did not relate to balance beam percentage correct scores either, and it has been found that balance beam and strategy development data are related.

The Kendall Tau correlations between the strategy development data and Mc rate can be seen in Table 62.

Table 62

Kendall Tau between first trial correct and consistency scores and Mc rate

	Mc rate 1	Mc rate 2	Mc rate 3
Mc rate 1			
Mc rate 2			
Mc rate 3			
2 conflict balance consistency	-.15, p = .34 (n = 27)	-.20, p = .20 (n = 27)	-.08, p = .61 (n = 24)
2 balance consistency	.07, p = .64 (n = 27)	.09, p = .52 (n = 27)	.04, p = .78 (n = 24)
4 balance consistency	.24, p = .13 (n = 27)	.15, p = .35 (n = 27)	-.07, p = .67 (n = 24)
2 conflict balance first correct	-.12, p = .41 (n = 28)	-.12, p = .43 (n = 26)	-.02, p = .91 (n = 24)
2 balance first correct	.01, p = .95 (n = 36)	-.01, p = .93 (n = 31)	.13, p = .41 (n = 29)
4 balance first correct	.12, p = .41 (n = 35)	-.04, p = .81 (n = 30)	-.26, p = .10 (n = 29)

Table 62 shows no significant correlations between strategy development and Mc rate were detected – as found with the balance beam performance data. This finding is supported by the very small and small effect sizes, indicating there is not a strong relationship between the variables, but as before, there is only power to detect large effect sizes or large effect sizes at just below 80% power. The analyses with Mc interview scores, controlling for BPVS scores, is next and can be seen in Table 63.

Table 63

Spearman partial correlations between first trial correct and consistency scores and Mc interview scores, controlling for BPVS scores

	Mc interview 1	Mc interview 2	Mc interview 3
Mc interview 1			
Mc interview 2			
Mc interview 3			
2 conflict balance consistency	-.02, p = .93 (df = 23)	.02, p = .91 (df = 24)	.07, p = .75 (df = 21)
2 balance consistency	.25, p = .24 (df = 23)	.38, p = .06 (df = 24)	.22, p = .32 (df = 21)
4 balance consistency	-.03, p = .88 (df = 23)	-.11, p = .61 (df = 24)	.24, p = .27 (df = 21)
2 conflict balance first correct	-.02, p = .91 (df = 24)	-.02, p = .93 (df = 23)	-.29, p = .18 (df = 21)
2 balance first correct	.07, p = .68 (df = 32)	.33, p = .08 (df = 28)	.16, p = .41 (df = 26)
4 balance first correct	-.30, p = .07 (df = 31)	-.31, p = .10 (df = 27)	.24, p = .22 (df = 26)

Table 63 shows no significant correlations were identified. The effect sizes range from very small to medium, but as some of these tests were underpowered and would not be able to detect large effect sizes or to only detect large effect sizes, these should be seen as trends in the data.

Next, the pattern of strategies children used will be examined in relation to EF and Mc, to see whether any connections exist. An overview of the patterns used by children for each problem type was presented earlier and the means and SDs for EF and Mc scores for each strategy pattern at each TP (for the 2 balance, 4 balance, and 2 conflict balance trials) can be seen in Appendix S.

Univariate (entering the strategy pattern classification as the independent variable and EF or Mc scores as the dependent variable) ANOVAs and ANCOVAs (entering BPVS as a covariate) were used to analyse whether there was a difference in performance between the different strategy patterns. A separate analysis was carried out for each TP and for each balance beam problem type, resulting in nine ANCOVAs for the EF data, nine ANOVAs for the Mc rate, and nine ANCOVAs for the Mc interview scores. They are presented next, grouped by balance beam problem type, and presented in tables for ease of reading. Effect size is presented as partial Eta squared, as discussed in Section 5.4.

The 2 conflict balance problems are presented first. These examine if there is a difference in EF performance (Table 64), Mc rate (Table 65), or Mc interview scores (Table 66) based on strategy classification for the 2 conflict balance problems.

Table 64

ANCOVAs examining the difference in EF scores based on strategy use classification for the 2 conflict balance beam problems

		df	*F*	*p*	η^2_p
TP1	BPVS	1, 29	12.38	**< .01**	.30
	2 conflict problems	4, 29	0.83	.52	.10
TP2	BPVS	1, 27	9.65	**< .01**	.26
	2 conflict problems	4, 27	0.63	.64	.09
TP3	BPVS	1, 29	20.65	**< .01**	.42
	2 conflict problems	4, 29	0.61	.66	.08

Table 65

ANOVAs examining the difference in Mc rate based on strategy use classification for the 2 conflict balance beam problems

		df	*F*	*p*	η^2_p
TP1	2 conflict problems	4, 28	0.82	.52	.11
TP2	2 conflict problems	4, 25	0.95	.45	.13
TP3	2 conflict problems	4, 23	1.50	.24	.21

Table 66

ANCOVAs examining the difference in Mc interview scores based on strategy use classification for the 2 conflict balance beam problems

		df	*F*	*p*	η^2_p
TP1	BPVS	1, 26	8.68	**<.01**	.25
	2 conflict problems	4, 26	1.24	.32	.16
TP2	BPVS	1, 24	16.47	**< .01**	.41
	2 conflict problems	4, 24	0.30	.87	.05
TP3	BPVS	1, 22	4.97	**.04**	.18
	2 conflict problems	4, 22	0.55	.70	.09

Tables 64-66 do not show any significant differences at any TP, although BPVS was significant at each TP for the EF scores and Mc interview scores. The tests did not have sufficient power, meaning the results should only be seen as indicative of trends in the data and not used to conclude whether differences exist. The small and medium effect sizes suggest some differences may exist in the data, but more power is required to detect significant differences.

This analysis was repeated for the 2 weight balance beam problems – see Tables 67-70.

Table 67

ANCOVAs examining the difference in EF scores based on strategy use classification for the 2 weight balance beam problems

		df	*F*	*p*	η^2_p
TP1	BPVS	1, 30	12.42	**< .01**	.29
	2 weight problems	3, 30	0.53	.67	.05
TP2	BPVS	1, 28	6.67	**.02**	.19
	2 weight problems	3, 28	0.09	.97	.01
TP3	BPVS	1, 30	10.23	**<.01**	.25
	2 weight problems	3, 30	0.73	.54	.07

Table 68

ANOVAs examining the difference in Mc rate based on strategy use classification for the 2 weight balance beam problems

		df	*F*	*p*	η^2_p
TP1	2 weight problems	3, 29	1.30	.29	.12
TP2	2 weight problems	2, 27	7.03	**< .01**	.34
TP3	2 weight problems	3, 24	0.76	.53	.09

Table 69

ANCOVAs examining the difference in Mc interview scores based on strategy use classification for the 2 weight balance beam problems

		df	*F*	*p*	η^2_p
TP1	BPVS	1, 27	6.53	**.02**	.20
	2 weight problems	3, 27	0.42	.74	.04
TP2	BPVS	1, 26	8.60	**<.01**	.25
	2 weight problems	2, 26	2.24	.13	.15
TP3	BPVS	1, 23	5.06	**.03**	.18
	2 weight problems	3, 23	0.35	.79	.04

These results again indicated BPVS scores were significant for EF and Mc interview scores. A further significant finding emerged for Mc rate at TP2 (Table 68) and pairwise comparisons (with the Bonferroni correction applied) identified this was because there was a significant difference between children who were classified as always using the correct strategy and those using trial and error (wrong) (***p* = .01**) and those using the correct strategy and using trial and error (correct) (***p* < .01**). Each time the significant difference was because those who used the correct strategy had a higher Mc rate. The significant difference for Mc rate at TP2 may have been detected due to the large effect size, meaning the difference between the groups was large enough to be detected even with low power. The low power makes it difficult to conclude whether it is a true result though and should be seen as a trend within the data only.

These analyses were repeated a last time for the 4 weight balance beam problems – see Tables 70-72.

Table 70

ANCOVAs examining the difference in EF based on strategy use classification for the 4 weight balance beam problems

		df	*F*	*p*	η^2_p
TP1	BPVS	1, 31	16.02	**< .01**	.34
	4 weight problems	2, 31	1.43	.26	.08
TP2	BPVS	1, 29	16.47	**< .01**	.36
	4 weight problems	2, 29	7.01	**< .01**	.33
TP3	BPVS	1, 31	26.41	**< .01**	.46
	4 weight problems	2, 31	5.80	**< .01**	.27

Table 71

ANOVAs examining the difference in Mc rate based on strategy use classification for the 4 weight balance beam problems

		df	*F*	*p*	η^2_p
TP1	4 weight problems	2, 30	1.09	.35	.07
TP2	4 weight problems	2, 27	1.80	.19	.12
TP3	4 weight problems	2, 25	0.73	.49	.06

Table 72

ANCOVAs examining the difference in Mc interview scores based on strategy use classification for the 4 weight balance beam problems

		df	*F*	*p*	η^2_p
TP1	BPVS	1, 28	13.27	**< .01**	.32
	4 weight problems	2, 28	0.07	.93	.01
TP2	BPVS	1, 26	18.03	**< .01**	.41
	4 weight problems	2, 26	0.11	.90	.01
TP3	BPVS	1, 24	12.62	**< .01**	.35
	4 weight problems	2, 24	0.83	.45	.07

BPVS was again significant for EF and Mc interview scores. A significant difference was seen between EF scores at TPs 2 and 3 for the different strategy classifications (Table 70). Pairwise comparisons (with the Bonferroni correction applied) identified that at TP2 this was

because there was a significant difference between those who were classified as always using the wrong strategy and those using trial and error (correct) (***p* < .01**). At TP3 a difference existed between children classified as always using the wrong strategy and those using trial and error (correct) (***p* < .01**). At TPs 2 and 3 the differences were due to the trial and error (correct) classified-children scoring significantly higher on EF. However, only 2 children were classified as always using the wrong strategy and 4 as using trial and error (correct), so the few children contributing to the significant differences should not be taken as strong evidence. The power of the tests fell below 80%, so the findings should be viewed as trends in the data. Significant differences likely emerged due to the large effect sizes, meaning the differences between the groups were more easily detected. The low power still means it cannot be concluded that there is a significant difference between the groups.

The pattern of strategies children used during the balance beam task were analysed to see whether differences existed in EF scores, Mc rate, or Mc interview scores depending on the strategy classification. A significant difference was seen when the 2 weight classifications were examined: a difference in Mc rate at TP2 was seen between those classified as always using the correct strategy and those using trial and error (wrong) and (correct) due to those who used the correct strategy having a significantly higher Mc rate. This provides some support to a significant link between Mc rate and classifications for the 2 weight problems, but only at one TP, so it should be considered a potential trend and not conclusive evidence that strategy use relates to Mc rate. When the 4 weight classifications were examined, a difference in EF was seen at TPs 2 (between using the wrong strategy and those using trial and error (correct)) and at TP3 (between using the wrong strategy and using trial and error (correct). However, as previously noted, the number of children classified for each strategy for the 4 weight trials was low, so these results should only be viewed as indicative of a pattern. As stated throughout these analyses, the power of the ANOVAs and ANCOVAs fell below 80%, meaning there is less certainty in differences being detected, should they exist in the data. The significant differences that emerged are likely due to large differences between the groups, as seen by the large effect sizes. The large effect sizes are positive and support that some differences may exist where found here.

6.3.1 Research question 1 summary

This section aimed to examine the potential links between EF scores and Mc measures and how children performed on the balance beam task. This was examined in terms of percentage correct at the different TPs and with reference to the strategy development data. The data indicated a potential link between Mc rate and strategy use, but only at TP2. As noted with all these analyses, the power was low due to the sample size and so the significance levels should be considered alongside the effect sizes, which are not dependent on sample size and more indicative of relationships within the data (Sullivan & Feinn, 2012). The low power likely contributed to the results and potentially masked significant differences from being detected. This means the null hypothesis that no difference exists cannot be rejected or fail to be rejected. The research question has been answered as best as possible with these data and it is concluded that the data here do not show any strong links between EF or Mc and balance beam performance, but the results should only be viewed as trends and areas for further research.

The next section of this chapter will examine whether differences exist between the DI and GP support groups on any of the measures.

6.4 Results: Research question 2: What impact does support type have?

Research question 2 examined the impact of the two support types, in terms of performance on the different measures at each TP. The two groups were compared using one-way univariate ANCOVAs for the EF and Mc interview scores analyses (entering group as the fixed factor, EF or Mc interview scores as the dependent variable, and BPVS as a covariate). One-way univariate ANOVAs (entering group as the fixed factor) will be used to compare the groups' Mc rate, balance beam performance, strategy development (consistency and first correct), and physics knowledge transfer task. Strategy classifications were analysed using Fisher's exact tests. Scores on each measure will be compared at each TP. It should be highlighted that separate Univariate ANOVAs have been carried out for each analysis rather than carrying out a mixed ANOVA (variable x group) over the three TPs, due to the drop in *n* if a mixed ANOVA was used. It is acknowledged that this is less powerful and will not highlight any interactions over the varying TPs. It is also noted that the sample size is lower than the 80% required to detect a large effect size, so the results are presented to show the trends in the data, rather than as conclusive evidence for or against the hypotheses. Effect

sizes are reported to aid in examining the differences between groups. (Holm-Bonferroni corrections have not been applied, but differences will be noted for comparative purposes, where appropriate.)

As stated in the methodology chapter, children were matched on EF scores at TP1. Note: the first EF measure was obtained before support type was implemented, so they are free of any influence from support type. All of the other measures were taken after the support type was implemented.

The analyses will be presented in the following order: EF scores, Mc rates, Mc interview scores, balance beam total performance scores, balance beam problem type performance scores, support type protocol, strategy classifications, strategy development consistency scores, strategy development first correct scores, and physics knowledge transfer (measured using performance of total correct and correct on the first trial, as well as Mc rate and interview scores).

6.4.1 Is there a difference in EF scores between the groups?

The means and SDs for each groups' EF scores at each TP can be seen in Table 73.

Table 73

EF mean scores and SDs at each TP for each group

	EF 1	EF 2	EF 3
GP	-0.04 (2.38)	0.02 (2.42)	0.13 (2.46)
DI	0.04 (2.20)	-0.02 (2.17)	-0.14 (2.31)

Notes. N = 19, 17, and 19 for GP TPs 1, 2 and 3. N = 19, 19, and 18 for DI TPs 1, 2 and 3.

The assumptions of the statistical test were checked for each group separately and no issues found – Appendix T. To examine whether there were any significant differences in the groups' EF scores, one-way ANCOVAs (entering group as a fixed factor, EF scores as the dependent variable, and BPVS scores as the covariate) were carried out. The Levene's test of equality was not significant at any TP. At TP1 there was no main effect of group on EF scores ($F(1, 35) = .01$, $p = .95$, $d = .04$), but there was a main effect of BPVS scores ($F(1, 35) = 13.34$, $\mathbf{p < .01}$). The same finding was observed at TP2: no main effect of group on EF

scores ($F(1, 33) = .15$, $p = .70$, $d = .02$), but a main effect of BPVS scores ($F(1, 33) = 12.35$, $\mathbf{p < .01}$), and again at TP3: no main effect of group on EF scores ($F(1, 34) = 0.51$, $p = .48$, $d = .01$), but a main effect of BPVS scores ($F(1, 34) = 20.15$, $\mathbf{p < .01}$).

These results indicate that BPVS scores are associated with EF scores. No significant differences were found between the groups and the effect sizes were very small, indicating there was little difference between the groups' data. As already noted, the tests were underpowered, so the results should be considered alongside the effect sizes and viewed as trends.

(For comparison, had the individual EF measures been used the results would have been the same.)

6.4.2 Is there a difference in Mc between the groups?

Mc rates will be presented first, followed by Mc interview scores. BPVS was entered as a covariate for the Mc interview scores only.

6.4.2.1 Is there a difference in Mc rate between the groups?

The Mc rate means and SDs for each group at each TP can be seen in Table 74.

Table 74

Mc rate means and SDs at each TP for each group

	Mc rate 1	Mc rate 2	Mc rate 3
GP	3.70 (1.39)	3.99 (1.67)	4.49 (1.61)
DI	2.27 (.98)	2.49 (1.10)	2.13 (1.01)

Notes. $N = 17$, 16, and 13 for GP TPs 1, 2 and 3. $N = 19$, 15, and 16 for DI TPs 1, 2 and 3.

The data were screened before carrying out the analyses (Appendix U). The Shapiro-Wilk tests were only significant for DI at TPs 3 ($\mathbf{p = .03}$), suggesting the data were non-normal at the last TP. One-way ANOVAs were carried out, entering Mc rate as the dependent variable and Mc rate as the dependent variable. At TP1 a significant main effect of group on Mc rate was seen ($F(1, 34) = 13.18$, $\mathbf{p < .01}$, $d = .38$). There was also a main effect of group on Mc rate at TPs 2 ($F(1, 29) = 8.57$, $\mathbf{p < .01}$, $d = .26$) and 3 ($F(1, 27) = 23.24$, $\mathbf{p < .01}$, $d = .59$). The significant differences between the two groups' Mc rates at each TP are due to GP obtaining

higher rates. The effect sizes for the significant results at each TP are small or medium, indicating there may be a difference. As already noted, the tests were underpowered so significant differences may not have been detected and the null hypothesis that no difference exists cannot be rejected nor fail to be rejected. The results should therefore be seen as trends.

6.4.2.2 Is there a difference in Mc interview scores between the groups?

The means and SDs for the Mc interview scores for each group at each TP can be seen in Table 75.

Table 75

Mc interview means and SDs at each TP for each group

	Mc interview 1	Mc interview 2	Mc interview 3
GP	41.91 (18.72)	46.88 (14.07)	49.04 (20.07)
DI	43.75 (18.32)	50.83 (23.84)	48.44 (19.30)

Notes. N = 17, 16, and 13 for GP TPs 1, 2 and 3. N = 18, 15, and 16 for DI TPs 1, 2 and 3.

The assumptions of the statistical test were checked for each group separately (Appendix V). The Shapiro-Wilk test for the GP interview scores at TP2 was not significant ($p > .05$), but it was for GP at TPs 1 and 3 ($p < .01$, $p = .04$). The Shapiro-Wilk tests for DI were not significant at any TP ($p > .05$).

BPVS scores were entered as a covariate in the Mc interview ANCOVAs, with group as the fixed factor. At TP1 no main effect of group on Mc interview scores was seen ($F(1, 32) = .01$, $p = .95$, $d = .01$), but it did show a main effect of BPVS score ($F(1, 32) = 14.05$, $\mathbf{p < .01}$). The same result was found at TP2: there was not a main effect of groups on Mc interview scores ($F(1, 28) = .33$, $p = .57$, $d = .20$), but there was a main effect of BPVS score ($F(1, 28) = 18.67$, $\mathbf{p < .01}$). The same finding was again seen at TP3: there was not a main effect of groups on Mc interview scores ($F(1, 26) = .20$, $p = .66$, $d = .03$), but there was a main effect of BPVS score ($F(1, 26) = 12.55$, $\mathbf{p < .01}$). The Levene's tests at TPs 1 and 3 were not significant ($p > .05$), but it was at TP2 ($p = .04$). This analysis has not detected any significant differences in interview scores between the two groups, but BPVS scores were linked to Mc interview scores. These tests were underpowered, but the effect sizes are very small or small,

indicating the groups' data did not show a large difference, but significant differences may not have been detected.

6.4.3 Is there a difference in balance beam performance between the groups?

First the balance beam total performance scores will be presented, then the performance scores for the different balance problems, then the instruction and feedback provided during the balance beam task will be examined to see whether the support type elements influence performance.

6.4.3.1 Is there a difference in balance beam scores between the groups?

The means and SDs for balance beam task performance (percentage correct) can be seen in Table 76.

Table 76

Balance beam performance means and SDs at each TP for each group

	Balance beam 1	Balance beam 2	Balance beam 3
GP	19.12 (10.93)	23.44 (11.47)	19.27 (10.85)
DI	24.56 (11.27)	24.45 (8.60)	28.43 (12.17)

Notes: $N = 17$, 16, and 16 for GP TPs 1, 2 and 3. $N = 19$, 15, and 17 for DI TPs 1, 2 and 3.

The assumptions of the statistical test were checked for each group separately (Appendix W). The Shapiro-Wilk tests for GP's balance beam performance scores at TPs 1 and 2 were significant (**$p = .02$**, **$p < .01$**), but not at TP3 ($p = .42$). The Shapiro-Wilk tests for DI were significant at each TP (**$p = .03$**, **$p = .01$**, **$p = .03$**). Due to the assumption of variance being violated, Mann-Whitney U tests were carried out, entering group (DI and GP) as the grouping variable. Balance beam performance scores for each group were compared and no significant differences were detected between the two groups at TP1 ($U = 117.00$, $p = .14$, $d = .49$) or TP2 ($U = 118.50$, $p = .95$, $d = .10$), but there was a significant difference at TP3 ($U = 80.00$, **$p = .045$**, $d = .80$). This difference is due to DI scoring significantly higher than GP.

It should be noted, had the Holm-Bonferroni correction been applied, a significant difference between the groups at TP3 would not have been detected. The tests are underpowered, as a

sample size of 54 is required to run these tests with 80% power to detect a large effect size. The effect size at TP3 is large, which is why this significant difference may have been detected. The effect sizes are TPs 1 and 2 are small, indicating there is less of a difference between the two groups' performance on the balance beam.

The means show that at TP3 DI scored higher than at previous TPs, but GP also scored lower than at previous TPs, suggesting regression in performance. The power of the tests is a limitation and it could be that although no significant difference was seen at TPs 1 and 2 it may be due to the small sample size. The insufficient power means that the null hypothesis that no difference exists cannot be rejected, but the large effect size at TP3 does lend support to a difference existing between the groups, making the finding noteworthy.

6.4.3.2 Is there a difference in balance beam problem types between the groups?

The different balance beam problem types are presented next to examine whether the groups performed differently on any of the problems posed. The performance means and SDs for each balance beam problem type for each group can be seen in Table 77. This is calculated using percentage correct for each child based on how many trials of each problem they attempted.

Table 77

Balance beam problem type performance means and SDs for each group

	2 conflict balance trials	2 balance trials	4 balance trials
GP	4.47 (11.17)	60.53 (30.99)	89.53 (24.93)
DI	20.16 (23.62)	69.89 (20.49)	93.00 (23.76)

Note: $N = 38$.

The assumptions of the statistical test were checked for each group separately and the same pattern as described earlier was again found: the 2 conflict and 4 balance data violated the assumption of normality ($p < .05$), but the 2 balance data did not ($p > .05$). The 2 conflict and 4 balance data were therefore analysed using Mann-Whitney U tests, entering group as the grouping variable.

No significant difference was detected between the groups for 2 balance trials performance ($F(1, 36) = 1.21$, $p = .28$, $d = .36$) or 4 balance trials performance ($U = 162.50$, $p = .41$, $d = .14$), but a significant difference between the groups for performance on the 2 conflict balance trials was detected ($U = 115.50$, $\mathbf{p = .02}$, $d = .85$). This difference is due to DI scoring significantly higher than GP. (Had the Holm-Bonferroni correction been applied, the significant difference would remain.) These tests did not have the power required to detect large effect sizes with 80% power. However, the effect size for the 2 conflict balance trials is very large, indicating there is between the groups' data, and the significant p value adds support. The other two problem types have medium and small effect sizes, which could mean that had the tests had enough power a statistically significant p value may have been detected. It is therefore suggested that these results be seen as indicative of trends in the data, while acknowledging the null hypothesis cannot be rejected or fail to be rejected.

It was previously found that the 2 balance and 4 balance trials were the easiest and the 2 conflict balance trials were much more difficult, so this is a worthy finding as it indicates that there could be something about the DI instruction that makes it easier to solve the difficult 2 conflict balance beam problems.

6.4.3.3 Is there a difference in the balance beam support type protocol between the groups?

Next, the instruction and feedback elements of the support types will be examined to see whether there is a difference between the groups. Chapter 4 presented data on the differences between the two groups' instruction and feedback, so it was analysed to see whether the differences influenced balance beam performance. Holm-Bonferroni-corrected Kendall Tau correlations between balance beam performance and the previously noted significant variables (length of instruction provided by the adult, "generic" information, all pieces of instruction obtained before the trials (including discoveries during the play for GP), and pieces of feedback provided) can be seen in Appendix X.

Overall, the results show the two support type protocols have mixed associations with performance. GP received a longer instruction than DI at TP1 and it was seen to significantly negatively relate to performance. At TP3 GP received more pieces of "generic" information and obtained more pieces of information before starting the trials compared to DI, but this did not significantly relate to balance beam performance. However, at TP2, the number of pieces

of "generic" information GP had and pieces of information before starting the trials significantly related to GP balance beam performance. No significant differences were detected between the groups for the protocol information at TP2, but a significant link to performance was seen, and the opposite pattern seen at TP3. Since the tests were underpowered it is difficult to conclude whether these results would remain with more power. The three significant correlations had large effect sizes, indicating a difference between the groups, and all of the non-significant correlations had very small or small effect sizes, indicating less of a difference existed between the groups.

6.4.3.4 Summary of balance beam performance

The analyses on differences in balance beam performance between the groups at each TP showed a significant difference at TP3 due to DI scoring significantly higher, supported by a large effect size. It was noted that had the Holm-Bonferroni corrections been applied the significant difference would not remain. There was also a significant difference in performance on the 2 conflict balance trials due to the DI children scoring significantly higher, supported by a large effect size. Had the Holm-Bonferroni corrections been applied this result would remain.

The other findings worth noting are the significant negative link between GP's length of instruction at TP1 and balance beam performance, and the mixed result with the number of pieces of information provided by the adult and the total number of pieces of information before the starting the trials at GP's performance. At TP3 a significant difference in these two measures was seen between the groups, but it did not show as a significant correlation with balance beam performance, and at TP2 no significant difference was detected between the two groups on these measures, but a significant correlation between the measures and balance beam performance was. As already noted, the tests were underpowered, so it is not possible to reject or fail to reject the hypotheses, and the findings should be viewed with some caution. The effect sizes support the significant findings, but the low power may have contributed to not detecting medium or small differences, should they exist in the data.

6.4.4 Is there a difference in strategy development between the groups?

These analyses start with figures showing the strategy development per group, then analyses of the strategy pattern classifications are presented, then consistency and first correct data.

6.4.4.1 Is there a difference in strategy patterns?

Figures displaying the percentage of each strategy used by each group to solve each problem type at each TP can be seen next. The 2 conflict trials can be seen in Figures 17 and 18, the 2 balance trials can be seen in Figures 19 and 20, the 4 balance trials can be seen in Figures 21 and 22, the 3 conflict trials in Figures 23 and 24, and the 3 conflict dissimilar trials can be seen in Figures 25 and 26.

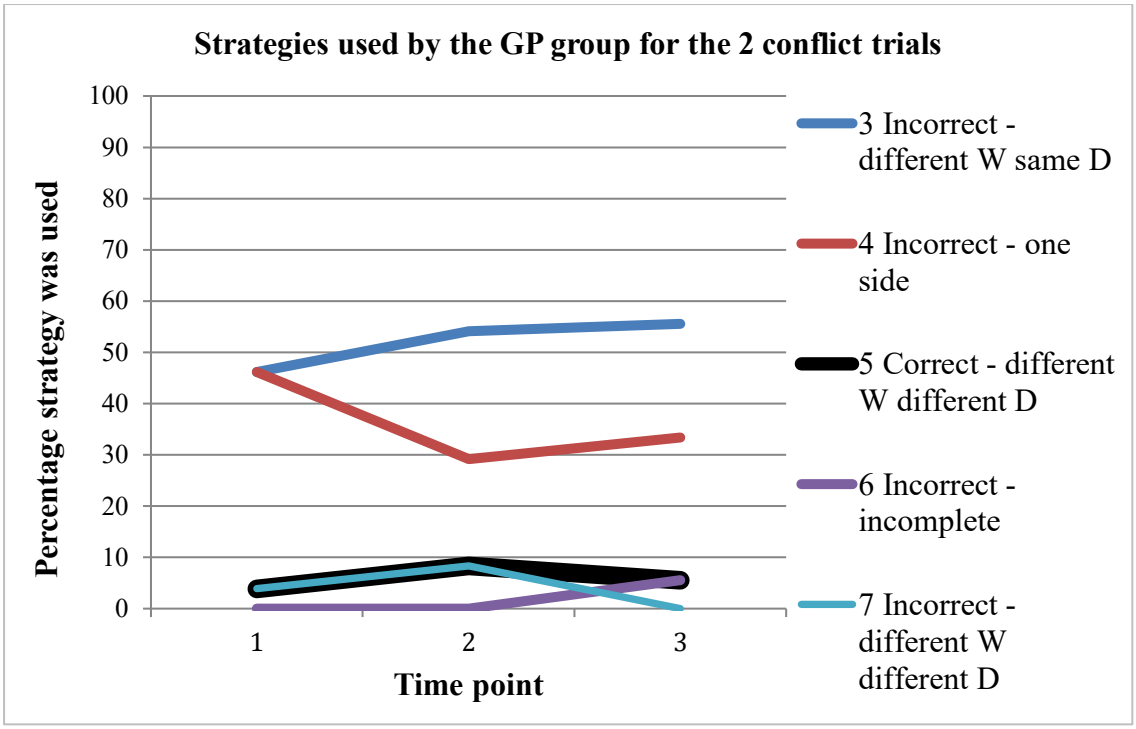

Figure 17. *Percentage of each strategy used by GP for the 2 balance trials.*

Note. N trials = 26, 24, and 18, at TPs 1, 2, and 3.

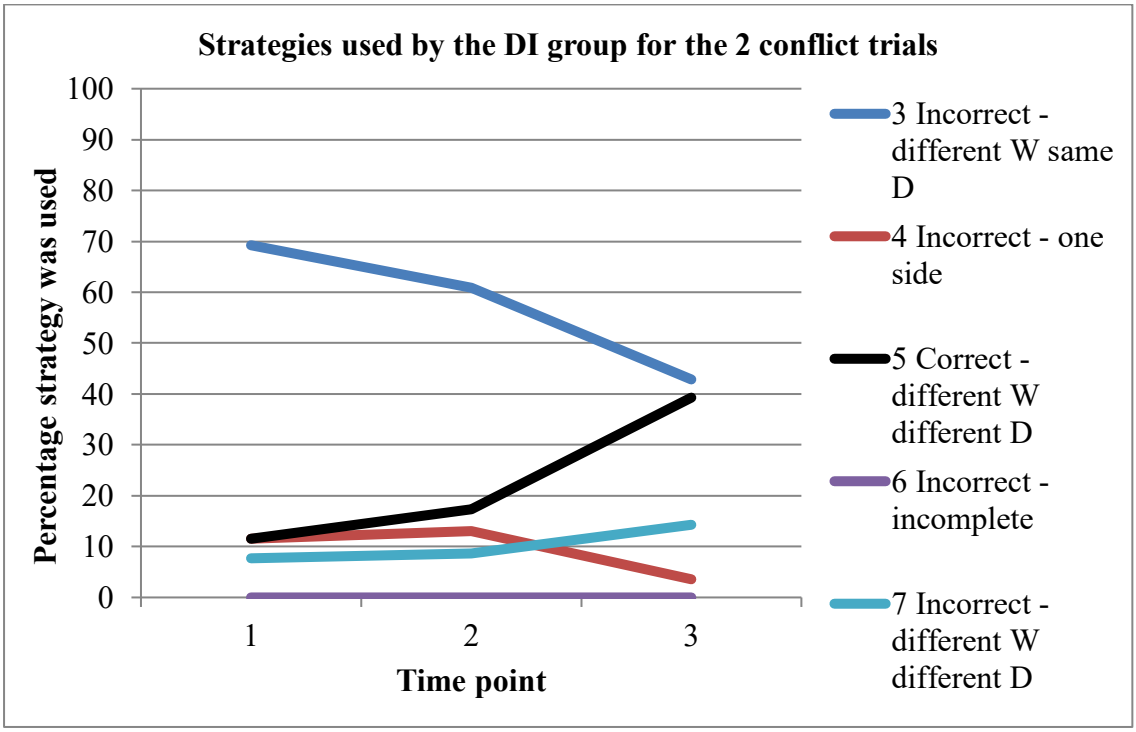

Figure 18. *Percentage of each strategy used by DI for the 2 balance trials.*

Note. N trials = 26, 23, and 28, at TPs 1, 2, and 3.

Comparing GP's and DI's performance there appears to be a difference in the strategies used. GP use the correct strategy less than 10% of the time at each TP, whereas DI show an increase from just over 10% to 39% by TP3. Use of strategy 7 indicates that children understand that the weights cannot be placed on the same pegs on each side of the beam (strategy 3), but have not discovered where they must go. GP's use of strategy 7 is quite constant, not going over 10%, and DI also show constant performance, although reaching 14%. Strategy 4 could be said to be the least informed – by placing weights on only one side the beam there is no chance the beam will balance, but GP frequently use this strategy – from 35 – 45% of the time. DI in comparison does not exceed 13%. It could perhaps be said DI use more informed strategies compared to GP.

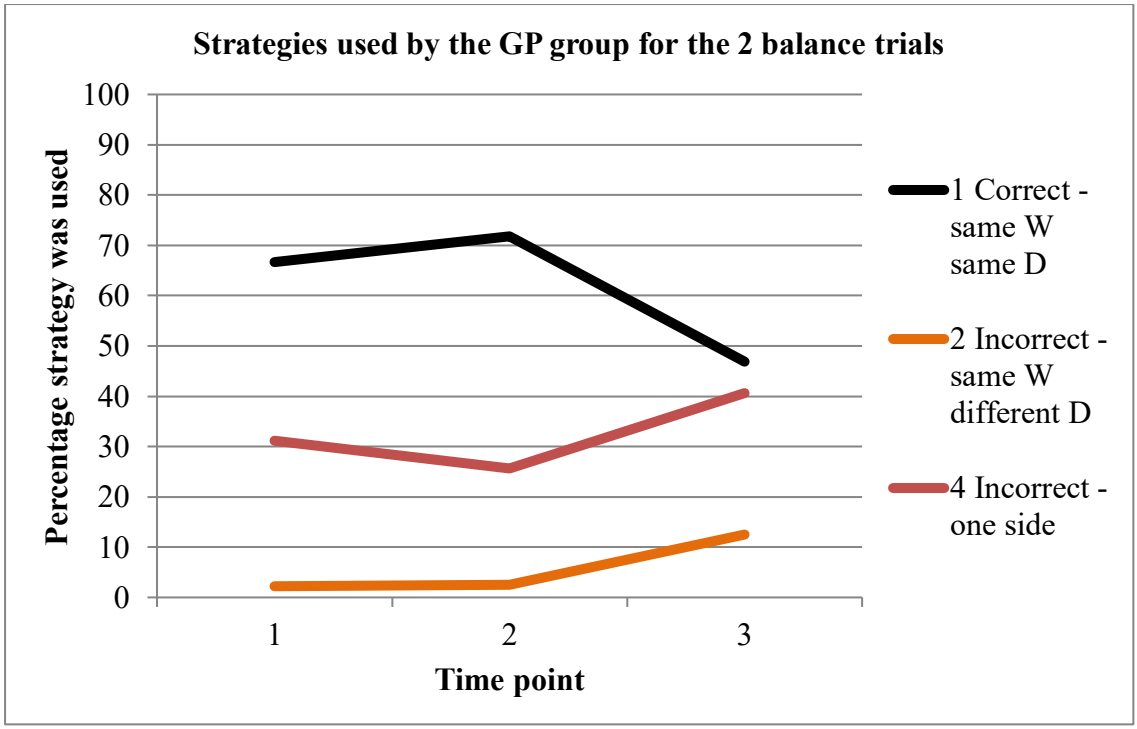

Figure 19. *Percentage of each strategy used by GP for the 2 balance trials.*

Note. N trials = 45, 39, and 32, at TPs 1, 2, and 3.

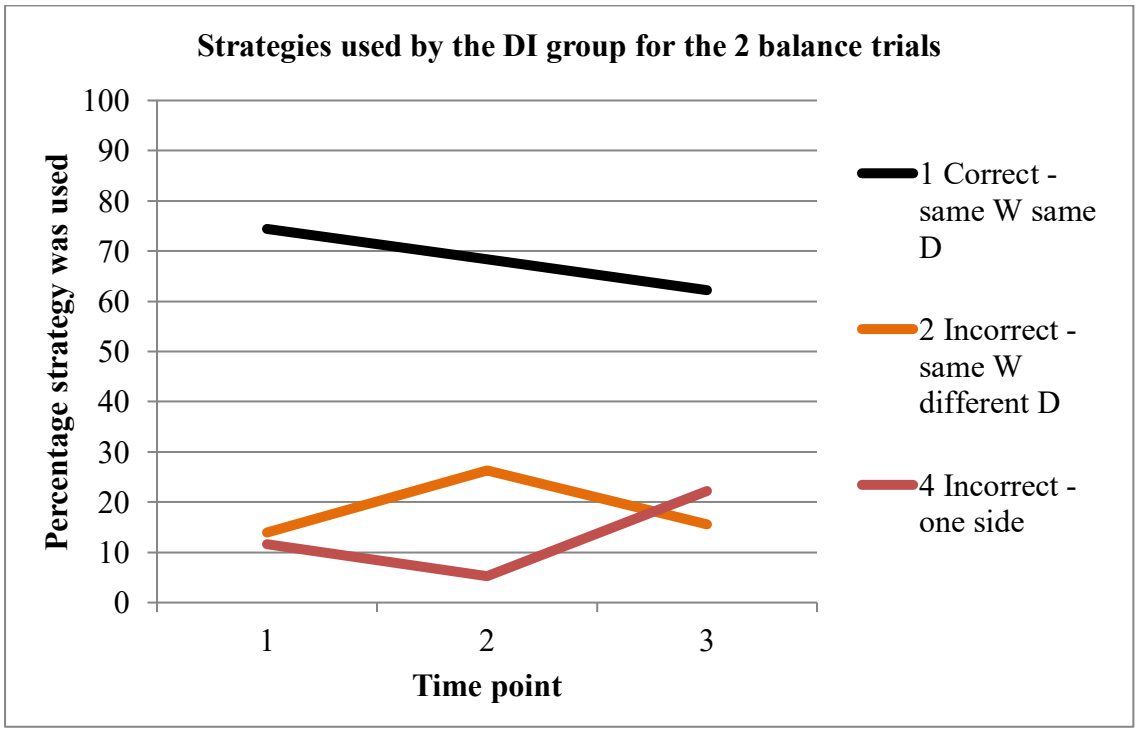

Figure 20. *Percentage of each strategy used by DI for the 2 balance trials.*

Note. N trials = 43, 38, and 45, at TPs 1, 2, and 3.

It can be seen from Figures 19 and 20 that each group's performance correct remains relatively constant at TPs 1 and 2, although DI is slightly better. Both groups show a drop in performance at TP3, which is unexpected, and at the same time, an increase in only placing weights on one side of the beam is seen.

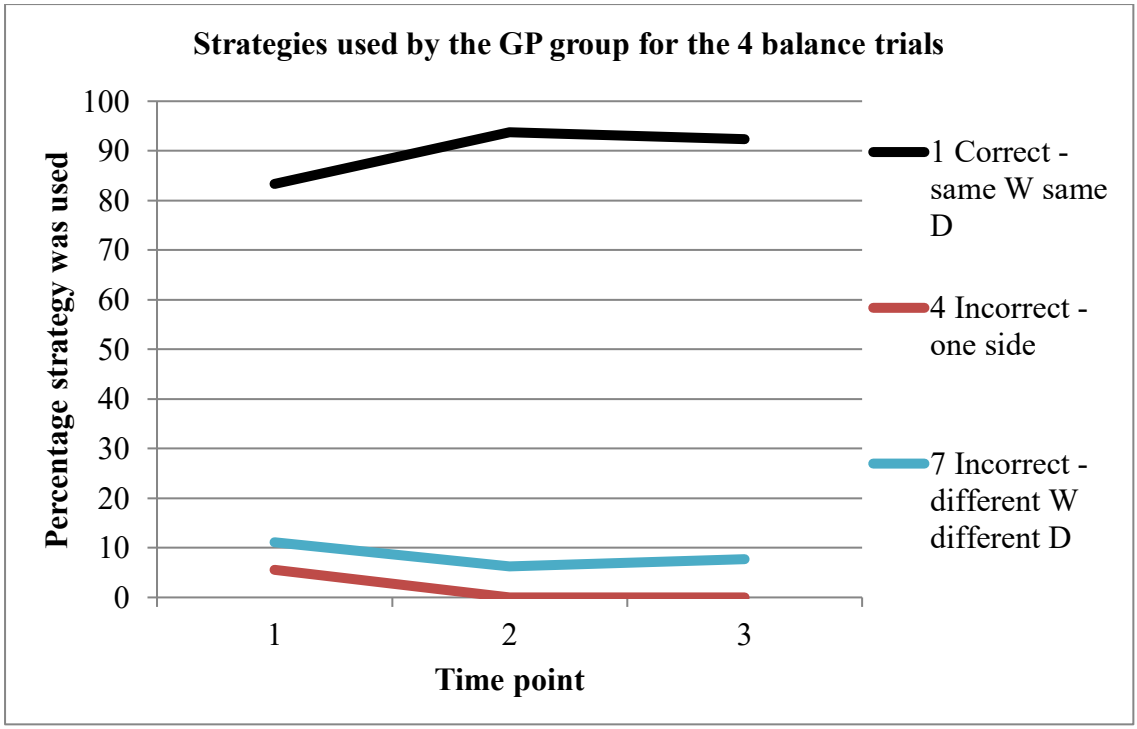

Figure 21. *Percentage of each strategy used by GP for the 4 balance trials.*

Note. N trials = 18, 16, and 13, at TPs 1, 2, and 3.

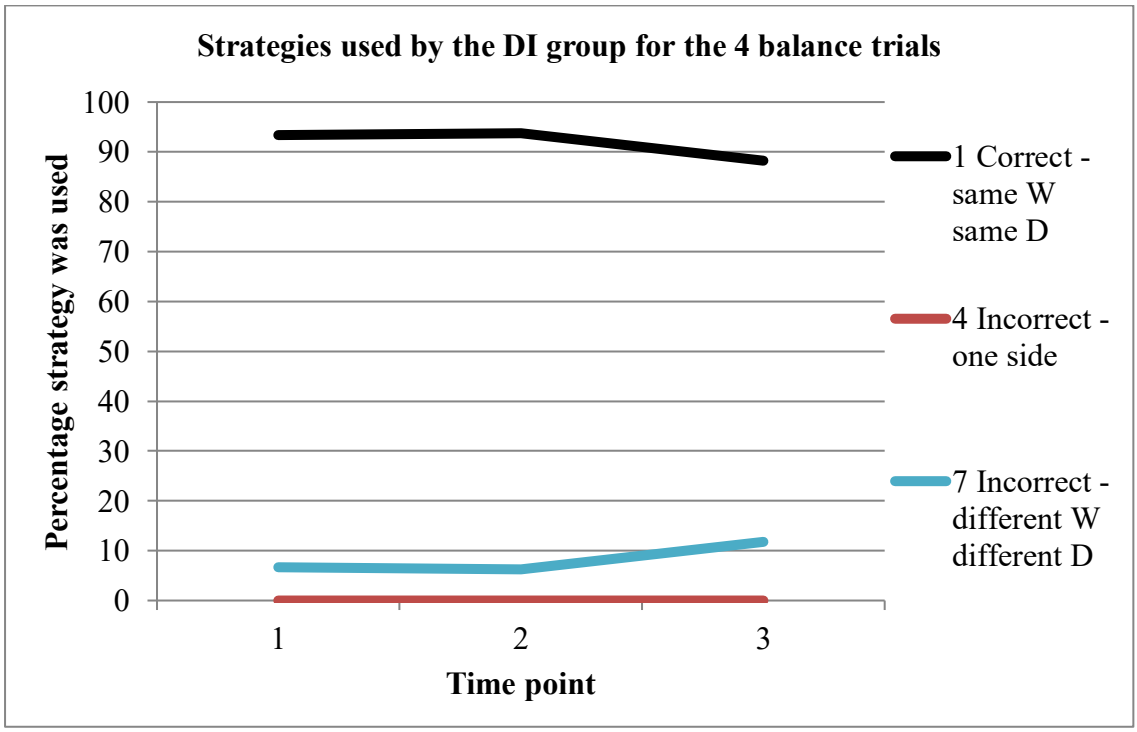

Figure 22. *Percentage of each strategy used by DI for the 4 balance trials.*

Note. N trials = 15, 16 and 17, at TPs 1, 2, and 3.

Both groups' performance for the 4 balance trials is similarly high for each TP, indicating these were easy to solve.

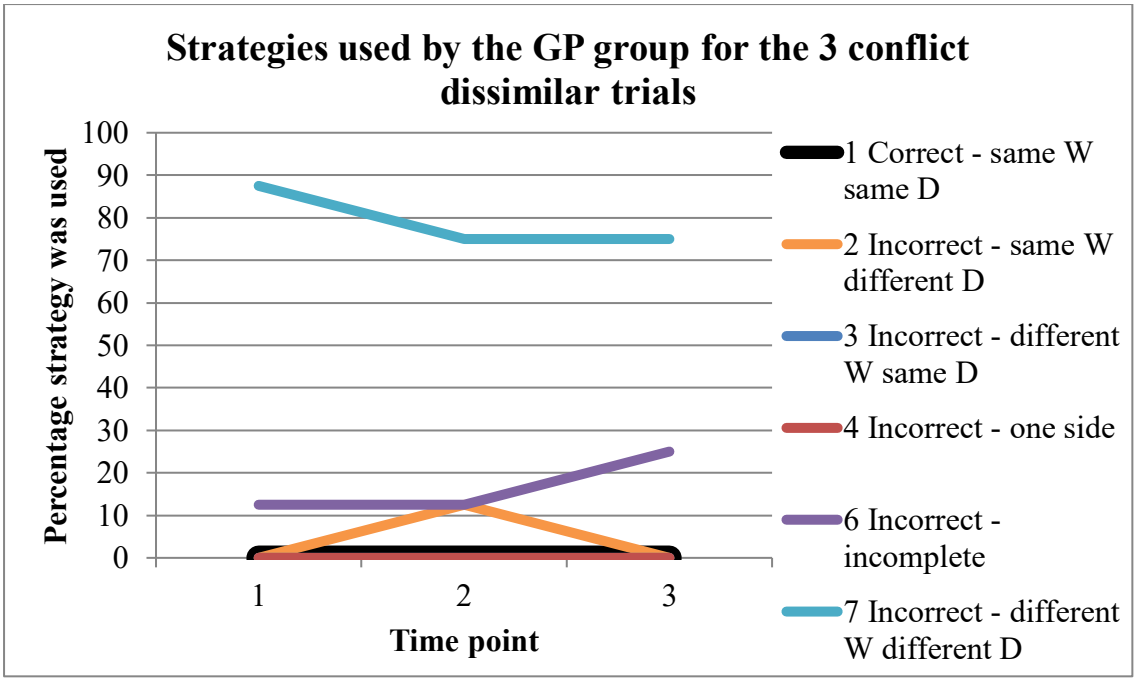

Figure 23. *Percentage of each strategy used by GP for the 3 conflict dissimilar trials.*

Note. N trials = 9, 9, and 11, at TPs 1, 2, and 3.

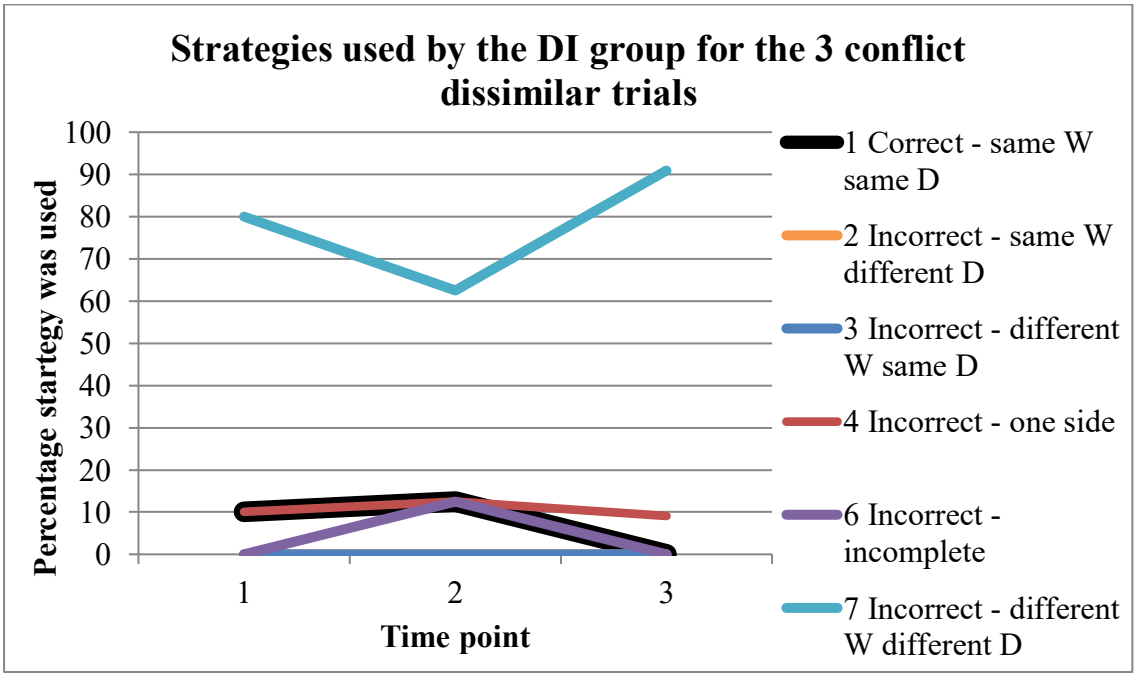

Figure 24. *Percentage of each strategy used by DI for the 3 conflict dissimilar trials.*

Note. N trials = 8, 8, and 4, at TPs 1, 2, and 3.

The 3 conflict trials appear quite challenging to solve. No child in GP correctly solved these trials and only around 10% of children in DI did solve them, however, the number of children who completed these trials is very small, so the finding should be taken with some caution. The most common strategy for both groups was to place the different weights at different distances. The other strategies were used less frequently and it can be seen there is an increase at TP3 showing GP not properly completing the trial (strategy 6).

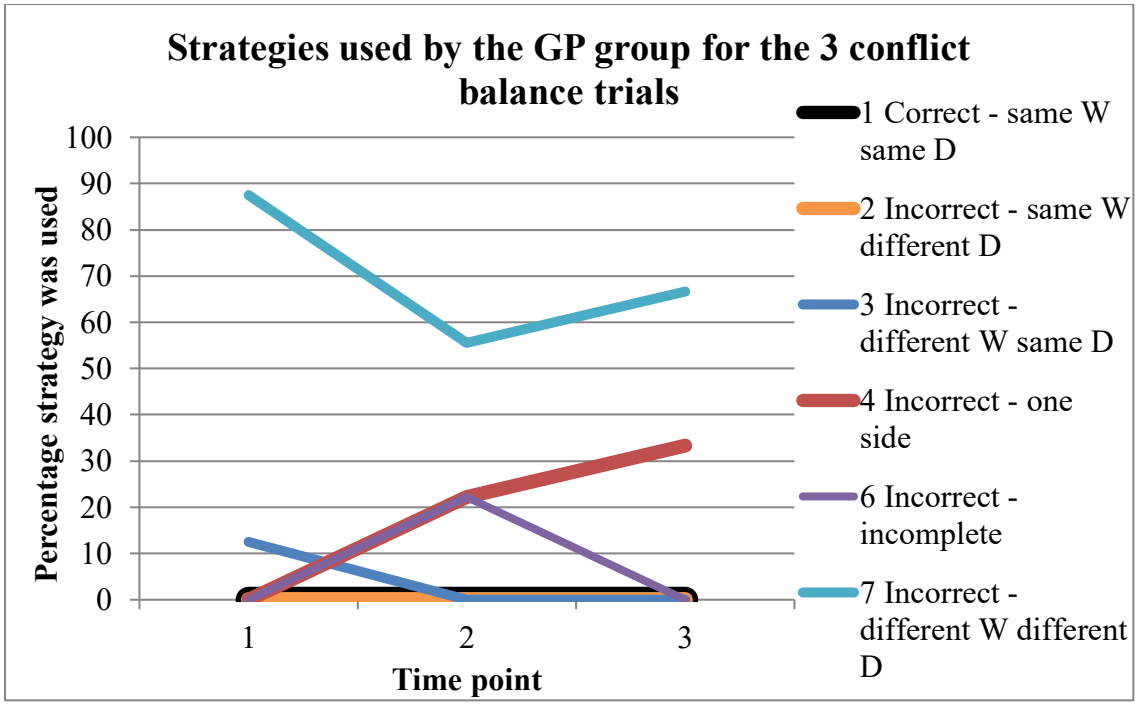

Figure 25. *Percentage of each strategy used by GP for the 3 conflict balance trials.*

Note. N trials = 8, 9, and 3 at TPs 1, 2, and 3.

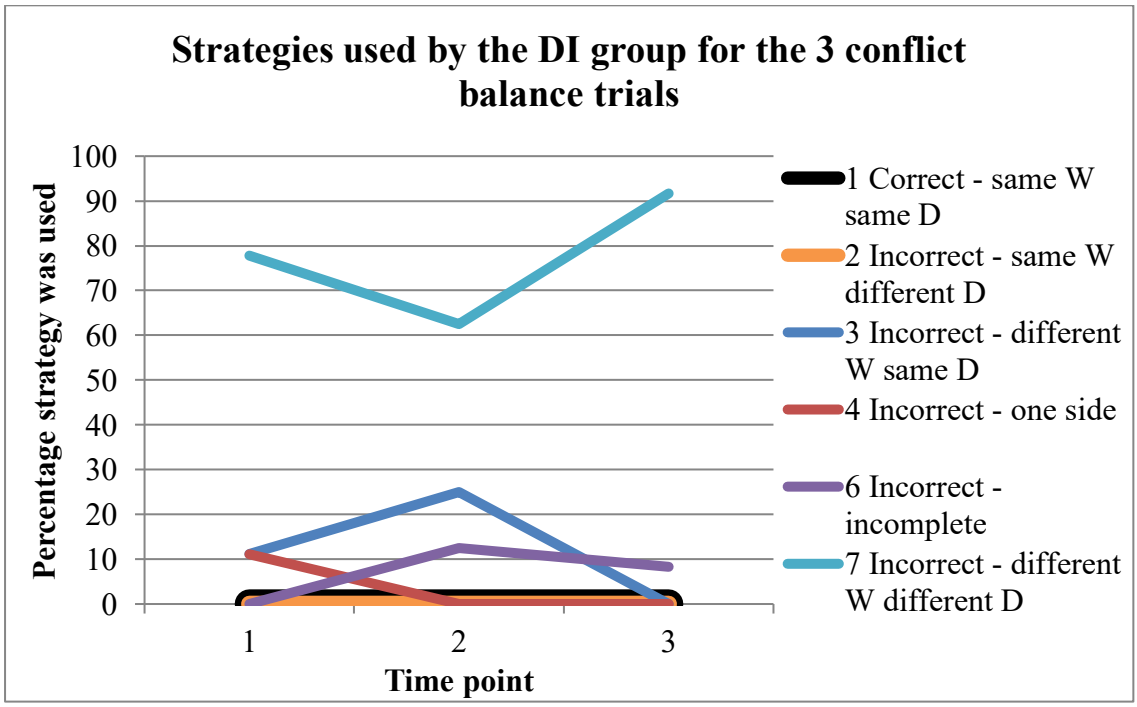

Figure 26. *Percentage of each strategy used by DI for the 3 conflict balance trials.*

Note. N trials = 9, 8, and 12, at TPs 1, 2, and 3.

Figures 25 and 26 show no child correctly solved the 3 conflict dissimilar trials and the most frequently used strategy was to place the different weights at different distances. The other strategies are used less and at similar rates over TPs for both groups, with the exception of GP at TP3 showing an increase in placing all the weights on one side of the beam, indicating no concept of weight.

Overall, these figures indicate the range of patterns used by the children in each group. Some differences were noted and these will be examined next. The pattern of strategies children used over the various trials for each problem were examined. The 35 children who completed at least two balance beam tasks were included and the classifications described earlier were applied. The data for each group for each problem type can be seen in Table 78.

Table 78

Strategy classifications for each group per problem type

	2 conflict trials		2 balance trials		4 balance trials	
Strategy	GP	DI	GP	DI	GP	DI
Correct	0	0	4	2	14	15
1 Wrong strategy	10	4	0	0	1	1
2 Wrong strategies	3	4	0	0	0	0
Wrong then correct	0	3	0	1	0	0
Trial and error (wrong)	5	4	8	3	0	0
Trial and error (correct)	0	2	6	11	3	1

Note. $N = 18$ for GP. $N = 17$ for DI.

It was hoped by examining whether children consistency used the correct or incorrect strategy or appeared to employ trial and error that some conclusions could be drawn about their knowledge. Continuously using the same wrong strategy indicates they have incorrect knowledge and so always use the wrong strategy believing it is correct. This can be seen in GP's data for the 2 conflict trials, where they were most often seen to always place the different weights at the same distance, perhaps indicating they do not have a true understanding of weight, despite doing well in the 4 balance trials. For DI, the 2 conflict trials contain three children who began the trials solving them incorrectly, but managed to find the correct solution more than a third of the time by the end, and two children who used trial and

error but used the correct solution more than a third of the time. The trial and error behaviour might indicate that children began with little knowledge but learnt. The number of children using trial and error is highest for the 2 balance trials, with GP appearing less successful than DI as seen by the difference between the number using the correct solution more than a third of the time ("correct") and those who do not ("incorrect").

Fisher's exact tests were carried out between the groups for the strategies used for each problem type. No significant findings were detected for 2 conflict trials (p = .09), 2 balance trials (p = .12), or 4 balance trials (p = .80). This finding suggests there is not a statistically significant difference in the number of children in each support type showing different patterns of strategy use, however, the three tests were underpowered, so the null hypothesis cannot be rejected or fail to be rejected. Although the frequency of children in each group show some differences, it cannot be interpreted as one group showing a significantly different pattern of learning, in terms of their strategy use. The two groups did not differ in whether they consistently used the correct or incorrect strategy, or used trial and error and solved some of the trials or solved fewer trials.

6.4.4.2 Is there a difference in consistency and first correct scores?

Next, the consistency and first correct scores are presented to see if differences exist between the groups. (Note: the consistency data only included the 27 children who completed all three balance beam sessions and the first correct data includes these 27 children plus any children who correctly solved the problem type. Any children who only completed one or two balance beam sessions and did not correctly solve that particular trial are therefore excluded from the first correct analysis.)

The consistency means and SDs for each group can be seen in Table 79.

Table 79

Consistency scores means and SDs for each group

	Mean number of trials the correct strategy was consistently used		
	2 conflict balance	2 balance	4 balance
GP	.25 (.45)	3.25 (2.73)	2.67 (.49)
DI	.80 (1.01)	3.67 (1.99)	2.67 (.82)

Notes: $N = 12$ for GP. $N = 15$ for DI.

The data were screened for each group separately and the Shapiro-Wilk tests for GP were significant for each problem type (***p* < .01**). For DI, only the 2 conflict problems violated the assumption of variance (***p* < .01**). Boxplots were created to check for outliers and three extreme outliers were seen in the GP 2 conflict problem data but were not removed due to the small sample size and the data points reflecting the abilities within the groups.

Due to the assumption of variance being violated, a Mann-Whitney U tests were carried out, entering the consistency data as the dependent variable and group as the grouping variable. No significant differences were detected between the two groups for the 2 conflict balance problems ($U = 64.50$, $p = .15$, $d = .70$), the 2 balance problems ($U = 73.00$, $p = .40$, $d = .18$) or the 4 balance problems ($U = 80.00$, $p = .52$, $d = 0.00$). However, these tests were underpowered due to not reaching at least 54 participants to detect a large effect size with 80% power. The 2 conflict balance problems show a medium effect size, indicating differences may exist between the groups, but the other tests only show very small effect sizes.

The means and SDs for when each group first solved each problem can be seen in Table 80. (As before, this is the reverse-scored data, whereby a higher number indicates better performance.)

Table 80

Mean trial when correct strategy was first used and SDs for each group

	Mean trial the correct strategy was first used		
	2 conflict balance	2 balance	4 balance
GP	0.73 (1.42)	1.58 (.84)	1.11 (.32)
DI	1.24 (1.52)	1.37 (.68)	0.94 (.23)

Notes: $N = 11$ for 2 conflict balance, 19 for 2 balance, and 18 for 4 balance for GP. $N = 17$ for 2 conflict balance, and 19 for 2 balance and 4 balance for DI.

The assumptions of the statistical test were checked for each group separately. The Shapiro-Wilk tests for both groups were significant for each problem type (***p* < .001**). Some extreme outliers were found, but will not be removed due to the already small sample size and because it will change the true pattern of results found.

Due to the assumption of variance being violated, Mann-Whitney U tests were carried out, entering group as the grouping variable. No significant difference was detected between the two groups for the 2 conflict balance problems ($U = 67.50$, $p = .17$, $d = .35$), for the 2 balance problems ($U = 158.50$, $p = .43$, $d = .27$) or the 4 balance problems ($U = 162.00$, $p = .56$, $d = .61$). As with the consistency data analyses, these tests were underpowered. The varying effect sizes indicate some differences may exist between the groups' data, but enough to reach statistical significance.

The results from examining strategy development in the two groups suggests support type did not have a statistically significant impact on strategy development in terms of consistency or when the correct strategy was first used. However, fewer children were included in the consistency and first correct analyses, which may have contributed to the statistically non-significant findings due to the tests being underpowered. Perhaps with more participants, significant differences could have been detected, so the hypotheses cannot be supported or rejected and instead the findings should only be viewed as trends in these data.

6.4.4.3 Summary of strategy development data

The data from the strategy development classifications were interesting and highlighted some differences between the groups, although not found to be significantly different. The findings

for the first correct and consistency data show there to be no statistically significant differences between the two support types, but the underpowered tests were noted as a likely limitation and it could be that differences do exist, but were not detected.

6.4.5 Transfer ramps task data

The transfer task was carried out in the last testing session. Fewer children completed this task, as explained in Section 4.3. The ramps task was used as the transfer task and various measurements were taken, including performance on the trials, Mc rate during the trials, and Mc interview score. The ramps data will first be examined to see how it relates to the background measures, to EF, to Mc during the balance beam, the balance beam data, and finally, it will be examined between the groups in order to answer research question 2.

The means, SDs, and range for total percentage correct (regardless of how many attempts it took), percentage solved on the first try (total of first tries for every trial), Mc rate (behaviours per minute), and Mc interview score (percentage coded as Mc) can be seen in Table 81.

Table 81

Means, SDs and ranges for total percentage correct, percentage of first trials correct, Mc rate, and Mc interview scores during the ramps task

	Mean (SD)	Range
Total percentage ramps trials correct	53.89 (34.49)	10.00 – 100
Percentage correct on first ramp trial	34.44 (28.54)	0 – 90.00
Mc rate during ramps task	2.29 (0.78)	1.01 – 3.85
Mc interview score after ramps task	46.03 (21.94)	14.00 – 86.00

Note: $N = 18$.

Table 81 shows over 50% of all the 10 trials were solved and about a third of the trials were correctly solved on the first try. The range indicates that every child solved at least one trial, as indicated by the non-zero minimum range. The mean Mc rate was low and a low range observed, although comparable to the balance beam data. The Mc interview received a vast range in scores and the mean was nearly 50%, indicating Mc. Next, this data will be

correlated to see how it relates. As this was exploratory analysis and not related to the hypothesis, the Holm-Bonferroni correction was applied.

The assumptions of the data were first checked and some issues found. Histograms for the total trials correct and first trials correct did not show normal distributions and the Shapiro-Wilk test of normality showed the data not to be normally distributed for each measure (***p* < .05**) (Appendix Y). Due to the issues regarding normality, the small sample size, and the varying strengths of linearity between variables, two-tailed Kendall Tau correlations were used to investigate the relationship between the different ramps measures – see Table 82.

Table 82

Kendall Tau correlations between the percentage of total trials correct, percentage of first trial correct, Mc rate, and Mc interview scores

	1	2	3
1. Total percentage ramps trials correct			
2. Percentage correct on first ramps trial	**.81, *p* < .008**		
3. Mc rate during ramps task	-.14 *p* = .44	-.17 *p* = .34	
4. Mc interview score after ramps task	.30 *p* = .11	.34 *p* = .07	.34 *p* = .06

Note. $N = 18$.

Table 82 shows one statistically significant correlation between total percentage correct and percentage of first trials correct, but no other significant correlations between the ramps data were found. These correlations are notably underpowered due to the low sample size, which means that the hypotheses cannot be supported either way. The effect sizes between the Mc interview scores and the other variables are medium, suggesting some differences may exist in the data, so perhaps with more participants more significant differences may have been detected. However, the significant finding is perhaps not surprising, as it likely that the more trials children got correct on the first try relates to total performance. This pattern of results is similar to the balance beam data found, as no significant correlations between Mc and

physics task performance were detected. Next, the ramps data are presented alongside the background measures.

The ramps task data were correlated with the background measures to check whether there was any influence on the data. Two-tailed Kendall Tau correlations were carried out and the Holm-Bonferroni correction applied– see Table 83.

Table 83

Kendall Tau correlations between the background measures and the ramps task data

	Age	BPVS	NEPSY
Age			
BPVS			
NEPSY			
Total ramps trials percentage correct	.38, $p = .04$	.04, $p = .82$	.39, $p < .05$
Percentage correct on first ramps trial	.33, $p = .07$	.04, $p = .82$	.35, $p = .06$
Mc rate during ramps task	-.03, $p = .88$	.49, $p < .01$	.08, $p = .67$
Mc interview score after ramps task	.30, $p = .10$	.45, $p = .01$	.53, $p < .01$

Note. $N = 18$.

Table 83 shows that no significant correlations between the background measures and the ramps task data were detected. This is different from the balance beam data, as BPVS scores were found to relate to Mc interview scores. These tests are more underpowered, which results in less power to detect statistical differences, should they exist The effect sizes here range from very small to large, so it could be that some significant differences exist, but have not been detected. Based on the findings here, the background measures will not be considered covariates for the ramps task data. Next, the ramps and balance beam data will be examined together.

The data from the ramps task were entered into two-tailed Kendall Tau correlations, with Holm-Bonferroni corrections applied. The results of the correlations with EF can be seen in Table 84, with balance Mc rate in Table 85, with balance Mc interview scores in Table 86, balance beam performance in Table 87, and with the balance strategy development data in Table 88.

Table 84

Kendall Tau correlations between the ramps task data and EF at each TP

	1	2	3	4
1. Ramps total trials percentage correct				
2. Ramps percentage correct on first trial				
3. Ramps Mc rate				
4. Ramps Mc interview score				
EF 1	.30, p = .10	.23, p = .19	.30, p = .08	.44, p = .01
EF 2	.10, p = .59	.09, p = .62	.41, p = .02	.46, p = .01
EF 3	-.03, p = .88	-.01, p = .94	**.50, p < .004**	.40, p = .03

Note. N = 18 for EF 1 and 3, and 17 for EF 2.

Table 84 shows that one significant relationship emerged between the ramps data and the EF scores: this was between the Mc rate during the ramps task and EF scores at TP3. However, the correlations are underpowered, so it is not possible to conclude whether significant associations exist, although the large effect size suggests there could be. There are some medium effect sizes in the other correlations, so perhaps with more power, significant correlations may have been detected.

Table 85

Kendall Tau correlations between the ramps task data and balance Mc rate at each TP

	1	2	3	4
1. Ramps total trials percentage correct				
2. Ramps percentage correct on first trial				
3. Ramps Mc rate				
4. Ramps Mc interview score				
Balance Mc rate 1	-.01, p = .97	.09, p = .62	.41, p = .02	.23, p = .21
Balance Mc rate 2	.02, p = .93	.16, p = .41	.06, p = .75	.10, p = .62
Balance Mc rate 3	.18, p = .39	.25, p = .21	.31, p = .11	.18, p = .36

Note. N = 18 for balance Mc rate 1, 16 for balance Mc rate 2, and 15 for balance Mc rate 3.

Table 85 shows that no significant correlations emerged from the ramps task data and the balance Mc rate. As with the other tests using the ramps data, these are underpowered, so it could be there was not enough power to detect significant differences. Based on this data, BPVS scores will not be considered a covariate in any ramps Mc rate analyses.

Table 86

Kendall Tau correlations between the ramps task data and balance Mc interview scores at each TP

	1	2	3	4
1. Ramps total trials percentage correct				
2. Ramps percentage correct on first trial				
3. Ramps Mc rate				
4. Ramps Mc interview score				
Balance Mc interview 1	.18, p = .36	.13, p = .52	.33, p = .09	.36, p = .07
Balance Mc interview 2	.37, p = .07	.22, p = .28	.26, p = .19	.33, p = .11
Balance Mc interview 3	-.09, p = .67	-.15, p = .49	.50, p = .02	.39, p = .07

Note. N = 17 for balance Mc interview 1, 16 for balance Mc interview 2, and 15 for balance Mc interview 3.

Table 86 shows no significant correlations emerged between the ramps task data and the balance Mc interview scores at each TP. Again, power is an issue and the range in effect sizes suggest there could be some differences in the groups' data here, but they perhaps could not be detected.

Table 87

Kendall Tau correlations between the ramps task data and balance beam performance scores at each TP

	1	2	3	4
1. Ramps total trials percentage correct				
2. Ramps percentage correct on first trial				
3. Ramps Mc rate				
4. Ramps Mc interview score				
Balance beam 1	.36, p = .07	.30, p = .12	-.05, p = .78	-.08, p = .69
Balance beam 2	-.25, p = .24	-.22, p = .31	.25, p = .23	-.22, p = .31
Balance beam 3	-.11, p = .59	-.19, p = .34	.02, p = .94	-.04, p = .84

Note. N = 18 for balance performance 1 and 3, and N = 16 for balance performance 3.

Table 87 indicates that balance beam performance at each TP did not significantly correlate with the any of the ramps task data, but the tests are underpowered, which will have contributed to the likelihood of detecting differences, should they exist.

Table 88

Kendall Tau correlations between the ramps task data and the balance strategy development data

	1	2	3	4
1. Ramps total trials percentage correct				
2. Ramps percentage correct on first trial				
3. Ramps Mc rate				
4. Ramps Mc interview score				
2 conflict balance consistency	-.06, p = .79	-.12, p = .59	-.17, p = .43	-.19, p = .39
2 balance consistency	-.04, p = .85	-.27, p = .18	.24, p = .22	-.18, p = .35
4 balance consistency	.03, p = .90	.27, p = .22	-.06, p = .76	.09, p = .67
2 conflict balance first correct	-.01, p = .96	-.11, p = .61	-.10, p = .61	-.10, p = .65
2 balance first correct	.51, p = .02	.40, p = .05	-.11, p = .58	.08, p = .71
4 balance first correct	-.32, p = .13	-.06, p = .78	-.10, p = .62	-.22, p = .29

Note. N = 18 for 2 balance first correct and 4 balance first correct. N = 16 for all other correlations.

Table 88 shows no significant correlations emerged between the ramps task data and the balance beam strategy development data. These tests are also underpowered, making it difficult to conclude whether a difference exists.

6.4.5.1 Summary of the ramps data

This section has examined the ramps data, how it relates to one another, how it relates to the background measures, and how it relates to the EF, balance Mc, and balance beam

performance. From all of these exploratory analyses, only two significant findings emerged: a positive correlation between total percentage correct on the ramps trials and the percentage of first trials correct on the ramps trials, and between the Mc rate during the ramps task and EF at TP3. No significant correlations were detected with balance beam Mc, or balance beam performance. It has been noted that *n* was much lower in these analyses, reducing the power of the test, and likely contributing to the non-significant findings. The range of effect sizes and potential trends in the data are still interesting, and if they had been pre-planned and the Holm-Bonferroni corrections not applied then there would have been more significant findings due to the large effect sizes.

Performance on the ramps task by each group will be examined next: by total percentage correct in the ramps trials and percentage of first correct trials. This will also consider whether differences in Mc rate and Mc interview scores between the groups exist.

6.4.6 Is there a difference in performance on the ramps task between the groups?

18 children completed the ramps task; the means, SDs, and ranges for each group for the percentage correct for the trials and the percentage correct for the trials solved on the first try can be seen in Table 89.

Table 89

Means and SDs for the ramps trials total percentage correct and percentage solved on the first try

	Percentage correct for all trials		Percentage correct for correct on the first try	
	Mean (SD)	Range	Mean (SD)	Range
GP	62.50 (37.32)	10.00 – 100	47.50 (30.59)	10.00 – 90.00
DI	47.00 (32.34)	10.00 – 100	24.00 (23.19)	0 – 80.00

Note: $N = 8$ for GP, $N = 10$ for DI.

Table 89 shows that GP scored considerably higher on both the percentage of all trials correct and the percentage correctly solved on the first try. The SDs for DI are smaller, indicating less variance from the mean. The range for both groups on the total percentage correct is the same, but the range for the first correct trials for GP is slightly higher than DI.

The assumptions of the statistical test were carried out for each group separately and an issue with DI's first trials correct data was seen, as it violated the assumption of variance (***p* = .03**) (Appendix Z).

One-way ANOVAs were carried out to examine the total percentage correct and percentage correct for the first try data, entering group as the between-subjects factor. No significant difference was detected between the groups for percentage correct total ($F(1, 16) = .89$, $p = .36$, $d = .44$) or the first correct scores ($F(1, 16) = 3.5$, $p = .08$, $d = .89$). The Levene's test of variance was not significant for either ($p > .05$). These tests were underpowered and did not reach 80% power to detect large effect sizes, deemed to be 52 participants. The effect sizes in these tests are small and large, indicating some differences may exist, but not to statistical significance in this sample, which is not surprising. This lack of power means it cannot be concluded that support type does or does not have a later influence on how well children perform on another physics task. Therefore, the null hypothesis cannot be rejected or fail to be rejected.

The strategies were examined, but due to the nature of the task could not be coded in detail, like the balance beam trials. The reason was that there were too many variables to consider and in total it would mean 24 potential strategies per trial, which would not provide meaningful results. For example, the coding would need to consider the incline of the ramps, the surfaces, whether the ramps were on the inclines, whether only one ball was rolled, whether two balls were rolled, whether they were rolled down one ramp or two, and which attempt solving the problem it was. The Mc measures were examined to see if a difference exists between the groups.

Next, the Mc measures between the groups will be examined to see whether the previous support type impacted Mc on the ramps task. The means, SDs, and ranges for the ramps task Mc rate and Mc interview scores can be seen in Table 90.

Table 90

Means, SDs and ranges for the ramps task Mc rate and Mc interview scores

	Ramps Mc rate		Ramps Mc interview scores	
	Mean (SD)	Range	Mean (SD)	Range
Guided play group	2.36 (0.76)	1.01 – 3.15	44.64 (26.66)	14.00 – 86.00
Direct instruction group	2.23 (0.83)	1.07 – 3.85	47.17 (18.81)	21.00 – 86.00

Note: $N = 8$ for GP, $N = 10$ for DI.

Table 90 shows that the groups scored similarly on both measures of Mc, although the ranges are slightly higher for DI. During the balance beam task, it was found that GP scored significantly higher at each TP, but the difference here looks less distinguished.

The assumptions of the statistical test were next checked for each group separately (Appendix AA). One-way ANOVAs were carried out on the ramps Mc data, entering group as the between-subjects factor. No significant difference was detected between the groups' Mc rate ($F(1, 16) = 0.12$, $p = .74$, $d = .16$) or Mc interview scores ($F(1, 16) = 0.05$, $p = .82$, $d = .11$), but the tests were underpowered. The Levene's test of variance was not significant for either measure ($p > .05$). Since the ramps Mc interview used seven questions instead of four, the analyses were also checked with only the same four questions as used in the balance beam and no differences were found. As already noted, these tests were underpowered, impacting the conclusions that can be drawn.

The results from this section examining the Mc measures in the ramps task indicate that there is not enough data to conclude whether support type received during the balance beam task has an impact on how children scored on the physics transfer task. The underpowered tests mean any findings could be due to chance and it could be that with a larger sample size the findings may not be supported. Instead, the data should be viewed as trends.

6.4.6.1 Summary on the transfer task

The results from this section examining the data collected during the transfer physics task indicates that the previous support type children received when completing the balance beam task is inconclusive. No significant differences emerged concerning how children performed

on the transfer physics task, but the low power means it cannot be concluded that this is not due to chance. No significant differences in performance or Mc were detected between the groups either. It was earlier found that Mc rate significantly differed between the groups during the balance beam task, but as that is not evident during the ramps task, it could be a result of the support type provided while the task is being carried out. Nonetheless, the means from the ramps task are interesting since GP appears to be performing much better during the trials, which is in opposition to what was found during the balance beam analyses, as DI was found to have scored significantly higher than GP at TP3. It is likely that the lower number of children who took part in the ramps task has impacted the findings and power is noted as a limitation here, which has resulted in being unable to reject the null hypothesis that no difference exists between the two groups.

6.4.7 Research question 2 summary

Research question 2 examined whether there was an impact of support type on EF, Mc, balance beam performance, strategy development, or the transfer physics task. No significant group differences were detected for EF scores or balance Mc interview scores. A difference in Mc rate was seen between the groups at each TP due to GP scoring significantly higher each time. A significant difference in balance beam performance at TP3 was found between the groups due to DI scoring significantly higher. No significant difference in strategy development was detected between the groups. No significant differences in the ramps task data were detected between the groups. A limitation with all these analyses, and particularly the ramps tasks, is that the sample size was too small to reach the necessary 80% power to confidently detect differences. This means that the null hypothesis cannot fail to be rejected, because it could be that with more power a significant difference could potentially have been found, should it exist. It is still worth noting the significant finding that emerged, the effect sizes, and trends in the data, and to follow them up in future work with a larger sample.

7 Chapter 7
Discussion

This chapter will present the findings of the study and address each research question in turn. The limitations of the work will be discussed, along with recommendations for future work. The theoretical, educational, and methodological contributions that the work has made will then be outlined.

7.1 Aims and hypotheses of the present study

This work aimed to investigate the role of EF, Mc, and support type (GP and DI, implemented during the balance beam task) on children's performance on physics tasks (including a transfer physics task).

Exploratory analyses were first carried out to establish the relationship between some of the measures. EF and Mc were expected to relate since they have been implicated as having an interactive relationship (Diamond, 2013), thus a positive link between the measures was expected. The structure of EF and Mc were also considered, but no hypotheses were made concerning the construct of each.

In response to research question 1, what role do EF and Mc have in children's physics task performance, EF and Mc were expected to positively relate to performance on the balance beam task. A positive link between EF and the balance beam task would lend support to Diamond's (2013) theory, which suggests EF has a role in problem-solving tasks. The GR account (Munakata, 2001) could perhaps account for EF links as well, through EF demands changing the graded representations.

Looking at balance beam performance, which relies on the strategy data, the data were expected to support one or some of the theories over others. If children mainly showed consistency in the strategies they used for each problem within a session, with little deviation except when going through times of improvement, this would support the staircase model of Halford et al. (2002). If some consistency was seen, but regressions from and back to the correct solution were seen then the RR model (Karmiloff-Smith, 1992) could account for this. If children were found to use several different strategies for solving a particular problem, both within a session and between different sessions, this could lend support to the OW

theory (Siegler, 1996), the connectionist model (Schapiro & McClelland, 2009), and the GR account (Munakata, 2001).

There were several predictions for research question 2, which focused on the impact of support type. It was suggested that if there is a benefit to EF it could be for GP, due to the nature of the support type. GP support allowed children more control in the way they approached the task. Considering Mc, it was again unclear, as both groups may show benefits, but for different reasons. Balance beam performance would be reflected in the strategies used and some predictions were made about performance rate and strategy development. Considering the elements of the GP support and the EF and Mc predictions, it was thought GP would show faster strategy development, but it was also thought that the instructional elements in the DI support could aid with better performance from TP1, as children will know how to solve the problems from the start, unlike the GP children who might instead show more improvement over time. It was unclear how support type might impact the transfer ramps task. Theoretical accounts tend to suggest experience on the task is the important aspect in performance, so performance transfer may not be expected.

The findings and decisions from the exploratory analyses, balance beam and strategy development data are presented next, then the research questions are addressed.

7.2 Exploratory analyses

The exploratory analyses were conducted to see how the EF data, the Mc data, and the background measures (vocabulary and visual-spatial skills) related within and between these measures. Based on these results, decisions were then made on how to treat each measure.

The three EF measures appeared to be somewhat linked, as seen by several significant correlations before BPVS scores were factored in. After BPVS scores were factored in a link was still seen, since each EF significantly correlated with the EF composite score at each TP, showing at least a medium effect, which supports the idea of the measures feeding into one overall EF component. The correlations were not the same strength as before since they did not all significantly correlate with one another, so it was concluded that there was some overlap, but dissociable. The correlations between each EF measures (at each TP) showed small and medium effect sizes but were not always statistically significant, perhaps due to the

sample size and the tests only having the power to detect large significant differences.

The idea that EF components are dissociable in young children, but are all linked, supports the findings of Garon et al. (2008). Garon et al. (2008) concluded after a review of the literature that inhibition and WM develop before shifting, and that inhibition and WM feed into shifting. Here, shifting was seen to not correlate at each TP (after BPVS was accounted for), which might indicate shifting is less stable or the task used was not the best measure. The effect sizes were medium in strength, so it could be an issue of underpowered tests, and if so it might be masking more significant associations between the EF components. Others have found the three core EF components to be unitary (Wiebe et al., 2011), for inhibition, WM, and planning to unitary (Hughes, Ensor, Wilson, & Graham, 2010), or the three core EFs to group into fewer than the three components (Monette et al, 2015). With adults, Miyake et al. (2000) found the EF components to be moderately linked and to link to an overall EF measure, which they suggest could be the central executive.

In the current study, the decision was made to use the EF composite score for each child at each TP, rather than three EF scores. It was noted that if the individual EF scores had been used in the analyses the conclusions would not have changed, therefore the use of a composite score is justified. The findings are not unsurprising since previous research with children is somewhat mixed as to whether there is one or several measurable EF components at this age, which could be influenced by EF development or by the selection of tasks used. The measures used here appeared somewhat stable and, overall, are believed to reflect the EF components they aimed to measure.

Mc rate was calculated from the individual Mc codes, which were combined and treated as one measure due to the small means and SDs, the reduced power and increased complexity in using 15 different codes throughout the analyses, and the belief that separate MK and MR analyses would not add to the conclusions drawn. It was therefore decided to use one Mc rate per TP (as done by Whitebread, et al., 2009b) and to keep the two Mc measures separate.

Mc rate significantly correlated at TPs 1 and 2, and 1 and 3, and showed a small effect size between TPs 2 and 3, but it did not reach statistical significance. No significant correlations were detected between Mc rate and BPVS, and the effect sizes were small, suggesting the

relationship was not strong, so BPVS was not considered a covariate in the Mc rate analysis. The Mc interview scores significantly correlated at all three TPs before BPVS scores were factored in as a covariate (due to the significant links seen). The Mc interview scores did not show significant relationships after BPVS was accounted for, which might suggest the Mc interview was a less reliable measure, in terms of independence from language, than the Mc rate. It is likely that the power reduced the tests' ability to detect medium and small differences in the data, and perhaps with more power more significant findings may have been detected.

As discussed in the introduction chapter, there is not a validated task to measure Mc in 3- and 4-year-olds, and since the two measures used rely on overt behaviour, it is possible that the recorded Mc for each child is not an accurate reflection of their Mc ability. However, the measures used were deemed the best available for this age range and by taking vocabulary into account, some variation could be controlled for. The fact that Mc rate and interview scores could be recorded and reliably coded by a second person indicates agreement that these did reflect Mc processes, although it could be the rate and interview did not utilise the same Mc process, since no strong links were found between them. Robson (2016) also reported using two measures of Mc and finding they did not completely correspond with one another, as the observation data recorded a higher MR score, but the reflective task recorded higher MK and ME scores. It may therefore be that different Mc tasks do not tap the same Mc components in the same way, providing varying results. It could be that the Mc interview here targeted MK (based on the questions asked), while the Mc rate appeared to record few instances of MK, which was thought to be related to the need for MK to be verbalised for it to be scored. Therefore, the rate here might have provided a better recording of MR and the interview a better recording of MK.

Age and vocabulary, and age and visual-spatial skills significantly correlated, suggesting a link between age and these scores, which might be expected since scores will likely increase with age. No significant correlation between BPVS and NEPSY was detected, but the tests were not sufficiently powered to detect medium and small effect sizes, so it could be that they do exist in the data but could not be detected. However, vocabulary was linked to EF and Mc interview scores, indicating language to be more strongly related to these tasks. Links between EF and language have previously been found in research with some concluding that

EF is important for language development (Weiland et al., 2014 found 4-year-olds' EF predicted language scores six months later, but the reverse relationship was not found), that language is important for EF development (Botting et al., 2017 found 8-year-olds' language was key in EF development), and that neither predicts performance on the other during the early primary years, but they are strongly related (Gooch, Thompson, Nash, Snowling, & Hulme, 2016). Another idea is that vocabulary can help performance on EF tasks to some extent, but beyond a certain level no further benefit of vocabulary is seen (Hughes et al., 2010). These mixed findings suggest that a relationship does exist between EF and language, although the direction is unclear, which the current study has not resolved.

The EF tasks used in the present study required knowing individual words (receptive language), but not comprehension or expression, the Mc interview required comprehension and expression, and the Mc rate benefitted from expression (for scoring). Although the vocabulary measure taken here was receptive, it may have provided a proxy for other aspects of language. The EF tasks and Mc interviews required comprehension of the task instructions and questions. Cragg and Nation (2010) have previously suggested that vocabulary is related to inner speech and inner speech is related to thinking through task demands and the accompanying actions. This idea would help explain why vocabulary played a role in EF and the Mc interviews, where instructions and questions were used, as perhaps language was required to hold the information in mind, to understand the task instructions, and to respond. This was not a requirement for Mc rate, so it could explain the lack of a link here. Vocabulary was thus included as a covariate in analyses using EF and Mc interview scores.

No significant correlations between visual-spatial skills and EF or Mc were seen here, but the tests only had sufficient power to detect large effects, so they may not have detected smaller differences. Diamond's (2013) model suggested that a link between EF and visual-spatial skills might be found, as visual-spatial skills could be seen as a form of non-verbal reasoning. This was not found, but perhaps due to low power. Some research has found links between visual-spatial skills and scientific reasoning (Mayer et al., 2014) and others have found some link between non-verbal reasoning and performance of physics tasks (van der Graaf et al., 2016), but significant findings could be driven by the particular tasks used and the overlap in specific cognitive skills required. Visual-spatial skills did not significantly correlate with any of the measures here, thus it was not included as a covariate in any analyses. It was

acknowledged that some of the measures might involve more than the component being targeted, so additional cognitive components were considered in case they helped explain any of the variance in performance on any of the tasks. As already noted, the power of some of the tests likely contributed to the findings and it is suggested the results be viewed as trends in the data while acknowledging that the null hypothesis that no differences exist cannot be rejected or fail to be rejected.

When the relationship between EF and Mc was examined, no significant correlations were detected at any TP. EF scores significantly correlated with the Mc interview scores at TP3, but when vocabulary was controlled for, this result became non-significant. As explained above, it is likely that language ability aided performance on the EF and Mc interview tasks, so including BPVS as a covariate would explain the variance, thus it was not surprising that the result turned non-significant. Past research has shown mixed findings concerning whether a link between EF and Mc exists (Bryce et al., 2015) or does not exist (Spiess et al., 2016), but overall there is a lack of work to be able to say whether a link definitely exists. Roebers (2017) provides a review of these fields and ultimately concludes there is not yet a definitive answer as to whether EF and Mc are related, due to a lack of research. Some studies have found links between some EF components (inhibition / WM / shifting / attention / etc.) and some types of Mc (procedural / declarative / monitoring / control / knowledge / etc.) (Roebers, 2017), but so far there is not convincing evidence to suggest EF and Mc are related in young children. Here the data did not find a significant link and so it cannot support Diamond (2013) who suggests a link between EF (inhibition) and self-regulation (Mc). However, the low power in this study means that the null hypothesis fails to be rejected, as it could be that a link does exist, but it has not been detected with these tests. This result, therefore, adds to the field of research that has found no significant link between EF and Mc in young children, while acknowledging the limitation of power here, potentially obscuring findings. The result could also perhaps be due to EF and Mc undergoing rapid development at this age, due to the tasks used, that the link is weak and only found when particular tasks are used/cognitive components assessed, due to language ability, or simply the relationship is very complex and difficult to assess in this age group.

7.3 Balance beam and strategy development

These analyses examined the relationship between the balance beam data (total correct and performance on the different problem types), the strategy development data, and whether they related to the background measures. These analyses are presented now, as the conclusions drawn were used to answer research question 1, presented in the next section.

Balance beam performance for the complete sample improved at each TP and significantly correlated between TPs 1 and 3 and between TPs 2 and 3, suggesting some consistency in what was measured. The different problem types were examined and from the means and the consistency and first correct data, the 2 balance trials were seen to be the easiest, followed by 4 balance trials, then the 2 conflict balance trials, the 3 conflict balance dissimilar trials, and finally the 3 conflict balance trials (which no child correctly solved). (The 2 conflict balance trials were solved only slightly earlier than the 4 balance trials.) These findings support previous research that balance trials are easiest and trials that involve conflict are more difficult (Halford et al., 2002; Siegler, 1976). Performance for the problem types did not significantly relate and showed only very small or small effect sizes, perhaps due to the varying difficultly levels or the different knowledge required to solve each, or perhaps the tests lacked the power required to detect significant differences, if they exist. However, all of the production trials used balance *and* distance, so it could be said that the balance trials here are more challenging than in prediction tasks, as the two variables must always be considered to correctly solve the trial, which is very complex for young children. It may be the difference in procedure here encouraged children to take account of both weight and distance or that the game-like set-up made children more focused on the goal (making the dinosaurs seesaw) compared to having to predict or produce results that could be seen as more abstract and without a goal for doing it.

Previous research has shown mixed findings concerning when children can solve balance beam problems. Schrauf et al. (2011) concluded from their study that 3-year-olds do not have an understanding of the role of weight, but 4-year-olds do, whereas Halford et al. (2002) concluded 2-year-olds can solve problems that involve only one variable. The current data suggest that 3- and 4-year-olds can successfully produce answers to balance beam problems that require more than one variable to be considered, although the different problems were seen to range in difficulty, resulting in a higher performance on some compared to others.

The current finding does not support Halford et al.'s (2002) relational complexity theory, as they state that it is not until around 5 years of age that children can consider both weight and distance, or Siegler (1976) who suggests young children cannot encode two pieces of information at once. Here, children did manage to consider two variables in many of the trials, and (although the analysis was not presented) it was sometimes above chance levels (see the 4 weight trials for example).

Some significant correlations were seen between balance beam performance and the consistency and first correct strategy development, which suggest there are links between some of the problem types and total performance. It could be the 2 balance and 2 conflict balance problems were more revealing and pertinent to the overall scores, due to the high rate of success on the 4 balance problems by nearly all the children and the low rate of success on the 3 conflict balance (and dissimilar) trials by nearly all the children. Thus it was decided to keep the strategy development data per problem separate, as different patterns of performance were seen and it could be certain problems would show significant findings.

The strategy development data gave a detailed overview of the strategies used to solve the different problems over all the trials. The strategy development data showed that for the 2 balance trials, the majority of children used the correct strategy. The next most common strategy was to only put one weight on the beam or both weights on the same side – indicating no concept of how the beam works. For the 4 balance trials the correct strategy was mostly used, with little deviation to other strategies, indicating this was an easy problem to solve. The 2 conflict balance trials resulted in a range of strategies being used: the most common was to put the two (different) weights at the same distance – possibly indicating an understanding of distance, but not of weight. However, this changed for the children who reached trial 5, as use of this strategy decreased and it was instead used as often as the correct strategy. No child correctly solved the 3 conflict balance trials, and a similar pattern was seen with the 3 conflict balance dissimilar trials, whereby the most common solution was to place the different weights at different distances, indicating both problem types were too complex. This information was needed to later classify children based on the strategies they employed.

The background measures were examined to see how they related to the balance beam data. Age was not significantly related to performance, despite previous research suggesting

children's balance beam performance improves with age, although the young and narrow age range and small sample size could explain this result. Vocabulary and visual-spatial skills were examined for links to physics performance, but no significant correlations were detected. Although language was not seen to factor into balance beam performance here, others have found some links, but the underpowered tests used here means it cannot be concluded that a link does not exist, only that none were detected in these data with these tests. Rhodes et al. (2016) found EF to predict science outcomes, but not beyond what was predicted by BPVS scores, however, as discussed earlier, this could be due to a potential interactive relationship between EF and language. Van der Graaf et al. (2016) found something similar, with links between science exploration and inhibition, which were again mediated by vocabulary, and other links that were mediated by non-verbal reasoning. Tolmie (2014, cited in Tolmie, Ghazali, & Morris, 2016) found language to be important in science learning and noted that language is likely essential when children are making links between their science observations and trying to form explanations for what they are learning. Philips and Tolmie (2007) also found that if children already hold language relevant to the task being examined this could benefit performance on the task. It was thought that if language played a role in balance beam learning here it could support the RR model through the idea that language drives redescription and knowledge, but this was not found. It was somewhat unexpected that language did not significantly relate to balance beam performance, having found it did link with EF and Mc interview scores, but as already stated, this could potentially be due to low statistical power. It was suggested that vocabulary may be important during the EF tasks and the Mc interviews due to instructions and questions being used and a better-developed vocabulary could help children process and hold information to solve the tasks, but that argument does not now hold up for the balance beam task. Perhaps no significant link was seen between vocabulary and the balance beam due to the concepts of weight, distance, and balance being explained to the children prior to beginning the balance beam task. That is, all children began with the necessary key vocabulary and as the children were not required to verbalise during the task vocabulary did not affect performance. The language used during the balance beam was planned out over much piloting in order to give children the best chance to understand the task. As long as children understood they had to balance and seesaw the dinosaurs they could potentially do well on the task even if they did not understand all of the information. However, if inner speech is really an important aspect of working through tasks and relates to vocabulary (Cragg & Nation, 2010) it would have been expected that

balance beam performance would relate to vocabulary and possibly Mc rate (through the need to plan, control, monitor, and evaluate actions during the trial), but this was not found. Vocabulary was therefore not considered a covariate in the balance beam analyses. The tests were not sufficiently powered to detect medium and small effect sizes, so it may be that they do exist, but there was not enough power to detect them here. This makes it difficult to reject or fail to reject the null hypothesis that no links exist and instead these results should only be viewed as trends in the data and potential future areas of interest.

The preliminary analyses found the different balance beam problems varied in difficulty with links between performance and the strategy development data and a variety of strategies used. It was found that 3- and 4-year-olds can solve balance beam problems that incorporate weight and distance. No significant correlations were detected between age, vocabulary, or visual-spatial skills and balance beam performance or strategy development, resulting in the background measures not being considered as covariates in these analyses, while acknowledging power to be a likely limitation. The analyses so far have revealed that vocabulary significantly correlated with EF and Mc interview scores, but not Mc rate or balance beam performance. This could be due to the EF and Mc interview requiring more facilitation from language in terms of understanding instructions, utilising inner speech to think through answers, and for the Mc interview, to express language as well. Language may not have been so facilitatory in Mc rate due to the coding also relying on actions. For the balance beam task, it may be due to the different nature of the task or the language and game set-up used before starting the task. It could also be because the effects were smaller, thus the statistical tests employed here did not detect them. These findings were incorporated when answering the main research questions, while noting that some of the statistical tests employed were underpowered and may have only had enough participants to detect large effect sizes with 80% power, resulting in the recommendation that the data be viewed as trends that require further exploration.

7.4 Research question 1: What role do EF and Mc have in children's performance on physics tasks?

The link between EF and Mc, and balance beam performance and strategy development were examined, but no significant correlations were detected. The findings could potentially be due to insufficient statistical power, thus significant differences may not have been detected.

It was expected that positive links would be found between EF and/or Mc and balance beam performance, based on some previous research and theoretical accounts (Diamond, 2013), but this was not seen. Previous research has found some links, although they have not been consistent, which indicates that the relationship between EF and Mc to physics performance could be very complex.

Some have found links between only some EF components or tasks and physics task performance (Baker et al., 2011), between EF and physics with the relationship being mediated by other factors (van der Graaf, et al., 2016), and some evidence that only some EF components can predict science performance (not specifically physics) (Latzman et al., 2010). Others have found no significant link between EF and physics (Mayer et al., 2014; Tolmie, 2014, cited in Tolmie et al., 2016). The mixed findings could be due to a variety of reasons, including the variety of age groups used in the referenced studies, the various different tasks used to test EF and physics, and the different EF and physics components tested. The lack of research in this area and lack in consistency in what is being examined means the current work can only add to some of the gaps in knowledge, but it cannot support previous findings of a significant link between EF or Mc and physics. The statistical power in the tests used could be a contributing factor in these conclusions, as they involved small samples sizes and lacked the necessary power to confidently detect significant relationships in the data. This means the null hypothesis that no significant relationships exist cannot be rejected or fail to be rejected, and instead, the findings must only be seen as trends that would benefit from further research with more participants.

A similar conclusion is drawn on the question of a link between Mc and physics performance. As discussed in the literature review, there is little research in this area, so the current work adds to this limited field of work, while acknowledging that measuring Mc in young children is complex, due to the nature of measures available and because Mc development is believed to be undergoing rapid change at this age (Kuhn, 2000; Whitebread et al., 2009b). The question of what is responsible for strategy use can also not be adequately addressed due to a lack of significant findings between strategy development and the potential variables, including Mc and visual-spatial skills.

The current study did not find a significant relationship between EF or Mc and physics performance, but this could be due to the reasons outlined above. Perhaps if different EF, Mc, or physics tasks had been used a different finding may have emerged or if a larger sample had been used the findings may have been different or more compelling. The lack of consistency between the tasks used in previous research makes it more difficult to tease apart the question of whether the tasks used are the key to finding or not finding statistically significant results. Some theoretical research suggests that there could be an interactive relationship between EF and Mc, which feeds into problem-solving ability (Diamond, 2013), but perhaps it is due to EF and Mc both developing at this age that mixed or statistically non-significant findings are often seen.

To further address research question 1, the strategy development data were examined alongside EF and Mc. The aim was to examine whether performance was potentially affected by a lack of knowledge or due to holding misconceptions. For the 4 balance trials, the majority of children always used the correct solution (29/35), which suggests children started the trials with the necessary knowledge to solve these problems (whether they knew before the study or found out during GP/DI). For the 2 balance trials, 17/35 children used trial and error and found the correct answer and used it consecutively for at least a third of the trials. The second most common strategy was to use trial and error and to not find or to not consistently use the correct answer (11/35). The high percentage of children using trial and error for the 2 balance trials indicates that they more likely had a lack of knowledge, which for most children was rectified since they discovered the solution and went on to repeatedly solve the problem type. For the 2 conflict balance trials, a number of children (14/35) consistently used the same incorrect strategy for all of their attempts. This might indicate that they believed this incorrect strategy was the way to solve the trial, despite it always being wrong. These strategies were either placing the (different) weights at the same distance (9/14) (which could indicate no understanding of weight, but an understanding of distance) or placing both weights on the same side of the beam or only placing one weight (5/14) (indicating no understanding of balance). These findings suggest that some children may have held a misconception concerning the correct solution, but others had a lack of knowledge. The other interesting finding with the 2 conflict balance trials were the number of children who appeared to use trial and error, sometimes finding the correct answer (2/35) and sometimes not (9/35). The majority of the remaining children were seen to consistently use

two wrong strategies (7/35) and to use trial and error and not find or consistently use the correct answer (3/35). These findings were explored further when answering research question 2 and whether the support groups differed in their strategy development.

These data, therefore, suggest that many individuals used various strategies for some of the problems, which would not easily lend itself to the staircase models, especially as the use of trial and error (wrong) was often seen. Halford et al. (2002) can only account for such a diverse range in strategies if children were going through periods of improvement, but it was not seen that the variations went on to lead to the correct answer in the time period examined, making it difficult for Halford et al. (2002) to account for. The RR model (Karmiloff-Smith, 1992) would not predict multiple strategies and regressions within a session unless children had found the correct strategy and were attempting other strategies to see if there is a better solution. The range of trial and error and incorrect strategy use despite not necessarily finding the correct solution makes it unlikely that this is what was happening and thus the data do not convincingly support the RR model. The data fit more easily to the OW theory (Siegler, 1996), the GR account (Munakata, 2001), and the connectionist model (Schapiro & McClelland, 2009), as they support the use of multiple strategies for the same problem within and between sessions, which was seen. The OW theory suggests children have various strategies available to them for use with each problem, so multiple strategies are accounted for, especially at times when a child is unsure of the correct answer. The GR account and connectionist model can account for the findings, as the strategies used on each trial (whether always consistency incorrect, correct, trial and error, or other) are selected based on the information available from the trial (e.g. through the graded representations and the activation of units related to the weighting of the connections). In the connectionist model knowledge is held in these weighted connections, so depending on how they are activated on each trial will influence which strategy is selected and used (Schapiro & McClelland, 2009). However, both the GR account and connectionist model would predict an overall improvement, as each would suggest the graded representation and weighted units would increase in strength when the correct solution is found.

Despite the strategy patterns displaying a variety of interesting data, few significant links between the pattern classifications and EF or Mc were detected. One significant finding was a difference in Mc rate at TP2 between the children who used the correct strategy and those

who used trial and error (correct) or (incorrect), with those using the correct strategy found to have higher Mc rates. There is previous research that indicates EF could aid in suppressing misconceptions (Masson et al., 2014; Brookman, 2015, cited in Tolmie, Ghazali, & Morris, 2016), which may have been supported if significant links between the strategy patterns showing children consistently using the wrong strategy and EF had been found. However, it could also be that the task in the present study made it easier to learn and to overcome misconceptions, thus EF played less of a role than would be predicted. Mc was expected to relate to the strategy development data, as it was thought strategy use and strategy development could be part of Mc (Zohar & Barzilai, 2013). It is also likely that power was a factor here, as the tests were underpowered, thus the trends and large effect sizes are an important indicator, while acknowledging significant differences may not have been detected due to a lack of power. This was an interesting aspect to consider and perhaps if a higher number of trials been completed by each child or a larger sample of children had taken part a different finding could have emerged.

In sum, the results of research question 1 found no significant correlations between EF or Mc and physics task performance or strategy use, but insufficient statistical power is acknowledged as a limitation. The balance beam strategy development data could be accounted for by the OW theory (Siegler, 1996), the GR account (Munakata, 2001), and the connectionist model (Schapiro & McClelland, 2009), but are less well accounted for by the RR model (Karmiloff-Smith, 1992) and Halford et al.'s (2002) work. Although no significant link between EF and Mc to physics was seen here, it could potentially still be accounted for if EF and Mc's role changes each trial, thus no stable link would be seen, which could be possible at this young age when EF and Mc are still developing. If this is true, it could potentially be accounted for by the RR model and the connectionist model. The lack of significant findings may also be accounted for by the small sample size and the low power of the statistical tests. Although no significant link was seen between EF or Mc and physics task performance or strategy use the power in the tests run may explain this, resulting in being unable to conclude whether any links exist between the variables. Instead, the findings should be seen as potential areas for further research.

7.5 Research question 2: What impact does support type have?

Research question 2 explored the impact of support type and whether there were any

differences between the two groups. Performance on the EF, Mc, balance beam, strategy development, and transfer physics tasks were examined and the results will be discussed next.

7.5.1 EF

It was hypothesised that GP might show more improvement in EF scores since they would have the chance to direct their own behaviours during the physics tasks, but this was not supported. The groups were matched on EF scores at TP1 and no differences were seen between the groups' performance at TPs 2 or 3, but the tests were insufficiently powered, so there could be differences that were not detected. At present, there appears to be very little research in this area, with only work by Barker et al. (2014) seemingly addressing the impact of support on EF. Barker et al. (2014) found that children who engaged in less-structured activities displayed stronger EF skills when it came to directing their own behaviour. However, they examined EF over much longer periods of time than in the current study, and here, the length of time the support type was provided and the frequency in which they received it was likely not long enough or frequent enough to impact EF in a measurable way. Although the data here indicate there was not a significant impact of support type on EF, the issue of sample size and power will have contributed to the findings. It makes it difficult to conclude whether support type impacted EF, as there is not enough power to reject the null hypothesis or fail to reject it, so the findings here should be seen as trends only.

7.5.2 Mc

It was hypothesised that both groups might show benefits to Mc: GP through the support type elements including questions, prompts, and opportunities for self-discovery, and DI through incorporating the feedback on why a trial was correct or incorrect and why. The results of the Mc rate analyses showed GP scored significantly higher than DI at every TP, although it should be noted that the tests were insufficiently powered. The trends suggest an additive effect, as GP benefitted from the support from the start, but did not go on to make any more gains than DI. There was no baseline Mc rate taken, but the significant difference between the two support groups at each TP suggests that different levels of Mc behaviours were likely due to support type. This finding indicates that support type had an impact, but the lack of power means it is not possible to say this with certainty, as it could potentially be due to chance, although the effect sizes support the idea of differences existing between the groups. The data suggest it is likely due to GP being questioned after the trials, prompting them to

verbalise, thus increasing their chance of scoring on the Mc rate. It was hypothesised the two groups might show benefits for different reasons: GP might benefit from the questions and prompts (as was found), but also through the chance to discover information themselves during the play time, and DI might benefit from the feedback on whether a solution was correct or incorrect and why. If DI did benefit, it was not to the same level as GP, as seen by the statistically significant difference. To follow up on the finding, one important question was whether there was a benefit to this increased Mc rate, that is, did GP's higher Mc rate reflect in performance differences on any of the other measures. From the other results, it appears that GP's higher Mc rate did not translate into any other tasks, as they did not score significantly higher than DI on any other measures. This was also seen in the Mc interview scores, to be discussed next.

Although a difference in Mc rate between the groups was seen, no significant difference in Mc interview scores was seen at any TP, but this finding again could be due to insufficient power and being unable to detect smaller differences. As stated in the exploratory analyses findings, vocabulary scores were seen to play a role in Mc interview performance at each TP. If language played a role in children's understanding of the questions or if they struggled to express themselves this could have led to poor scores on Mc interview, regardless of their "true" Mc. No baseline measure of Mc was taken, so there may be an element of individual abilities here, beyond the impact of support type.

One issue with the two Mc measures is that only overt behaviour was measured. Thinking processes and thoughts cannot be measured and recorded beyond children translating into overt actions, which is unlikely to always happen. This may also be one reason why vocabulary was related to Mc interviews, but not Mc rate, as the support type caused a change in language use for GP, but this did not carry on into the Mc interviews, which could be why language was found to be significantly related to interview scores, but not Mc rate. It may also be, as suggested earlier, that Mc rate is a better measurement method for MR and the Mc interview a better measure of MK. The Mc measures used were the best available for the age group, although they have been used and reported here acknowledging that they are unlikely to be a "true" representation of children's Mc. This may be one reason for not finding a statistically significant relationship between Mc and EF and to physics task performance when addressing research question 1, but power is another, more likely, reason.

However, as previous research in young children has not established strong, consistent links between Mc and either EF or physics task performance, this data instead add to this emerging field of research and should be seen as highlighting areas of research others can build on.

To conclude, a potential significant impact of support type on Mc rate was seen, with GP showing more benefit than DI, but no significant impact of support type on Mc interview scores was detected. As stated throughout, statistical power is a limitation of these analyses and the significant findings should be noted, but viewed with caution. The significant difference in Mc rate was not found to significantly relate to other measures or to show benefits, such as being reflected in balance beam performance or strategy development, suggesting that although the GP support increased Mc rate it did not benefit the children as measured by the various tasks here. The sample size will have contributed to the findings, and the data, therefore, cannot reject or fail to reject the null hypothesis, as there was insufficient power. The Mc data trends should be considered further in future research.

7.5.3 Balance beam task

It was hypothesised that GP might show more improvement over time through the chance to discover information themselves in each session, but also that DI might score higher from TP1 but show less progress, as they will have been provided with all the necessary information to begin with. A significant difference between the groups' balance beam scores was found, as DI scored significantly higher than GP at TP3. It should be noted that although DI's scores increased, GP's scores decreased, which will have added to the significant difference in scores between the groups. It should also be noted that the tests lacked sufficient power, so the findings should be considered as trends only. (For comparison, these results were found both when all children who completed the balance beam task at TP3 were included in the analyses and when only children who completed all three balance beam tasks were included in the analyses, supporting the finding that DI showed a higher performance score at TP3, but GP's performance regressed.)

It may be that DI support provided a better opportunity to retain and build on the information and their new knowledge, which is why an increase was seen by TP3. It is interesting to note that DI showed no improvement between TPs 1 and 2, but did between TPs 2 and 3, whereas GP showed an improvement between TPs 1 and 2, but regressed between 2 and 3. In

opposition to the DI support, perhaps the GP support did not provide a way for children to solidify their learning. There are no significant links to the protocol analyses that could explain why the GP would regress (such as instruction or feedback provided). The RR model can only account for the data if children are undergoing redescription in order to improve on the correct strategy they already know, but that is not representative of the data here. A pattern of regression in strategy use can be accounted for by the OW theory, the GR account, and the connectionist model. The OW theory would suggest children have a range of strategies children can rely on, but it would not account for why GP (and not DI) showed such a regression – to the mean of TP1. The GR account and connectionist model may suggest it was dependent on the information available via the graded representations or connections for the trials at TP3, so perhaps the progress made at TP2 did not transfer into strengthening the representations or connections, meaning they performed as they did at the first TP.

Some research has found GP to be better for children's learning (Fisher et al., 2013), but research in the physics field tends to find DI to be better for children's learning (Chen & Klahr, 1999). The current finding supports the idea that when it comes to learning the type of physics tasks explored here, DI is a better support type for teaching children. This claim is supported by the finding that DI scored significantly higher than GP on the 2 conflict balance problems. As discussed earlier, these problems are more difficult than the 2 or 4 balance problems, as they require placing different weights at different distances, and not hold one variable constant. This is a noteworthy finding as it indicates that children can learn how to solve these difficult problem types, which require considering and manipulating weight and distance, before the age many would predict (Halford et al., 2002; Siegler, 1976; Karmiloff-Smith, 1992). Schrauf et al. (2011) found 3-year-olds struggled with during a production task, but a key difference is that the children here were given weights and told to make them balance, whereas in Schrauf et al.'s (2011) study children had to select the correct weight to put on one side of the beam to make it tip.

Halford et al. (2002) claim that only at 5 years of age can children consider both weight and distance and Siegler (1976) claimed that children could only achieve this at age 12. Some have also found that considering distance alone is challenging for young children, such as Siegler and Chen (1998) who found that only after completing trials, receiving feedback from

an adult, and having to explain the outcomes, could some 4-year-olds solve distance trials, and even then it was only a small percentage of children (6%). Li, Xie, Yang, and Cao (2017) also found that when feedback was provided (as in the DI condition), 4- to 6-year-olds' performance on distance trials improved. Here, 4% of GP correctly solved a 2 conflict balance trial, compared to 20% of DI. The difference in the present study's results could be the task type used (production), and the instruction, demonstration, and feedback provided to DI. Looking at the previous studies' methodology for support type, some used mixed DI and GP elements, but the key seems to be an adult explaining why something works or does not work, although Fisher et al. (2011) would disagree and claim instruction is not enough. The RR model's final level (E3) requires that knowledge can be verbalised, which implicates language as a key element in making knowledge explicit. It could be the instruction provided during the DI support aids with redescription of knowledge, and within the GR account and the connectionist model, this could be likened to modifying the graded representations or the weighted connections, making knowledge more explicit.

One issue to consider in the current work is whether children truly considered weight and distance when solving the problems, or simply solved the trials by "matching" where the dinosaurs sat on each side of the seesaw. There were times children referenced the heavy and light dinosaurs and were seen to weigh them in their hands, so some notion of weight appeared to be there. How much this really translates into an understanding of weight and distance is difficult to unravel, however, at the same time, it could be said that how adults calculate weight and distance raises a similar question. As has previously been noted (by Siegler, 1976), adults are not always capable of solving problems that involve calculating weights and distances in order to make a beam balance. Nevertheless, knowing that the same weights must be at the same distance shows an understanding that both are important for making it balance – which is shown in the performance scores, the first correct scores, and the consistency scores (discussed in the next section). Overall, there appears to be some evidence showing young children can solve these problems to be able to say that they do not have some understanding of weight and distance.

Although the results do show a significant difference at TP3, the test was underpowered due to the sample size. It is possible that if there had been more data the results could have been different or further significant differences could have emerged. The difference at TP3 shows a

large effect size, indicating that differences exist in the data, which is likely the reason a significant difference was detected. The same was seen with the difference in performance on the 2 conflict problems, as a large effect size was detected. In sum, it can be concluded that the data here indicate that the support type did have some impact on balance beam performance, with DI showing significantly higher scores than GP at TP3 and on the 2 conflict problems.

7.5.4 Strategy development

Kuhn (2000) suggested that through feedback on strategy use, strategy selection can become more conscious and strategy development can be seen. This idea was considered and the impact on both support types examined. It was thought GP would have opportunities to consider their verbalisations concerning the strategies they used and why they were correct, and DI could consider the feedback from the adult on whether their strategies were correct and why. The data here do not lend support to one support type impacting strategy development more so than the other support type. As previously stated, it was expected that Mc would be linked to strategy development and a relationship be seen between these factors since Mc was suggested as being responsible for strategy use, but no statistically significant link was seen. It was thought that Mc might impact how strategies are considered, developed, and applied over time, and thus support type might impact Mc, which might impact strategy development. It was hypothesised that GP would show more progress in strategy development, but DI show better strategy use from the start. Even though a significant difference was found between the groups' balance beam performance at TP3, between the groups' 2 conflict balance problems performance, and between the groups' Mc rate, these have not been found to significantly impact strategy development. No differences were seen between the groups' strategy development – first correct, consistently correct, or strategy classifications, however, as already stated, the power of the tests employed may have contributed to these findings. A lack of power means that the tests are less likely to detect differences, especially small and medium effect sizes, should they exist.

Some qualitative differences in strategy classifications for each problem were seen between the groups, but they did not reach the level of statistical significance. For the 2 conflict trials, interestingly, the only children who showed improvement in their strategy use were from DI: 3 children began using an incorrect strategy, but consistently used the correct strategy by the

end of the trials, and 2 children were using trial and error (correct), which suggests the DI children were showing some learning of this problem type. The majority of the GP children always used the same wrong strategy, which suggests a misconception on how to balance these weights and shows no learning. For the 2 balance trials, 4 GP children always used the correct strategy, compared to 2 DI children, but 8 GP children used trial and error (wrong) compared to 3 DI children, and 6 GP children used trial and error (correct) compared to 11 DI children. These data suggest DI showed more learning through more correct trials using trials and error. For 4 balance trials, the majority of children in both groups always used the correct strategy, which supports the idea that children perhaps already had knowledge of weight concepts before starting the task. Although the data do not reach statistical significance there are some trends that may suggest qualitative differences between the two support groups, with DI showing some more learning over the trials compared to GP, based on their strategy classifications. If this is the case it could be the instruction before the balance beam highlighted the ways to solve the trials, and the feedback after the trials enforced it, as children were told if they were correct and why. This may then have encouraged children to either use the correct strategy on the next same problem or consider a different strategy. The data trends do suggest some change in learning, thus in children's understanding of physics concepts concerning balance, but as before, power is a likely limitation of these analyses and may have masked significant differences.

As discussed in relation to research question 1, the data do not support Halford et al.'s (2002) staircase theory of learning and are difficult to be adequately accounted for by Karmiloff-Smith's (1992) RR model. Instead, the strategy classifications are better accounted for by the OW theory (Siegler, 1996), the GR account (Munakata, 2001), and the connectionist model (Schapiro & McClelland, 2009), as they allow for a variety of strategies to be used, and these can be influenced by experience and feedback. The theories explain the qualitatively different patterns and the trial and error patterns frequently seen rather than discussing them in terms of strategies of learning, as in the staircase accounts. Overall, the results here do not show any trends to indicate that support type has an impact on strategy development, but statistical power is a limiting factor in what can be concluded.

7.5.5 Transfer physics task

No solid prediction was made on whether a transfer effect would be seen. The ramps task was

included to examine whether Mc related to physics task performance, if Mc related between the two tasks, if performance on the two physics tasks was related, and see if support type had a lasting impact. When the ramps task data were examined, no differences were seen between the two groups on any of the measures, however, only around half of the sample completed this task due to space and time restraints within the nurseries, so this will have impacted the results. The low power would have made it difficult to detect significant relationships between the variables and significant differences between the groups, so the data should be viewed as trends only and not conclusive of whether significant links exist.

Like in the balance beam task, no statistically significant links between Mc and performance or between balance Mc and ramps Mc were seen in the data. The small sample size makes it difficult to reach a conclusion, as it is likely that the lack of data is a limiting factor here. Besides a lack of power to be able to detect differences, other reasons to consider include that the Mc measured during each physics task either did not capture the same component or the two physics tasks elicited different Mc scores from children. Interestingly, the opposite relationship to what was seen during the balance beam task concerning vocabulary and Mc was seen here: the ramps task data revealed vocabulary significantly correlated with Mc rate, but not Mc interview scores. Perhaps, as suggested earlier, the support type somehow overshadowed the role of language in Mc rate, whereas during the ramps task that was not the case. It is unclear why language would not continue to play a role in Mc interview scores, although a drop in participants and statistical power could be a reason. Considering the data as trends, it could be due to children having a better understanding of the questions to be used, although the measurements were only taken a week after the last balance beam task, so a significant difference would not be expected. However, the week delay may also be a contributor to the difference. The Mc and balance beam tasks were each taken 5-6 weeks apart, but the ramps task was conducted just one week later, so maybe there was a lasting effect from the previous week's session. This is difficult to test, but it does raise the question of a potential lag effect.

Performance on the balance beam did not significantly link to performance on the ramps task. As before, statistical power is likely playing a role, but the trends would suggest these skills are not transferable, or each task targeted different skills. This was seen through examining individuals' performance on the balance beam against the ramps data, as well as comparing

the two support groups' data. This finding does not support Klahr and Nigam's (2004) finding of a link between physics tasks, where they found performance on a CVS ramps task related to performance on a transfer task involving rating science posters. Their finding indicates that perhaps the same skills were being utilised in each task, but this was not clear here. The effect sizes between performance on the balance beam and the ramps task showed small and medium effect sizes, but some were positive relationships and some were negative. The findings mean the null hypothesis cannot be rejected or fail to be rejected since there is not enough power to confidently conclude whether relationships exist in the data.

Looking at the ramps task performance data between the groups, the mean total scores show GP performed better than DI during the trials, with a small effect size seen, and for being correct on the first try there was a large effect size, indicating some difference existed between the groups, but it was not statistically significant. The small number of participants is likely a problem and one can only speculate what may have been found had more children participated. However, the finding does lend some support to Klahr and Nigam (2004) who found that support type did not impact performance on a transfer task. They found their DI group (who received instruction while seeing the ramps set up) performed better than their discovery learning group (who had to set up their own ramps tests without instruction), similar to what was found here. Their transfer task was somewhat different though, as it involved rating science posters, although the key concepts of designing experiments and implementing CVS was still included. They did not find support type to play a role in the transfer task, but as stated earlier, a link between individuals' performance on the two tasks. During the balance beam task GP scored significantly higher than DI on the Mc rate, and DI was found to score significantly higher than GP on the balance beam problems at TP3 (as well as on the complex 2 conflict balance problems), but these performance differences have not appeared to have carried over to any ramps task measures, or least to be detected with the power of the tests employed here. For Mc rate, it suggests that the support type was the driving force in the different rates displayed by the two groups, as the difference disappeared during the ramps task when no specific support type was employed.

It is concluded that the low power likely affected the ramps task analyses, resulting in not being able to reject or fail to reject the null hypothesis. The trends in the data show that support type did not significantly impact performance on the physics transfer task, but there

were some differences between the groups, as seen by the effect sizes. Performance on the two physics tasks did not significantly relate, possibly due to the tasks requiring different skills and/or experience, but more likely due to the small sample size. It could be that training in one physics task does not aid performance on another physics task, but more research is required. The findings, therefore, do not provide any conclusive evidence as to whether support type had a lasting or transferable effect.

The theories discussed in this work would suggest that experience is an important factor in performance and learning, so this could be one reason no strong transfer effect was seen in the present study. In the current work, the balance beam and ramps task both test forces, and although there is little overlap in the physics concepts tested (weight and distance versus friction and incline), there is overlap in the scientific enquiry aspect of the tasks, such as observation, answering questions, and testing variables. In Klahr and Nigam's (2004) work both tasks tapped CVS skill, and their sample was older (8- to 10-year-olds), so drawing links between the two tasks may have been easier for these children. Children here may have had more knowledge of the ramps' variables, as the mean scores for first try and total percentage correct appeared to be much higher than for the balance beam task. Some theories support experience being an important aspect of learning, so perhaps children's knowledge level of balance and ramps task concepts are too different for significant links to be found. It may also be that the change in support type from the balance beam task to the ramps task meant there was not much overlap in the scientific enquiry aspects in each task, as DI and GP support employed different scientific enquiry elements, which differed to the ramps task where the support type was neither DI nor GP. The smaller sample size should be noted when considering this finding, as the effect sizes and trends show some differences, but unfortunately, the tests were underpowered, possibly leading to being unable to detect significant differences. As stated, the mean scores showed a notable difference between the groups, which could be considered worthy of future research.

7.6 Limitations of the present study and recommendations for future studies

There were some limitations to the study, which will be addressed here alongside recommendations for future studies. One limitation was the small sample size and missing data (resulting in low statistical power), which are common issues when working in an educational setting and relying on children attending on set days over a number of sessions.

Other limitations relate to the tasks used, particularly due to the age of the children and reliance on language.

The small sample size resulted in the statistical tests having low power to detect significant effects, should they exist. The simplest statistical tests were used in order to increase the power as much as possible, and some of the tests (primarily the correlations) had enough power to detect large effect sizes, but the other statistical tests did not. The low power means that the null hypothesis cannot be rejected when the result is non-significant, as it could be that a difference does exist in the data but there was not enough power to detect it. A large effect is easier to detect and reflects less overlap in the data (Sani & Todman, 2005), so effect sizes have been reported throughout to give an indication of the relationships within the data being tested, as recommended by Sullivan and Feinn (2012). The effect sizes here should be seen as trends within the data and are particularly noteworthy when the effect size is large but the test does not reach statistical significance.

The issue of power is linked to an increase in type II errors – incorrectly failing to reject the null hypothesis when there is a significant difference in the data. The multiple comparison corrections applied may also contribute to this, as the *p* value was lowered (based on the Holm-Bonferroni correction method) before a result could be declared significant. Lowering the *p* level makes it more difficult to reach significance, potentially increasing type II errors (Sani & Todman, 2005). However, as stated earlier, the number of statistical tests performed, and the exploratory nature of those that were corrected, required corrections to be applied to avoid type I errors – incorrectly rejecting the null hypothesis when there was not a true significant difference in the data. It is hoped that employing Holm-Bonferroni corrections reduced type I errors, while still allowing significant findings to be detected.
The low power means the null hypothesis cannot be rejected or fail to be rejected, as there is not enough evidence that the result is not due to chance (Sullivan & Feinn, 2012).

The missing data could have been addressed through changes to study design and/or changes to the tasks used. The study design meant that the balance beam was carried out in the last session of TPs 1 and 2, so if a child was absent they would not complete the task, and the ramps task was carried out during the last visit to the nursery, so again if the child was absent they could not complete the task. Due to the nurseries' locations, I could only visit one

nursery per day and I could not add in extra visits due to the set dates I could visit. The number of absent participants was not anticipated when designing the schedule, but adding in extra visits would have reduced the number of nurseries that could be visited. It could be if this change was made there would be more complete datasets per participant, but fewer participants, thus the power in the study may still not have been sufficient.

The tasks were selected after piloting and deemed the most appropriate and reliable for testing the measure it apparently tests. The range in scores for all the tasks indicates the tasks were suitably challenging for the participants in the study, even after repeated testing. The variability in scores supports the claim that the tasks were appropriate for this age range and for multiple uses six weeks apart. This provides support for the tasks not being the reason for the results found, although some aspects of some tasks could potentially be improved for future work. For example, a baseline for the physics task would have been helpful when considering how much the children improved to try to and unravel how much learning took place. As stated in the methodology chapter, including a baseline was considered here, but it was deemed problematic, as it would be difficult to develop a baseline method that did not employ either GP or DI elements, which could potentially influence the result and not be a true baseline test that did not incorporate GP or DI elements. It would have been useful to have a baseline, but at the same time, it would not be useful for children to have the chance to learn during the baseline session, as it would affect being able to measure the impact of support type on learning. Taking a baseline from a sample of children matched on other measures (such as age, EF, and Mc) could be a helpful comparison to an experimental group. However, it is believed here that since the two groups matched on other variables there will likely be a mix of abilities in the groups.

It may be worth future work considering whether discontinuation rules are required for the balance beam tasks. I employed the balance beam task in this study based on previous work that suggested it was of no benefit to ask a child to continue to complete trials after they had reached a point where they could not successfully solve the trials, but also based on the piloting I carried out, which suggested children became bored and/or disheartened when they had to complete too many trials and when they were often wrong. By having all children complete the same number of trials it would allow for a more direct comparison to be made concerning total performance scores and performance on each of the problem types. The trial

numbers likely impacted the power of the data and if every child had completed more trials a clearer picture of the findings may have emerged.

The EF tasks showed varying strengths with one another and across TPs, although with the shifting task appearing less stable than the other two EF tasks, although the effect sizes were still noteworthy, but the tests did not reach statistical significance (after the Holm-Bonferroni correction was applied). This may have somehow contributed to the relationships examined but should not have compromised the composite score. As stated earlier, the Mc rate relies on language and overt behaviours, so the coding may not capture the cognitive processes involved with Mc and only measures visible Mc. This is a difficult problem to overcome, but the use of more than one type of Mc in a study should help, as should the use of an independent measure of Mc out-with support type, such as Mc rate during another task, which could then be used as a comparison. This was attempted here, but the two Mc measures did not significantly correlate, suggesting they were perhaps not measuring the same Mc components or the relationship was not strong enough to be detected in the current data.

The EF and Mc interview scores were significantly related to BPVS, and Mc rate partly relied on verbalisations for scoring, suggesting vocabulary was important. BPVS was factored into the EF and Mc interview analyses to control for some of the variance in data due to BPVS, but a similar approach could not be taken with Mc rate. This likely resulted in a measure that only partly measured Mc and so it is possible that some findings related to Mc may not have been detected. It was hoped by using two Mc measures that something would emerge from the data, but it instead seems that the tasks may have measured different Mc components. The EF and Mc tasks available for use with young children was very limited and although the data collected using the tasks here did not always present as expected, it is still thought that they were the most appropriate tasks to use.

When selecting EF and Mc tasks for future studies it is recommended to use the same tasks and instructions as others have used within a research study, so direct comparisons can be made. As every study selects their own tasks it can add variance concerning whether the task really measures what it is said to measure, meaning comparisons between studies will always need to consider whether the tasks used measured the same component as others attempted to measure with another task. For example, considering the literature that examined EF in young

children: some found EF components to be unitary and others found them to group into two or three components, but each study used different EF tasks, so it is unclear whether this is a contributor to mixed findings in the field.

Receptive language was used as the measure of vocabulary here, and although it is a widely used task with young children, a measure of expressive vocabulary could have helped unravel whether Mc rate and interview scores were impacted by children's ability to express themselves. Thus, future studies should consider using measures of both types of vocabulary. Vocabulary was found to be an important factor here, so being able to account for expressive language and comprehension is worthwhile. Longitudinal studies would be recommended to use a second measure of language at the end of the study to see whether the links between language and other measures can be examined for directionality, for example, does EF impact language or vice-versa. This would help unravel the direction of the variables' relationships and examine what drives performance.

Despite the limitations outlined here, it is thought that the same or at least similar pattern of results would have been found, although the conclusions could have been stronger with more power. The results that were significant tended to show large effect sizes, indicating there was a noteworthy trend in the data, such as the difference in groups' balance beam performance at TP3 and performance on the 2 conflict trials. If there were more data and statistical power more significant associations may have been found between some of the EF components or some of the Mc components, but the trends and effect sizes do not indicate that a link between the balance beam task and EF or Mc would likely have emerged with more data – which was research question 1. The trends and effect sizes in research question 2 also indicate that with more data and power a difference could have emerged between the groups for the EF tasks or the Mc interview scores. The trends and effect sizes for the difference in groups' performance on the balance beam task would have benefitted from more data, as it is less clear if a significant difference could have emerged, particularly at TP1 when the effect size is just below the cut-off for medium. Additional data would have increased the power of the tests, allowing for medium and small effect sizes to be detected, thus strengthening the findings. The statements made here are based on the visual data trends and effect sizes reported and it is acknowledged that more data may alter the trends resulting in different conclusions. The ramps transfer task was interesting because GP had a noticeably

higher mean but did not reach statistical significance, again impacted by low power, so more participants would have beneficial.

In sum, the identified limitations of the study are not thought to have impacted the overall findings and conclusions made. The limitations highlighted here and the suggested recommendations should be considered for future research, as they could add further control over some of the issues identified and aid in the analyses, particularly through increasing statistical power to be able to the null hypotheses in favour of the alternative.

7.7 Contributions of the present study

This study has contributed to the theoretical fields examined, to educational aspects, and to methodologies. Each will be discussed next, alongside some recommendations.

7.7.1 Theoretical contributions

This work was influenced by frameworks and research that addressed the structure and function of EF and Mc, how they relate to physics performance, strategy development, support type, language, and visual-spatial skills. The work aimed to explain the data with use of the theories available.

This work found evidence to support the theory that EF components in young children are separate, but show links, supporting the findings of Garon et al.'s (2008) work with children and Miyake et al.'s (2000) work with adults. The data here suggested the components were not unitary, but also not completely dissociable, suggesting some linkage. More statistical power may have strengthened the relationships between the EF components, so it is recommended that future work consider this further and employ tasks others have used in order to form direct comparisons of the measures. The structure of Mc was more difficult to unravel: Mc rate and Mc interview scores were not significantly correlated in the balance beam or the ramps task, thus it was thought they either did not assess the same Mc component or because low power meant small and medium effects could not be detected. This finding could also perhaps be due to individual differences in how much overt Mc is displayed (rate) and differences in comprehension or expressive language ability. The methods did, however, show that some measure of Mc can be taken in young children and so it can be said the data support the claim that Mc has started to develop by around age 3,

supporting the work of Kuhn (2000) and Whitebread et al. (2009b). However, others (Roderer & Roebers, 2014) would argue that Mc, in particular control and monitoring, do not develop until much later, but the data here dispute that since Mc rate was coded for control and monitoring, albeit the rates were low. This work has not provided enough evidence to support the idea that EF and Mc are statistically correlated, at least in young children, so cannot support Diamond (2013). It could be due to the age of participants, the tasks used, that the link is only measurable with certain tasks, or it can vary due to the development EF and Mc is likely undergoing. The power of the tests used should be seen as a limitation here, but trends and effect sizes indicate little association here, but it is believed this is still worth further research to examine these areas in order to add further knowledge to the field.

The literature review addressed the question of what component might be responsible for strategy use and development, with vocabulary and visual-spatial skills (Roberts et al., 2007), EF (Diamond, 2013), and Mc (Zohar & Barzilai, 2013) being suggested as having possible roles. If links between strategy development and EF were found it could support the idea that EF is responsible and if links to Mc were found it could support the idea that Mc is responsible for strategy use. Not enough evidence was found to support these ideas, with low power playing a role in the likelihood of detecting significant differences, although one result in the strategy development data indicated that the use of the correct strategy was linked to a higher Mc rate. Rozencwajg (2003) found evidence to suggest that strategy selection could be linked to Mc, so although the evidence here is limited, it could be highlighted as an area for future research.

The theories and accounts to explain balance beam performance were used here in an attempt to explain the present findings. The data here suggest that children could consider 2 variables –weight and distance when solving problems, but do not necessarily get every trial correct or all problem types correct. This does not support Halford et al. (2002), Siegler (1976), or Karmiloff-Smith (1992) who say children either cannot consider more than one variable at once during the balance beam task, or performance does not improve until older. The strategy development data showed that for some problem types the solutions used often changed per trial and/or per session, which does not support the staircase accounts (Halford et al., 2002; Siegler, 1976) and difficult to account for by Karmiloff-Smith's (1992) RR model. The data instead support Siegler's OW theory (1996), Schapiro and McClelland's (2009) connectionist

model account, and Munakata's (2001) GR account. The connectionist model and GR account would explain the strategy selected for a particular trial as being influenced by the information available to the child at that time (which could include the materials provided, language, and support) as it would change the information available through the weighted connections or through the representations available to the child.

Although the GR account and connectionist model offer the best account for the data here, the problem with these types of accounts is that they can potentially always explain all outcomes seen, which makes it difficult to disprove. This study would conclude that children cannot be neatly classified as consistently using correct or incorrect strategies, especially over different problem types and instead continuous learning seems a better way to account for the current data. If children were classified by the staircase accounts a fair amount of the qualitative differences would be lost. The use of staircase accounts can be acknowledged as existing to try and determine which 'stage' or 'level' of knowledge children have, but the data suggest it is not that simple or useful. It is instead recommended that when examining strategy development data, that the pattern of strategies used is examined in order to see whether children rely on a few select strategies or multiple different strategies, and whether it is within one session only or it changes over time. Considering how children's strategy pattern can be classified could also aid in answering the question of whether misconceptions are being displayed, whether children are relying on trial and error, and ultimately how successful children are. The in-depth strategy development data has been seen here to be incredibly insightful in children's balance beam performance.

Overall, this study has made theoretical contributions to the field of EF, Mc, and physics. It is hoped this work will aid others in forming new research questions and drive more research in the area.

7.7.2 Educational contributions

This work set out to examine two different support types to examine if and how children's ability to learn can be influenced and what role individual differences play. The literature review identified conflicting fields of work that suggested GP support is more beneficial versus work that suggested DI support is more beneficial for learning. The two support groups in this work did show some differences in their learning, as seen by the strategy

development data, by the significant difference in performance on the balance beam at TP3, and the difference in performance on the 2 conflict problems. Although there was some evidence to support the idea that DI support is a better support type for children to learn about balance concepts, there was a trend in the ramps task data to suggest GP performed better on the concepts of incline and friction. As has been stated throughout this work, there was a lack of power in the statistical tests carried out to be able to confidently say whether one support type aided learning more than the other, but the trends and effect sizes indicate there was a difference. The difference in balance beam performance at TP3 appeared to be partly due to the DI group improving over time, but the GP group also showed a decrease in scores. It could be that the structure of the DI support allowed for learning to progress, but GP support relied heavily on children either remembering how to solve the different problems or re-discovering the solutions each time. DI always began the trials with instruction on how to solve the different trials, so it could be the element of repetition in the DI support reinforced the solutions over time and aided learning.

The strategy development data suggested a change in knowledge was captured, as seen by the trend in changing from consistently using the wrong strategy to using the correct strategy, particularly for DI. This could be said to show that some children did show learning over these TPs, but the small sample size makes it difficult to conclude whether this finding would likely be found again in another sample. Previous research has suggested EF could play a role in suppressing misconceptions (Masson et al., 2014; Brookman, 2015, cited in Tolmie, Ghazali, & Morris, 2016), but this was not supported by the findings here since significant associations between consistently using the wrong strategy and EF were not detected. However, a trend from consistently using the wrong strategy to using the correct strategy was seen, and more so for DI, which may be explained through these children learning over time, resulting in the expected trend not emerging. However, it could also be that the task in the present study made it easier to learn and to overcome misconceptions, thus EF played less of a role than would be predicted. There was also a difference in the 2 conflict trials between the groups and it was only DI children who showed progress in these trials, again suggesting something about DI support perhaps aided learning more than GP support. The qualitative differences between the groups could indicate the support type is playing a role in learning and it is therefore recommended that this be investigated further in future work.
The findings in this work may provide support for the feasibility of teaching physics concepts

and scientific enquiry to young children and in a relatively short timeframe. It could prompt early teaching of topics where misconceptions are often seen, such as how the seasons work (Dunbar, 2007) or the trajectory of a ball leaving a curved tube (Kaiser, 1986). It would also be recommended that tasks are presented in a game-like format, as done here. It may be this contributed to performance here. Although the mean scores on the balance beam were not beyond what could have been predicted, the fact children could solve problems while considering two variables is impressive, as it would not be expected from this age group, based on previous research. Making the aims of the task clear and tangible to the children may encourage them to do better.

The balance beam data indicated that children as young as 3 years of age may already have some knowledge of balance beam concepts (as seen by the 4 balance trials' high performance). It showed that balance concepts can be taught to and learnt by young children over a relatively short space of time, and adult-led support has an advantage over child-led support. The support types did not focus on teaching only one particular problem type or to include numerous repetitive trials, yet children still learnt about different problems. Based on previous work and the finding here that children in DI perform better than GP over time, some elements from the support type seem to emerge. Looking back at some of the physics studies carried out with children, and as discussed earlier, different elements of instruction seemed to relate to children's performance, e.g., Halford et al. (2002) found 2-year-olds could solve problems after a familiarisation session, but Schrauf et al. (2011) found 3-year-olds were not be able to solve balance problems, although children received no instruction. Schrauf et al. (2011) noted that the children in their study had no prior experience with the balance scale, which could be why the children performed so poorly. As has been discussed here, experience is important for developing knowledge, so perhaps this impacted their findings, as in the current study children started the trials with some knowledge, whether through DI or GP support. In the other work children were provided some information, such as being shown how to solve trials, being told how solve trials, and being told if the trials' outcome was correct. It could be that some of these elements are the important aspects involved in the difference found between GP and DI here. It could be interesting for future studies to consider using a combination of GP and DI elements to examine whether children benefit further from all the support elements. For example, in Halford et al.'s (2002) study they provided demonstrations and explanations on how to solve trials (such as in the DI

support) and then asked children to have some goes making the beam tip (somewhat like the GP support) (all before the test items) and they found that 2-year-olds could solve balance beam problems. Perhaps it is the combination of the different support elements that aided the 2-year-olds' learning. Thus it is suggested that a support type whereby children are both explicitly taught and allowed play may be better than DI alone. In terms of applying this in nurseries and schools, it is feasible to imagine children being taught as a class first, whereby they are told and shown how to solve problems, and then be allowed to work by themselves or with others on the task posed. The Department for Education (2013) already suggest elements of each support type for teaching children and scientific reasoning, such as observation (DI), answering questions (GP), and testing (more like DI, but both test solutions).

Since support type has become a more popular area of research in recent years, it is hoped this work will draw the two fields of GP and DI support together and encourage more work in the field of physics with young children. It is hoped that future research in the area will examine different support types and different elements within each support type in order for a protocol that benefits all children can be established.

7.7.3 Methodological contributions

Measuring balance beam knowledge and strategy use in a production task over several TPs has not been carried out with young children before. The addition of measuring EF and Mc alongside physics performance to examine the links between these variables has not been conducted before, despite some indications that there should be links. The use of several TPs allowed for an in-depth analysis of strategy development to be carried out. By recording each strategy used on each trial for each problem type at each TP the changes could be observed and the children's strategy use could then be categorised by the strategy pattern displayed. This qualitative analysis allowed for more inferences to be made as to whether children believed they knew the solution or whether they lacked the knowledge and whether they learnt over time. These analyses provided a detailed account of knowledge to be made, leading to the conclusion that learning balance beam problems does not show a staircase pattern of results and instead children often attempt multiple ways to solve a problem, even regressing after finding the correct solution. The use of production tasks should be encouraged in order to examine strategy development in the ways addressed here.

It has been shown that Mc in young children is measurable and two measures have been used with some success here. The C.Ind.Le did provide some interesting results, but the length of time taken to code the observational videos would be highlighted as a downfall of utilising this method. Observations take a long time to code and to double-code, so it is only recommended for work with few observations or hours of videos to code.

The longitudinal design here allowed for children's progress to be tracking, and as noted, was crucial for examining strategy development. If further data points had been included it could perhaps be shown an interesting picture at TP4 and revealed whether DI continued to progress or whether GP caught up. Future research designs should make use of longitudinal designs where possible when measuring children's knowledge, rather than rely on one data point.

This work will end with the 8, the conclusion.

8 Chapter 8
Conclusion

This study focused on whether EF, Mc, and support type have a role in young children's physics task performance. EF has previously been linked to physics performance in some research, and to Mc in other work, thus the connections between all three were examined, including how EF and Mc are structured and related. Some research has suggested GP support is better for teaching children, but research in the field of physics has suggested DI support is better, thus these two research fields were brought together to compare GP and DI within physics tasks.

This study concludes that EF components in young children are separable, but connected, but no firm conclusion was made regarding Mc components, although it did find Mc in 3-year-olds to be measurable and provides support for observational work and interviews for assessing Mc. The study cannot support the idea that EF and Mc are related, at least in 3- and 4-year-olds, or as measured with the tasks and data here, as the sample size was a notable limitation.

The research has also added to work that has examined children's understanding of balance beam concepts and to methodologies advocating for strategy development to be examined, rather than only considering performance at a given time. It has been found that 3- and 4-year-olds do understand balance beam concepts from this early age and can learn complex problems involving two variables. No significant associations to a transfer physics task were seen, likely impacted by low statistical power, so support for or against the idea that these skills are domain-general cannot be provided. DI was found to be better a support type for children to learn about balance beam concepts, as seen by data trends and effect sizes for performance on some problem types.

Language was seen to play a role in EF and Mc rate during the balance beam, and Mc interviews during the ramps task, so included as a covariate, as it was deemed to play a role in these tasks. Language has previously been linked with EF, although the direction of the relationship is still unclear, support for it was found here. The need to understand instructions and questions in some of these tasks could explain why language was linked. Inner speech

could also have played a role in children processing the information given to them (such as task goals), to plan task actions, or when responding to questions. Language is therefore seen as an important individual factor in performance, which future studies should account for.

It is hoped this work will add to the research examining EF, Mc, physics, strategy development, and support type, and also provide encouragement to others to continue to assess the variables examined here to improve on the limitations of this work and employ the recommendations presented.

9 References

Anderson, P. (2002). Assessment and Development of Executive Function (EF) During Childhood. *Child Neuropsychology, 8*(2), 71–82. http://doi.org/10.1076/chin.8.2.71.8724

Armstrong, R. A. (2014). When to use the Bonferroni correction. O*phthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists), 34*(5), 502–508. http://doi.org/10.1111/opo.12131

Bacon, A. M., Handley, S. J., Dennis, I., & Newstead, S. E. (2008). Reasoning strategies: The role of working memory and verbal-spatial ability. *European Journal of Cognitive Psychology, 20*(6), 1065–1086. http://doi.org/10.1080/09541440701807559

Baillargeon, R. (2008). Innate Ideas Revisited For a Principle of Persistence in Infants' Physical Reasoning. *Perspectives on Psychological Science, 3*(2), 2–13. http://doi.org/10.1111/j.1745-6916.2008.00056.x

Baker, S. T., Gjersoe, N. L., Sibielska-Woch, K., Leslie, A. M., & Hood, B. M. (2011). Inhibitory control interacts with core knowledge in toddlers' manual search for an occluded object. *Developmental Science, 14*(2), 270–279. http://doi.org/10.1111/j.1467-7687.2010.00972.x

Barbey, A. K., & Barsalou, L. W. (2010). Reasoning and Problem Solving: Models. *Encyclopedia of Neuroscience, 8*, 35–43. http://doi.org/10.1016/B978-008045046-9.00435-6

Barker, J. E., Semenov, A. D., Michaelson, L., Provan, L. S., Snyder, H. R., & Munakata, Y. (2014). Less-structured time in children's daily lives predicts self-directed executive functioning. *Frontiers in Psychology, 5*(JUN), 1–16. http://doi.org/10.3389/fpsyg.2014.00593

Beck, D. M., Schaefer, C., Pang, K., & Carlson, S. M. (2011). Executive Function in Preschool Children: Test–Retest Reliability. *Journal of Cognition and Development, 12*(February 2015), 169–193. http://doi.org/10.1080/15248372.2011.563485

Berhenke, A., Marulis, L. M., & Neidlinger, N. (2012). *Hot and Cold Executive Functioning, Motivation, and Metacognition: Disentangling Young Children's Approaches to Learning.* Paper presented at the annual meeting of the American Educational Research Association Vancouver, BC.

Best, J. R., & Miller, P. H. (2010). A developmental perspective on executive function. *Child*

Development, 81(6), 1641–1660. http://doi.org/10.1111/j.1467-8624.2010.01499.x

Botting, N., Jones, A., Marshall, C., Denmark, T., Atkinson, J., & Morgan, G. (2017). Nonverbal Executive Function is Mediated by Language: A Study of Deaf and Hearing Children. *Child Development, 88*(5), 1689–1700. http://doi.org/10.1111/cdev.12659

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: evidence from observational analysis of young children's behavior during problem-solving. M*etacognition and Learning, 7*(3), 197–217. http://doi.org/10.1007/s11409-012-9091-2

Bryce, D., Whitebread, D., & Szucs, D. (2015). The relationships among executive functions, metacognitive skills and educational achievement in 5 and 7 year-old children. *Metacognition and Learning, 10*(2), 181–198. http://doi.org/10.1007/s11409-014-9120-4

Brydges, C. R., Reid, C. L., Fox, A. M., & Anderson, M. (2012). A unitary executive function predicts intelligence in children. *Intelligence, 40*(5), 458–469. http://doi.org/10.1016/j.intell.2012.05.006

Case, R. (1985). *Intellectual development: Birth to adulthood.* New York, NY: Academic.

Chen, Z., & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development, 70*(5), 1098–1120. http://doi.org/10.1111/1467-8624.00081

Chetland, E., & Fluck, M. (2007). Children's performance on the "give x" task: a microgenetic analysis of "counting" and "grabbing" behaviour. *Infant and Child Development, 16*, 35–51. http://doi.org/10.1002/icd.499

Cheyne, J. A., & Rubin, K. H. (1983). Playful precursors of problem solving in preschoolers. *Developmental Psychology, 19*(4), 577–584. http://doi.org/10.1037/0012-1649.19.4.577

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*(3), 341–9. http://doi.org/10.1016/j.cognition.2011.03.003

Cornu, V., Schiltz, C., Martin, R., & Hornung, C. (2018). Visuo-spatial abilities are key for young children's verbal number skills. *Journal of Experimental Child Psychology, 166*, 604–620. http://doi.org/10.1016/j.jecp.2017.09.006

Cragg, L., & Nation, K. (2010). Language and the Development of Cognitive Control. *Topics in Cognitive Science, 2*(4), 631–642. http://doi.org/10.1111/j.1756-8765.2009.01080.x

Department for Education. (2017). *Statutory framework for the early years foundation stage. Setting the standards for learning, development and care for children from birth to five.* London, UK.

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

Dunbar, K. N., Fugelsang, J. A., & Stein, C. (2007). Do Naïve Theories Ever Go Away? Using Brain and Behavior to Understand Changes in Concepts. In M. C. Lovett & P. Shah (Eds.), Carnegie Mellon symposia on cognition. *Thinking with data* (pp. 193-205). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Dunn, L. M., Dunn, L. M., Whetton, C., & Burley, J. (1997). *British Picture Vocabulary Scale (2nd Edition).* Windsor: NFER-Nelson.

Eichstaedt, K. E., Kovatch, K., & Maroof, D. A. (2013). A less conservative method to adjust for familywise error rate in neuropsychological research: The Holm's sequential Bonferroni procedure. *NeuroRehabilitation, 32*(3), 693–696. http://doi.org/10.3233/NRE-130893

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G * Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. http://doi.org/10.3758/BRM.41.4.1149

Fisher, K. R., Hirsh-Pasek, K., Newcombe, N., & Golinkoff, R. M. (2013). Taking shape: Supporting preschoolers' acquisition of geometric knowledge through guided play. *Child Development, 84*(6), 1872–1878. http://doi.org/10.1111/cdev.12091

Flavell, J. H., Miller, P. H., & Miller, S. A. (2002). *Cognitive Development.* New Jersey: Upper Saddle River.

Garćia, T., Rodríguez, C., González-Castro, P., Álvarez-García, D., & González-Pienda, A. (2016). Metacognition and executive functioning in Elementary School. *Anales de Psicología, 32*(2), 474-483. http://dx.doi.org/10.6018/analesps.32.2.202891

Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: a review using an integrative framework. *Psychological Bulletin, 134*(1), 31–60. http://doi.org/10.1037/0033-2909.134.1.31

Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. http://doi.org/10.5812/ijem.3505

Gooch, D., Thompson, P., Nash, H. M., Snowling, M. J., & Hulme, C. (2016). The

development of executive function and language skills in the early school years. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 57*(2), 180–187. http://doi.org/10.1111/jcpp.12458

Gopnik, A. (2012). Scientific thinking in young children: theoretical advances, empirical research, and policy implications. *Science, 337*, 1623–1627. http://doi.org/10.1126/science.1223416

Gray, C.D. & Kinnear, P.R. (2012). *IBM SPSS statistics 19 made simple*. New York: Psychology Press.

Gropen, J., Clark-Chiarelli, N., Hoisington, C., & Ehrlich, S. B. (2011). The Importance of Executive Function in Early Science Education. *Child Development Perspectives, 5*(4), 298–304. http://doi.org/10.1111/j.1750-8606.2011.00201.x

Halford, G. S., Andrews, G., Dalton, C., Boag, C., & Zielinski, T. (2002). Young children's performance on the balance scale: the influence of relational complexity. *Journal of Experimental Child Psychology, 81*(4), 417–445. http://doi.org/10.1006/jecp.2002.2665

Hast, M., & Howe, C. (2013). Towards a Complete Commonsense Theory of Motion: The interaction of dimensions in children's predictions of natural object motion. *International Journal of Science Education, 35*(10), 1649–1662. http://doi.org/10.1080/09500693.2011.604685

Hood, B., Carey, S., & Prasada, S. (2000). Predicting the Outcomes of Physical Events : Two-Year-Olds Fail to Reveal Knowledge of Solidity and Support. *Child Development, 71*(6), 1540–1554.

Howe, C., Taylor Tavares, J., & Devine, A. (2012). Everyday conceptions of object fall: explicit and tacit understanding during middle childhood. *Journal of Experimental Child Psychology, 111*(3), 351–366. http://doi.org/10.1016/j.jecp.2011.09.003

Hughes, C., Ensor, R., Wilson, A., & Graham, A. (2010). Tracking executive function across the transition to school: A latent variable approach. *Developmental Neuropsychology, 35*(1), 20–36. http://doi.org/10.1080/87565640903325691

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York, Basic.

Jacques, S., & Zelazo, P. D. (2010). The Flexible Item Selection Task (FIST) : A Measure of Executive Function in Preschoolers. *Developmental Neuropsychology, 20*(3), 573–591. http://doi.org/ 10.1207/S15326942DN2003_2

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on

the balance scale task. *Journal of Experimental Child Psychology, 81*(4), 383–416. http://doi.org/10.1006/jecp.2002.2664

Jurado, M. B., & Rosselli, M. (2007). The elusive nature of executive functions: A review of our current understanding. N*europsychology Review, 17*(3), 213–233. http://doi.org/10.1007/s11065-007-9040-z

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition, 14*(4), 308–312. http://doi.org/10.3758/BF03202508

Karmiloff-Smith, A. (1992). *Beyond Modularity: A developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Karmiloff-Smith, A., & Inhelder, B. (1974-1975). If you want to get ahead, get a theory. *Cognition, 3*(3), 195-212. https://doi.org/10.1016/0010-0277(74)90008-0

Kim, I. K., & Spelke, E. S. (1999). Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science, 2*(3), 339–362. http://doi.org/10.1111/1467-7687.00080

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: effect of direct instruction and discovery learning. *Psychological Science, 15*(10), 661–7. http://doi.org/10.1111/j.0956-7976.2004.00737.x

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II: Clinical and interpretive manual.* San Antonio, TX: Harcourt Assessment.

Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristics: dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review, 8*(3), 439–53. http://doi.org/10.3758/BF03196179

Kuhn, D. (2000). Metacognitive Development. Current Directions in Psychological *Science, 9*(5), 178–181. http://doi.org/10.1111/1467-8721.00088

Latzman, R. D., Elkovitch, N., Young, J., & Clark, L. A. (2010). The contribution of executive functioning to academic achievement among male adolescents. *Journal of Clinical and Experimental Neuropsychology, 32*(5), 455–462. http://doi.org/10.1080/13803390903164363

Lee, V., & Kuhlmeier, V. a. (2013). Young children show a dissociation in looking and pointing behavior in falling events. *Cognitive Development, 28*, 21–30. https://doi.org/10.1016/j.cogdev.2012.06.001

Li, F., Xie, L., Yang, X., & Cao, B. (2017). The effect of feedback and operational

experience on children's rule learning. F*rontiers in Psychology, 8*(APR), 1–8. http://doi.org/10.3389/fpsyg.2017.00534

Marusteri, M., & Bacarea, V. (2009). Comparing groups for statistical differences: How to choose the right statistical test? *Biochemia Medica*, *20*(1), 15–32. http://doi.org/10.11613/BM.2010.004

Masson, S., Potvin, P., Riopel, M., and Foisy, L.M.B. (2014). Differences in Brain Activation Between Novices and Experts in Science During a Task Involving a Common Misconception in Electricity. *Mind, Brain, and Education, 8*(1), 37-48 http://doi.org/10.1111/mbe.12043

Mayr, S., Erdfelder, E., Buchner, A., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 51–59.

McClelland, M. M., Cameron, C. E., Duncan, R., Bowles, R. P., Acock, A. C., Miao, A., & Pratt, M. E. (2014). Predictors of early growth in academic achievement: The head-toes-knees-shoulders task. *Frontiers in Psychology, 5*(June), 1–14. http://doi.org/10.3389/fpsyg.2014.00599

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 636–49. http://doi.org/10.1037/0278-7393.9.4.636

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, *46*(1), 38–41. http://doi.org/10.1016/j.bjoms.2007.09.002

Messer, D. J., Pine, K. J., & Butler, C. (2008). Children's behaviour and cognitions across different balance tasks. *Learning and Instruction, 18*, 42–53. http://doi.org/10.1016/j.learninstruc.2006.09.008

Miyake, A., & Friedman, N. P. (2012). Executive Functions : Four General Conclusions. *Current Directions in Psychological Science, 21*(1), 8–14. http://doi.org/ https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, a H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology, 41*(1), 49–100. http://doi.org/10.1006/cogp.1999.0734

Monette, S., Bigras, M., & Lafrenière, M. A. (2015). Structure of executive functions in typically developing kindergarteners. *Journal of Experimental Child Psychology, 140*,

120–139. http://doi.org/10.1016/j.jecp.2015.07.005

Muentener, P., & Schulz, L. (2012). What Doesn't Go Without Saying: Communication, Induction, and Exploration. *Language Learning and Development, 8*(September), 61–85. http://doi.org/10.1080/15475441.2011.616455

Munakata, Y. (2001). Graded representations in behavioral dissociations. T*rends in Cognitive Sciences, 5*(7), 309–315. http://doi.org/10.1016/S1364-6613(00)01682-X

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

Paulus, M., Proust, J., & Sodian, B. (2013). Examining implicit metacognition in 3.5-year-old children: An eye-tracking and pupillometric study. *Frontiers in Psychology, 4*(145), 1–7. http://doi.org/10.3389/fpsyg.2013.00145

Peters, L., Davey, N., Messer, D., & Smith, P. (1999). An investigation into Karmiloff-Smith's RR model: The effects of structured tuition. *British Journal of Developmental Psychology, 17*(2), 277–292. http://doi.org/10.1348/026151099165276

Philips, S., & Tolmie, A. (2007). Children's performance on and understanding of the balance scale problem: the effects of parental support. *Infant and Child Development, 16*, 95–117. http://doi.org/10.1002/icd.504

Ponitz, C. C., McClelland, M. M., Matthews, J. S., & Morrison, F. J. (2009). A structured observation of behavioral self-regulation and its contribution to kindergarten outcomes. *Developmental Psychology, 45*(3), 605–19. http://doi.org/10.1037/a0015365

Rasch, D., Kubinger, Klaus, D., & Yanagida, T. (2011). *Statistics in Psychology Using R and SPSS*. United Kingdom: John Wiley & Sons, Ltd.

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. http://doi.org/10.1111/bjdp.12129

Roberts, M. J., & Erdos, G. (1993). Strategy Selection and Metacognition. *Educational Psychology, 13*(3–4), 259–266. http://doi.org/10.1080/0144341930130304

Roberts, M. J., Taylor, R. J., & Newton, E. J. (2007). Explaining inappropriate strategy selection in a simple reasoning task. *British Journal of Psychology, 98*(4), 627–644. http://doi.org/10.1348/000712607X173763

Robson, S. (2016). Are there differences between children's display of self-regulation and

metacognition when engaged in an activity and when later reflecting on it? The complementary roles of observation and reflective dialogue. *Early Years, 36*(2), 179–194. http://doi.org/10.1080/09575146.2015.1129315

Roderer, T., & Roebers, C. M. (2014). Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. *Metacognition and Learning, 9*(1), 1–23. http://doi.org/10.1007/s11409-013-9109-4

Roebers, C. M. (2017). Executive function and metacognition : Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Roebers, C. M., Cimeli, P., Röthlisberger, M., & Neuenschwander, R. (2012). Executive functioning, metacognition, and self-perceived competence in elementary school children: An explorative study on their interrelations and their role for school achievement. *Metacognition and Learning, 7*(3), 151–173. http://doi.org/10.1007/s11409-012-9089-9

Roebers, C. M., & Feurer, E. (2016). Linking Executive Functions and Procedural Metacognition. *Child Development Perspectives, 10*(1), 39–44. http://doi.org/10.1111/cdep.12159

Rothman, K. J. (1990). No Adjustments Are Needed for Multiple Comparisons. *Epidemiology, 1*(1), 43–46. http://doi.org/10.1097/00001648-199001000-00010

Rozencwajg, P. (2003). Metacognitive factors in scientific problem-solving strategies. *European Journal of Psychology of Education, 18*(3), 281–294. http://doi.org/10.1007/BF03173249

Sani, F. & Todman, J. (2006). *Experimental Design and Statistics for Psychology*. Oxford: Blackwell Publishing.

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition, 110*(3), 395–411. http://doi.org/10.1016/j.cognition.2008.11.017

Schrauf, C., Call, J., & Pauen, S. (2011). The Effect of Plausible Versus Implausible Balance Scale Feedback on the Expectancies of 3- to 4-Year-Old Children. *Journal of Cognition and Development, 12*(4), 518–536. http://doi.org/10.1080/15248372.2011.571647

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking.* Oxford: University Press.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 8*(4), 481–520. http://doi.org/10.1016/0010-0285(76)90016-5

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: a microgenetic analysis. *Cognitive Psychology, 36*(3), 273–310. http://doi.org/10.1006/cogp.1998.0686

Siegler, R. S., & Stern, E. (1998). Conscious and unconscious strategy discoveries: a microgenetic analysis. *Journal of Experimental Psychology: General, 127*(4), 377–397. http://doi.org/10.1037/0096-3445.127.4.377

Siegler, R. S., & Svetina, M. (2002). A Microgenetic/Cross-Sectional Study of Matrix Completion: Comparing Short-Term and Long-Term Change. *Child Development, 73*(3), 793–809. https://doi.org/10.1111/1467-8624.00439

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of Knowledge. *Psychological Review, 99*(4), 605–632. http://doi.org/10.1037//0033-295X.99.4.605

Spiess, M. A., Meier, B., & Roebers, C. M. (2016). Development and longitudinal relationships between children's executive functions, prospective memory, and metacognition. *Cognitive Development, 38*, 99–113. http://doi.org/10.1016/j.cogdev.2016.02.003

Streiner, D. L., & Norman, G. R. (2011). Correction for multiple testing: Is there a resolution? *Chest, 140*(1), 16–18. http://doi.org/10.1378/chest.11-0523

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the *P* Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. http://doi.org/10.4300/JGME-D-12-00156.1

Tolmie, A. K., Ghazali, Z., & Morris, S. (2016). Children's science learning: A core skills approach. *British Journal of Educational Psychology, 86*(3), 481–497. http://doi.org/10.1111/bjep.12119

van der Graaf, J., Segers, E., & Verhoeven, L. (2015). Scientific reasoning abilities in kindergarten: dynamic assessment of the control of variables strategy. *Instructional Science, 43*(3), 381–400. http://doi.org/10.1007/s11251-015-9344-y

van der Graaf, J., Segers, E., & Verhoeven, L. (2016). Discovering the laws of physics with a serious game in kindergarten. *Computers and Education, 101,* 168–178. https://doi.org/10.1016/j.compedu.2016.06.006

van der Sluis, S., de Jong, P. F., & van der Leij, A. (2007). Executive functioning in children, and its relations with reasoning, reading, and arithmetic. *Intelligence, 35*(5), 427–449. http://doi.org/10.1016/j.intell.2006.09.001

Vandenbroucke, L., Verschueren, K., & Baeyens, D. (2017). The development of executive functioning across the transition to first grade and its predictive value for academic achievement. *Learning and Instruction, 49,* 103–112. http://doi.org/10.1016/j.learninstruc.2016.12.008

Veenman, M. V. J., Van Hout-Wolters, B. H. M., & Afflerbach, P. (2006). Metacognition and learning: *Conceptual and methodological considerations. Metacognition and Learning, 1*(1), 3–14. http://doi.org/10.1007/s11409-006-6893-0

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology, 24*(4), 535–585. http://doi.org/10.1016/0010-0285(92)90018-W

Wagensveld, B., Segers, E., Kleemans, T., & Verhoeven, L. (2015). Child predictors of learning to control variables via instruction or self-discovery. *Instructional Science, 43*, 365–379. http://doi.org/10.1007/s11251-014-9334-5

Watson, P. (2016, September 21). *How do I produce nonparametric Spearman partial correlations using SPSS?* Retrieved from http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/partsp

Watson, P. (2018, December 4). *Rules of thumb on magnitudes of effect sizes.* Retrieved from http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/effectSize

Weiland, C., Barata, M. C., & Yoshikawa, H. (2014). The Co-Occurring Development of Executive Function Skills and Receptive Vocabulary in Preschool- Aged Children: A Look at the Direction of the Developmental Pathways. *Infant and Child Development, 23*, 4–21. http://doi.org/10.1002/icd

Weisberg, D. S., Hirsh-Pasek, K., & Golinkoff, R. M. (2013). Embracing complexity: rethinking the relation between play and learning: comment on Lillard et al. (2013). *Psychological Bulletin, 139*(1), 35–39. http://doi.org/10.1037/a0030077

Weisberg, D. S., Kittredge, A. K., Hirsh-Pasek, K., Golinkoff, R. M., & Klahr, D. (2015). Making play work for education play. *Phi Delta Kappan, 96*(8), 8–13. http://doi.org/10.1177/0031721715583955

Whitebread, D., Coltman, P., Jameson, H., & Lander, R. (2009a). Play, Cognition and Self-Regulation: What exactly are children learning when they learn through play? *Educational and Child Psychology, 26*(2), 40–52.

Whitebread, D., Coltman, P., Pasternak, D. P., Sangster, C., Grau, V., Bingham, S., Almeqdad, Q., & Demetriou, D. (2009b). The development of two observational tools

for assessing metacognition and self-regulated learning in young children. *Metacognition and Learning, 4*(1), 63–85. http://doi.org/10.1007/s11409-008-9033-1

Wiebe, S. A., Sheffield, T., Nelson, J. M., Clark, C. a C., Chevalier, N., & Espy, K. A. (2011). The structure of executive function in 3-year-olds. J*ournal of Experimental Child Psychology, 108*(3), 436–452. http://doi.org/10.1016/j.jecp.2010.08.008

Willoughby, M., & Blair, C. (2011). Test-retest reliability of a new executive function battery for use in early childhood. *Child Neuropsychology, 17(*6), 564–579. http://doi.org/10.1080/09297049.2011.554390

Zohar, A., & Barzilai, S. (2013). A review of research on metacognition in science education: current and future directions. *Studies in Science Education, 49*(2), 121–169. http://doi.org/10.1080/03057267.2013.847261

10 Appendices

10.1 Appendix A

C.Ind.Le codes selected for assessing Mc

Mc codes taken from Whitebread et al.'s (2009b) C.Ind.Le.

MC component and code	Overview of component	Examples
MK: knowledge of persons	A verbalization demonstrating the explicit expression of one's knowledge in relation to cognition or people as cognitive processors.	Refers to his/her own strengths or difficulties in learning and academic working skills Refers to his/her own strengths or difficulties in learning and academic working skills. Talks about general ideas about learning.
MK: knowledge of tasks	A verbalization demonstrating the explicit expression of one's own long-term memory knowledge in relation to elements of the task.	Compares across tasks identifying similarities and differences. Makes a judgment about the level of difficulty of cognitive tasks or rates the tasks on the basis of pre-established criteria or previous knowledge.

MC component and code	Overview of component	Examples
MK: knowledge of strategies	A verbalization demonstrating the explicit expression of one's own knowledge in relation to strategies used or performing a cognitive task, where a strategy is a cognitive or behavioural activity that is employed so as to enhance performance or achieve a goal.	Defines, explains or teaches others how she/he has done or learned something. Explains procedures involved in a particular task. Evaluates the effectiveness of one or more strategies in relation to the context or the cognitive task.
MS: planning	Any verbalization or behaviour related to the selection of procedures necessary for performing the task, individually or with others	Sets or clarifies task demands and expectations. Sets goals and targets. Decides on ways of proceeding with the task. Seeks and collects necessary resources.
MS: monitoring	Any verbalization or behaviour related to the ongoing on-task assessment of the quality of task performance (of self or others) and the degree to which performance is progressing towards a desired goal	Self- commentates. Reviews progress on task (keeping track of procedures currently being undertaken and those that have been done so far). Rates effort on-task or rates actual performance. Rates or makes comments on currently memory retrieval. Checks behaviours or performance, including detection of errors. Self-corrects.

MC component and code	Overview of component	Examples
MS: control	Any verbalization or behaviour related to a change in the way a task had been conducted (by self or others), as a result of cognitive monitoring	Changes strategies as a result of previous monitoring. Suggests and uses strategies in order to solve the task more effectively. Applies a previously learnt strategy to a new situation. Repeats a strategy in order to check the accuracy of the outcome. Seeks help. Uses nonverbal gesture as a strategy to support own cognitive activity. Copies from or imitates a model.
MS: evaluation	Any verbalization or behaviour related to reviewing task performance and evaluating the quality of performance (by self or others).	Reviews own learning or explains the task. Evaluates the strategies used. Rates the quality of performance. Observes or comments on task progress. Tests the outcome or effectiveness of a strategy in achieving a goal.

10.2 Appendix B

Examples of coding used during the physics tasks

Examples of Mc coded in the current study.

Mc component	Example in current study's coding
MK: knowledge of persons	N/A
MK: knowledge of tasks	Child 600: ""and if you put a light one over here it tips over". Child 602: "that one's too heavy it tips over".
MK: knowledge of strategies	Child 625: "because they're both in the same seats". Child 636: "if you put this one over there and this one will be the same seat".
MR: planning	Changing where they decided to put the weights was a frequent occurrence. Commenting where the weights were going to go was also a frequent occurrence.
MR: monitoring	Looking back and forth between the weights either on the beam/in hand/on table and the pictures was often seen. Commenting on progress was also seen, child 617: "that's not right".

Mc component	Example in current study's coding
MR: control	Changing strategy and repeating strategies between trials was often seen. After seeing the beam tipping some children tried to fix their errors.
MR: evaluation	Commenting on the outcome was often seen. Child 636: "it looks like that one" comparing it to the picture on the table. Child 629: "oh, it's not working".

10.3 Appendix C

Balance beam support type instruction used in the main study

Instruction for the GP and DI support can be seen below.

Guided play:

"Now we are going to look at a science game. Scientists test things to see why or how something happens. We will be scientists, so let's put our science coats on before we start. This game involves a seesaw- have you ever been on a seesaw? …

This is the seesaw (show). The best people to play together make it balance like this (show + pictures), so they can then bounce the seesaw up and down instead of it just tipping over. So in this game you are going to test the dinosaurs by weighing them on the seesaw to see which dinosaurs make it balance, so you know who can seesaw and play together.

…couple of things to tell you before we start…

Balance means it is straight like this (show + pictures) and not tipping over like this (show + pictures) or this (show + pictures), so this is it balancing (show + pictures). Weight is like how easy or hard something is to pick up – light things are easy to pick up and heavy things are more tricky to pick up.

Remember, to be able to play they must be able to balance the seesaw – like these two here (show + pictures). If the two dinosaurs can't balance the seesaw then they can't play on it as it will tip over – like these two here (show + pictures) – and they can't make it seesaw.

So in this game you need to weigh the different dinosaurs to see which balance together. There are two kinds of dinosaurs – light ones, which are blue, and heavy ones, which are purple (let child hold + weigh).

These are the seats – there are ones near to the middle (show) and far from the middle (show) and the dinosaurs go on the seats like this (show).

Can you have a go at weighing the dinosaurs on the seesaw to see who can make it balance? I will let you do that first and then I will ask you to show me if some of them can play together."

Direct instruction:

"Now we are going to look at a science game. Scientists test things to see why or how something happens. I will be the scientist, so let me put my science coat on before we start. This game involves a seesaw- have you ever been on a seesaw? …

This is the seesaw (show). The best people to play together make it balance like this (show + pictures), so they can then bounce the seesaw up and down instead of it just tipping over. So in this game I am going to test the dinosaurs by weighing them on the seesaw to see which dinosaurs make it balance, so I know who can seesaw and play together.

…couple of things to tell you before we start…

Balance means it is straight like this (show + pictures) and not tipping over like this (show + pictures) or this (show + pictures), so this is it balancing (show + pictures). Weight is like how easy or hard something is to pick up – light things are easy to pick up and heavy things are more tricky to pick up.

Remember, to be able to play they must be able to balance the seesaw – like these two here (show + pictures). If the two dinosaurs can't balance the seesaw then they can't play on it as it will tip over – like these two here (show + pictures) – and they can't make it seesaw.

So in this game I need to weigh the different dinosaurs to see which balance together. There are two kinds of dinosaurs – light ones, which are blue, and heavy ones, which are purple (let child hold + weigh). Two of these light dinosaurs are the same as one heavy dinosaur. So I need to think about how heavy the dinosaurs are and which seats they sit in.

These are the seats – there are ones near to the middle (show) and far from the middle (show) and the dinosaurs go on the seats like this (show). I need to think really hard about which seats the dinosaurs sit in, because the ones far from the middle (show) are more likely to tip it over, but the heavy dinosaurs will also try to make it tip over.

I will have a go at weighing the dinosaurs on the seesaw to see who can make it balance. I will do that first and then I will ask you to show me if some of them can play together.

I will use two light dinosaurs to see what happens … they weigh the same
In the same seats on each side (show) they balance – because they weigh the same and are in the same seats
In different seats (show) they don't balance – they weigh the same, but this (show) dinosaur far from the middle makes it tip over … dinosaurs who sit far from the middle are more likely to make it tip over
So when I use dinosaurs who weight the same, they have to be in the same seats on each side to make it balance so they can play together

But now I'll see what happens when I have two dinosaurs who are different weights – so now I'll use a heavy dinosaur and a light dinosaur

In the same seats (show) they don't balance – because they don't weigh the same and the heavy dinosaur makes it tip over (show)
(heavy far, light near, show) In different seats they don't balance – the heavy dinosaur in the far seat makes it tip over – remember, dinosaurs in this seat are more likely to make it tip over (heavy near, light far, show) But when I change their seats – they do balance because the heavy dinosaur in the near seat is trying to tip it over, but the light dinosaur in the far seat is also trying to tip it over, so it balances, like this (show)

Now I have two light dinosaurs and one heavy dinosaur. Two light dinosaurs weigh the same as one heavy dinosaur so they need to sit in the same seats on each side before they can balance and can play together (show)

So what have I learnt? I learnt that to make the dinosaurs balance, the dinosaurs should weigh the same and be in the same seats on each side (point), OR you have to put heavier dinosaurs near the middle and lighter dinosaurs far from the middle (point), OR if you have two light dinosaurs and one heavy dinosaur they need to sit in the same seats on each side. So, I need to think about how much the dinosaur weighs and the seats they sit in to make it balance."

10.4 Appendix D

Comparison of length of instruction in the two groups

The time of the GP and DI support was meant to be equal in order to make sure the length of time introducing the task before the trials did not influence performance. In GP, the time allowed for play is included in the instruction length, while for DI the time taken for the adult to explain and show the problems was included. The means (minutes:seconds) and SDs (in parentheses) for each TP can be seen in the table below.

Comparison of mean and SDs (in parenthesis) instruction length for each group at each TP (minutes:seconds)

	TP1	TP2	TP3
GP	6:56 (00:20)	5:25 (00:09)	5:17 (00:11)
DI	5:54 (00:11)	5:01 (00:09)	4:54 (00:06)

The assumptions of the data were checked: boxplots revealed no extreme outliers. The normality of the data were checked using histograms and p-plots, which indicated no problems. The Shapiro-Wilk tests showed the data on the length of the instruction to be normally distributed at TPs 1 (p = .07), 2 (p = .27), and 3 (p = .42). Levene's tests of homogeneity of variances were not significant at TP2 (p = .94) or 3 (p = .12), but it was at TP1 (p = .01), therefore variances are not assumed and the non-parametric Mann-Whitney independent samples t-tests were used, entering group as the grouping variable and time as the test variable.

This showed a significant difference between the groups due to GP receiving a longer instruction time than DI (U = 89.50, p = .02, d = 0.89). This was not seen at TPs 2 (U = 80.00, p = .12 d = 0.70) or 3 (U = 64.00, p = .08 d = 0.66). The difference in length of time at TP1 is likely due to refining the GP procedure over time. Despite piloting, there was still some learning over the first TP and as indicated by the means and SDs at TPs 2 and 3, the instruction shortened over time. These data are examined later to check if there is a link to performance.

10.5 Appendix E

Comparison of how much information was provided by the adult

To ensure the amount of information provided to each group matched and differed accordingly (therefore making the groups qualitatively different), how many pieces of information each child received during the instruction was calculated. The groups were to receive the same amount of information during the "generic" instruction at the start, but DI was to then receive the "DI" information on how to solve the different trials depending on the weights provided – this information in the forms of means and SDs (in parentheses) can be seen in the table below. Children were to receive the same information in the same format, however some deviances occurred due to individual children. Examples include children sharing what they remembered (thus reducing how much information the adult said) and children being unable to focused (so the adult at times had to repeat information).

Means and SDs for how much information was provided to each group at each TP

	TP1		TP2		TP3	
	GP	DI	GP	DI	GP	DI
"Generic" information	13.59 (1.42)	13.16 (1.46)	12.19 (1.83)	12.33 (1.35)	13.46 (0.78)	12.50 (1.21)
"DI" information	N/A	9.79 (1.40)	N/A	9.67 (1.11)	N/A	9.13 (0.62)

Examples of a "piece" of information include: showing and explaining the picture of the characters balancing the seesaw (given to both support types), weighing the dinosaurs in our hands to feel the difference (given to both support types), explaining and showing that when dinosaurs weigh the same they need to sit in the same seats (given to DI), and showing two light dinosaurs weigh the same as one heavy dinosaur (given to DI). If the information was repeated in the session it was recorded as twice (or the number of times given).

The table shows that the two groups received a similar amount of "generic" information at each TP and DI received additional information on solving the trials. The assumptions of the

data were checked separately for each group. The normality of the data were checked for each group separately. Boxplots revealed three extreme outliers for the DI data in the "generic" information they received. The average was 9.13, but these three received 8, 10, and 11 pieces of information. They were kept in the sample due to the low participant number and to ensure the data reflected the true procedure.

It should be noted that the Shapiro-Wilk tests were significant for most of the DI checks, suggesting non-normally distributed data. Mann-Whitney independent *t*-tests were thus carried out in order to compare the two groups, entering number of "generic" pieces of information as the test variable and group as the grouping variable. No significant difference was detected between the groups for the amount of "generic" information received at TPs 1 ($U = 132.50$, $p = .36$, $d = 0.30$) or 2 ($U = 114.00$, $p = .83$, $d = 0.09$), but there was at TP3 ($U = 53.00$, $p = .03$, $d = 1.12$) due to GP receiving more pieces of information. The "DI" instruction provided to DI was always more due to GP receiving zero pieces of information.

For reference, had the data been calculated as "unique" pieces of information (i.e., excluding the times information was repeated) the same results would have been reported.

10.6 Appendix F

Comparison of how much information each group had before starting the trials

The Appendix E analyses were carried out to compare the amount of information provided to each group by the adult, but how many pieces of information the GP children discovered during the play time is also important. Examples of discoveries include: successfully balancing two dinosaurs of the same weight (therefore seeing that if they weigh the same they must sit in the same seats) and finding that two light dinosaurs weigh the same one heavy dinosaur. The comparison was to see whether the groups differed in how much information they had before starting the trials.

It is expected that even when the number of discoveries the GP children made is added to the number of pieces of information given that the two groups will still be different overall. DI is the same as in Appendix E, as nothing changes for that group, but displayed again for ease of reading – see below.

Means (SDs) for how much information was discovered and provided to GP and how much provided to DI at each TP

	TP1		TP2		TP3	
	GP	DI	GP	DI	GP	DI
"Generic" information provided	13.59 (1.42)	13.16 (1.46)	12.19 (1.83)	12.33 (1.35)	13.46 (0.78)	12.50 (1.21)
"DI" information provided	N/A	9.79 (1.40)	N/A	9.67 (1.11)	N/A	9.13 (0.62)
GP information "discovered"	4.09 (2.22)	N/A	3.88 (1.67)	N/A	4.31 (1.89)	N/A
Total information before trials	17.65 (2.12)	22.95 (2.20)	16.06 (3.04)	22.00 (1.73)	17.77 (2.24)	21.63 (1.41)

As can be seen in, even with the addition of the play discoveries, GP did not have as much information as DI before starting the trials. The number of pieces of information "discovered" during the GP play ranged from 0 – 8 at TP1, 1 – 8 at TP2, and 1 – 7 at TP3.

The normality of the data were checked for each group separately and no issues found. Boxplots revealed no extreme outliers. Histograms and p-plots indicated the data to be normally distributed at each TP for each group. Shapiro-Wilk tests indicated the variance was normal for GP at TP1 ($p = .20$), 2 ($p = .25$), and 3 ($p = .53$), and for DI at TPs 1 ($p = .17$), 2 ($p = .28$), and 3 ($p = .14$). Levene's test of homogeneity of variance was not significant at TP1 ($p = .51$) or 2 ($p = .12$), but it was at TP3 ($p = .03$), therefore non-parametric Mann-Whitney independent *t*-tests were used, entering the total number of pieces of information as the test variable and group as the grouping variable. The tests showed there to be a significant difference between the total information each group obtained at TP1 ($U = 11.50$, $p < .01$, $d = 2.45$), TP2 ($U = 15.50$, $p < .01$, $d = 2.40$) and TP3 ($U = 13.50$, $p < .01$, $d = 2.06$). The difference at each TP was due to DI having more information. The results show that there is a difference between how much information each group obtained during the introduction, even when this included GP's discoveries made during the play times.

10.7 Appendix G

Feedback after the balance beam trials

Besides the instruction, the feedback after each trial was the other central difference between the two support types. The GP children were asked if their strategy worked/if it seesawed and why/why not. Depending on the response given (if there was one) there were sometimes other questions. In DI children were told whether it worked/seesawed and why it did/did not work. Essentially, the difference was whether they were asked questions (GP) or told something (DI).

In order to ensure the feedback given to the two groups differed (therefore making the conditions different) a comparison was carried out. Examples of feedback in GP include the adult asking: "is that one balancing?" (or "seesawing") and "why does that one work?" (or "not work"). Examples in DI include the adult saying: "it balances because they weigh the same and they're sitting in the same seats" and "you always need to have dinosaurs on each side of the seesaw to make it balance". If the feedback was repeated during any of the trials' feedback it was recorded as happening more than once. Scores are calculated from how many pieces of feedback they received divided by how many trials the child completed (since the more trials they completed the more feedback they received). Therefore the score is the average number of pieces of information per trial – see the table below. Scores differ between children based on individual differences in children, such as how responsive the child was to the feedback, if they did not answer a question, or if they answered incorrectly.

Means and SDs for average number of pieces of information per trial provided to GP and DI at each TP

	TP1		TP2		TP3	
	GP	DI	GP	DI	GP	DI
Feedback provided	1.90 (.51)	2.54 (.46)	2.37 (.38)	2.73 (.48)	2.48 (.61)	2.56 (.46)

The table shows DI received more feedback than GP at each TP. The feedback given to each group was expected to be different, therefore the normality of the data were checked for each

group separately, but no issues were seen. Boxplots revealed no extreme outliers. Histograms and p-plots indicated the data to be normally distributed at each TP for each group. Shapiro-Wilk tests showed the data to be normally distributed for GP at TPs 1 (p = .37), 2 (p = .71), and 3 (p = .95), and for DI at TPs 1 (p = .11), 2 (p = .44), and 3 (p = .41). Levene's test of homogeneity of variance was not significant at TP1 (p = .58) or 2 (p = .20), but was at 3 (p = .02). Therefore, Mann-Whitney independent *t*-tests were carried out entering the number of pieces of feedback as the test variable and group as the grouping variable. The tests showed there to be a significant difference between how much feedback was given to each group at TP1 (U = 57.50, p < .01, d = 1.30) and TP2 (U = 69.00, p = .04, d = .83), but not TP3 (U = 86.00, p = .45, d = .31). The significant differences are due to the DI receiving more feedback than GP.

10.8 Appendix H

Ramps task instructions and trials

The instructions and ramps used during the raps trials can be seen here.

"We are going to play a game with these ramps and balls now. Before we start I want to show you some things, then you'll get to have a go, then we'll have a go together.

There are two heights, two bits they can sit on – high and low (point). To set up the ramps they need to sit on one of them, like this (show). There are two surfaces carpet and wood, can you feel they are different? (Make child feel the two surfaces.)

On the other side of the ramps is the other…show
Can you show me how to put this one here (point) and this one here (point).

I want you to have a go at rolling the balls down the different ramps for a few minutes to see what happens to how far the balls goes when you change the, then we'll have some goes together. So you can move the ramps up and down, and turn them over.

LET THEM PLAY FOR 3-4 MINUTES

In this game there is a nice caterpillar and a bad alien (show). Do you know the story of the hungry caterpillar? Caterpillars start out as little eggs and then they eat some leaves and turn into a small caterpillar and then they eat some more leaves and turn into a big caterpillar and then they eat even more leaves and make themselves a cocoon and turn into a butterfly (read along with the picture). In this game you need to help Caterpillar by making sure he gets closer to the leaves (point out the leaves now sitting at the end of the ramp) than Bad Alien. If Caterpillar gets closer to the leaves he can eat them and turn into a butterfly, but if Alien gets closer to the leaves then he will take them all away from Caterpillar because he's a bad alien.

So each time you need to set up the ramps and make sure Caterpillar goes down the best ramp so he ends closer to the leaves than Bad Alien does.

Closer means nearest to the leaves (show – here he is near/close vs. far).

I will tell you some things you need to do each time and you need to set the ramps up so Caterpillar makes it closer to the leaves than Alien.

Have you got that? So what do you need to do?"

“Use XXX and have a go at setting up the ramps so that Caterpillar ends closer to the leaves than Alien. Tell me when you’ve found out how to do it.”

Give 2 minutes. Give encouragement if not doing anything.
Each time ask: “Did that work?” If wrong, ask: “How can you fix it so Caterpillar ends closer to the leaves?”

At 45 sec and 90 sec: remind them of variables(s) and what the aim is.

The trials are seen on the next page

Trial	Variables	Condition
Practice	**SET: TWO LOW** ***"Using one carpet and one wood ramp…"***	Surface
1	SET: ONE HIGH ONE LOW "Using two wood surfaces…"	Incline
2	SET: TWO HIGH "Using one carpet and one wood surface…"	Surface
3	SET: ONE HIGH ONE LOW "Using two carpet surfaces…"	Incline
4	SET: TWO LOW "Using one carpet and one wood surface…"	Surface
Continue if 3 or 4 correct		
5	ONE HIGH ONE LOW "Using two carpet surfaces…"	Incline
6	TWO HIGH "Using one carpet and one wood surface…"	Surface
7	ONE HIGH ONE LOW "Using two wood surfaces…"	Incline
8	TWO LOW "Using one carpet and one wood surface…"	Surface
Continue if 3 or 4 correct		
9	"Using one high and one low ramp AND one carpet and one wood ramp…"	Both
10	"Using one high and one low ramp AND one carpet and one wood ramp…"	Both

10.9 Appendix I

Comparison of the ramps data between the two support groups

The table below displays the means and SDs (in parentheses) for GP and DI on the ramps task. The pieces of information "given about the ramps" before starting was provided before the children were allowed to use the ramps for three to four minutes (these data are displayed as minutes and seconds). The "information about the game" was provided after this and outlined the aim of the game/task. The total feedback given and as a rate per try are also displayed.

Mann-Whitney *t*-tests were carried out to compare the two groups and no significant differences were detected (see the table below), which suggests there were no differences between the groups on any of the listed variables.

Test variable	GP	DI	Mann-Whitney *t*-tests
Time given to use ramps (minutes:seconds)	03:29 (00:39)	03:24 (00:31)	$U = 39.00$, $p = .97$, $d = .14$
Information given about ramps before starting	11.13 (0.99)	11.10 (1.20)	$U = 39.50$, $p = .97$, $d = .03$
Information given about game	9.50 (1.60)	9.40 (1.43)	$U = 38.00$, $p = .90$, $d = .07$
Total information given	20.63 (2.07)	20.50 (1.78)	$U = 39.50$, $p = .97$, $d = .07$
Feedback given (rate per try)	1.72 (0.21)	1.52 (0.46)	$U = 30.00$, $p = .41$, $d = .56$
Total feedback given	25.88 (6.40)	27.30 (9.08)	$U = 40.00$, $p = 1.00$, $d = .18$

Notes. GP $n = 8$. DI $n = 10$.

10.10 Appendix J

Mc interview questions used after the balance beam task

The four Mc interview questions used after the balance beam task are listed below:

Q1. Zebra thought that the dinosaur seesaw game was easy but Penguin thought that was hard…who are you like? Did you think it was easy like Zebra or hard like Penguin?

Q2. Can you tell zebra what you thought was easy about the game?

Q3. Can you tell penguin what you thought was hard about the game?

Q4. Is there anything that would have made that game easier for you?

The seven Mc interview questions used after the ramps task are listed below:

Q1. Dragon thought that the Caterpillar game was easy and Puppy thought the Caterpillar game was tricky…who are you like? Did you think it was easy or tricky?

Q2. Can you tell Dragon what was easy about the game?

Q3. Can you tell Puppy what was hard about the game?

Q4. Is there anything that would have made that game easier for you?

Q5. What did you learn in the game?

Q6. What was the best surface for Caterpillar? The wood surface or the carpet surface? Why?

Q7. What was the best ramp for Caterpillar? The high ramp or the low ramp? Why?

10.11 Appendix K

Information sheet and consent form used in the main study

The information sheet and consent form used in the main study are below.

The role of executive function, metacognition and support type in children's ability to solve physics tasks

I would like to invite you and your child to take part in a research study. Please read this information sheet before deciding whether to take part. Please get in touch if anything is unclear or if you have any questions about what is involved.

What is the study about?

I am looking at the role of executive function, metacognition and support type in how children develop problem-solving skills.

Who can take part?

Children who are aged 3- or 4-years-old by April 2016 are eligible to take part in this project, but children may be selected based on their date of birth. Unfortunately, only children who have English as their first and main language can take part, due to some of the activities they will be completing.

What will happen if I give permission for my children to take part?

I would visit your child in nursery 8 times between January and July, and each time the session would last around 15-20 minutes. This will allow your child to get to know me over the time we work together and it will allow me to follow their progress over a longer period of time instead of a one-off observation. Over these sessions your child will complete some hands-on puzzles, as well card games, memory games, spoken games and other puzzles.

I will videotape your child completing these activities so I can make detailed observations about their behaviour during the puzzles and games. This allows me to draw a more complete picture of the learning process.

Your child will receive a small gift for taking part.

Will the results be kept confidential?

Every child is allocated a unique code that will be used on all of their data, so they cannot be identified. Only myself and my supervisors will have access to the database which links the child's name to the code, but no data from the tests will be saved in this database. The databases and video-recordings will be stored securely and only myself and my supervisors will have access. Authorised colleagues may also have access to the data, such as for reliability checking, but they will be kept securely and for use in this research project only.

What happens if there is a problem?

If you or your child have any questions or concerns about the study please let me know. If you feel you cannot approach me then please contact my supervisor, Dr Sara Baker on stb32@cam.ac.uk.

What to do next

If you would like for your child to take part please complete and sign the consent form accompanying this information sheet.

How to contact me

If you would like to discuss any aspect of the research, or if you have questions afterwards, please email me on eg453@cam.ac.uk. If you would like to discuss this via telephone please give me a contact number to reach you on and any days/times when I can reach you.

This project has undergone ethical approval by the Faculty of Education, University of Cambridge.

Elaine Gray
1st year PhD student in the Faculty of Education, Cambridge University
Funded by the LEGO Foundation and the faculty of Education, 2014-2017
Supervisors: Dr Sara Baker and Prof Christine Howe

The role of executive function, metacognition and support type in children's ability to solve physics tasks

Please initial in box

1. I confirm I have read the accompanying information sheet and have had the opportunity to ask questions about the project. ☐

2. I understand that having my child take part is voluntary and that we can withdraw from the study at any time without giving any reason. ☐

3. I agree for my child to be videotaped during the sessions. ☐

3. I understand that the data collected during the study may be seen by authorised individuals from Cambridge University. ☐

4. I agree for my child to take part in the above study. ☐

_____	_____	_____
Name of child	Child's Date of birth	girl/ boy

_____	_____	_____
Name of parent/guardian	Today's date	Signature of parent/guardian

ELAINE GRAY	_____	_____
Name of person taking consent	Date	Signature of person taking consent

Please complete:

Child's first language:

Other languages your child can speak/understand:

Any known medical conditions:

Any known special educational needs:

Days child attends nursery (please tick):

	AM	PM
Monday		
Tuesday		
Wednesday		
Thursday		
Friday		

10.12 Appendix L

Data screening for age, BPVS, and NEPSY data

Boxplots identified no extreme outliers. Histograms appeared largely normal, except for NEPSY scores, due to a slight positive skew. The Shapiro-Wilk test of normality was not significant for age (W = .97, p = .43) or BPVS scores (W = .97, p = .45), but it was for NEPSY scores (W = .90, p < .01). Skewness and kurtosis values were converted to Z scores and the data were all found to be normally distributed, as seen within the +/-1.96 range (Field, 2013). Therefore the NEPSY data were not transformed and the remaining analyses will be carried out using the original data. Scatterplots showed positive linear relationships between the variables.

10.13 Appendix M

Data screening for the individual and composite EF measures

Boxplots were created to examine each EF measure and the composite and no extreme outliers were seen. The data were checked for normality and some assumptions were violated. Histograms showed some of the individual EF measures (not the composite scores) were skewed. The Shapiro-Wilk test of normality showed the data not to be normally distributed for all of the individual EF variables at each TP (all nine tests $p < .05$). EF composite scores at TPs 1 and 2 did not violate the assumption of normality, but they did at TP3 ($W = .93$, $p = .03$), likely because the scores were positively skewed (i.e., performance was higher at TP3). Skewness and kurtosis Z scores indicated the data were within a normal distribution, except for inhibition at TP1, thus were not transformed. Scatterplots of the different EF variables showed monotonic relationships of varying strengths, however all the fit lines showed positive linear relationships.

10.14 Appendix N

Data screening for the EF and background measures

Boxplots were created to check for extreme outliers in the data, but none were identified. Scatterplots were created between the background measures and the EF measures to check the pattern of the data. There were varying degrees of linear relationships between the measures, as seen by the fit lines. For age these ranged from R^2 = .05 to .10, for BPVS they ranged from R^2 = .11 to .37, and for NEPSY they ranged from R^2 = .02 to .11.

10.15 Appendix O

Data screening for Mc rate

Boxplots identified no extreme outliers. Scatterplots showed differing degrees of linear relationships between the Mc rates. Fit lines on the scatterplots between the three total Mc rates ranged from $R^2 = .18$ to .35. This indicates there is some consistency between each measure of the three TPs. The histograms for total Mc rate at TP 1 and 2 showed a normal distribution and the Shapiro-Wilk tests were not significant ($p > .05$). Total Mc rate at TP3 showed a positive skew and the Shapiro-Wilk test was significant ($W = .93$, $p = .04$). However, Z scores of the skewness and kurtosis fell within the +/-1.96 range, so it was decided to use the data in their current format, rather than try to transform all of the Mc rate data.

10.16 Appendix P

Data screening for Mc interview scores

Boxplots were created and no extreme outliers were found. The Mc interview data were positively skewed at TP2 and the Shapiro-Wilk tests were significant at each TP ($p < .01$, $p = .02$, $p = .01$). The data were log10 transformed and root squared transformed to investigate whether this would help with the distribution of data. The histograms did not improve and the spread of data and all the Shapiro-Wilk tests that were originally significant were still significant. Due to non-normality still existing after trialling two different transformations, it was decided not to use transformed data and instead use the original data. Scatterplots showed differing degrees of linear relationships between the three Mc interview scores, ranging from $R^2 = .18$ to .33.

10.17 Appendix Q

Data screening for the balance beam performance data

The data were checked for outliers using boxplots, but no outliers were identified. The balance beam data were checked for normality using histograms and it was seen that the distributions were acceptable. However, the Shapiro-Wilk test of normality showed the data not to be normally distributed at each TP ($p < .05$), but the Z scores for skewness and kurtosis were within the acceptable level (+/- 1.96). Scatterplots of the balance beam data were quite varied, but tended towards positive relationships, as confirmed by the fit lines, which ranged from $R^2 = .12$ to $R^2 = .25$.

10.18 Appendix R

Data screening for the strategy development data

The consistency and first correct data were checked for normality and it was found that all the Shapiro-Wilk assumptions were violated ($p < .01$), with the exception of 2 balance consistency. However, this is not unexpected due to the limited range of trials and the skewed results (such as many children getting trial 1 correct), therefore the data are analysed acknowledging this.

Holm-Bonferroni-corrected Kendall Tau correlations between the balance beam consistency and the first trial correct scores

	1	2	3	4	5
1. 2 conflict balance consistency					
2. 2 balance consistency	.38, p =.02 (n = 27)				
3. 4 balance consistency	.05, p =.77 (n = 27)	.10, p = .56 (n = 27)			
4. 2 conflict balance first correct	**.78, p <.0036** (n = 26)	.34, p = .04 (n = 26)	-.05, p = .78 (n = 26)		
5. 2 balance first correct	.24, p = .19 (n = 27)	.45, p = .01 (n = 27)	.18, p = .32 (n = 27)	.07, p = .68 (n = 28)	
6. 4 balance first correct	-.03, p = .86 (n = 27)	-.10, p = .56 (n = 27)	**.64, p = .001** (n = 27)	-.11, p = .55 (n = 28)	.03, p = .84 (n = 37)

10.19

Means and SDs for the EF and Mc scores for each strategy pattern at each TP

An overview of the means and SDs for EF and Mc scores per each strategy pattern at each TP can be seen below. The 2 conflict balance problems are presented first, then the 2 balance problems, and then the 2 balance problems.

2 conflict balance problems: EF, Mc rate and Mc interview scores.

Means and SDs per strategy classification for the EF scores for the 2 conflict balance trials

	1 wrong strategy	2 wrong strategies	Wrong then correct	Trial and error (wrong)	Trial and error (correct)
EF 1	-0.25	0.25	2.02	-0.91	0.34
	(1.97)	(2.39)	(2.46)	(2.37)	(0.36)
	n = 14	*n* = 7	*n* = 3	*n* = 9	*n* = 2
EF 2	0.21	0.33	1.77	-0.85	-0.43
	(2.36)	(2.611)	(0.67)	(2.41)	(1.11)
	n = 13	*n* = 6	*n* = 3	*n* = 9	*n* = 2
EF 3	0.26	-0.08	0.95	-0.48	1.38
	(2.27)	(3.15)	(0.40)	(2.47)	(0.02)
	n = 14	*n* = 7	*n* = 3	*n* = 9	*n* = 2

Means and SDs per strategy classification for the Mc rates for the 2 conflict balance trials

	1 wrong strategy	2 wrong strategies	Wrong then correct	Trial and error (wrong)	Trial and error (correct)
Mc rate 1	2.88	3.50	2.81	2.79	1.49
	(1.64)	(1.58)	(0.58)	(1.18)	(0.42)
	n = 12	*n* = 7	*n* = 3	*n* = 9	*n* = 2
Mc rate 2	3.80	3.12	2.99	3.19	1.44
	(1.96)	(1.59)	(0.63)	(1.37)	(0.10)
	n = 11	*n* = 6	*n* = 2	*n* = 9	*n* = 2
Mc rate 3	4.28	2.82	2.57	2.39	3.89
	(2.42)	(1.79)	(0.94)	(0.89)	(0.88)
	n = 8	*n* = 7	*n* = 3	*n* = 8	*n* = 2

Means and SDs per strategy classification for the Mc interview scores for the 2 conflict balance trials

	1 wrong strategy	2 wrong strategies	Wrong then correct	Trial and error (wrong)	Trial and error (correct)
Mc interview 1	48.86	44.64	58.33	30.56	43.75
	(20.50)	(18.90)	(7.22)	(12.67)	(8.84)
	n = 11	*n* = 7	*n* = 3	*n* = 9	*n* = 2
Mc interview 2	50.00	52.08	56.25	44.44	50.00
	(12.50)	(30.02)	(8.84)	(22.63)	(17.68)
	n = 11	*n* = 6	*n* = 2	*n* = 9	*n* = 2
Mc interview 3	53.13	51.79	58.33	34.38	62.50
	(16.02)	(20.95)	(7.22)	(17.36)	(35.36)
	n = 8	*n* = 7	*n* = 3	*n* = 8	*n* = 2

2 balance problems: EF, Mc rate and Mc interview scores.

Means and SDs per strategy classification for the EF scores for the 2 balance trials

	Correct	Wrong then correct	Trial and error (wrong)	Trial and error (correct)
EF 1	0.69	-0.61	-0.86	0.35
	(1.88)	(N/A)	(2.31)	(2.20)
	n = 6	*n* = 1	*n* = 11	*n* = 17
EF 2	1.44	1.00	-0.72	-0.01
	(1.58)	(N/A)	(3.09)	(1.68)
	n = 6	*n* = 1	*n* = 11	*n* = 15
EF 3	2.26	1.26	-0.98	-0.20
	(0.67)	(N/A)	(2.66)	(2.05)
	n = 6	*n* = 1	*n* = 11	*n* = 17

Means and SDs per strategy classification for the Mc rate for the 2 balance trials

	Correct	Wrong then correct	Trial and error (wrong)	Trial and error (correct)
Mc rate 1	1.30	3.48	2.54	2.81
	(4.09)	(N/A)	(1.57)	(1.30)
	n = 4	*n* = 1	*n* = 11	*n* = 17
Mc rate 2	1.43	N/A	2.98	2.79
	(5.35)	N/A	(1.48)	(1.43)
	n = 5	N/A	*n* = 10	*n* = 15
Mc rate 3	1.81	1.88	2.74	3.13
	(4.11)	(N/A)	(1.41)	(1.81)
	n = 5	*n* = 1	*n* = 7	*n* = 15

Means and SDs per strategy classification for the Mc interview scores for the 2 balance trials

	Correct	Wrong then correct	Trial and error (wrong)	Trial and error (correct)
Mc interview 1	62.50	50.00	36.25	42.65
	(10.21)	(N/A)	(17.13)	(18.25)
	n = 4	*n* = 1	*n* = 10	*n* = 17
Mc interview 2	70.00	N/A	37.50	50.00
	(18.96)	N/A	(10.21)	(18.90)
	n = 5	N/A	*n* = 10	*n* = 15
Mc interview 3	62.50	50.00	35.71	50.00
	(8.84)	(N/A)	(11.25)	(22.66)
	n = 5	*n* = 1	*n* = 7	*n* = 15

4 balance problems: EF, Mc rate and Mc interview scores.

Means and SDs per strategy classification for the EF scores for the 4 balance trials

	Correct	1 Wrong strategy	Trial and error (correct)
EF 1	-0.07	-1.28	1.10
	(2.21)	(2.66)	(1.88)
	n = 29	*n* = 2	*n* = 4
EF 2	-0.05	-3.10	2.28
	(2.06)	(2.66)	(1.80)
	n = 27	*n* = 2	*n* = 4
EF 3	0.02	-3.28	1.65
	(2.22)	(2.57)	(1.55)
	n = 29	*n* = 2	*n* = 4

Means and SDs per strategy classification for the Mc rate for the 4 balance trials

	Correct	1 Wrong strategy	Trial and error (correct)
Mc rate 1	3.05	1.73	2.43
	(1.45)	(0.76)	(1.13)
	n = 27	*n* = 2	*n* = 4
Mc rate 2	3.48	1.31	3.01
	(1.49)	(0.29)	(2.38)
	n = 24	*n* = 2	*n* = 4
Mc rate 3	3.01	3.26	4.35
	(1.58)	(N/A)	(3.47)
	n = 24	*n* = 1	*n* = 3

Means and SDs per strategy classification for the Mc interview scores for the 4 balance trials

	Correct	1 Wrong strategy	Trial and error (correct)
Mc interview 1	43.75	37.50	43.75
	(18.11)	(17.68)	(23.94)
	n = 26	*n* = 2	*n* = 4
Mc interview 2	48.96	50.00	50.00
	(21.15)	(17.68)	(10.21)
	n = 24	*n* = 2	*n* = 4
Mc interview 3	50.00	37.50	41.67
	(20.52)	(N/A)	(14.43)
	n = 24	*n* = 1	*n* = 3

10.20 Appendix T

Data screening for the EF performance data for each group

The assumptions of the statistical test were checked for each group separately. The Shapiro-Wilk tests for EF scores at each TP were not significant for either group ($p > .05$), the Z scores for skewness and kurtosis were acceptable, and boxplots identified no extreme outliers.

10.21 Appendix U

Data screening for the Mc rate data for each group

The assumptions of the statistical test were checked for each group separately. The Shapiro-Wilk tests were not significant for GP at any TP ($p > .05$) or for DI at TPs 1 or 2 ($p > .05$), but it was significant at TP3 ($p = .03$), suggesting the data were non-normal at the last TP. However, Levene's test of variance between groups was not significant ($p > .05$), suggesting the groups' variance were not different. Boxplots were created for each groups' Mc rates data to check for outliers and no extreme outliers were identified.

10.22 Appendix V

Data screening for the Mc interview scores data for each group

The assumptions of the statistical test were checked for each group separately. The Shapiro-Wilk test for the GP interview scores at TP2 was not significant ($p > .05$), but it was for GP at TPs 1 and 3 ($p < .01$, $p = .04$). The Shapiro-Wilk tests for DI were not significant at any TPs ($p > .05$). Boxplots were created to check for outliers and no outliers were seen in the data.

10.23 Appendix W

Data screening for the balance beam performance data for each group

The assumptions of the statistical test were checked for each group separately. The Shapiro-Wilk tests for GP's balance beam performance scores at TPs 1 and 2 were significant (p = .02, p < .01), but not at TP3 (p = .42). The Shapiro-Wilk tests for DI were significant at each TP (p = .03, p = .01, p = .03). Boxplots were created to check for outliers and no outliers were seen in the data. This is the same pattern as described earlier, when the groups' data were examined together.

10.24 Appendix X

Is there a difference in the balance beam support type protocol between the groups?

Kendall Tau correlations between length of instruction and balance beam performance

	1	2	3
1. Balance beam 1			
2. Balance beam 2			
3. Balance beam 3			
GP	**-.58, $p < .01$**	.23, $p = .26$	-.29, $p = .90$
DI	.23, $p = .20$	-.27, $p = .20$	-.05, $p = .80$

Notes. $N = 36$ for TP1. $N = 29$ for TP2. $N = 27$ for TP3.

It was previously reported that the length of instruction at TP1 significantly differed between the groups due to GP receiving a longer instruction time.

It was found that the length of instruction at TP1 negatively correlated with performance for GP, as seen in the table above. No other significant correlations were found for either group for length of instruction. It seems that the longer the instruction at TP1, the poorer the children performed.

Kendall Tau correlations between "generic" pieces of instruction (that I provided) and balance beam performance

	1	2	3
1. Balance beam 1			
2. Balance beam 2			
3. Balance beam 3			
GP	-.10, $p = .62$	**.64, $p < .01$**	-.14, $p = .58$
DI	.28, $p = .14$	-.15, $p = .51$	.08, $p = .72$

Notes. $N = 36$ for TP1. $N = 29$ for TP2. $N = 27$ for TP3.

It was previously reported in the methodology chapter that GP received more pieces of "generic" information (from the adult) than DI at TP3, however, no significant correlation with balance beam performance was detected. The table above shows a significant positive correlation was found between the number of "generic" pieces of instruction given to GP and their balance beam performance at TP2. However, this did not differ between the groups and was not seen for the other TPs, suggesting this is not the driving factor in performance.

Kendall Tau correlations between pieces of instruction obtained before the trials (including discoveries during the play for GP) and balance beam performance

	1	2	3
1. Balance beam 1			
2. Balance beam 2			
3. Balance beam 3			
GP	.05, p = .83	**.52, p < .01**	.12, p = .61
DI	.28, p = .14	.15, p = .51	.08, p = .72

Notes. N = 36 for TP1. N = 29 for TP2. N = 27 for TP3.

It was reported in the methodology chapter that DI had a significantly higher total number of pieces of information before starting the trials than GP (this includes the information GP discovered during the play time) at every TP. However, this did not translate into any significant correlations with balance beam performance (see above table). As in the above analysis, a significant correlation was found between the pieces of instruction obtained before the trials by GP and their balance beam performance at TP2, likely due to the same reason as the previous analysis.

Kendall Tau correlations between pieces feedback and balance beam performance

	1	2	3
1. Balance beam 1			
2. Balance beam 2			
3. Balance beam 3			
GP	.11, $p = .57$	.04, $p = .84$	-.10, $p = .65$
DI	.02, $p = .91$	-.28, $p = .20$	-.21, $p = .32$

Notes. $N = 36$ for TP1. $N = 29$ for TP2. $N = 27$ for TP3.

It was reported in the methodology chapter that DI received more feedback than GP at TP1, but no significant correlations between feedback and balance beam performance were detected at any TP for either group (see table above).

10.25 Appendix Y

Data screening for complete sample's ramps data

The assumptions of the ramps data (performance and Mc) were checked and boxplots were created for each measure to check for outliers, but none were found. The data were checked for normality and some assumptions were violated. Histograms for the total trials correct and first trials correct did not show normal distributions and the Shapiro-Wilk test of normality showed the data not to be normally distributed for each measure ($p < .05$). Mc rate and Mc interview scores did not violate the assumption of normality ($p > .05$). Scatterplots of the different measures showed relationships of varying strengths and most fit lines showed positive linear relationships, except Mc rate and the two trials measures (total percentage correct and percentage first correct).

10.26 Appendix Z

Data screening for each group's ramps task trials data

The assumptions of the statistical test were carried out for each group separately. The Shapiro-Wilk tests for GP were not significant ($p > .05$). For DI, the first trials correct data violated the assumption of variance ($p = .03$). Boxplots were created to check for outliers and no extreme outliers were found.

10.27 Appendix AA

Data screening for each group's Mc data during the ramps

The assumptions of the statistical test were next checked for each group separately. The Shapiro-Wilk tests for both groups were not significant on either measure ($p > .05$). Boxplots indicated no outliers existed in the ramps Mc data.