Approaches to Evaluating Teaching for Mastery in Secondary Mathematics Education

SHEPHERD, REBECCA

Approaches to Evaluating Teaching for Mastery in Secondary Mathematics Education

Rebecca Shepherd

Masters of Arts by Research in Education (MA)

School of Education, University of Durham

September 2022

Abstract

Teaching for Mastery (TfM) in mathematics education is a pedagogical approach which seeks to develop students' depth of understanding to ensure that they 'master maths' and develop a deep, secure and adaptable understanding of the subject. The term was coined by the NCETM following the England-Shanghai Teacher Exchange Programme of 2015 in a bid to address mathematics underperformance of students in England following transnational assessments.

This research explores observational methods for assessing the impact of TfM on secondary mathematics outcomes, and is the first of its kind in moving away from experimental approaches to attempt to ascribe cause to TfM at the individual, student level. The move away from an experimental approach comes from the need to overcome the risk of trial effects and bias that are inherent to approaches such as Randomised Controlled Trials (RCTs), and the wish to explore the longer-term impact of embedded TfM. This thesis explores the research field to date as well as other viable methods for ascribing cause, before justifying the use of propensity score matching methods. Three individual school cases are considered, and propensity score matching methods applied and analysed, showing that observational methods may, to some extent, help evaluate curriculum interventions.

Key words: *Teaching for Mastery, propensity score matching, randomised controlled trials, 'maths for all', depth, conceptual understanding, mathematical connections, variation, mathematical language, problem-solving*

Contents

List of Figures 1
List of Tables 2
Acknowledgements 4
Chapter One: Introduction 6
1.1 Background 6
1.2 Research Initiatives 6
Chapter Two: Research Literature 8
2.1 What is Teaching for Mastery? 8
2.2 Mastery Defining Features and Programmes 13
2.4 Overview of Defining Features 28
Chapter Three: Existing Evidence Base 30
3.1 The Existing Evidence Base 30
Chapter Four: Methodology 35
4.1 Introduction and Research Question 35
4.2 Experimental Methods 35
4.3 Longitudinal Evaluation 37
4.4 Multiple Regression Methods 38
4.5 Interrupted Time Series Methods 40
4.6 Propensity Score Matching Methods 42
4.7 Conclusion 48
Chapter Five: Methods 50
5.1 Introduction 50
5.2 Design 50
5.3 Steps for Propensity Score Matching 54
5.4 Methods Adoption 66
5.5 Ethical Considerations 67
5.6 Conclusion 68
Chapter Six: Implementation of Methods 69
6.1 Overview 69
Chapter Seven: Implementation of Methods – Pilot Study 71
7.1 Introduction 71
7.2 Objectives 71
7.3 Interview and School Context 72
7.4 Sample Data Preparation 74
7.5 Analysis A - Characteristics before Matching 76

7.6 Analysis A - Propensity Score & Matching 77
7.7 Estimation of Treatment Effect 87
7.8 Sensitivity Analysis 88
7.9 Analysis B - Characteristics before Matching 89
7.10 Analysis B - Propensity Score & Matching 90
7.11 Estimation of Treatment Effect 95
7.12 Sensitivity Analysis 95
7.13 Analysis C - Characteristics before Matching 96
7.14 Analysis C - Propensity Score & Matching 97
7.15 Estimation of Treatment Effect 103
7.16 Sensitivity Analysis 103
7.17 Lessons for the Main Study 104
7.18 Objectives Realisation 104
7.19 Steps for the Main Study: Pre-Registration 105
Chapter Eight: Implementation of Methods – Main Study (Part 1) 108
8.1 Introduction 108
8.2 Interview and School Context 108
8.3 Sample and Data Preparation 110
8.4 Analysis A: Characteristics before Matching 113
8.5 Analysis A: Propensity Score & Matching 114
8.6 Analysis A: Treatment Effects and Sensitivity Analysis 118
8.7 Analysis B: Characteristics before Matching 119
8.8 Analysis B: Propensity Score & Matching 120
8.9 Analysis C: Characteristics before Matching 124
8.10 Analysis C: Propensity Score & Matching 124
8.11 Analysis C: Treatment Effects and Sensitivity Analysis 129
8.12 Conclusion 130
Chapter Nine: Implementation of Methods – Main Study (Part 2) 131
9.1 Introduction 131
9.2 Interview and School Context 131
9.3 Sample and Data Preparation 134
9.4 Analysis A: Characteristics before Matching 136
9.5 Analysis A: Propensity Score & Matching 137
9.6 Analysis A: Treatment Effects and Sensitivity Analysis 143
9.7 Analysis B: Characteristics before Matching 143
9.8 Analysis B: Propensity Score & Matching 144

9.9 Analysis B: Treatment Effects and Sensitivity Analysis 150
9.10 Conclusion 150
Chapter Ten: Discussion 152
10.1 Summary of Main Findings 152
10.2 Methodological Lessons and Limitations 155
10.2.1 Lessons and Limitations 156
10.2.2 PSM Lessons and Limitations 157
10.3 Key Implications 159
10.4 Conclusion 160
10.5 Future Research 160
References 162
Appendices 175
Appendix A 175
Appendix B 176
Appendix C 177

List of Figures

2.1	NCETM's 5 Big Ideas	p.13
2.2	Example and non-example of one quarter	p.18
2.3	Example of a problem-solving question	p.27
2.4	Example of a goal-free problem	p.27
4.1	Propensity to go to Catholic school	p.45
4.2	One-to-one matching	p.46
4.3	Overlap of propensity scores	p.47
5.1	KS2 Attainment and GCSE Maths Grade	p.57
5.2	Matching with Calipers	p.63
5.3	Flowchart of the research methods orchestrated in this study	p.67
7.1	Pilot study analysis A propensity scores	p.78
7.2	Pilot study analysis A propensity score distribution after version 1 of matching	p.80
7.3-7.6	Pilot study analysis A covariate balance before and after version 1 of matching	p.80-81
7.7	Pilot study analysis A propensity score distribution after version 2 of matching	p.82
7.8-7.11	Pilot study analysis A covariate balance before and after version 2 of matching	p.83
7.12	Pilot study analysis A propensity score distribution after version 6 of matching	p.86
7.13-7.16	Pilot study analysis A covariate balance before and after version 6 of matching	p.86-87
7.17	Pilot study analysis B propensity scores	p.90
7.18-7.22	Pilot study analysis B covariate balance before and after version 1 of matching	p.91
7.23	Pilot study analysis B propensity score distribution after version 3 of matching	p.93
7.24-7.28	Pilot study analysis B covariate balance before and after version 3 of matching	p.93
7.29	Pilot study analysis C propensity scores	p.97
7.30	Pilot study analysis C propensity score distribution after version 2 of matching	p.98
7.31-7.35	Pilot study analysis C covariate balance before and after version 2 of matching	p.99
7.36	Pilot study analysis C propensity score distribution after version 4 of matching	p.100
7.37-7.41	Pilot study analysis C covariate balance before and after version 4 of matching	p.101
7.42	Pilot study analysis C propensity score distribution after version 5 of matching	p.102
7.43-7.47	Pilot study analysis C covariate balance before and after version 5 of matching	p.102
7.48	Pre-registration steps	p.106
8.1	Main study 1 analysis A propensity scores	p.116
8.2	Main study 1 analysis A propensity score distribution after version 2 of matching	p.116
8.3-8.11	Main study 1 analysis A covariate balance following version 2 of matching	p.117/8
8.12	Main study 1 analysis B propensity scores	p.122
8.13	Main study 1 analysis B propensity score distribution after version 3 of matching	p.122
8.14-8.17	Main study 1 analysis B covariate balance following version 3 of matching	p.123
8.18	Main study 1 analysis C propensity scores	p.127
8.19	Main study 1 analysis C propensity score distribution after version 4 of matching	p.127
8.20-7.25	Main study 1 analysis C covariate balance following version 4 of matching	p.128
9.1	Main study 2 analysis A propensity scores	p.140
9.2	Main study 2 analysis A propensity score distribution after version 2 of matching	p.140
9.3-9.8	Main study 2 analysis A covariate balance following version 2 of matching	p.141
9.9	Main study 2 analysis A propensity score distribution after version 1 of matching	p.142
9.10-9.15	Main study 2 analysis A covariate balance following version 1 of matching	p.142
9.16	Main study 2 analysis B propensity scores	p.147
9.17	Main study 2 analysis B propensity score distribution after version 5 of matching	p.148
9.18-9.21	Main study 2 analysis B covariate balance following version 5 of matching	p.148
9.22	Main study 2 analysis B propensity score distribution after version 3 of matching	p.149
9.23-9.28	Main study 2 analysis B covariate balance following version 3 of matching	p.149

List of Tables

5.1	Matching Algorithms	p.61
5.2	Matching Structures	p.62
6.1	Overview of data from each school	p.69
7.1	Pilot study analysis A pre-matching covariate distribution	p.76
7.2	Pilot study analysis A version 1 matching results	p.79
7.3	Pilot study analysis A version 2 matching results	p.81
7.4	Pilot study analysis A version 3 matching results	p.84
7.5	Pilot study analysis A version 4 matching results	p.84
7.6	Pilot study analysis A version 5 matching results	p.85
7.7	Pilot study analysis A version 6 matching results	p.85
7.8	Pilot study analysis B pre-matching covariate distribution	p.89
7.9	Pilot study analysis B version 1 matching results	p.91
7.10	Pilot study analysis B version 2 matching results	p.92
7.11	Pilot study analysis B version 3 matching results	p.92
7.12	Pilot study analysis B version 4 matching results	p.94
7.13	Pilot study analysis B version 5 matching results	p.94
7.14	Pilot study analysis C pre-matching covariate distribution	p.96
7.15	Pilot study analysis C version 1 matching results	p.97
7.16	Pilot study analysis C version 2 matching results	p.98
7.17	Pilot study analysis C version 3 matching results	p.99
7.18	Pilot study analysis C version 4 matching results	p.100
7.19	Pilot study analysis C version 5 matching results	p.101
8.1	Main study 1 analysis A characteristics before matching	p.113
8.2	Main study 1 analysis A propensity score variants and result balance	p.115
8.3	Main study 1 analysis B characteristics before matching	p.119
8.4	Main study 1 analysis B propensity score variants and result balance	p.121
8.5	Main study 1 analysis C characteristics before matching	p.124
8.6	Main study 1 analysis C propensity score variants and result balance	p.126
9.1	Main study 2 analysis A characteristics before matching	p.137
9.2	Main study 2 analysis A propensity score variants and result balance	p.139
9.3	Main study 2 analysis B characteristics before matching	p.144
9.4	Main study 2 analysis B propensity score variants and result balance	p.146

Acknowledgements

I would like to thank my two supervisors, Prof Adrian Simpson and Dr Linda Wang, for their never-ending support and enthusiasm for this thesis. Their passion and dedication for my research has been unwavering and I am extremely thankful to have two supervisors who share my passion for education research.

I would also like to thank my fiancée, Nicholas Betts, whose support and patience with me whilst learning to code has been a stabilising influence.

Finally, to Stephen Betts, for the hours spent proof reading and reviewing the thesis.

Chapter One: Introduction

1.1 Background

Teaching for Mastery (TfM) has become a mathematics education phenomenon since the inception of regional Maths Hubs in 2014 and the England-Shanghai Teacher Exchange Programme of 2015. The phrase 'Teaching for Mastery' describes elements of classroom pedagogy that seeks to give pupils the best chance of mastering mathematics. Some educationalists believe that to 'master maths' means pupils acquire a deep, long-term, secure and adaptable understanding of the subject (NCETM, 2016). The National Centre for Excellence in the Teaching of Mathematics (NCETM), through regional Maths Hubs, support state-funded primary and secondary schools in improving mathematics education and embedding a TfM approach. According to the NCETM, the essential features of TfM in Maths are: working to develop understanding, keeping the class together working on the same content and believing that every pupil is capable of success (NCETM, 2016).

1.2 Research Initiatives

Whilst TfM has become increasingly popular across primary and secondary schools, there is little research to date on the effects of the approach on student attainment. There is some limited research in the field, which are discussed in the literature review of this thesis, but a gap remains where there has been little exploration of the longer-term impact of the approach. Where research has been conducted, it is predominantly of an experimental nature which is open to risks of bias through trial effects and of a narrow scope. The seminal research to date is the Education Endowment Foundation's (EEFs) randomised controlled trial which explored the impact of Ark's Mathematics Mastery programme, but in looking at only one type of 'mastery' and for a one year 'dose', the findings do not address a desire to explore the long-term impact of a mastery approach. The gap that remains poses the opportunity to explore the impact of TfM using observational approaches, where data collected over a longer period of time can be utilised.

There is minimal documented research to date on the impact of TfM using observational approaches. As such, this research trials such methods to see if they can work effectively as an alternative to experimental approaches as a means to ascribe cause. Specifically, propensity score matching (PSM) methods are trialled and supplemented by semi-structured interviews to gain contextual insights of individual school cases. PSM is a statistical technique that seeks to reduce bias

due to confounding variables by comparing outcomes among units in two treatment groups following matching on propensity scores. The technique was first introduced by Rosenbaum and Rubin (1983) and is used in the context of secondary mathematics education in this research.

This research is particularly significant as mathematics education and outcomes has been at the forefront of political attention for some time. Following the identification of a wider attainment gap between pupils from different backgrounds in England than in some East Asian countries, as well as the realisation that pupils in Shanghai and Singapore are 'ahead' of their counterparts in England, interest in East Asian mathematics grew (OECD, 2012). The Department for Education funded the England-Shanghai Teacher Exchange programme in 2015 which saw teachers from England learning from Shanghai pedagogy. Following this, in 2016, the Schools Minister pledge over £40 million of funding to support schools in adopting 'mastery' approaches and local Maths Hubs which has started to establish themselves from 2014, have worked on the ground with classrooms teachers ever since, aiming to drive up standards of mathematics education.

It is clear, therefore, of the large-scale investment nationally into mathematics education and therefore it is important to investigate the impact of a 'mastery approach' on pupils' outcomes. The thesis begins with an extensive review of the research literature, exploring what makes a 'mastery approach' and culminating with a tentative definition for mastery in secondary mathematics education for the purpose of this thesis. Then, the existing evidence base is explored, arguing that there is little evidence to answer the question of pupil impact following long-term embedded mastery approaches. Following this, a variety of different methods for ascribing cause are considered, arguing that propensity score matching is the most suited for research of an observational design that seeks to measure student impact. The results chapters, detailing the findings from three different participating schools, leads to the conclusion that observational methods may, to some extent, help evaluate curriculum interventions – in this case, Teaching for Mastery.

The research question is structured to compare student outcomes of those taught using 'mastery approaches' to a previous pedagogical approach so that the difference in curricula can be quantified:

"How might non-experimental methods be adapted to address whether Teaching for Mastery plays a causal role in improved assessment scores, compared to previous pedagogical approaches?"

Chapter Two: Research Literature

2.1 What is Teaching for Mastery?

2.1.1 Introduction

Since the England-Shanghai Teacher Exchange in 2015, further discussed in section 2.1.4, Teaching for Mastery (TfM) has attracted significant attention from policy makers, professionals in the education sector, and academics alike. The first aim of this section is to identify the key factors which make up the various definitions of TfM and to gain a clear understanding of how the pedagogical approach is understood across various communities. From this exploration, a tentative definition of TfM is established for the purpose of this research. The second aim of the literature review is to evaluate the existing evidence base into the efficacy of TfM, leading to a rationale for this research study as a whole.

2.1.2 Early Origins of Mastery

Whilst TfM has become central to mathematics education since 2015, the term 'mastery' is not unique to the twenty-first century. Benjamin Bloom, Professor of Education at the University of Chicago, proposed a 'mastery' learning strategy in 1968, arguing that most students can 'master' taught content so long as instruction is approached sensitively and systematically (Bloom, 1968).

According to Bloom, 'mastery' is exemplified when students achieve at least 90% in a knowledge test before moving on to new learning. If a student does not achieve 'mastery', then Bloom's curriculum approach suggested that they should receive additional support until they can. The teaching and curriculum approach that Bloom formulated was underpinned by high-quality of instruction, perseverance, and an investment of time. It was suggested that all students be given the time required to learn the same material and achieve the same level of 'mastery' as their peers. Bloom found that by varying teaching strategies to meet the needs of all learners, there was less variation in learning outcomes. As such, it was the teacher's role to ensure their pedagogical strategies allowed all students to achieve the same level of learning since students differ in their learning style and aptitude (Bloom, 1981; Levine, 1985).

The idea of varying teaching and learning practices and allowing students enough time to 'master' taught content underpinned the notion of 'mastery' throughout the late twentieth century (Anderson, 1975). However, compared to 'mastery' in the present today, Bloom's theory has evolved

(Watson, Geest, and Prestage, 2003). There are similarities with the view that all students can achieve, however, TfM is largely informed by East Asian education practices, such as Shanghai and Singapore, which attracted attention due to their successes in transnational assessments (Boylan et al., 2018). Therefore, whilst the word 'mastery' is not new, this section will explore how the term has evolved into the twenty-first century.

2.1.3 Transnational Assessments

The rationale for a focus on TfM in mathematics education stemmed from an observation of overall higher achievement and a smaller attainment gap in mathematics in East Asian countries, compared to England (Boylan et al., 2018). Results of transnational assessments in 2012 raised concerns in both areas. The Organisation for Economic Co-operation and Development (OECD) found that in some East Asian countries, there was clearly a smaller gap in attainment levels between pupils, compared to England (OECD, 2012). In Shanghai, the share of 'low achievers' in mathematics, according to Programme for International Student Assessment (PISA) assessments, was just 3.8% (OECD, 2012). In England, the same measure was 21.8%, suggesting that there is a larger attainment gap across pupils (OECD, 2012). PISA also argued that pupils in countries such as Shanghai and Singapore are up to three years ahead of UK pupils in their mathematical ability by the age of 15 (OECD, 2012). It was also identified that England had become accustomed to seeing around one fifth of their children fall below national average by the end of primary school, with twice as many by the end of secondary school (Drury, 2014). Thus, the discourse around TfM developed from a desire to address the evidence that highlighted underachievement against international comparators, a wide attainment gap, and a high proportion of students not meeting national expectations (Drury, 2014; Clapham and Vickers, 2018).

2.1.4 England-Shanghai Exchange

Following concerns with English mathematics performance in comparison to East Asian countries, the National College for School Leadership (NCSL) led two study visits to Shanghai in exploration of mathematics education (NCSL, 2013, 2014). In 2014, it was agreed with the Shanghai Municipal Education Commission to hold a Teacher Exchange with the aim of learning from Shanghai practices to hopefully improve student outcomes in mathematics education (Boylan et al., 2019).

The Mathematics Teacher Exchange programme, launched in 2014, was co-ordinated through national Maths Hubs, led by the NCETM. The first cohort saw teachers from 48 English

schools visit Shanghai for one week as well as hosting a return visit for Shanghai teachers for two weeks. The aim of the programme was to share teaching practices and experiences. A second cohort, involving 70 schools, was similarly set up in 2016-17 with the aim of expanding the exchange programme to get more schools involved. Approximately 830 teachers from England and Shanghai have taken part in the exchange programme since its inception in 2014 (Chen, 2019).

During the exchanges, it was noted that mathematics teaching in Shanghai was different in many ways to mathematics teaching in England. In the Final Report published by the Department for Education (DfE) in 2019, these differences were summarised (Boylan et al., 2019):

> *"Shanghai whole-class interactive teaching aims to develop conceptual understanding and procedural fluency. This is achieved through lessons designed to be accessible to all, through skilful use of teacher questioning and incremental progression. Teaching is supported by well-crafted mathematical models and exemplar problems, as well as practice materials that focus on critical aspects of mathematical learning. To ensure pupils progress together, tasks are designed to allow for extension by deepening understanding of concepts and procedures, and daily intervention is used to support those needing extra tuition" (Boylan et al., 2019, p.26).*

The DfE's longitudinal report evaluated the impact of the exchange programme across the two teacher cohorts. Following the first exchange, it was found that participants had developed their understanding of Shanghai mathematics education. In the subsequent exchange, once the TfM programme had been established, participants were encouraged to implement mastery approaches into their schools. The Final Summary report[1], published by the Department for Education, concluded that the exchange visits had been important to the development of TfM and acted as a catalyst for change in participating English schools. However, it also concluded that there was little quantifiable evidence to suggest that there was any significant increase in student attainment at the time, other than KS1 students (Boylan et al., 2019).

Nonetheless, the Mathematics Teacher Exchange contributed to the establishment of a specific 'Teaching for Mastery' programme, coordinated by regional Maths Hubs. Through funding granted by the Department for Education in 2016, the 'TfM programme' has sought to reach 9,300

[1] Boylan, M., Wolstenholme, C., Demack, S., Maxwell, B., Jay, T., Adams, G., & Reaney, S. (January 2019). Longitudinal evaluation of the Mathematics Teacher Exchange: China-England - Final Report. *Department for Education*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/773320/MTE_main_report.pdf

primary schools and 1,700 secondary schools by 2023 (Boylan et al., 2019). The government funded initiatives that make up the 'TfM programme' consists of:

- A professional development course to train a team of mastery specialists who will promote TfM in their schools
- Funding to support the use of high-quality mastery-aligned textbooks
- Materials and resources to support TfM
- Mathematics Teacher Exchange programme between Shanghai and England.

Thus, the Shanghai-England teacher exchange was integral in developing professional understanding of the East Asian 'mastery' approach and also acted as a catalyst for a government funded programme that supported the development of TfM across English schools.

2.1.5 Political Support

Politicians have played a considerable role in ensuring that TfM has attracted attention since 2015, not least because Schools Minister, Nick Gibb, pledged £41 million of funding to support schools in adopting mastery approaches in 2016 (Department for Education, 2016). Boylan et al. (2019), in the Final Evaluation Report of the Mathematics Teacher Exchange, stated that "this is arguably the largest policy innovation in mathematics education since the introduction of the National Numeracy Strategy in the late 1990s" (p. 149). Just before Gibbs' pledge, the 2014 National Curriculum was reformed and redesigned to raise standards in mathematics (Department for Education, 2014). It was stated that, "…pupils will move through the programme of study at broadly the same pace…" (p. 42) and that "pupils who grasp concepts rapidly should be challenged through being offered rich and sophisticated problems before any acceleration through new content…" (p. 42). Simultaneously, "those who are not sufficiently fluent should consolidate their understanding…before moving on" (p. 42). Whilst the National Curriculum document does not explicitly attribute these statements to TfM, keeping the class together and challenging depth of understanding through problem solving are upheld as crucial facets of TfM by members of the professional community, as explored in section 2.1.8.

2.1.6 Education Sector

Since the England-Shanghai exchange, education professionals have also played a fundamental part in seeking to define the approach as well as advocating it across many primary and

secondary schools. The NCETM established the Teaching for Mastery Programme in 2014 which includes a number of initiatives, including a Continuing Professional Development (CPD) course to train 'mastery specialists', financial support for schools to engage with the programme, as well as the distribution of high-quality textbooks (Boylan et al., 2019).

The NCETM also coined the 'Five Big Ideas' in 2017 that they thought should underpin TfM in both primary and secondary schools, as depicted in figure 2.1. Mastery Professional Development material has since been published to support teachers in planning for and embedding these 'Five Big Ideas' in their practice.

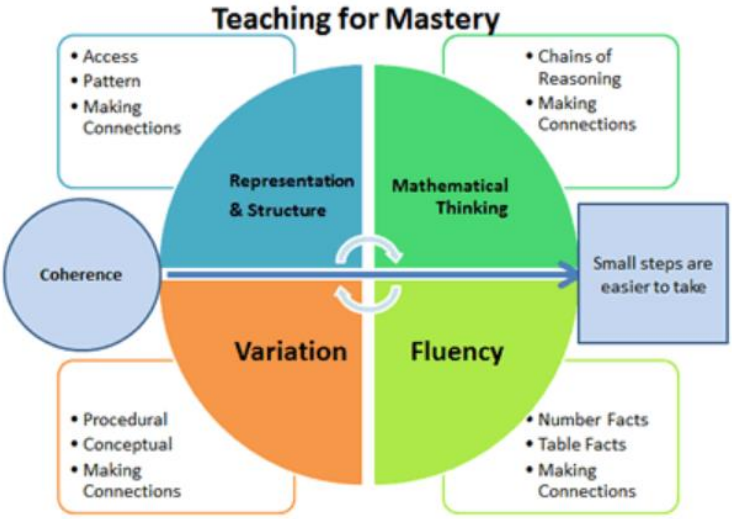

Figure 2.1: *Five Big Ideas in TfM* (NCETM, n.d.)

The Five Big Ideas have been drawn from research evidence and the diagram above helps to bind them together. 'Coherence' runs through the centre of the diagram and in practice, would be evident where lessons are broken down into small, connected steps to ensure that all children can access and generalise a new concept.

'Representation and Structure' is seen as crucial to develop pupils' depth of understanding and helps to develop conceptual understanding (Stripp, 2014). In mathematics lessons, representations are used to expose the mathematical structure with the aim that students can in time do the maths without relying on the representation (NCETM, 2016).

'Mathematical Thinking', according to Charlie Stripp, director of the NCETM, is about mathematical debate and reasoning since they are thought to be crucial to developing deep understanding (Stripp, 2014). In a classroom which promotes mathematical thinking, students would not passively receive taught ideas but, instead, those ideas would be thought about, reasoned, and discussed (NCETM, 2016).

'Fluency' is best defined as the ability to recall facts and procedures quickly and efficiently as well as being capable of flexibly applying mathematical skills to different contexts and representations.

Finally, 'variation' comprises of two principal strands: how teachers represent the concept being taught and the sequencing of exercises. In practice, teachers would draw attention to critical aspects of new content and would give students practice questions that draw attention to mathematical relationships (NCETM, 2016).

Like Bloom's early philosophy of 'mastery', the NCETM upholds the notion that TfM "rejects the idea that a large proportion of people 'just can't do Maths'" (NCETM, 2016). Through whole-class interactive teaching, with all pupils working together on the same content at the same time, it is suggested that 'no pupil is left behind'.

The NCETM's work is supplemented by regional Maths Hubs, each of which is a network of schools and colleges across England, with one leading institution. The core purpose of the Maths Hubs Programme is to lead improvement in mathematics education across the country. By running several projects each year, they seek to share best practice by harnessing maths leadership and expertise for the benefit of all pupils. Maths Hubs work collaboratively with neighbouring schools, colleges, universities, CPD providers, maths experts and employers (Boylan et al., 2019). At the start of the England-Shanghai exchange there were 32 Maths Hubs in England. This has now grown to 40 Hubs serving the whole of England.

The education sector, therefore, plays an instrumental role in embedding TfM across primary and secondary schools. At the core of the NCETM's philosophy is TfM and the 'Five Big Ideas' help teachers to consider *how* to embed mastery learning into mathematics lessons. Regional Maths Hubs promote the NCETM's notion of TfM and continue to support schools and colleges in delivering mastery education.

2.2 Mastery Defining Features and Programmes

2.2.1 Introduction

Work by the NCETM and Maths Hubs is supplemented by many 'mastery' programmes that have been developed over time, providing schools with schemes of learning, lesson resources, and professional training. These mastery programmes share the fundamental belief that all students should be exposed to all content and that, to ensure depth of understanding, topics should be

taught for longer periods of time than they were before the implementation of 'mastery'. The programmes also encourage teaching through variation, developing conceptual understanding and forming connections, as well as the use of key mathematical language and problem-solving.

This section explores and details the elements which are often taken to be the key features of TfM, whilst considering how each feature appears in the original international context and how it appears in the UK TfM curricula. Each feature is exemplified with some context-based examples.

2.2.2 Success for All

Whilst acknowledging that 'Maths for All' is a key feature of some East Asian curricula, Stripp (2014) proposed that it is one of the key ingredients for their transnational success for mathematics education. According to Stripp, a key distinguishing feature between top-performing countries for mathematics education and other countries is the extent to which differentiation is achieved through preserving some topics for 'more-able' students. Rather than preserving such topics, top-performing countries instead differentiate through questioning children at an appropriate level to support them to succeed, whilst allowing high-attaining students to explore concepts in greater depth without accelerating onto new content (Jackson, 2020).

The idea of 'Maths for All' was noted by Jerrim and Vignoles (2016) in an analysis of East Asian curricula, where it was suggested that the curricula aim was to ensure that every child reaches a certain level for all topics, with the expectation that all children 'master' the curriculum before the class progresses onto the next topic (Jerrim and Vignoles, 2016). To enable this, it was noted in the evaluation of the Mathematics Teacher Exchange that, "in order to keep the whole class moving through the curriculum broadly together, the pace was slowed to support accessibility, differentiation was more likely to be focused on depth rather than acceleration…" (Boylan et al., 2017, p. 23).

Following the Teacher Exchange, some schools adopted this notion. School Standards Minister Nick Gibb noted that there had been a marked increase in the number of schools using 'whole-class teaching' rather than splitting pupils by attainment, meaning that schools had started to move away from the idea of preserving particular areas of maths for higher attaining students (Department for Education, 2019).

Many of the developed UK mastery programmes share the belief that all students can succeed, similar to Bloom's late twentieth-century philosophy. Ark's Mathematics Mastery

Programme[2], founded by Helen Drury, is built around a set of key beliefs, one of them being: "every child can succeed, regardless of background" (Ark Curriculum Plus, n.d.). Similarly, White Rose Maths[3], another mastery programme offering resources and support to schools, is centred around the belief that every child can do maths, despite their prior attainment or background (White Rose Maths, 2020). Through their programme, White Rose Maths claims to build a "culture of deep understanding, confidence, and competence in maths" (White Rose Maths, 2021) regardless of a student's starting point.

Thus, a defining feature of TfM is 'Success for All'. It was noted that, in East Asian pedagogy, certain topics are not preserved for higher attaining students and this idea has been adopted by UK mastery programmes. To support this, there has been a noted move away from splitting pupils by attainment for mathematics lessons.

2.2.3 Structure

Linked to the idea of 'Success for All' is the structure of mastery curricula. By spending longer on each topic, the class is able to progress through a series of small steps, meaning that each mathematics lesson is focused on one small key learning point. It is thought that, by ensuring teaching is carefully structured and planned in small steps, the necessary scaffold for all to achieve is provided, ensuring that classes progress through the curriculum at the same pace (NCETM, 2021).

Research found that East Asian curricula typically covers fewer topics in greater depth, ensuring that each subject is broken down into small steps to ensure deep understanding (Jerrim and Vignoles, 2015). According to Singapore's Ministry of Education, the Singaporean mathematics curriculum is carefully designed so that higher concepts and skills build upon foundational ones and learnt in sequence (Ministry of Education, Singapore, n.d.). A specific example helps to exemplify this. In grades 7-8, at the start of secondary school, students work with prime numbers, highest common factors and lowest common multiples, which builds upon content learnt in primary school on factors and multiples. Thus, the idea of teaching topics for a longer period, with careful sequencing to ensure depth of understanding, has become associated with TfM. Prior to TfM, from

[2] Ark Curriculum Plus (2021) https://www.arkcurriculumplus.org.uk/
Ark's Mathematics Mastery programme offers lesson resources, schemes of learning and professional training to primary and secondary schools.
[3] White Rose Maths (2021) https://whiterosemaths.com/
The website offers resources from early years to post-16 in line with 'mastery' and support schools with change through in-depth training programmes.

personal experience, it was more common for students to spend shorter periods of time on seemingly disconnected areas of mathematics.

Observations during the Mathematics Teacher Exchange stated that lessons in Shanghai appeared 'pacey' without an association of 'covering' a lot of material. The 'pace' observed in Shanghai is not associated with rushing through the curriculum, but instead reflects the high levels of engagement in lessons where purposeful work that focuses on depth of understanding helps to ensure pupils master the mathematics. It was noted that, "compared to a typical lesson in England, in terms of curriculum coverage the pace was slow" (Charlie's Angels, 2016, para 8). Instead, "constant teacher-pupil interaction and continual engagement in mathematical thinking" is what made the lesson appear 'pacey' (Charlie's Angels, 2016, para 8). Thus, the concepts of breaking topics down into small steps and teaching topics for a longer period of time are directly associated with East Asian mathematics education.

The idea of teaching less topics over a period of time but in much more depth has been adopted by UK mastery curricula. Ark's curriculum is organised into 'mastery half terms' where extended time is spent within a single area of mathematics, allowing teachers to spend more time developing students' conceptual understanding and providing opportunities to explore concepts in greater depth (Mathematics Mastery, n.d.). One of Ark's curriculum principles is 'depth before breadth' and the programme encourages 'dimensions of depth' as a way to deepen pupils' understanding and ensuring high expectations of what pupils learn and understanding.

A review of White Rose Maths' Schemes of Learning shows that each topic[4] is taught for a significant length of time to ensure students' depth of understanding, but that topics within each strand or area of mathematics are spread over a programme of learning. Across the secondary curriculum, a strand such as 'algebra', is broken down into manageable objectives for each year group so that prior knowledge and skills are built upon. For example, in year seven, students are taught how to use algebraic notation. In year eight, students work with brackets, equations, and inequalities, but do not form and solve equations until year nine. In year ten, these skills are built upon through the topic of simultaneous equations. Finally, in year eleven, the culmination of algebra is seen with changing the subject and rearranging formulae (White Rose Maths, 2020).

A similar pattern is seen in the strand of geometry, with students starting by constructing and measuring in year seven, followed by working with trapezia, circles, parallel lines, and polygons

[4] A 'topic' of study is a section of a 'strand' of mathematics. For example, the topics of area, perimeter, volume, transformations, similarity and congruency all make up 'geometry' but the topics are generally not taught together but split up over a few years in a programme of study.

in year eight. The progression continues into years ten and eleven where students work with bearings (White Rose Maths, 2020). By sequencing each mathematical strand carefully and breaking each topic down into smaller steps, it is believed that students can progress together (White Rose Maths, 2020). Thus, teaching topics for a longer period of time is considered a defining feature of TfM.

2.2.4 Teaching with Variation

In seeking to draw out features of Chinese mathematics instruction that might contribute to their high levels of performance, the presence of 'bianshi' has been noted (Gu, Yang, and He, 2015; Clarke, Keitel, and Yoshinori, 2006). 'Bianshi' is an approach developed in Shanghai, meaning 'teaching with variation'. It involves generalising from examples using conceptual and procedural variation to promote deeper understanding in mathematics. Teachers use 'pudian', or sequences of problems, to build relational understanding of a mathematical concept. Bianshi involves the architect of learning, whether that is the teacher or textbook author, devising opportunities to gain a deeper understanding of a mathematical concept by distinguishing variant and invariant properties (Jacques, 2018).

Bianshi has drawn attention of researchers as a potential explanation for high mathematics achievement in China compared to their international counterparts in comparative tests (Clarke et al., 2006). Whilst other factors, such as teacher's subject knowledge or use of mathematical language may contribute to Chinese high performance, evidence suggests that effective variation in Chinese classrooms has a positive impact on pupil learning (Gu et al., 2004; Bao et al., 2003; Gu et al., 2017). In a large longitudinal study conducted in the Shanghai district of Qingpu, the use of variation was trialled in a group of experimental schools. As a result of the trial, the pass rate for entrance to the junior high schools in the region rose from 16% in 1979 to 85% in 1986 (Gu et al., 2015).

According to some academics, variation aims to help students develop profound understanding of a concept through multiple perspectives, focusing on helping students to develop an interconnected knowledge structure by varying examples and exercises (Gu, Huang and Gu, 2017). The central idea of teaching with variation is to highlight the essential features of a particular concept through varying the non-essential features (Gu, Huang and Marton, 2004).

In adopting this feature into UK classrooms, variation has been split into two categories: conceptual and procedural (Gu, Huang, and Marton, 2004). For both categories, variation is where

teachers draw attention to critical aspects and when sequences of activities and exercises are considered carefully so that students appreciate mathematical relationships and structure (Solent Maths Hub, 2019). Both types of variation lead pupils to generalise new concepts; a goal expressed strongly by Chinese teachers (Cai and Hwang, 2002).

Conceptual variation involves learners experiencing a concept from multiple perspectives where examples are devised to offer deliberately varied contexts of representations (Jacques, 2018). For example, when comparing three-quarters with one-quarter, number lines, bar models, and area models can all be used to draw out pupil generalisations of the concept. Within a set of examples presented to the class, careful planning ensures that attention is drawn to the similarities and differences between questions so that students can make generalisations about the key features of a new concept. An effective collection of examples for a new concept should consist of standard examples, non-standard, and non-concept (Ballentine, 2018).

For example, when introducing the concept of one quarter, three examples used to cover these categories might be as follows:

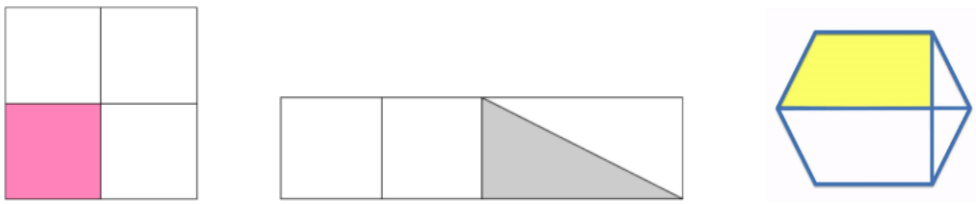

Figure 2.2: Example and Non-Example of One Quarter (Ballentine, 2018)

On the left is the standard representation of a square cut into four equal parts. In the middle is a non-standard representation, which challenges students to think about the similarities and differences between the standard and non-standard representation (Ballentine, 2018). The non-standard aspect of the middle representation is that the four parts are not congruent, despite the shaded region still representing one quarter. The representation on the right is an example of a non-conceptual example, where students are exposed to a 'non-example'. The shaded yellow section does not represent one quarter since it is not one out of four equal pieces. The three examples together provide the opportunity for students to make connections and challenges them to think to a greater depth (Cambridge Maths Hub, 2016).

Ballentine's (2018) categories for conceptual variation provide a structure for the introduction of a topic to help students to make generalisations through comparing and contrasting different examples. The NCETM has broken down conceptual variation into two theoretical strands: *positive* variation and *negative* variation (NCETM, n.d.). According to the NCETM, *positive* variation is when students are presented with a series of questions that allow the same content to be seen in

different contexts. For example, when asking students to use Pythagoras' Theorem to calculate the length of the hypotenuse, the triangle may be presented in varying orientations so that students are required to think about which side is the hypotenuse. *Negative* variation aligns with Ballentine's non-conceptual examples as students are exposed to what a concept is *not*. For example, sticking with the topic of Pythagoras, students may be presented with a non-right-angled triangle or with a triangle where there is not enough information to use Pythagoras' theorem (e.g., a side length and two angles). Variation in this sense supports pupils' ability to reason and to generalise as they compare and contrast examples or questions.

Closely linked to conceptual, procedural variation aims to draw students' attention to key mathematical structures; however, this time, it is manifested by the richness of varying problems and the variety of transferring strategies (Gu et al., 2004). In a bank of questions that are designed with procedural variation in mind, some aspects are varied whilst others will stay the same, allowing learners to think carefully about mathematical structure (Watson and Mason, 2006). Procedural variation provides the opportunity for intelligent practice[5] rather than mechanical repetition so that students can focus on relationships and not just procedure (NCETM, 2016; Gu, 1991). Below is a set of questions on expanding single brackets that are planned with intelligent practice in mind:

Expand $3(x + 4)$
Expand $3(4 + x)$
Expand $3(x - 4)$
Expand $3(4 - x)$
Expand $3x(x + 4)$
Expand $3x(4 + x)$
Expand $-3x(x + 4)$
Expand $-3x(4 + x)$
Expand $-3x(x - 4)$
Expand $-3x(4 - x)$

In the above set of questions, whilst at first glance they may look easy and repetitive because of the frequent recurrence of the numbers 3 and 4, the exercise encourages discussion about the effect that the negative sign or the unknown variable outside the bracket has on each question, and therefore encourages an appreciation of mathematical relationships. Further examples of intelligent practice can be seen on Craig Barton's[6] website called 'Variation Theory'

[5] Intelligent practice means setting carefully designed questions which help students focus on specific elements of a topic, rather than only regular mechanical practice. Practice is considered 'intelligent' when students can discuss the rationale behind the structure of the questions that form the exercise.
[6] Craig Barton's mathematical education websites include: https://variationtheory.com/, https://mathsvenns.com/about/, https://www.dqaday.com/, https://diagnosticquestions.com/,

which provides teachers with banks of questions for topics that are well chosen so that students can be encouraged to appreciate the similarities and differences between questions (Barton, 2021).

The two elements of variation – conceptual and procedural – work in tandem to achieve deep, sustained learning. Both strands encourage learners to engage with mathematical structure whilst also ensuring that they actively involve themselves in the process of learning (Kullberg, Runesson, & Marton 2017; Gu, Huang & Marton, 2004). If mathematical examples and tasks are not presented in this way, then students do not build sustained and deep conceptual understanding, which in turn undermines mathematical confidence (Watson and Mason, 2006). When using variation theory, teachers ensure that careful attention is paid to both the examples being used and the questions that are posed to students so that they see common and unusual representations as well as being exposed to misconceptions (Kullberg et al., 2016; Nilsson, 2014; Vikstrom, 2014).

The NCETM holds 'Variation' as of its '5 Big Ideas' for TfM and in line with this, White Rose Maths, offers "Thinking Through Variation" online video training as professional development for teachers, with the intent of equipping teachers to carefully plan topics to reveal the underlying structure of mathematics (White Rose Maths, 2021). Thus, it is deemed a crucial aspect of a 'mastery approach'.

2.2.5 Depth of Understanding and Mathematical Connections

Deep, conceptual understanding reflects a student's ability to understand and reason mathematical concepts and relationships (Mutawah et al., 2019). Students with conceptual understanding know more than isolated facts and methods. They understand the importance of each mathematical idea and understand why certain strategies or formulae work. Students must learn mathematics with understanding, actively building new knowledge from experience and prior knowledge (Cummings, 2015). The notion of deep understanding can be traced back to the late twentieth century when Skemp (1976) preached the importance of 'relational' understanding in mathematics, rather than 'instrumental' understanding. Relational understanding is where students understand the structure behind mathematical concepts, whereas instrumental understanding is taken as students simply recalling a procedure.

Research into East Asian teaching practices has found that conceptual understanding is a key feature and is correlated with increased student performance. A study that compared selected

https://eedi.com/. His influential book titles are: "How I wish I'd taught Maths" and "Reflect, Expect, Check, Explain".

textbooks from England and Shanghai concluded that Shanghai textbooks encourage a greater depth of understanding, compared to those used in England (Wang, Barmby and Bolden, 2017). By analysing examples used for linear functions, it was clear that there were different emphases on conceptual understanding levels in each region. The study speculated that students in Shanghai might be encouraged to move towards more abstract levels of understanding, which in turn may lead to better performance (Wang, Barmby and Bolden, 2017). Bao (2002) suggested that the requirement of the mathematics curricula in East Asian countries are more difficult than those in Western countries due to the emphasis on conceptual understanding.

It has been suggested that forming mathematical connections is also key to East Asian mathematics education. According to Huang and Leung (2005), meaningful learning takes place when pupils are guided to establish "a substantial and non-arbitrary connection" (p. 349) between their prior knowledge and the new knowledge. Following a mathematics classroom study in Hong Kong and Shanghai of nineteen lessons, it was found that the forming of mathematical connections was a common feature, alongside teaching with variation (Huang and Leung, 2005). In a study on composite difficulty between new and old Chinese mathematics textbooks, topic coverage was analysed and a comparison to a UK-based textbook was given (Bao, 2004). It was found that in the UK Mathematics Enhancement Programme textbooks, more than 64% of chapters in year 8 only have a "single topic" and less than 1% have more than "three topics" (Bao, 2002). Therefore, within these textbooks, there is little evidence of connections between different areas of maths (Bao, 2004). In contrast, of the Chinese textbooks analysed, around 28% of chapters have a "single topic" with higher percentages of chapters having two, three, or more topics integrated within (Bao, 2004). Whilst this study only analysed one series of UK textbooks, it has been suggested that Chinese textbooks have historically placed more importance on the relationships between different mathematical concepts (Bao, 2004).

Thus, in adopting a 'mastery approach' in the UK, there has been an emphasis on developing conceptual understanding and forming mathematical connections. Ark's Mathematics Mastery programme is underpinned by the 'dimensions of depth' which aims to help pupils develop understanding of mathematics. The first principle of the dimensions of depth is 'conceptual understanding' which enables pupils to make rich connections between mathematical ideas (Ark Academy Plus, n.d.). Ark's curriculum is underpinned by the belief that "mathematics tasks are about constructing meaning and making sense of relationships" (Ark Academy Plus, n.d., para. 8). Ark also considers mathematical connections as a key principle of its programme, stating that making explicit links between different areas of maths allows students "to construct a comprehensive conceptual framework that can be used as the foundation for future learning" (Ark Academy Plus, n.d., para. 9).

White Rose Maths also holds conceptual understanding and mathematical connections as being important underlying principles. The programme offers professional support and training on the 'Concrete, Pictorial, Abstract' (CPA)[7] teaching approach since it is thought that the use of manipulatives can help to develop conceptual understanding (White Rose Maths, 2019). The mastery programme emphasises the need for teachers to address common misconceptions in all areas of maths in order to help students to make accurate generalisations, conclusions and connections (White Rose Maths, 2018).

Lessons that strive to develop students' conceptual understanding are carefully designed so that pupils are able to see why they are applying a particular method to a question. For example, when learning how to calculate the area of a triangle, rather than immediately being presented with the formula $A = \frac{b \times h}{2}$ and asked to apply it, classroom discussion and teacher questioning would guide students to the formula, by making connections to their prior knowledge of the area of a rectangle.

By creating connections to different areas of maths, conceptual understanding helps students to organise their knowledge as a coherent whole. In a classroom where students are encouraged to make mathematical connections, they would be tasked to answer questions that require them to apply skills from more than one area of maths. For example, when working on volume, a question might require students to find the maximum and minimum volume of a cuboid using rounded dimensions, thus allowing them to connect their skills of bounds and volume.

Research into conceptual understanding and forming mathematical connections pre-dates TfM, and although a full overview of the research to date is beyond the scope herein, it is clear that both features positively impact student learning and therefore little doubt as to why the NCETM references the two elements under the category of 'Mathematical Thinking' in its '5 Big Ideas' (NCETM, n.d.). Researchers largely agree that conceptual understanding and procedural knowledge work in tandem to develop in depth understanding of mathematics (Rittle-Johnson et al., 2001; Desimone et al., 2005; Hiebert et al., 2005). According to Nahdi and Jatisunda (2020), conceptual understanding and understanding the 'why' supports learning, since connections between prior and new knowledge can be formed. In a study that investigated student's understanding of area, it was found that those who had a good understanding of the concept of area as well as using formulae displayed competence in identifying geometric shapes, using formula for determining areas, and self-correcting mistakes (Huang and Witz, 2011). In contrast, students who misunderstood the

[7] CPA is a teaching approach that uses physical and visual aids to build learner's understanding of abstract topics.

concept of area, but understood multiplication, showed only some ability to use area formulas. As such, conceptual understanding of why a particular formula should be applied, as well as making connections between different areas of maths, are important in ensuring student success (Mutawah et al., 2019).

2.2.6 Mathematical Language

According to the NCETM, the insistence on using precise mathematical language and terminology in the classroom supports pupils' ability to think mathematically, and therefore helps to develop a deeper understanding (NCETM, 2016). Investigation into East Asian teaching practices has revealed that significant emphasis is placed upon mathematical language in mathematics teaching.

In a study that reviewed the characteristics of mathematics teaching in five Shanghai schools, Lim (2007) revealed that an emphasis on 'precise and elegant mathematical language' was one of the key characteristics of effective teaching, alongside others such as teaching with variation, reasoning, discipline, teacher collaboration and strong student-teacher rapport. In one specific classroom observation, a student attempted to define a 'perpendicular bisector' and whilst the given answer was deemed reasonably correct, the teacher was not satisfied with the explanation and challenged another student to define it again using the least number of words. When another student was able to define it precisely, the teacher praised him for his aptitude in using precise mathematical language. This was not an isolated incident in the study as other cases were also reported where teachers were seen to stress the importance of precision when reading the unit of speed, producing written reasons for stages of working, and precise format of writing an algebraic expression. It has been claimed, therefore, that some East Asian teachers place greater emphasis on the role of precise mathematical language compared to English teachers (Axbey, 2020).

Jerrim and Vignoles (2015) explored the link between East Asian 'mastery' teaching methods and English pupils' mathematical skills and found an increase in pupil attainment following the use of precise mathematical language. Furthermore, a study into the 'mastery' method of teaching multiplication found that explicit student and teacher language, alongside manipulatives, helped to ensure an in-depth understanding of multiplication rules such as multiplying by ten (Gurganus and Wallace, 2005). Moreover, Ding et al. (2017) found that when teachers worked on developing their own precise mathematical language, students were able to focus on the key learning point and therefore develop a deeper understanding of a given concept.

Research into mathematical language predates any exploration of the 'mastery approach' and is beyond the scope of this research. However, a brief synthesis of the research shows that key mathematical language has a positive impact on student learning. Two research projects conducted in Germany in the late 1990s found that a classroom culture which encouraged student participation and interaction, centred around using key mathematical language, helped to develop their conceptual understanding since they were provided a platform to negotiate meanings of key terms (Krummheuer, 1999). Very earlier research also showed that student problem-solving ability improved following instruction in key mathematical vocabulary since students could understand the meaning of more complex questions (Dresher, 1934; Johnson, 1944).

Furthermore, a study focused on students' understanding of written mathematics problems found that the use of precise language led to increased understanding and enhanced communication between students and their peers (Kranda, 2008). A study that examined the relationship between use of precise language and mathematical justification also found increased student understanding (Adams et al., 2016).

Thus, mathematical language is emphasised as a key element in UK mastery approaches. Ark states 'language and communication' as one of its key aims in its curriculum intent (Ark Academy Plus, n.d). Within a 'mastery' lesson, time is dedicated to developing confidence with specific mathematical vocabulary as well as verbal reasoning. 'Talk tasks' are part of every lesson where students may start off by using informal language to describe their thought process, which leads to formal and precise mathematical language using teacher scaffolding. Ultimately, it is proposed that mathematical language strengthens conceptual understanding since it enables pupils to explain and reason (Ark Academy Plus, n.d.). It is thought, therefore, that language and communication should be developed alongside conceptual understanding and mathematical thinking as part of a 'mastery' education (McCourt, 2019).

Mathematical language has been interpreted as 'formal mathematical terminology', something which the National Numeracy Strategy cited when suggesting that students should "explain their methods and reasoning using correct mathematical terms" (DfEE, 1999, p.4). Specific mathematical terms could include any from the following, but not exhaustive, list: rectangle, triangle, parallelogram, bisector, perpendicular, congruent, similar, right-angle, vertically opposite, quadratic, polynomial, commutative, associative, sum, product, quotient, divisor, subtract, dividend, numerator, denominator. According to Adams et al., (2016), precise use of mathematical language will help to engage students in justification and reasoning. It is thought that in classrooms where students are encouraged to explain and justify their thinking, learners demonstrate greater

achievement and more in-depth complex thinking (Boaler and Staples, 2008; Kazemi and Stipek, 2001).

School-based research, conducted by the NCETM, found that the use of precise mathematical language was one of the reasons schools had shown improved classroom practice and lesson design in the 2018-19 academic year (NCETM, 2019). The report stated an association between correct use of mathematical language, enthusiasm of reasoning and increased depth of understanding, claiming that "precise mathematical language and full sentences" should be considered a main characteristic of a mastery lesson (NCETM, 2019, p. 8). Schools that participated in the research have identified the development of mathematical language and vocabulary as having a positive impact on pupil progress (NCETM, 2019).

2.2.7 Problem Solving

In mathematics, problem-solving is where pupils are challenged to find a way to apply knowledge and skills to answer unfamiliar types of problems (Almond, 2020). To try and suggest an exhaustive definition of problem-solving would in fact negate the complexity of the term.

Mathematical problem-solving has been the subject of substantial debate for decades since it has long been considered an important aspect in the teaching of mathematics, either as an aim or a mechanism (English, 2010; Liljedahl and Santos-Trigo, 2019). Nevertheless, in the broadest sense, problem-solving is a process which involves systematic observation and critical thinking to find an appropriate solution to achieving a particular goal (Rahman, 2019). In the context of mathematics education, it has been described as a 'goal-directed activity', a 'complex endeavour', and a 'variety of cognitive actions' (Lesh and Zawojewski, 2007; English, 2010; Lester, 2013). Early research agrees with the claim that problem-solving incorporates tasks that are not straight-forward and require resilience to solve (Pólya, 1965; Mason et al., 1982; Liljedahl, 2008). It was the emergence of the cognitive learning theories, credited to Educational psychologist Jean Piaget, that led to the acknowledgement of problem-solving as a complex mental activity (Rahman, 2019).

Chinese educators uphold the ability to solve problems as the ultimate aim of their curriculum (Gu, 2017). Huang and Leung (2005) observed, during a mathematics lesson, procedural variation being used as a scaffold for problem solving to enable students to be successful. Therefore, there has been a long history of interest in integrating these skills into mathematics education in countries such as China and Singapore (Siu, 2004; Stanic & Kilpatrick, 1988). Problem solving is seen as fundamentally essential in mathematics learning in both south-eastern countries, so much so that

the National Council of Supervisors of Mathematics upheld it as “the principal reason for studying mathematics” (National Council of Supervisors of Mathematics, 1977, p. 2).

Ark’s mastery programme is largely influenced by the Singaporean mathematics curriculum, despite many professionals associating TfM with Shanghai pedagogy (Boylan, 2018). In Singapore, problem solving is the clear aim of the mathematics curriculum. Their National Curriculum documents state that: “The learning of mathematics should focus on understanding, not just recall of facts or reproduction of procedures…Only with understanding can students be able to reason mathematically and apply mathematics to solve a range of problems. After all, problem solving is the focus of the mathematics curriculum.” (Ministry of Education, 2012, p. 21). The idea of problem solving being central to the curriculum is further seen in Singapore’s first principle of mathematics teaching which states that “Teaching is for learning; learning is for understanding; understanding is for reasoning and applying and, ultimately problem solving” (Ministry of Education, 2012, p. 21). Singapore’s pentagonal framework for mathematics education also places problem solving as the central element (Pittard, 2018).

Fluency, reasoning, and problem-solving became the key aims of the English Mathematics National Curriculum in 2014. It was stated that the aim of mathematics education is to ensure that all students “can solve problems by applying their mathematics to a variety of routine and non-routine problems with increasing sophistication, including breaking down problems into a series of simpler steps and persevering in seeking solutions (Department for Education, 2014, p. 40).

According to Helen Drury, founder and director of Ark’s Mathematics Mastery, problem-solving is the purpose of TfM. Regardless of a student’s background or prior attainment, problem-solving is the ultimate aim of learning mathematics since every student can learn to solve complex problems in unfamiliar contexts (Mathematics Mastery, 2019). Teacher educator, Rachel Jackson, concurs with Drury, arguing that challenge and problem-solving are defining factors of ‘mastery’ (Jackson, 2020). Problem-solving should not be reserved for a Friday-morning activity or seen as a bolt-on at the end of a topic; instead, it should be an intrinsic part of every lesson, alongside fluency and reasoning (Leighton, 2020).

Within the context of a lesson where problem solving is an integral component, students would be encouraged to present a mathematical skill in multiple ways and given time to practice applying it independently to a new problem in an unfamiliar situation (Drury, 2014). Frequent exploration of problems, alongside practice and application help to build student confidence in solving problems. An example of problem solving in practice might be the problem below:

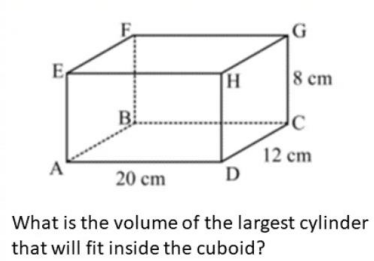

Figure 2.3: Example of a problem-solving question (SSDD Problems, 2021)

This problem could be presented to students at the end of a unit of work on volume and 3D shapes and classified as a problem-solving question requiring students to draw upon their knowledge of three different mathematical areas: volume of a cuboid, volume of a cylinder, and capacity. There are also a series of ways in which the question can be answered, therefore encouraging procedural variation. An example of a volume question which would not be problem solving is as follows:

Find the volume of a cuboid measuring 11 cm by 12 cm by 5 cm

One of the fundamental misconceptions of problem solving is that if the question is given in worded form, rather than as a diagram, it is classified as problem solving. This question is not a problem-solving question because students are able to apply a standard algorithm (length x width x height) to successfully solve it. If a child already has a readily available method to solve a problem, problem solving has not occurred (Almond, 2020). Problem solving in maths is finding a way to apply knowledge and skills to answer unfamiliar types of problems and where reasoning and higher-order thinking must occur (Lester & Kehle, 2003).

Another example of problem solving are 'low threshold, high ceiling' questions, allowing all students from a range of attainment to access the problem. The question below invites pupils to apply their learnt skills to an unfamiliar context – being that there is no set question to answer:

Jay is paid £2000 each month.

He saves 6% of the £2000 each month.

Work out what you can from this information.

Figure 2.4: Example of a Goal-Free Problem (Mattock, 2022)

Teachers could expect a range of answers from this question. A previously low attaining student may simplify find 6% of £2000 whilst a previously high attaining student might express 6% of £2000 as a fraction of £2000, or they might write the amount saved to the amount spent as a

simplified ratio – thus drawing on their mathematical skills learnt from other topics.

Whilst problem-solving is not a new concept, its prevalence in East Asian curricula means that is has become inextricably linked to 'mastery'. Problem-solving helps to ensure that pupils develop a deep and sustained understanding of mathematics due to the challenge of applying learnt techniques to unfamiliar contexts.

2.2.8 Conclusion

The exploration of East Asian pedagogy and UK mastery programmes in this section have shown that there are some key distinguishing features of 'mastery'. First, is the idea that all students should be exposed to all content to ensure 'maths for all' and 'success for all'. Second, is that topics should be taught over a longer period of time through small step objectives to ensure that the class is kept together, and that depth of understanding is developed. Teaching with Variation, or bianshi, is another defining feature that is integral to East Asian practices and has been adopted through UK mastery programmes. Conceptual understanding, mathematical connections and mathematical language are the remaining distinguishing features which are considered as key features of mastery and have been found to be inherent in East Asian mathematical pedagogy. All of these features were also noted during the Mathematics Teacher Exchange programme and as such have become intertwined with the idea of TfM. Although problem-solving is an aim of the National Curriculum for Mathematics, since Ark's Mathematics Mastery programme was built around elements of the Singaporean curriculum, it is also important to consider problem-solving as a distinguishing feature of mastery.

2.4 Overview of Defining Features

The above section has shown that there are several key characteristics that are thought to define TfM in mathematics education, each of which have been adopted from East Asian practices. For this research, TfM is defined as the following:

- 'Maths for All' – where all students are exposed to the same mathematical content and some content is not preserved for higher attaining students
- Where topics are taught for a longer period to ensure depth of understanding

- Where there is an emphasis on teaching for conceptual understanding and forming mathematical connections
- Where variation is embedded in chosen examples and exercises
- Where there is an insistence on key mathematical language usage
- Where problem solving is considered a key aim of the curriculum.

Whilst these key attributes provide an understanding of what is meant by TfM for this research, participating schools may have slightly different understandings. For this reason, the above definition is tentative and not exhaustive. Not all schools may agree that each characteristic is a defining factor of their version of 'mastery', and careful consideration will need to be paid as to whether a school is delivering a 'mastery' curriculum, according to my definition. Nevertheless, structure needs to be provided for the purpose of this study to be able to evaluate a 'mastery' curriculum against a 'non-mastery' curriculum, not least because 'mastery' is a flexible term that is employed in a variety of ways (Boylan et al., 2018).

Fundamental to evaluation is comparison and therefore it is important to consider what makes something *not* mastery. In the methodology section of this thesis, the discussion around the conducted interviews gives personal perspectives of teachers who have moved to a 'mastery' curriculum and the key differences between 'mastery' and 'non-mastery' curricula and pedagogy are considered. From personal experience, before TfM, students were often set by ability with some content preserved for higher-attaining students. Mathematical reasoning and deeper thinking through problem-solving were not an integral part of each lesson in the way that the NCETM promotes through their 'Five Big Ideas'. Moreover, less time was spent on each topic, often resulting in the learning not being broken down into small steps that help to keep the class together.

Chapter Three: Existing Evidence Base

3.1 The Existing Evidence Base

The previous chapter helped to unpick how various communities define TfM, concluding with a tentative definition of the approach for the purpose of this research. This chapter analyses existing evidence into the efficacy of the pedagogical approach, arguing that there is a lack of research into longer term, embedded TfM.

3.1.1 DfE Longitudinal Evaluation of Mathematics Teacher Exchange (2019)

Following the England-Shanghai Teacher Exchange, the Department for Education sponsored a longitudinal evaluation in 2019 to assess its impact. According to the evaluation, the Teacher Exchange was important in developing teacher perceptions on Shanghai whole-class interactive teaching methods, and continues to be crucial today, to the development of TfM across primary and secondary schools in England (Boylan et al., 2019). Hence, it is crucial to examine the evaluation's findings.

The evaluation aimed to examine the impact of the exchange on both cohort 1 and 2 schools. For cohort 1 schools, who participated in the exchange in 2014/15, attention was paid to changes in teaching practice, the impact on pupil attainment, and how other schools had been influenced since the implementation of the exchange programme. This analysis was extended to capture findings on the implementation of change in cohort 2 schools, who participated in the programme in 2016/17, and how teachers' perceptions of TfM had been influenced. Since this thesis is concerned with the impact of TfM on student progress, it is the evaluation findings from cohort 1 schools that are of most interest.

In the 2019 final report, which followed preceding interim reports on the exchange, it was claimed that there was some evidence of positive impact on pupils' KS1 mathematics attainment. A sub-sample of 16 cohort 1 schools that had implemented Shanghai mastery pedagogy were found to be more likely to attain KS1 threshold compared with pupils in the comparison schools (Boylan et al., 2019). A positive finding was therefore claimed, indicating a potential for increased attainment through adoption of Shanghai informed mastery practices. However, it is worth noting that other findings within the same report indicated no evidence of increased attainment.

When considering the whole sample of cohort 1 schools, rather than the sub-sample mentioned above, six impact analyses were conducted which also compared participating schools to a set of matched comparison schools through a quasi-experimental study, using limited propensity score matching. It was concluded that, overall, the analyses did not indicate any positive effect on attainment, at KS1 or KS2 (Boylan et al., 2019). Whilst there was evidence to suggest that individual features of Shanghai pedagogy have the potential to improve attainment within the participating schools, overall policy ambitions of increasing pupil attainment may not have been realised (Boylan et al., 2018). Further, the conducted analyses found that there was no evidence of differential impact relating to students' prior attainment, gender and free school meal status (Boylan et al., 2019).

Additionally, changes in relative attainment identified in the sub-sample may be due to other features of the school, that were not identified in the evaluation. Moreover, the assessment measure was undertaken by teachers in schools who had participated in the Teacher Exchange programme and therefore opens the study up to potential sample bias. It has been acknowledged that, through interviews, participating teachers perceived a positive effect on student attainment at KS1 and KS2 following the implementation of Shanghai pedagogy, despite the lack of quantitative evidence to support the perception (Boylan et al., 2019). Thus, whilst the report found that the exchange visits positively impacted English teachers' beliefs about mathematics teaching, resulting in them showing a commitment to learning from Shanghai teaching methods, there was no quantifiable evidence to support a claim in increased student attainment. As such, the evaluation concluded that further evaluation is needed to provide an evidential base for TfM, especially since the evaluation was limited to the Mathematics Teacher Exchange that did not extend to the whole of the mastery initiative (Boylan et al., 2018, 2019). Boylan was the first to use propensity score matching methods in an attempt to assess the impact of the mastery approach, but this was at the school level rather than the individual level and was limited to the impact of the England-Shanghai exchange; thus, there is still a gap in the existing evidence to use such methods to explore the longer-term impact of more embedded TfM.

3.1.2 Education Endowment Foundation (EEF) Randomised Controlled Trial

Perhaps the most seminal research was conducted by the EEF in the form of large-scale Randomised Controlled Trials (RCTs). The EEF funded two RCTs between 2011 and 2014 to investigate the impact of Ark's Mathematics Mastery Programme in terms of maths attainment for all pupils, as well as the attainment gap between lower and higher attaining pupils. Mathematics Mastery, founded

by education charity Ark, aims to improve the quality of mathematics education by deepening pupil's conceptual understanding, as summarised earlier in section 2.2.

To assess the impact of the Mathematics Mastery programme, two trials were funded: one across primary schools, and another across secondary schools. Both trials were set up to evaluate the programme's impact in its first year of adoption. For the primary phase, 5108 Year 1 pupils across 90 schools were involved. In the secondary phase, 7712 pupils across 40 schools participated. For the year 1 pupils, when comparing progress to students in comparison schools that had not adopted the Mathematics Mastery programme, it was claimed that they made two months' additional progress on average (EEF, 2015). For the year 7 pupils, it was one months' additional progress on average.

However, this 'positive treatment effect' was only deemed statistically significant at the five percent level when results from the Year 1 and Year 7 trials were combined. Through a 'meta-analysis'[8] approach, the EEF combined the primary and secondary findings, to estimate the effect of the intervention. Without the combination, the intervention in Year 7 was only associated with a small increase in average mathematics test scores (effect size = 0.06) which did not, on its own, show statistical significance. Thus, when considering the impact of Mathematics Mastery on secondary students alone, little evidence was found to support an effect over the 'business as usual' alternative on the outcome measure.

3.1.3 Limitations to EEF Research

Whilst the research had a large sample size of over 12,000 pupils and RCTs are often perceived as the 'gold standard' of educational research, there are limitations to the EEF's study. First, the experiment only looked at one very specific notion of mastery, Ark's Mathematics Mastery, with a short-term focus of a one year 'dose'. The alternative to the intervention was labelled 'business as usual', but without specific details, there is little possibility for direct comparison. As such, the research was very narrow and offered no room to investigate other mastery programmes over a longer period of time. It has been suggested that it would be worthwhile to track the medium and long-term impact of a mastery approach (Jerrim and Vignoles, 2015).

Another shortfall of the study is that participating schools were new to the Mathematics Mastery curriculum and therefore may not have been fluent with the TfM approach, compared to

[8] A meta-analysis is an examination of data from two or more independent studies of the same subject, in order to determine overall trends. The EEF combined the results of Year 1 and Year 7 trials to conclude a positive treatment effect.

teachers with more experience of it. Since the study looked at such a short time period, teachers may have had little time to develop their understanding of mastery and therefore could have impacted pupil progress.

On the contrary, participating schools in the experimental group received considerable support throughout the programme. Education charity, Ark, provided schools with training and resources to support the adoption of Mathematics Mastery. Headteacher, Maths co-ordinators, and class teachers received up to two days of 'launch training' and two in-school development visits. On top of this, teachers attended cluster workshops for collaboration. Furthermore, teachers had access to an online toolkit which included mastery-aligned lesson designs, continuous professional development resources, and assessments. As such, it may be that a stronger performance was found than would have been had Ark not provided as much support.

3.1.4 Limitations of Experimental Research

There are wider issues associated with using a RCT to assess the impact of TfM. As an experimental method, RCTs are naturally exposed to limitations such as the 'John Henry' effect (Saretsky, 1975). It is possible that the control group in the study may have introduced other changes following their assignment to the control arm and therefore the results may not be reliable. Moreover, the 'Hawthorne effect' may have impacted the results, with subjects in both groups specifically changing their behaviour as a result of being observed (Sedgwick and Greenwood, 2015). Teachers in the EEF's study may also have modified their teaching practice as a result of knowing they were being observed, regardless of the particular intervention being tested. Within the EEF's report, it was acknowledged that following intervention, some teachers opted to not continue to use manipulatives to support their teaching (Jerrim and Vignoles, 2015). Whilst RCTs are the usual methods used for assessing the impact of an intervention, experimental approaches are open to biases and trial effects that observational studies might not encounter.

The experimental nature of the study means that there was no scope to evaluate the longer-term impact of the programme following children having had extended exposure to the programme. RCTs work with changes in curricula, naturally looking at a relatively short time after implementation and evaluation of changes in student attainment over longer periods of time are expensive, thus rare. For questions that seek to establish if changes in student attainment can be associated with longer-term curricula changes, alternative approaches are needed.

3.1.5 Conclusions

In summary, the existing evidence base to date suggests that TfM may have a very small positive impact on student attainment if any, however the quantitative evidence available is of narrow scope and does little to assess impact over time. As such, this research will aim to investigate if it is possible to evaluate TfM where the approach has been embedded for longer, with pupils having had TfM as a coherent approach for a more substantial part of their schooling, while addressing some of the other concerns about RCTs.

Chapter Four: Methodology

4.1 Introduction and Research Question

The previous chapter highlighted quantitative evidence currently available that has sought to assess the efficacy of TfM is of narrow scope, evaluating only one type of 'mastery' over a short period of time and may be impacted by a number of biases and trial effects. This research investigates if alternative observational methods of quantitative data can be used to evaluate TfM when the approach has been embedded over a longer period, with pupils having been taught using TfM for a more substantial part of their schooling, rather than for a short period of time for the purpose of an experiment. In this instance, TfM will no longer be novel for the pupils and their teachers alike.

Since such methods have not been used before in educational research to assess the impact of embedded TfM, the focus is on establishing the validity of a method to answer a question which an experimental approach is unable to adequately address. Where the methods used are able to evaluate TfM through observational approaches using quantitative analysis of existing data, then the central research question is:

"How might non-experimental methods be adapted to address whether Teaching for Mastery plays a causal role in improved assessment scores, compared to previous pedagogical approaches?"

Since the research aim is about establishing cause, this methodology chapter will explain the viable options for answering the research question, before leading to the justification for the final choice. For each viable option, an insight into how the method works to ascribe cause is provided, as well the logical and practical problems for the research question. The intention of this research is to see if it is possible to argue that the difference between the TfM and non-TfM approaches "caused" any difference in assessment outcomes. Therefore, it is important to explore ways of establishing cause.

4.2 Experimental Methods

RCTs are often deemed the 'gold-standard' of educational research and have been argued to be the most rigorous and robust research method in determining whether there is a cause-and-effect relationship between an intervention and an outcome (Bhide et al., 2018). For this reason, RCTs are increasingly popular in clinical research and, where they are appropriate, in educational

research too. An RCT is performed under controlled conditions with random allocation of interventions to comparison groups, which helps to account for any systematic differences between the two groups at the point of allocation due to randomisation (Bhide et al., 2018).

Random assignment to the 'experimental' and 'control' groups is thought to prevent selection bias by distributing the characteristics of individuals that may influence the outcome randomly between the groups, so that a sufficiently large difference in outcome can be explained with high likelihood by the difference in treatment (Akobeng, 2005). The experimental group go on to receive the treatment or intervention, and the control group do not receive the treatment or intervention, or receive a 'placebo' instead.

Once randomisation is achieved, the experimental and control groups may be given a baseline assessment so that progress can be measured following the implementation of the treatment or intervention. Then, the experimental group receives the treatment or intervention for a fixed period of time whilst the control group receives the alternative treatment or placebo. After the fixed period of time elapses, both the experimental and control groups may be given a post-test against a specific set of measures to establish if the intervention has had an effect. Subsequent analysis can then explore the magnitude and significance of the treatment effect using t-tests, p-values and confidence intervals. P-values and confidence intervals are discussed in-depth in the results chapter of this thesis.

The seminal piece of research to date of such experimental nature is the EEF's randomised controlled trial of 2015, discussed in the literature review. An RCT was employed to ascribe cause, but in seeking to explore the efficacy of TfM, only one specific type of 'mastery' was analysed – Ark's Mathematics Mastery programme (EEF, 2015). Whilst robust experimental methods were used, the results of the trial are not generalisable to other notions of 'mastery' and therefore cannot be used to answer the research question that this paper is seeking to address.

Furthermore, educational experiments are inherently exposed to the risk of experimental or trial effects. Whilst RCTs are good at accounting for differences between groups at the point of allocation, in order to assign cause to the difference in intended treatments, one must be certain that the only post-allocation difference is the one intended (in this case, the curricula).

Schools that participated in the EEF's randomised controlled trial were given considerable support throughout the programme in implementing Mathematics Mastery, and therefore the research may have yielded a stronger performance than there might have been had they had not received as much support, irrespective of any causal effect of the change of curriculum. It is also

possible that the John Henry effect may be at play in experimental studies, where control groups implement other changes, thus impacting the reliability of any results. Moreover, the Hawthorne effect is another common risk of an experimental approach, where subjects can attempt to change their behaviour since they are aware they are being observed. Therefore, it is not certain that the only post-allocation difference between the two groups is the different curricula. This research is interested in alternative approaches to explore the effectiveness of TfM, where experimental bias effects are not of concern.

4.3 Longitudinal Evaluation

Also explored in the literature review were Boylan's interim research reports following the England-Shanghai Teacher Exchange Programme in 2016 and 2017. The longitudinal evaluation, sponsored by the DfE, sought to assess the impact of the exchange programme. The evaluation employed a longitudinal multiple case study design, comparing a sample of participating schools, and comprised of four strands. The first strand of the evaluation examined changes in teacher pedagogy; the second strand analysed the impact on student attainment; the third strand assessed early evidence of change and impact, whilst the fourth strand extended the evaluation to consider schools that formed 'cohort 2' of the programme. For strand two, national pupil data were retrieved and analysed, whilst pupil attitudes were also surveyed.

Whilst the findings of the evaluation have been deemed to be robust since propensity matching methods were used to identify sub-samples for sensitivity analyses and to increase confidence, the longitudinal evaluation only investigated schools that had been involved or related to the exchange programme (Boylan et al., 2019). Therefore, like the EEF's randomised controlled trial, a narrow scope of 'mastery' was investigated. In this instance, it was Shanghai pedagogy that had been employed in the schools that had participated in the exchange programme (Boylan et al., 2018; 2019). As outlined in the introduction to this chapter, this research investigates ways to evaluate TfM when the approach has been embedded for much longer and on the individual student level, so that TfM is no longer novel for the pupils and their teachers. As such, whilst Boylan did use propensity score matching methods, the evaluation does not provide an answer to this central research question.

4.4 Multiple Regression Methods

Multiple regression is another method that can be used to ascribe cause when it is believed that all other causal factors are accounted for. It is a statistical technique that uses several independent variables to predict the outcome of a dependent variable (Moore et al., 2006). The ultimate aim of multiple regression is to model the linear relationships between the independent and dependent variables and is an extension of simple linear regression such as ordinary least-squares (OLS) regression since it involves more than one independent variable.

Using statistical packages, each of the multiple independent variables are examined in their relationship with the dependent variable and predictions are subsequently made on the level of effect they each have on the outcome (dependent) variable. The regression model creates a linear relationship that best approximates all the data points to allow an analyst to make predictions. A line of best fit is calculated that minimises the variances of each of the variables included as they relate to the dependent variable.

An example that explored the relationship between the number of computers at a school and their mathematics test scores helps to give an insight into how multiple regression analysis works (Holm, 2021). The referenced example used existing data from California schools to test the prediction that the number of computers at a school would predict higher mathematics test scores. Using multiple variables in the regression, such as the number of students at each school, number of teachers, school size and parent wealth, a positive and significant relationship was found between the number of computers at a school and students mathematics test scores. Throughout the multiple regression analysis, line graphs tested many other explanation for the relationship before making any conclusions.

However, a drawback of a multiple regression method is that, whilst linear effects that were in the model can be excluded (in this case, school size and parental wealth), potential causes not in the model cannot be excluded, such as teacher experience and other school resources. For example, it might be that school with more IT are richer in other resources and it is those that play the causal role. Thus, multiple regression is perhaps best considered as showing association rather than causation.

Whilst it has been suggested that multiple regression can be used to estimate treatment effects, the models are underpinned by two fundamental assumptions (Zanutto, 2021). First, is the assumption that there is a linear relationship between the dependent and independent variables. The linear equation below highlights this notion:

$$y_i = \propto_o + \beta_1 x_i + \beta_2 z_i + \varepsilon$$

Where:

y_i is the outcome of each individual

$\propto_o$ is a baseline measure

x_i is the treatment on the individual

z_i is a group of covariates

ε is "noise" or "error"

β_1 the amount y_i changes when x_i changes from 0 to 1 when z_i is constant.

For the model to work, it is reliant upon the assumption that z_i captures all the important factors and that their impact is linear, which may not always be the case nor possible. The second assumption is that there are no major correlations between the independent variables or covariates; again, which may not be the case.

Furthermore, the analysis method does not provide any matching mechanism on variables that are associated with the outcome measure, and therefore it cannot be guaranteed that the control and treated groups are comparable on key covariates. The regression assumption of a linear relationship between an outcome measure and the independent covariates may not always hold if the covariate distributions differ between the two groups.

Thus, it is no surprise that Morgan et al. (2008) concluded that multiple regression alone was insufficient to establish cause. When seeking to quantify the effectiveness of special education services in US schools, regression analyses were not able to adequately account for selection bias (Heckman, Ichimur & Todd, 1997; Winship & Mare, 1992). Selection is the bias introduced by the selection of groups for analysis. It is where there is a systematic difference between the characteristics of those selected for the study and those who are not. When selection bias is present, the sample obtained is not representative of the population intended for analysis.

In seeking to quantify the effectiveness, Morgan et al. (2008) examined whether children receiving special education services displayed: greater reading or mathematics skills; more frequent learning-related behaviours; and less frequent problem behaviours than matched peers that did not receive such services. In the study, regression was used as one of four analytical strategies and the results from the model were only considered as a benchmark to compare results obtained through PSM techniques, which were deemed more reliable (Morgan et al., 2008). When the analyses were considered collectively, selection bias was reduced, resulting in the conclusion that the receipt of

special education services had a negative impact on pupils' learning or behaviour whilst having a small, positive effect on learning-related behaviours.

Multiple regression is not an appropriate method for answering the central research question of this paper since it is not possible to establish with certainty that key covariates (which are outlined in the next chapter) are not highly correlated. As well as this, it is difficult to determine if the impact of each is linear. Moreover, the research question seeks to ascribe cause through a comparison of approaches and therefore methods which can ensure there are no systematic differences between two groups are more fitting.

4.5 Interrupted Time Series Methods

Interrupted time series (ITS) methods can also be used to analyse observational data. ITS is a quasi-experimental design that can evaluate an intervention effect without randomisation, using longitudinal data that is collected consistently before and after an interruption. The design is advantageous for 'natural experiments' in real world settings as it can make full use of the longitudinal nature of the data and account for pre-intervention trends (Kontopantelis, 2015). It has been suggested that ITS is a strong method to use to estimate effects of an intervention when it is certain that the only 'interruption' is the one to which it is intended to ascribe cause. ITS is sensitive to differences in the effects of the intervention since the same participants are compared pre- and post- intervention (UK Health Security Agency, 2020).

In practice, ITS can be implemented in different ways and encompasses a wide range of modelling approaches. It can be implemented as a removal or reversal design, where the intervention is added and then removed. In this instance, continuous data collection allows the researcher to assess if the outcome measure changes when the intervention is added and whether the effect diminishes after it is removed. By alternating the presentation and removal of the program over time, the analyst can ascertain with confidence if a casual role can be identified. ITS could also be implemented using a multiple baseline design, where the start of the intervention is staggered across participants, and since treatment is started at different times, changes can be confidently attributed to the treatment rather than to a chance factor. Despite different implementation methods, to be able to draw conclusions through analysing trends and patterns, a large data set spanning a substantial time period is needed.

A research paper which used a range of ITS models to examine the impact of the introduction of the Quality and Outcomes Framework (QOF) pay for performance scheme in UK primary health care demonstrates how the method works (Kontopantelis, 2015). The QOF was introduced in the 2004-05 financial year to reward general practices for achieving clinical targets across a range of chronic conditions. The intervention was national and therefore large-scale; it was adopted almost universally by general practitioners because of the offered financial rewards. The central research aim relating to the QOF was to ascertain if the national intervention had a positive effect on the quality of care.

Performance data on asthma, diabetes, and coronary heart disease was collected from 42 general practices for four time points: 1998 and 2003 (pre-intervention) and 2005 and 2007 (post-intervention). ITS was modelled using a model whose regression coefficients estimated the pre-intervention slope, the change in level at the intervention point, and the change in slope from pre-intervention to post-intervention. In this example, regression modelling found that the intervention had an effect on quality of care in the three conditions of interest. Kontopantelis (2015) commented that a strength of ITS is the ability to account for the all-important pre-intervention trends so that the level change at the intervention point can be attributed to the intervention. However, this assumes that, without the intervention, the pre-intervention trend would continue unchanged into the post-intervention period which relies on there being no external factors, other than the included intervention, that systematically affect the trends.

As with any research method, there are limitations. ITS models assume that the characteristics of the populations remain unchanged throughout the study period. In the context of educational research, there are a multitude of factors that can impact student outcomes. Additionally, the demographics of student population within a given school can change over time. Therefore, methods that can control for covariates are more appropriate.

Kontopantelis (2015) acknowledged that ITS is not appropriate when trends are not linear or when the intervention is introduced gradually or at more than one time point. For some schools, the change to TfM occurred gradually, with some teachers trialling strategies before whole departmental shifts. In this instance, it would be difficult to identify one clear definitive point in time where the “interruption” occurred. Furthermore, and perhaps the key issue, is that for ITS to be effective, the researcher must be able to show that the only ‘interruption’ is the intended one and where school demographics naturally change over time (for example, catchment area or school policies), it is difficult to be able to state this with certainty when analysing a long time period.

Finally, for ITS methods to be successful, longitudinal data is required. In the context of this research into the efficacy of TfM, time series data is not appropriate. Whilst participating schools may be able to trawl back through student data; school policies, procedures and ethos are regularly subject to change over time and therefore student outcomes may fluctuate irrespective of the implementation of an intervention. These confounding factors add another complexity when trying to ascribe cause where covariates cannot be controlled for. Propensity Score Matching is a method which helps to overcome this shortfall of ITS.

4.6 Propensity Score Matching Methods

4.6.1 Summary of Propensity Score Matching (PSM)

PSM, first introduced by Rosenbaum and Rubin (1983), is another analysis method of observational data that seeks to estimate the effect of a treatment, policy, or intervention. When random assignment is not feasible, PSM can be used to create matched groups that are equal on propensity scores (Robinson et al., 2014, Henson, Hull, and Williams, 2010). According to Robinson et al., (2014), "a propensity score is defined as the conditional probability that an individual would be assigned to the treatment condition, given a set of relevant covariates" (p. 344).

By matching on a composite score, PSM accounts for the covariates that predict the treatment group and that relate to the outcome. Therefore, the method endeavours to reduce the bias due to confounding variables that could otherwise be found. In observational studies, since randomisation is not possible, the treated and control groups are systematically different on covariate distributions. Therefore, it is likely that, when directly comparing an outcome measure of two groups who have different characteristics, biased conclusions would be produced, especially if the different characteristics are likely to affect the outcome. These characteristics are called confounders or covariates and PSM offers a method of addressing this by creating two groups who do not differ on factors that matter.

In a study that compared logistic regression[9] with propensity scores, it was found that propensity score matching methods exhibited more empirical power than regression methods

[9] Logistic regression is used in statistical software to understand the relationship between the dependent variable and one or more independent variables by estimating probabilities using a logistic regression equation. The analysis can help to predict the likelihood of an event happening.

(Cepeda, 2003). The study, which used Monte Carlo[10] simulations, focused on bias, precision, empirical power, and robustness to compare the two analytical methods. Propensity scores were found to produce estimates that were less biased, and more robust and precise than regression estimates, as they more effectively controlled for imbalances when presented with confounders. Moreover, previous research which sought to estimate causal effects using school-level data concluded that the key to drawing accurate inferences from data is to "compare treatment and control groups with similar distributions of covariates so that any difference in the outcome can be attributed to the treatment, not to pre-existing differences between the groups" (Stuart, 2007, p. 197). PSM, therefore, is a useful method to compare student outcomes when other confounders (or covariates) are at play. Boylan et al. (2019) used propensity score matching methods as part of the evaluation of the England-Shanghai Teacher Exchange but looked at the school level, rather than the individual level, and explored the impact of TfM in the immediate short-term. This research is concerned with the impact at student level over a longer period of time.

It has been claimed that results obtained from quasi-experiments[11] using PSM methods are closely aligned to those from RCTs (Becker & Ichino, 2002). Weidmann and Miratrix (2020) since endorsed this stance when assessing whether unobserved factors substantially bias education evaluations. Through a meta-analysis of 42 estimates across 14 studies, it was found that there was no evidence of substantial selection bias due to observed characteristics. It was strongly suggested, therefore, that in educational settings, non-experimental approaches could play a more significant role than they currently do in generating reliable causal evidence for school policy and intervention (Weidmann and Miratrix, 2020). That said, like RCTs, the matching in PSM only ensures that the two groups under analysis are similar at the point of allocation; to ascribe cause, the researcher must be sure that the only difference between the two treatment groups is the intended difference (in the case of this research, the curricula). An example of propensity score matching in practice in section 4.6.2 helps to illustrate how a researcher might overcome this issue.

4.6.2 Example of PSM

To provide an overview of how propensity score matching methods work to ascribe cause and establish causal estimates, an example that sought to analyse the effect of going to a Catholic

[10] Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot be easily predicted due to the intervention of random variables (Kenton, 2021).
[11] A quasi-experiment is an empirical intervention study used to estimate the causal impact of an intervention on a population without random assignment.

school, as opposed to public school, on student achievement is considered here (Ejdemyr, 2020). They used PSM methods to create a balanced dataset, allowing a direct comparison of baseline covariates between treated and untreated participants before examining the effect of attending a Catholic school. The analysis was conducted in six systematic steps:

1. Pre-analysis: examine the difference-in-means between the treatment and control group for the outcome variable and the pre-treatment covariates
2. Estimate the propensity score (the probability of being treated given a set of pre-treatment covariates)
3. Examine the region of common support
4. Nearest neighbour PSM
5. Examine covariate balance after matching
6. Estimate treatment effects.

In order to give a clear understanding of how PSM methods were employed in this Catholic vs Public School example, each of the six stages above are described below.

Stage one was the pre-analysis stage that sought to determine the natural differences between the treated and the control group in terms of mathematics achievement and covariate balance. The pre-determined outcome variable, a standardised mathematics assessment score, was the measure used to assess mathematics achievement. Results showed that the third-grade Catholic school students' average mathematics score was more than 20% of a standard deviation higher than that of public-school students. T-tests were used to verify that the difference-in-means was statistically significant at conventional levels. The comparison of assessment outcomes alone cannot be used to answer the research question since the two groups differed systematically. For example, one might expect pupils with better resources to disproportionately attend Catholic schools.

The pre-analysis stage also analysed the differences in covariate balance between the treated and control groups. Five covariates were chosen for this research: race, mother's age, family income, number of places the student has lived for at least four months, and mother's education level. These covariates were chosen since the researcher deemed them to be related to both the treatment assignment and potential outcomes (Ejdemyr, 2020). The mean for each covariate was calculated according to treatment status, and it was found that there was an imbalance in the covariates across the two groups.

Stage two of the process was the propensity score estimation. Based on the set of covariates listed above, the propensity score for each student was calculated. A propensity score is the student's predicted probability of being assigned to the treatment group, given the set of covariates.

Stage three was to examine the region of common support. This is commonly used in PSM to determine where there is overlap of probabilities between the treated and control groups. That is, people who were just as likely to attend a Catholic school as a public school on the basis of the factors chosen. For example, consider student A and B who went to a Catholic school and a public school respectively. Because their propensity scores are similar and thus both are in the region of common support, it can be considered equally likely that each of them might have attended the other school. As such, student B's outcome acts as the outcome that student A would have had if they had attended a Catholic school (and vice versa). In Ejdemyr's research (2020), histograms were plotted to provide visual representations of the propensity to attend a Catholic school for the two groups. Figure 4.1 shows that, without any matching, the distribution of propensity scores differed between the two groups.

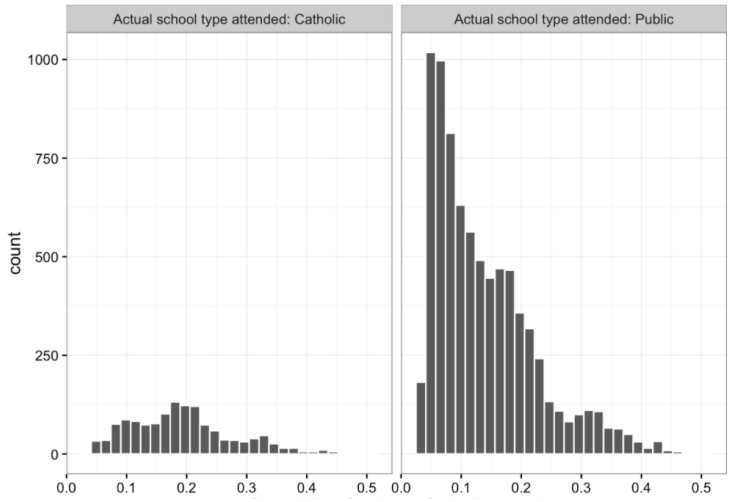

Figure 4.1: Propensity to go to Catholic school (Ejdemyr, 2020)

The next step, stage four, of the process was to execute a matching algorithm. In order to effectively estimate the treatment effect of attending a Catholic school on mathematics attainment, the sample was restricted to observations within the region of common support, where students had comparable propensity scores. Whilst there are many different types of matching algorithms, some of which are discussed in the subsequent sections of this chapter, the researchers in this instance opted to find pairs of observations that had very similar propensity scores, but that differed in their treatment status. Figure 4.2 below helps to exemplify how this might work in practice:

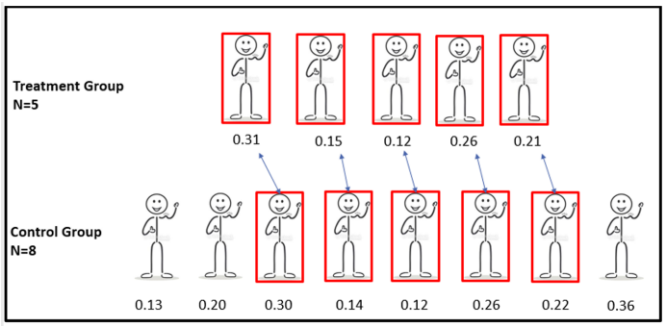

Figure 4.2: one-to-one matching example (Gant and Crowland, 2017)

Students that have connecting arrows between them are students who have similar propensity scores but are different in their treatment status. As the propensity scores are calculated based on covariate profile and treatment allocation, students that are connected can be deemed as "matched" and therefore they can be compared on their outcome measure. The three students in the control group from this image that have not been matched are disregarded from the sample since there are no matches found in the treatment group.

This is a simplified demonstration of how the matching process works in practice, and there are further technical considerations when deciding on how to match, not least how the researcher determines what a 'good match' is. These technical considerations are explored in the next chapter. In the context of the Catholic school example, following the matching process, the final dataset was smaller than the original due to statistical relevance.

Following the execution of a matching algorithm, covariate balance in the matched sample was then assessed through visual inspection and t-tests of difference-in-means (stage five). For the visual inspection, the mean of each covariate was plotted against the estimated propensity score, separately by treatment status. When matching had been successful, the treated and control groups had similar means of each covariate at each value of the propensity score. Figure 4.3 shows the overlap between propensity scores for both groups for the race covariate where 0 indicates students that did not attend Catholic school and 1 indicates students that did.

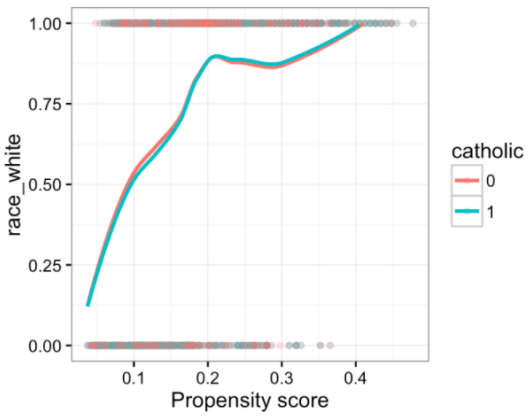

Figure 4.3: overlap of propensity scores (Ejdemyr, 2020)

To supplement visual inspection, the example analysed the difference-in-means for the newly matched group. It was found that matching had attained a high degree of balance on the five covariates in the model, since the means for each covariate, according to treatment status, were much more aligned than they were before matching.

Once it was clear that covariate balance had been achieved, stage six estimated the treatment effect of attending a Catholic school on mathematics performance. T-tests and ordinary least squares (OLS) were used to analyse the mean mathematics score for the treated and the control group with and without the covariates. The result suggested that, for those with a propensity to go to either Catholic school or not, the effect of Catholic school was negative (~ 0.14 of a standard deviation). This particular example showed that quantitative methods of observational data worked to ascribe cause.

4.6.3 Cause and Assumptions

The logic of PSM is that based on a composite score (the probability of belonging to a group), student A from treatment X can be matched with student B from treatment Y who can act as their 'counterfactual' because to all intents and purposes, they could have been assigned to the other group based on the basis of the key covariates. As such, the outcome of student A who had treatment X can act as the counterfactual for student B who had treatment Y because on the basis of the covariates which are strong predictors of treatment type, student A and B were equally likely to receive each treatment.

Propensity Score Matching therefore establishes cause through counterfactual analysis following the implementation of a rigorous and effective matching algorithm. Counterfactual analysis, as described above, enables analysts to attribute cause and effect between

treatment/intervention and outcome. The 'counterfactual' measures what would have happened to participants if they had been assigned to the other treatment. The impact is estimated by comparing counterfactual outcomes across the two treatment groups.

PSM is reliant on several assumptions, the key one being that, conditional on some observable characteristics, participants from different treatments can be compared, as if there has been randomisation. It has been claimed that PSM, when done effectively, mimics the attributes of an RCT by creating a counterfactual group so that the impact of a treatment or intervention can be measured (Abaasa et al., 2021). However, it is important to acknowledge that propensity scores are not a substitute for randomisation (Sainani, 2012). Propensity scores only balance measured, not unmeasured, confounders. The researcher can never be completely sure that they have controlled for all possible confounders so that any difference in outcome measure is definitely due to the difference in treatment.

Whilst this research plans to use a range of factors that have been found to be predictive of student attainment in the propensity score model (these are discussed in the next chapter), the results may still be affected by omitted variable bias. For example, mathematics self-concept[12] has been found to influence achievement in mathematics but is not a variable that can be accounted for in the PSM model for this research since the data is not readily available to schools (Wang, 2007). As such, it is important to consider various approaches to ensure as many confounding variables are accounted for as possible and ensure results are not biased. One way to overcome this is to use interviews to gain contextual insight into the participating schools. It was acknowledged earlier in this section that like RCTs, PSM only discounts differences between groups at the point of allocation. Interviews can be used to check for other post-allocation differences. The use of interviews is discussed and detailed in the next chapter.

4.7 Conclusion

This chapter has outlined a variety of possible methods to ascribe cause. Given the concerns regarding trial effects for an experimental RCT, and given that there already is one for TfM, an observational approach was pursued in an attempt to answer the research question. Of the observational approaches: multiple regression, interrupted time series, and propensity score matching, the latter is the most appropriate for answering the central question. ITS models that rely on longitudinal data are not appropriate when trends are not linear or when the intervention is introduced gradually or at more than one time point. Multiple regression similarly assumes that the

[12] Self-concept can be described as student ratings of their skills and ability

relationship between the independent variables and the dependent variable is linear, whilst also assuming that there is no correlation between the independent variables. A method which can ensure there are no systematic differences on key observed covariates is much more appropriate for this research. Thus, PSM is the best approach for the research question that this paper seeks to address.

Chapter Five: Methods

5.1 Introduction

The previous chapter explored different methods that could possibly be used to effectively ascribe cause. The chapter concluded that PSM was the most appropriate method to answer the following research question that seeks to establish cause since participants can be matched on propensity scores that account for key covariates:

"How might non-experimental methods be adapted to address whether Teaching for Mastery is associated with improved assessment scores, compared to previous pedagogical approaches?"

This chapter begins by outlining the chosen design for this research before discussing the practicalities of implementing PSM methods in this particular case, outlining six clear steps, whilst also describing the technical considerations associated with the method. Therefore, it will serve as an overview for this research and concludes with a visual flow-chart to help illustrate the steps involved in the research.

5.2 Design

Since this research seeks to investigate the effectiveness of TfM with methods that are not impacted by trial effects, an observational approach is needed as it is not feasible to run an RCT. As Kontopantelis (2015) remarked, "well designed observational studies can contribute greatly to the knowledge base, albeit with careful attention required to assess potential confounding and other threats to validity" (p.4). Since student assessment outcomes can be determined by a number of factors, the observational approach chosen will need to isolate and control variables by matching TfM and non-TfM individuals from data sets.

An observational design eliminates the risk of experimental effects (since there is no experiment) and since the focus is on existing data on student outcomes, there is no risk of subjects attempting to change their behaviour for the purposes of an experiment. That said, it is important to acknowledge that there is always risk of a 'change effect' when schools transition to a new curriculum.

The observational research design of this study predominantly utilises quantitative methods, seeking to evaluate the efficacy of TfM, compared to a previous alternative, by using PSM to control

for confounding variables. There has been limited research to date of such type, and therefore, the chosen design and methods help fill a gap in the existing evidence base.

The quantitative methods are complemented by qualitative methods which seek to gain contextual insight into the participating schools and the view of teachers from those schools, which provide a background to the quantitative data. The interviews were to ensure that participating schools can act as a case of non-TfM to TfM transition and would reveal the potential impact of other post-allocation causes. There were three participating secondary schools for this study, and an overview of these schools can be found in the results chapters of this thesis.

5.2.1 Qualitative Methods

The qualitative methods comprised semi-structured interviews that aimed to explore the school's understanding of TfM and their transition to the pedagogical approach, as well as its previous teaching approaches. The interviews also gained insight into the school's demographics and sought to ascertain if there were any other contemporaneous events, such as change in catchment area, that may have impacted student performance around the time of the transition to TfM. Whilst these interviews do not directly address the research question, the contextual insight is crucially important because there is no universal definition of TfM, and the approach may differ by school. Additionally, it is important to be able to discuss the type of data needed from the school and what possible outcome measures they may be able to provide.

The interviews were semi-structured; left open but guided by a list of key questions the researcher intended to ask. A benefit of semi-structured interviews is that they are flexible and allow new questions to be generated since the interviewee is able to determine the direction of the conversation (Sekaran, 2003). The selected interview participants were the Head of Maths or TfM lead at the participating schools. They were recruited via email and were given an overview of the purpose of the study in the first contact (see Appendix A).

The interviews were conducted via Zoom, lasting around 30-40 minutes, and were recorded, with the participant's approval. A summary of the interview findings are given in the results section of this thesis.

The interviews had four key aims, and the questions planned that provided a framework can be found in Appendix B. The key purposes were:

- To gain contextual insight into the school and mathematics department, particularly focused on their transition to TfM
- To understand the rationale behind their transition to TfM, how and when they undertook it
- To gain insight into what they perceive TfM to mean, and how their curriculum and departmental pedagogy differs to their previous curriculum
- To ascertain if the participating school had the required data needed for propensity score matching and to compare assessment scores of cohorts pre- and post- TfM implementation.

A limitation of an observational approach is the inability to control other contemporaneous events that may otherwise explain any changes in student outcomes. Therefore, it was important to use the interviews to gain contextual insight and to determine whether there had been any other major departmental or school changes since their implementation of TfM that could otherwise account for student outcome variation, such as change in leadership, ethos, or additional continuous professional development opportunities.

It was important that the participating school(s) were ones for which the transition to TfM was the main change since it is impossible to disaggregate change in curriculum from other school changes. If, during the interviews, it came to light that a participating school had very substantial simultaneous changes like the ones described, they would not have been suitable for the research. At the time of this research, Covid was a contemporaneous event that had impacted student outcomes. The only way to disaggregate the effects of Covid for this research was to look at data for TfM and non-TfM groups that existed pre-2020.[13] The details of the data provided by the participating schools is discussed in the results chapters.

Since there is not an established universal definition for TfM, each of the three participating schools had implemented the approach differently and, as such, the transition to a mastery approach differed from school-to-school, and depended on their unique contextual background. As such, it was not reasonable to do a meta-analysis that combined findings from the participating schools. Instead, each school has formed its own individual and separate case, yielding its own results. From this, it has been possible to draw suggestions from causal conclusions at a very local level for each individual institution. In order to build up a picture of each school's individual case, it is important to build up a depth of contextual understanding by gathering the views and experiences of teachers relating to each case (Hamilton, 2011). Therefore, whilst this research is predominantly concerned with analysis through quantitative methods, part of the design is qualitative to help bring

[13] Schools closed in March 2020 due to the Covid-19 crisis and subsequent national lockdowns between then and 2022 has caused disruption in the education sector.

together a holistic description and analysis of each school in question (Merriam, 1985). Following this research, if it is deemed that the non-experimental approaches trialled can assess the impact of the transition to TfM, then the research can pave the way for future studies where a meta-analysis would bring together results from individual separate cases.[14] The ultimate aim of this research is to explore if and how such non-experimental approaches can be adapted to assess the impact of TfM.

Using a methodology which treats each school as a separate individual case necessitates confronting the issues of validity, reliability, and generalisability (Merriam, 1985). According to Ekanayake (2015), data collection and analysis methods are important aspects in maintaining these facets when exploring distinct cases. Literature around "case studies" in educational research acknowledge the common criticism that there is little evidence for scientific generalisation and therefore little external validity (Burns, 2000). However, by ensuring good levels of validity and reliability of the data collection methods and data analysis techniques, as well as using multiple cases, it has been suggested that analytical generalisations can be made, to expand an explored theory (Merriam, 1985; Yin, 2003; Burns, 2000).

5.2.2 Quantitative Methods

To reiterate, the aim of this research is to explore methods for establishing if the transition to TfM is associated with improved assessment scores, compared to previous pedagogical approaches. The focus is on embedded TfM approaches and therefore the methods employed must be observational to consider the long-term impact. Whilst part one of the methods are qualitative to gain contextual insight into participating schools, a significant part of the methods must be quantitative to ensure strong internal validity. The two parts supplement one another to provide an insightful and in-depth case for each participating school.

Since PSM methods have been found to effectively match on covariates that predict treatment allocation and outcome measures, these methods are the most suitable for the quantitative part of this research. As argued in the previous chapter, propensity score methods can help to reduce bias in treatment effect estimates from observational studies when participants are not randomly assigned to conditions in the same way that they are in experimental studies (Leite, 2016).

[14] An in-depth exploration of this idea can be found in the discussion section of this thesis.

5.3 Steps for Propensity Score Matching

Although section 3.6 detailed an example of PSM in practice, it is important to outline the key stages of a PSM approach that are used as the framework for this research. According to Leite (2016), there are six main steps in propensity score analysis:

Step 1: Data Preparation and Missing Data

In the first step, the researcher must examine the data available, identify missing data and determine how to react. Propensity score analysis, according to Choi et al. (2018), is a popular method to control for confounding, but a significant challenge is dealing with missing values in those confounders. Choi et al. compiled guidance for researchers in choosing the best method for dealing with missing data. The four methods compared were: complete case analysis, missing indicator method, multiple imputation and a combination of multiple imputation and missing indicator methods. The study showed that complete case analysis (cases are deleted if the data are incomplete) provides an effective balance between simplicity and unbiasedness, albeit that it comes at a cost of reduced precision. Multiple imputation, where the researcher would make a 'best guess' about the missing value on the basis of the values that do exist and the estimated relationships between variables on the remaining data. This allows missing data to be put back in the data set and thus improves precision, but at the risk of bias of the imputation method. Choi et al. outlined that multiple imputation may fail when data are missing not at random. Therefore, there is a need to correctly specify the imputation model which can be a complex procedure. Where the missing data are deemed to be small or moderate (i.e., < 25%), it seems reasonable to argue that the potential bias is too high a price to pay for the rather moderate increase in precision. As such, in the results section of this research, where the level of missing data is small or moderate, complete case analysis is orchestrated, removing any units that have any missing covariate data from the sample.

As well as considering how best to deal with missing data, in the first stage of propensity score analysis, the researcher must also select the covariates by identifying variables that are 'true confounders'. A true confounder is a variable that is related to both treatment assignment and the outcome measure.

Within this first step, pre-registration takes place. Pre-registration in this context consists of specifying the steps that are taken as part of the propensity score methods ahead of observing the data. The practice of pre-registration in many disciplines is currently expanding so that researchers

gain a clear understanding of research goals and processes (Nosek et al., 2018). For this research, it is important to have a clear outline of the steps involved in PSM including matching methods and calipers[15] to avoid any bias. Gelman outlined the danger of 'the garden of forking paths' when conducting propensity score matching. Because there are so many opportunities open to the researcher to adjust their analysis, they might not be able to avoid landing on the approach that best suits their viewpoint even when trying to remain 'unbiased' (Gelman and Loken, 2013). As such, there is a need to pre-register the analytical steps to be taken before any PSM is undertaken. Pre-registration refers to specifying and recording study plans before observing the data.

Step 2: Propensity Score Estimation

In the second step, statistical methods are used to determine an estimate for the propensity score to check the initial imbalance before any matching takes place across the two comparison groups. Each individual receives a unique score that summarises the relationship between the covariates and the treatment assignment.

Which covariates to include is a key consideration when using propensity score methods. The covariates must be ones that predict the treatment allocation as well as being associated with the outcome measure. Thus, a covariate is a possible predictive or explanatory variable. According to this, any variable that is measurable and considered to have a statistical relationship with the dependent variable quantifies as a potential covariate (Salkind, 2010). In the context of this research, since different cohorts within the same school were of interest, the predictors were characteristics that were likely to change year-on-year naturally, or ones that were specific to a school which may have shifted demographics.

Research has suggested that to reduce the potential for selection bias, many covariates should be included in the model predicting propensity to receive treatment, even those that only weakly predict the treatment (Shadish et al., 2002; Rubin, 1997). Given that this research is concerned with pupil assessment data from schools, the covariate data depended on what each school could provide. Participating schools were asked to provide as much relevant data as possible on the cohorts of students. The list below was given to schools as a guideline, but not as an exhaustive list:

- Prior attainment scores (e.g., KS2 scores that can be used as a benchmark measure)

[15] Calipers can be used when performing PSM to define how close the paired units can be on their propensity scores in order to be matched.

- Gender
- Ethnicity
- Pupil Premium status[16]
- Free School Meal status
- SEND status
- EAL status.

The selected covariates listed above have been guided by a report published by the DfE which analysed students' GCSE results at the end of Year 11 in relation to a wide range of data relating to individual differences, family and home environments (Sammons et al., 2014). Those listed are ones for which schools have readily available data. Whilst Sammons et al. (2014) has identified further variables to have large influences on GCSE attainment, such as parental income or education level, schools are not likely to have these on record.

The DfE's report found that students eligible for Free School Meals (FSM) had lower average academic attainment compared to those who are not (Sammons et al., 2014). Specifically, it found that students eligible for FSM achieved, on average, one grade lower in GCSE Maths than those students not eligible. With regards to Special Educational Needs (SEND), it was found that students with a full SEN statement achieved the lowest average Maths results, as well as being entered for the lowest number of GCSE exams (Sammons et al., 2014). It was also found that there were, at the time, differing attainment levels for different ethnicity groups. Students of mixed heritage obtained the lowest average total GCSE score, whilst students of Bangladeshi heritage had the highest average results (Sammons et al., 2014). It was students of Pakistani heritage that obtained on average the lowest grades in GCSE English and Maths. Hattie's (2017) list of factors that impact on student learning rank ADHD, a special education need, as having the highest effect, with parental income, gender and ethnicity also featuring.

The issue of gender and attainment needs to be considered separately, since it is not clear cut. The DfE's report found that, whilst female students on average obtained a higher total GCSE points score, there were no statistically significant gender differences in the average grade achieved in GCSE Maths (Sammons et al., 2014). An analysis of the 2019 GCSE results revealed a 9.8% gender gap, with 71.7% of females achieving a grade 4 or above, compared to only 62.9% of males (Ofqual, n.d.). Some educational experts have proposed that the gender gap is less prominent in STEM

[16] Whilst there are different categories for 'pupil premium', in this research, a student was marked as 'pupil premium' if they fall into at least one of the categories. Full details of pupil premium eligibility can be found here: https://www.gov.uk/government/publications/pupil-premium/pupil-premium#pupil-eligibility-and-funding-rates-2021-to-2022

subjects, such as Maths. However, 'gender' will still be considered for the propensity score since a gender gap has been found in GCSE outcomes. Furthermore, in a study that investigated mathematical performance, gender was found to be one of the influencing factors on student achievement (Kenney-Benson, Pomerantz, Ryan, & Patrick, 2006).

KS2 prior attainment is the final confounding variable on the list above and is important to consider in the PSM process since research has found that a student's KS2 mathematics attainment level is a predictor of their GCSE mathematics grade (Department for Education, 2018). PSM is able to quantify how much of a predictor KS2 prior attainment is on treatment allocation and outcome measure, along with the other covariates.

Whilst assessing students at KS2 with levels is now something of the past, the data published by the DfE (2018) showed that it was students who achieved a level 4 or 5 at KS2 went on to achieve the top grade 9 at GCSE. Furthermore, for students who achieved the lower level 1 at KS2, none of them went onto achieve anything higher than a grade 4 at GCSE (Department for Education, 2018). Figure 5.1 shows the correlation between KS2 attainment level and GCSE mathematics grade for students that sat their GCSEs in 2016 and 2017.

		GCSE mathematics grade[1]									
		9	8	7	6	5	4	3	2	1	No entry
Key stage 2 mathematics attainment level	W	0	0	x	3	4	5	6	10	16	82
	1	0	0	x	x	0	12	18	53	232	476
	2	x	14	30	54	141	398	786	2,243	5,613	3,648
	3	x	16	81	180	1,472	6,372	12,163	16,412	13,051	3,415
	4	179	1,837	7,058	15,527	49,020	74,076	42,651	22,275	8,444	3,722
	5	16,856	30,484	39,989	38,587	41,871	23,074	4,395	1,014	306	1,077
	No valid KS2 level	1,283	1,933	2,611	2,728	4,470	4,865	3,205	2,649	2,160	2,397

Figure 5.1: KS2 Attainment and GCSE Maths Grade (Department for Education, 2018)

The data collected from participating schools required details of the above confounding variables for each anonymised student. These data were collated into an excel data collection sheet provided to the schools as a template.

Using the covariate data, statistical methods are employed to calculate a propensity score for each individual student. To see the variety of propensity scores for each treatment, a visual evaluation for both groups through boxplots and histograms will prove useful.

The visual evaluation will help the researcher to see if there is a region of "common support" between the two treatment groups. It is important to ensure there is overlap in the range of propensity scores across the two comparison groups before any matching takes place (Garrido et al., 2014). Treatment effect inferences cannot be made if there is not a comparison individual with a similar propensity score (Garrido et al., 2014). Common support is subjectively assessed by visual

evaluation of graphs of propensity scores for the two comparison groups and observations outside the range of common support are discarded since they cannot be matched.

Following this, the propensity score matching process will ensure that cohorts of students that are to be compared on assessment outcomes are similar on the measured covariates. This is to ensure that any change in assessment scores cannot be attributed to confounding variables, and instead can be credited to the transition to TfM.

A key consideration with PSM methods is how one can be sure they have addressed all the confounds (Morgan, 2018). As Hattie (2017) has shown, there are more than 250 factors that impact student learning and achievement. These factors include aspects that teachers, school leaders, students or parents can influence and control, as well as ones that are outside of their control. Some of the factors in Hattie's list include boredom, sleep and how much television a student watches. Furthermore, in a study conducted to identify the factors affecting mathematics achievement of students through an opinion poll, it was instructional strategies, teacher competency, along with motivation and concentration that were found to be the most influential factors (Saritas and Akdemir, 2009).

With any study that seeks to analyse and determine how independent variables can predict a dependent variable, multi-collinearity is a risk to be aware of. It is a statistical concept whereby several independent variables in a model are correlated and, where high intercorrelations occur, results can be misleading and therefore less reliable when considering statistical inferences. It has been suggested that it is better, therefore, to use independent variables that are not correlated when building regression models that use two or more variables (Hayes, 2022). The factors mentioned above from Hattie (2017) and Saritas and Akdemir (2009) are difficult to quantify for each individual student. For this research, the covariates used were ones for which the data is most commonly collected by schools, basically those listed at the start of this section.

The propensity score is estimated by running a logit or probit model using the `glm()` function in R where the outcome variable is a binary variable indicating treatment status. Any covariate that is related to both the treatment assignment and the potential outcomes is included. Once the model is set up, the propensity score for each student is calculated. The score represents each student's predicted probability of being 'treated', given the estimates from the logit model. In R, the `predict()` function is used on the date frame that is created from the first propensity score estimation (see Appendix C, p. 178, lines 1-7 for the code).

Step 3: Propensity Score Method Implementation

In this third step, the researcher will seek to balance the propensity scores between the treated and untreated group. There are a few ways of doing this, and the most widely accepted methods are matching, stratification and weighting. In stratification, the observations are divided into 'strata' that have similar propensity scores, with the objective of balancing the observed variables between the two treatment groups. The treatment effect is subsequently estimated by combining stratum-specific estimates of treatment effect. For weighting, subjects are assigned different 'weights' which weigh them up or down to ensure the subjects in each treatment group are similar to each other. This way, all subjects are included and none are excluded from the sample; which is particularly important in the instance of small sample sizes.

It is the matching method that was used, and specifically which matching method, such as a matching ratio or a matching algorithm, varied depending on the data available from each participating school. Some of the matching methods, as outlined by Leite (2016) include: 1:1, greedy matching, paired matching, fixed ratio matching, variable ratio matching, nearest neighbour matching, genetic matching, optimal matching, and full matching. For many of these methods, a researcher must decide whether to match with or without replacement and whether to use specified caliper widths (Austin, 2012).

Matching

In a comparison of matching methods, Gu and Rosenbaum (1993) compared two main matching algorithms: nearest neighbour and greedy matching, as well as different matching structures such as 1:1 and 1:k matching. Austin (2012) acknowledged that before choosing a matching algorithm, it is important to consider the differences between matching without replacement and matching with replacement.

As such, the tables below give an overview of the differences between the matching algorithms and the different structures, along with further 'considerations' of each where any advantages, disadvantages, or any circumstances that may mean one is better than the other, are highlighted.

Matching Algorithms

Gu and Rosenbaum (1993) stated that the selection of a matching method involves three choices. The first of these choices is of an algorithm that assigns units to matched sets. Although their comparison only included greedy and optimal matching, nearest neighbour

matching is included here since Austin (2012) deemed it to be an effective algorithm for matching. Additionally, since Austin (2012) stated the importance of considering whether to match with or without replacement, the two approaches will also be outlined below.

Matching Algorithm	How it works	Considerations
Greedy matching	A treated subject is first selected at random. The untreated subject whose propensity score is closest to the above is chosen for matching. This process is repeated until untreated subjects have been matched to all treated subjects or until the list of treated subjects has been exhausted. For each step, the nearest untreated subject is selected for matching to the given treated subject, even if that untreated subject would better serve as a match for a subsequent treated subject.	The greedy algorithm does not generally minimise the total distance within pairs. Examples have shown that the greedy algorithm's distance can be much larger than the minimum attainable distance (Rosenbaum, 1989). The order in which the treated subjects are matched may change the quality of the matches. If the goal is to simply find well-matched groups, greedy matching may be sufficient.
Optimal Matching	Matches are formed to minimise the total within-pair difference of the propensity score. Takes into account the overall set of matches when choosing individual matches.	It has been found that optimal matching did no better than greedy matching in producing balanced matched samples (Gu and Rosenbaum, 1993). More effective in producing *close* matches than greedy matches (Gu and Rosenbaum, 1993). If the goal is to find well-matched pairs, then optimal matching is preferable over greedy matching.

Nearest Neighbour Matching	Selects for matching to a given treated subject that untreated subject whose propensity score is closest to that of the treated subject. If multiple untreated subjects have propensity scores that are equally close to that of the treated subject, then one of these untreated subjects is selected at random.	A method to select untreated subjects whose propensity score is "close" to that of a treated subject. No restrictions are placed upon the maximum acceptable difference between the propensity scores of two matched subjects.
Matching without replacement	Once an untreated subject has been selected to be matched to a given treated subject, the untreated subject is no longer available for consideration for future matches. Each untreated subject is included in at most one matched set.	When matching without replacement, the order in which the treated individuals are matched matters. When matching *with* replacement, inference becomes more complex since matched controls are often in the matched sample more than once. Matching without replacement makes inference simpler. The treatment effect estimate will be based on a larger number of controls than if subjects were replaced.
Matching with replacement	Allows a given untreated subject to be included in more than one matched set. Matched controls are not independent – some are in the matched sample more than once. The number of times each control is matched should be monitored.	When matching with replacement, variance estimation must account for the fact that the same untreated subject may be in multiple matched sets (Hill and Reiter, 2006). When matching with replacement, the order in which the treated individuals are matched does not matter. Can decrease bias because controls that look similar to many treated units can be used multiple times. Helpful when there are few control individuals comparable to the treated individuals.

Table 5.1: Matching Algorithms

Matching Structures

The second[17] of Gu and Rosenbaum's choices for matching is of structure; that is, whether each treated unit will have one or many controls, or a more flexible arrangement.

Matching Structure	How it works	Advantages/Disadvantages
1:1 or Pair Matching	Pairs of treated and untreated subjects are formed. Matched subjects have similar values of propensity scores.	Most common method used with the nearest neighbour algorithm (Austin, 2012). Each treated unit has one control (notion of the counterfactual). Can discard a large number of observations.
Many-to-one Matching (M:1)	M untreated subjects are matched to each treated subject.	Improved bias reduction has been found when matching with a variable number of controls compared to matching with a fixed number of controls (Ming and Rosenbaum, 2000).
One-to-many Matching (1: k)	Each treated unit has k controls.	Forcing every treated unit to have k controls can create some poor matches (Gu and Rosenbaum, 1993).
Full Matching	Involves forming matched sets consisting of either one treated subject and at least one untreated subject, or one untreated subject and at least one treated subject.	A more flexible structure in which a matched set may contain either a single treated unit and several controls or a single control and several treated units. In some applications, all controls are matched to treated units, whereas in other applications, some potential controls are discarded. Considered the "optimal structure" (Gu and Rosenbaum, 1993). It can increase balance without discarding any units.

Table 5.2: Matching Structures

<u>Calipers</u>

When conducting the matching method, it is important to consider how close a treated and untreated subject can be on propensity scores for them to be deemed a suitable match. Calipers can be used when performing propensity score matching to define how close the paired units can be on their propensity scores. Figure 4.2 in section 4.6.2 gave an example where the caliper was 0.01. A

[17] Gu and Rosenbaum's (1993) third choice of matching is about the distance.

caliper is a distance restriction where the absolute difference in the propensity scores of matched subjects must be below some pre-specified threshold which acts as a maximal distance (Austin, 2012). For a given treated subject, one would identify all the untreated subjects whose propensity score lay within a specified distance from that of the treated subject. From this restricted set of untreated subjects, the one whose propensity score was closest to that of the treated subject would be selected as the match. When calipers have been used in matching methods, they have been found to produce less biased results when estimating the treatment effect of an intervention, compared with matching methods that did not use a caliper (Austin, 2012).

The use of a caliper depends on the relative size of the treated and control samples. Where the number of treated subjects is significantly less than the number of control subjects, and where there is a large area of common support and intervention group sitting inside the control group in terms of the range of propensity scores, a caliper may not be needed since nearest neighbours are always "nearby". But, if the number of subjects in the two samples are approximately equal, there is a greater risk of matching subjects with very different propensity scores and therefore, a caliper is required.

The image below shows an example of how using a caliper can help to ensure that treated and untreated subjects are only considered a match when their propensity scores are close. The difference in propensity scores between each matched treated and untreated subject is a maximum of 0.2. Had a caliper not been defined, a treated subject with a propensity score of 0.65 could have potentially been matched with an untreated subject with a propensity score of 0.2 if the aim of the matching method were to match all subjects.

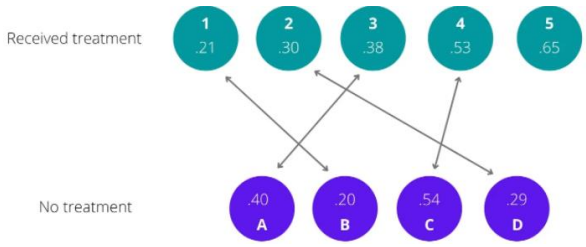

Figure 5.2: Matching with Calipers (Freitas, 2021)

It has been argued that selecting an appropriate caliper is essential for achieving good balance when using PSM methods (Lunt, 2013). However, there is no uniformly agreed definition of what constitutes an acceptable caliper distance (Austin, 2012). In a study that used simulations to explore the effects of decisions made surrounding matching methods and caliper distances, it was found that a 'tighter' caliper can greatly reduce bias, but can leave more subjects without a match

since matches are only made when the subjects propensity scores are very similar (Lunt, 2013). When using calipers, it is possible in principle that no matched sample exists. Therefore, there must be a balance between getting valid matches of treated and untreated subjects, whilst also not disregarding too much data that the sample size becomes too small to make any statistical inferences. According to Lunt (2013), a tight caliper is preferred when matches are easy to find (for example when there is little difference between treated and untreated subjects on the observed covariates), whilst a looser caliper should be used when matches are harder to find.

The specifics of the matching method used, and the caliper are discussed for each participating school in the subsequent chapters.

Step 4: Covariate Balance Evaluation

It is this step which serves as the main measure of success of the propensity score method as the quality of matches made in the previous stage are assessed. Within this step, the covariate balance is evaluated and if adequate balance of covariate distribution between treated and untreated samples is achieved, then the propensity score can be considered successful (Ho, Imai, King, and Stuart, 2007).

Balance can be evaluated through visual inspection, difference-in-means and t-tests. For visual inspection, the mean of each covariate is plotted against the estimated propensity score, separately by treatment status (Ejdemyr, 2020). Where matching has been effective, near identical means of each covariate at each value of the propensity score should be seen. Assessing balance through difference-in-means involves analysing the mean of each covariate for both treatment groups and assessing how close they are. A two-sample t-test formalises these inspections and the aim is to not be able to reject the hypothesis of no mean difference for each of the covariates.

If substantial balance between the two treatment groups is not achieved, then the researcher must return to step 3 and try a different matching method and/or a different caliper until there is adequate balance of propensity scores.

Step 5: Treatment Effect Estimation

Once the researcher has determined that the treated and untreated groups are balanced on the observed covariates, the treatment effect can be estimated. In this research, the treatment effect is TfM, and the outcome measure is the assessment scores for both groups. As discussed earlier, the assessment scores must be a consistent outcome measure, meaning that both groups

will have been subjected to an identical assessment. The need for a consistent outcome measure can be restrictive when recruiting potential schools for participation as some school refine their assessments over time.

Since the assessments for both treatment groups should be the same, the aim in estimating the treatment effect is to identify counterfactual instances for each individual in one of the two treatment groups (TfM and non-TfM). If student X who received treatment A and student Y who received treatment B are matched on their propensity scores under the prescribed caliper, then the assessment scores can be compared to determine if TfM can indeed be associated with improved performance for that assessment.

To estimate the treatment effect, different statistical methods can be used such as weighted mean differences and weighted regression, as well as more complex models such as multilevel models or structural equation models. Specifics on how the treatment effect is estimated will depend on the form of matching that is performed (Leite 2016; Griefer, 2022). For example, after 1:1 matching without replacement, paired t-tests or a simple regression of the outcome on the treatment in the matched sample can estimate the treatment effect. Regardless of how it is estimated, the treatment effect in this research is the average treatment effect on the treated (ATT); looking at the subset of the sample that received the 'treatment'.

Step 6: Sensitivity Analysis

The final stage, which was initiated by Cornfield et al. (1959) following a study on cigarette smoking and lung cancer association, determines the level of robustness of treatment effects to hidden bias. The aim is to show that significance tests would not change even with large levels of hidden bias from omitted covariates so that the confidence on the treatment effect is strengthened and consequently, the findings can be deemed valid. This is a critical stage since propensity score methods only remove selection bias due to observed confounders. In the context of this research, sensitivity analysis is carried out by conducting a further treatment effect estimation using a different matching specification to assess if the results are robust to different matching specifications. If the results are robust, then similar results are yielded from more than one version of matching.

5.4 Methods Adoption

To reiterate, this research follows an observational design that uses mixed methods. The first part of the research is to conduct interviews with participating schools to gain contextual background and an in-depth understanding of the transition to TfM. The more sizeable part of the methods is quantitative, using PSM to create matched groups on a set of covariates in an attempt to ascertain if a transition to TfM can be associated with an improvement in mathematics assessment scores.

R is the chosen statistical package used to conduct the quantitative methods for this research. There are many additional packages that can be used for propensity score analysis, such as *Matching*, *MatchIt, Twang, Survey,* and *Mice*. Through exploration, *Matching* and *MatchIt* have proved user-friendly and since *MatchIt* aggregates functionality from several other packages. R provides access to many alternative methods for propensity score estimation and matching.

It is important to recognise that matching is an iterative process, and although matching on the propensity score is mostly effective at eliminating differences between the treatment groups and achieving covariate balance, sometimes there remains an imbalance and in this case, a different matching specification must be tried (Griefer, 2022). As such, the researcher has not committed to one type of matching for each case, but instead, has trailed multiple specifications.

Following the stages above, results are reported for each individual school, detailing the matching procedure and decisions made throughout the process. For each case, the propensity scores pre- and post- matching are communicated as well as the method of estimating treatment effect, along with a standard error or confidence interval (Thoemmes and Kim, 2011). The exploration of each case culminates in the estimate of the treatment effect for the outcome measure specific to the induvial school.

The quantitative methods outlined above help to answer the research question since it allows an analysis of student outcomes overtime. PSM helps to disaggregate other causes of potential changes in student outcomes, the model will enable a comparison of student performance. By analysing pre-existing data, under observational conditions, there is no “trial” and therefore no risk of bias or trial effects since the participating schools will already have the data to be analysed. Whilst the methodology is primarily quantitative, it is important to gain contextual insight into the participating school(s) through interviews in the initial stages of the research to understand how and when TfM has been implemented. The flow-chart below helps to visualise the research methods orchestrated in this study.

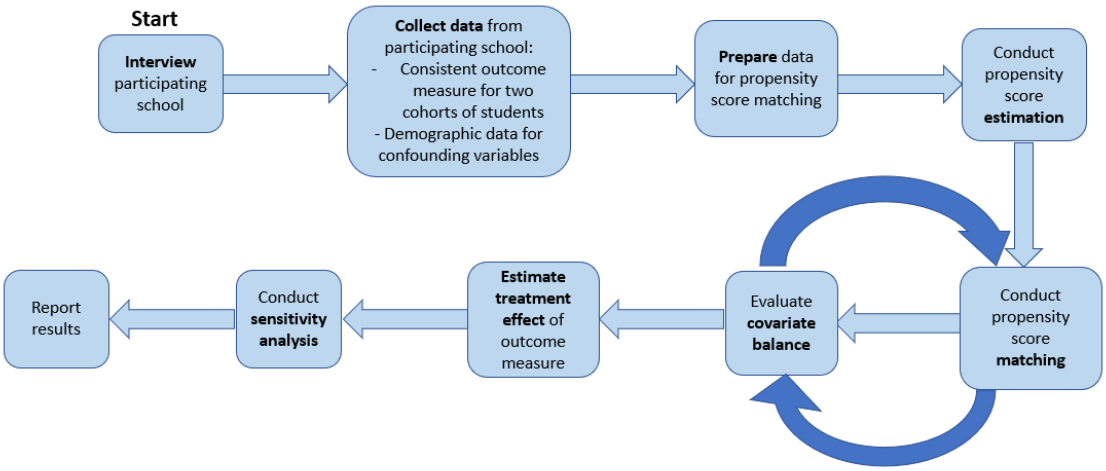

Figure 5.3: flow-chart of the research methods orchestrated in this study

This process was repeated for each participating school to build up the individual school cases.

5.5 Ethical Considerations

Research ethics are crucial components of any research process, which cannot be conducted without ethical compliance approval (Shawa, 2017). Therefore, as part of this research, a range of ethical issues have been considered.

The British Educational Research Association (BERA) highlighted that the first step to obtaining consent is for researchers to ensure that participants understand the process that they are engaging in, why their participation is required, who will use the research findings, and how they will be reported (BERA, 2011). Alongside informing participating schools and teachers of the purpose of the study and being transparent with the format of the research, the participating schools were informed that they had the right to, at any stage, withdraw from the study.

Due to the quantitative nature as part of the research, potentially personal identifiable data was anonymised. The participating schools were also anonymised, and the data sent for analysis had a file name that kept the school's name anonymous.

Data dissemination addresses a range of aspects such as confidentiality, anonymity, and the extent to which data can be reported back and future usage. Descriptive narratives of the schools involved were not written up to ensure that the participants remained anonymous, and that the data could not be linked back to the specific school. Whilst the interviews were recorded via zoom,

the name of the interviewee or the participating school were not discussed and the file name of the recording was named in a way that anonymity remained.

As part of the ethical considerations, two applications were sent to Durham University for approval at both the interview and the data collection stage. The first approval was granted on 19/07/2021 by the Education Ethics Committee, and the research supervisor. At this stage, approval to conduct the interviews was granted. The second approval was granted on 15/11/2021 and the data collection was categorised as ethically low risk.

5.6 Conclusion

These methods will give an insightful analysis to the research question, primarily because the innovative approach allows an investigation of change over time, where embedded TfM is the subject of analysis. Previous research into the efficacy of TfM has been centred on a short-term dose of 'mastery' which cannot be disaggregated from possible experimental effects and therefore, little has been done to investigate the longer-term impact of embedded TfM. Where observational PSM methods have been used, following the England Shanghai Exchange, they restricted to the short-term impact.

Part one of this research is the semi-structured interviews and part two is propensity score matching. Part one will supplement part two; the contextual data gathered from interviews gives the quantitative data of student assessment scores meaning.

Chapter Six: Implementation of Methods

6.1 Overview

The next three chapters take each of the three individual school cases and their data, using the methods outlined above, to see if a change in assessment outcomes can be associated with a transition to a mastery approach. It is important to look at multiple schools, in this case three, since this research is predominantly a methodological one, where alternative observational methods have been trialled for the first time since Boylan's study as a way of analysing the effects of TfM on student outcomes. The first case is used as the 'pilot study' for the subsequent two cases, since it is important to pre-register the steps intended to be taken in a propensity score analysis to avoid bias.

Before embarking upon the results chapters, it is useful to view the key differences outlined below between each of the three school cases and to be aware of what the intended learning from examining each set of results is. Table 6.1 illustrates the differences and the cohorts highlighted in yellow are the ones which were taught using TfM approaches:

	School X	School Y	School Z
Sample	660 students	602 students	1144 students
Covariates – binary	Gender FSM PP SEND Ethnicity	Gender FSM PP SEND EAL	Gender FSM PP SEND EAL
Covariates - continuous	KS2 Scaled Maths Score	MidYis assessments: - Vocabulary - Maths - Non-verbal - Skills - Overall KS2 Scaled Scores: - Grammar - Maths - Reading	Y7 Entry Assessment
Outcome measure(s)	Y7 Unit 1 Assessment Y7 Unit 2 Assessment	Y9 Unit Assessment Y10 GCSE Paper	Y7 Unit Assessment Y10 Half-Term Assessment
Cohorts	2017-18 Y7 cohort 2020-21 Y7 cohort	2018-19 Y10 cohort 2019-20 Y10 cohort 2020-21 Y10 cohort	2013-14 Y7 cohort 2014-15 Y7 cohort 2015-16 Y7 cohort 2016-17 Y7 cohort 2018-19 Y10 cohort 2019-20 Y10 cohort

Table 6.1: Overview of data from each school

The covariate data provided by school X and Z is predominantly binary, whereas the data provided by school Y is a mixture of continuous and binary. When taken through the pilot study which uses school X, it is evident that there is a need to include as many continuous covariates as possible to ensure that the propensity scores are well distributed between 0 and 1. School Z provided a larger sample size than school X and Y and with the year 10 cohorts being the same students as the 2014-15 and the 2015-16 year 7 cohorts, this analysis allowed for an exploration of the longer term impact of TfM.

The structures of the next three chapters is as follows:

- Chapter Seven: Implementation of Methods – Pilot Study

This chapter takes school X with the sole purpose of trialling the chosen methods. The chapter culminates with a pre-registration of the steps to be taken for the following two chapters.

- Chapter Eight – Implementation of Methods – Main Study (Part 1)

This chapter uses the data from school Y to investigate the effects of TfM and is split into three different analyses.

- Chapter Nine – Implementation of Methods – Main Study (Part 2)

This chapter uses the data from school Z to investigate the effects of TfM and is split into two different analyses.

The number of analyses conducted for each participating school was determined by the data provided, and this is detailed in each of the chapters.

Chapter Seven: Implementation of Methods – Pilot Study

7.1 Introduction

It has been highlighted that with PSM methods there is enormous flexibility in its implementation and therefore it is appropriate to use one set of data from the participating schools as a 'pilot study' where different matching methods, combination of covariates, and calipers are tried and tested using the MatchIt package in R Studio. Following the pilot study, it was then possible to pre-register the statistical process that is used for the remaining participating schools.

It was also acknowledged in earlier chapters that PSM methods are only effective when all known differences between the two treatment groups are controlled for. For school X, the treatment group was subjected to a period of interrupted schooling due to the Covid pandemic. Since this cannot be controlled for through matching, this chapter should not be considered as a TfM evaluation. Instead, it was useful as a test of PSM methods that were used to evaluate TfM in chapters eight and nine. With PSM, as well as with an RCT, pre-allocation between-group differences are accounted for by matching on propensity or random allocation, but post-allocation differences which are not changes in curricula cannot be ruled out as potential causes for differences in student outcomes. Whilst interviews are used in this research to gain contextual insight and rule out the possibility of unknown confounders, the effects of the disruption due to Covid are ones which cannot be disaggregated. This chapters' sole purpose therefore is to explore PSM methods using real outcome and covariate data to inform the plans for the main study. Any results derived cannot be taken as an evaluation of the transition to TfM in school X.

7.2 Objectives

The purpose of the pilot study is to confirm the efficacy of the steps to use for the main study. Since there is enormous flexibility with PSM, it is also important to determine a direction for the main study, setting out a series of steps for the decision making involved in choosing which covariates to include, which matching methods to try and which calipers to use. Section 7.19 will discuss and detail the pre-registration following the pilot study and an example of the code used is provided in the Appendices.

7.3 Interview and School Context

School X[18] is a comprehensive secondary academy school and is the lead school to its academy trust, catering for students from age 11 to 16 for boys and girls. There are approximately 1100 students on roll, and the most recent Ofsted inspection, which was in 2009, graded the school as 'Outstanding' in all areas. DfE performance tables show an overall progress 8[19] score for the school as being above average.

The interview for School X was conducted with Teacher X who has worked at the school for 7 years and is Assistant Headteacher and Curriculum Area leader. The teacher was responsible for embedding TfM in her department from 2019.

The semi-structured interview with can be summarised in five main sections: their understanding of TfM and what they consider to be the defining features; how the school transitioned to TfM; resources and changes to schemes of work; pedagogical change and available data for analysis. These elements of the interview are summarised under two main headings: understanding of TfM and implementation of TfM.

7.3.1 Understanding of Teaching for Mastery

When asked about their understanding of TfM and what they consider to be the defining features of the pedagogical approach, it was the depth of understanding and a chance to engage in mathematical discussions that were particularly apparent. According to the interviewee, TfM is about "making sure the students understand the structure of the maths and do not rush to sticking to a rule or algorithm". By ensuring that students understand the structure behind the maths they are working with, they should be able to apply their skills to unfamiliar questions in unfamiliar contexts. According to the interviewee, TfM is also about ensuring students become fluent and can make mathematical connections between different concepts such as simultaneous equations with area or perimeter. The interviewee also held mathematical discussion as a key defining feature of TfM. They commented that students should be encouraged and be able to explain and justify their answer to a question, and give a reason why, using key mathematical language. Within each lesson, there should be opportunities for discussion as well as opportunities to stretch and challenge

[18] For the purpose of this study, the school used in the pilot school is referred to as 'School X'
[19] The Progress 8 benchmark is an accountability measure used by the government to measure the effectiveness of secondary schools in England.

students by exposing mathematical links across different concepts. In chapter 2, it was concluded that TfM was to be defined in this thesis as having the characteristics listed below, and it is clear in the commentary above that the interviewee also values each aspect as an integral part of TfM.

- *'Maths for All' – where all students are exposed to the same mathematical content and some content isn't preserved for higher attaining students*
- *Where topics are taught for a longer period to ensure depth of understanding*
- *Where there is an emphasis on teaching for conceptual understanding and forming mathematical connections*
- *Where variation is embedded in chosen examples and exercises*
- *Where there is insistence on key mathematical language*
- *Where problem solving is considered a key aim of the curriculum.*

7.3.2 Implementation of Teaching for Mastery

School X transitioned to TfM gradually, following the interviewee's involvement in the local Maths Hub where they trained to be a Mastery Specialist. Following the training, where the interviewee upskilled themselves on the principles behind TfM and the NCETM's 5 Big ideas, the department collaboratively planned some year 7 lessons on fractions in April 2019. These lessons were delivered to the year group in the last half-term of the 2018-2019 academic year. In the 2019-2020 academic year, the year 7 cohort were taught using TfM for the whole year, which was the preferred departmental choice since time was needed to refine resources before moving the approach into year 8. Then, in the 2020-2021 academic year, years 7 and 8 were both taught using TfM.

School X have planned their TfM lessons from scratch for years 7 and 8, but used White Rose schemes of learning to guide their thinking. According to the interviewee, the main changes to the resources in comparison to pre-TfM is the thought that goes into the question design. Previously, it was common practice in school X to put a list of questions from an exercise on a board for the students to work through quietly without giving a purpose to each question. Now, teachers across the department think carefully about the purpose of each question and the order of questions within an exercise, to ensure mathematical structure is exposed. There is much more thinking about the choice of numbers in each question and the changes that can be made to reveal structure. As well as this, the interviewee commented that another change since the transition is the opportunities for mathematical discussion. There are many more opportunities embedded within the lesson.

When asked to compare the year 7 and 8 mastery schemes of work to pre-TfM schemes, the interviewee said that the main change has been that they now spend a lot longer on units of work. Before TfM, students would be taught fractions for two weeks in each school year. Now, they spend a half-term on it in year 7 and do not build upon it until year 9. The interviewee acknowledged that this has been a big shift and they have thought carefully, as a department, about how to embed opportunities for students to recap their learning from year 7. Whenever they have the opportunity to bring fractions into other topics, they do. The interviewee also commented that in years 7 and 8, students are no longer "streamed" or set by ability, and therefore all students are taught all areas of maths, aligning with the principles of TfM.

The main teacher pedagogical change, according to the interviewee, has been assessment for learning (AfL) and how it is orchestrated. Through departmental learning walks[20], it has been noted that the use of diagnostic questions has grown dramatically since TfM has been embedded. AfL is now used before students start a task to diagnose which students need support when they start their task to ensure they can be supported to be successful.

School X introduced TfM very steadily over the last three academic years, and therefore, the data available for analysis is from KS3 assessments. The KS3 baselines remained unchanged for the last three years and therefore it is possible to compare year 7 assessments from June 2021 (who have been taught using TfM) with the year 7 cohort from 2018 (who were not taught using TfM). The consistent outcome measure is the year 7 assessment, and all covariate information was available for both cohorts of students.

The changes in which school X have implemented in their transition to TfM align with the researcher's understanding of TfM. Throughout the implementation, teachers thought carefully about variation within chosen examples and exercises, they adapted their schemes of work so that topics are taught for longer periods of time, and place emphasis on deep, conceptual understanding.

7.4 Sample Data Preparation

The sample size provided by School X was 660 students: 330 of these students were those that were in year 7 in the 2017-18 academic year; whilst the other 330 were those that were in year

[20] A learning walk is a school process whereby typically a middle or senior leader visits multiple lessons one after the other, each for around 5-10 minutes.

7 in the 2020-21 academic year. The 330 students that were in year 7 during the 2017-18 academic year are the students not taught using TfM, whereas students that were in year 7 during the 2020-21 academic year are those that were taught using TfM.

For all 660 students, the following covariates were provided:

- Gender (Male or Female)
- Free School Meal status
- Pupil Premium status
- SEND status
- Baseline assessment percentage which was a KS2 paper that the year 7s sat in-house upon entry to the secondary school
- Ethnicity (although this was only provided for the 2020-21 cohort and not for the 2017-18, so was omitted from the data set)
- Outcome measure 1 percentage which was an assessment on number skills, including negative number, fractions and place value, taken after one term
- Outcome measure 2H/2F percentage which was an assessment on the same number skills above as well as further fractions, area and perimeter. Each student either sat the higher or foundation assessment for outcome measure 2.

As two different outcome measures were provided for both cohorts of students, the pilot study was split into three separate analyses. Initially, all students were considered together since they all sat the same assessment (outcome measure 1). The second and third analyses took the higher and foundation students respectively and worked with outcome measure 2, since the assessments were written and sat in two different tiers.

It is important to acknowledge that outcome measure 1 happened early in year 7 and therefore is only an assessment of a short-term difference in curricula, specifically one term. Outcome measure 2 happened later in the academic year, after almost a full year of the two alternative curricula so is a better evaluation of a longer-term effect.

Some students in the sample of 660 had gaps in their data; 3 students had blank ‘gender’ cells and therefore, these 3 students were removed from the sample. There are other methods of dealing with missing data in statistics, as set out in the methodology chapter, but since the level of ‘missingness’ here was substantially small, complete case analysis seemed the most effective approach to avoid the risk of imputation bias.

The aim was to conduct a propensity score matching analysis of the (ATT) impact of changing from non-TfM (coded 0) to TfM (coded 1) on the outcome assessment score, matching on propensity based on gender (coded 0 for female, 1 for male), pupil premium (coded 0 for non-PP, 1 for PP), SEND (coded 0 for no additional need, 1 for any additional need), FSM (coded 0 for non-FSM, 1 for FSM) and prior attainment (taken as percentage scores). Once the PSM process was complete, the outcome measures (taken as percentage scores) were considered for the two matched groups to look at the average treatment effect on the treated (ATT).

7.5 Analysis A - Characteristics before Matching

As mentioned, the pilot study was split into three separate analyses since more than one outcome measure is available; and that the second outcome measure is split into higher and foundation tiers. The first analysis uses outcome measure 1 (an assessment on number including negative number, fractions and place value) where all 660 students sat the same assessment.

Before any matching took place for this analysis, the distribution of baseline characteristics for the non-TfM and TfM groups were analysed to determine the difference between the two groups for each covariate and the outcome measure. Using two-sample z-tests and unpaired t-tests as well as difference-in-means analyses, the following results were found:

	Non-TfM group	TfM group	
Number of students	330	322	
Covariates	Mean	Mean	p-value
Prior Attainment % (mean, standard deviation)	65.3%, 0.2	66.3%, 0.23	0.59
Gender	0.524	0.522	0.95
Pupil Premium	0.0606	0.0901	0.15
SEND	0.221	0.13	0.002
~~FSM~~	~~0.024~~	~~0~~	n/a
Outcome Measure			Unpaired t-test result
Mean outcome % (mean, standard deviation)	60.3%, 0.2	57.9%, 0.21	t(649) = 1.5, p = 0.132, 95% CI [-0.7%, 5.6%]

Table 7.1: Pilot Study Analysis A pre-matching covariate distribution

The p-values were found through two-sample z-tests for the binary values and unpaired t-test for the continuous variables. A two-sample z-test tests the null hypothesis that there is no difference between the means of two independent populations. If the p-value comes out as being <0.05, it is implied that the difference of the means of the two populations is statistically significant.

For the SEND covariate in table 7.1, the non-TfM group had a mean of 0.221 and the TfM group had a mean of 0.13. Since SEND was coded 0 for non-SEND and 1 for SEND for each study, the

non-TfM group had, on average, more SEND pupils. A p-value of 0.002 suggested that the difference in the means for the two treatment groups is statistically significant, indicating that the same difference would be unlikely seen if random allocation to the two groups was the only adjustment made.

For the gender covariate in table 7.1, the non-TfM group had a mean of 0.524 and the TfM group had a mean of 0.522. Since gender was coded 0 for female and 1 for male, the non-TfM group had, on average, more male students. However, the difference in the means is minimal (0.002) and as such, the p-value is large (0.95), suggesting that the difference is not statistically significant. This means that with random allocation, a difference at least this large would be likely seen.

The FSM average of 0 for the TfM group indicates that there were no students in the entire cohort that were entitled to FSM which seemed unlikely to be true and therefore the covariate was omitted from the data.

On the outcome measure, the non-TfM group did better (by 2.4%) on average before any matching or adjustment. An unpaired t-test was used to determine if the average outcome measures of the two groups was significantly different. The t-test suggested that there was no significant difference between the groups on the outcome measure before matching (t(649)=1.5,p=0.132,95% CI [-0.7%, 5.6%]).

7.6 Analysis A - Propensity Score & Matching

In the methodology, Leite's (2016) six steps to propensity score matching were set out. The first of these steps was to prepare the data, which can be seen earlier in this chapter. The second step was to estimate the propensity score. Thus, following the pre-matching analysis, the propensity score for each student was estimated. Four of the five given covariates (prior attainment, gender, pupil premium and SEND) were used in the estimate of the propensity score; FSM was not used in the propensity score as the TfM average of 0 suggested that data was corrupted or invalid. The propensity score was estimated by ordinary linear regression using linear terms for each covariate, as set out in chapter 5. Following the estimation, specific to school X, pupil premium and SEND were revealed as the most influential covariates as predictors for group allocation due to the p-values, more than prior attainment and gender. The histogram below shows the distribution of propensity scores for the non-TfM group (curriculum type: other) and the TfM group (curriculum type: TfM).

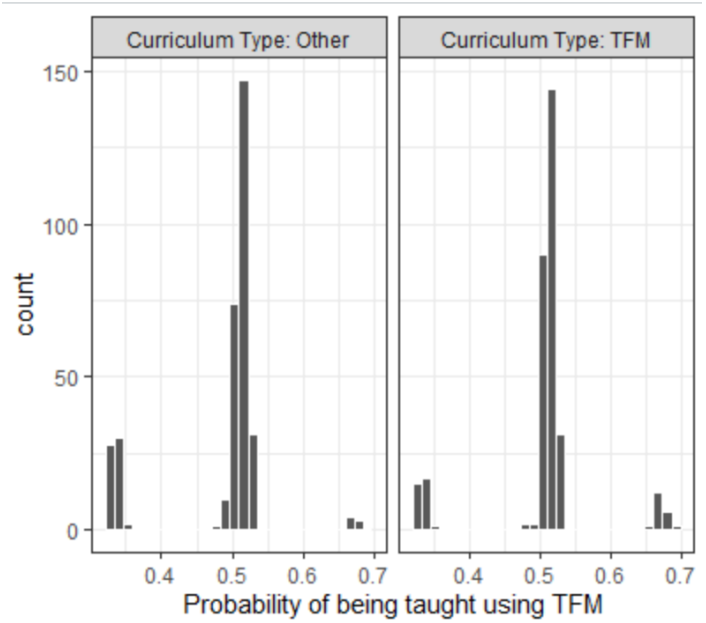

Figure 7.1: Pilot Study Analysis A Propensity Scores

Figure 7.1 displays the propensity scores across the horizontal for both treatment groups, and the vertical 'count' indicates the frequency for each propensity score. The lack of continuous covariates has led to a seemingly 'clustered' or 'lumpy' propensity score distribution. This is a critical consideration to thinking about further analysis of this case going forwards.

Persisting with considering all students as one study for analysis A, the third and fourth of Leite's (2016) steps is to conduct matching methods until sufficient balance was achieved and to evaluate covariate balance. It is at this point when one needs to consider what defines 'sufficient balance'. Balance can be deemed to be sufficient when covariates that were initially imbalanced (before matching) are balanced without unbalancing any covariates that were initially balanced.

Four versions of matching were conducted within analysis A until a sufficient match was achieved before the treatment effect was estimated. Morgan et al (2008) used three matching methods as a form of sensitivity analysis to eliminate the possibility of hidden bias. Whilst each matching technique used a different function, each yielded a treatment effect attributable to the intervention under scrutiny. This is important to allow the matching process to be iterative and not pre-determined. Even when a matching process has resulted in a seemingly well-matched sample, a second matching method was used as sensitivity analysis check as it was in the work conducted by Morgan.

7.6.1 Version 1

In the first version of matching, 1:1 nearest neighbour without replacement was the method used, with a caliper of 0.1. By default, the propensity score was used as the distance measure, resulting in 596 observations being matched. Nearest neighbour matching involves running through the list of treated units and selecting the closest eligible untreated unit to be paired with each treated, provided they are no further apart than the caliper width. It can be regarded as greedy matching in that pairing occurs without reference to how other units have been paired, and therefore does not aim to optimise any criterion.

Table 7.2 shows the covariate distributions in the original and matched sample, as well as the balance improvement. The balance improvement is based on the difference-in-means inspection, outlined in chapter 5, but in comparison to the balance before matching and given as a percentage.

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	298	298	66.9	65.1	-95.5	0.52	0.53	-34.1	0.06	0.05	88.6	0.14	0.14	96.3

Table 7.2: Pilot Study Analysis A version 1 matching results

Table 7.2 shows that after matching, 32 students from the non-TfM and 24 students from the TfM groups were not matched. It suggests that the balance for pupil premium and SEND improved with the balance improvement score being between 0-100. Any value between 0 and 100 indicates that balance has improved after matching; values less than 0 indicate more imbalance after matching. In this case, prior attainment and gender became more imbalanced. That said, when covariates are already well balanced before matching, any slight change can result in a statistic that implies dramatic imbalance after matching. Therefore, when looking at the balance statistics one must consider how well the covariates were balanced before matching. The percent balance improvement is computed as $100\frac{|\theta_M|-|\theta_U|}{|\theta_U|}$, where θ_M is a given balance statistic in the matched sample and θ_U is the same balance statistic in the unmatched sample.

Figure 7.2 shows the propensity score distribution for unmatched and matched units, which indicates that there is a region of 'common support' for the matched units but also shows that the propensity scores are 'lumpy' due to the lack of continuous covariates (prior attainment is the only continuous covariate). The points in figure 7.2 are clustered and not spread out across the horizontal range of propensity score values.

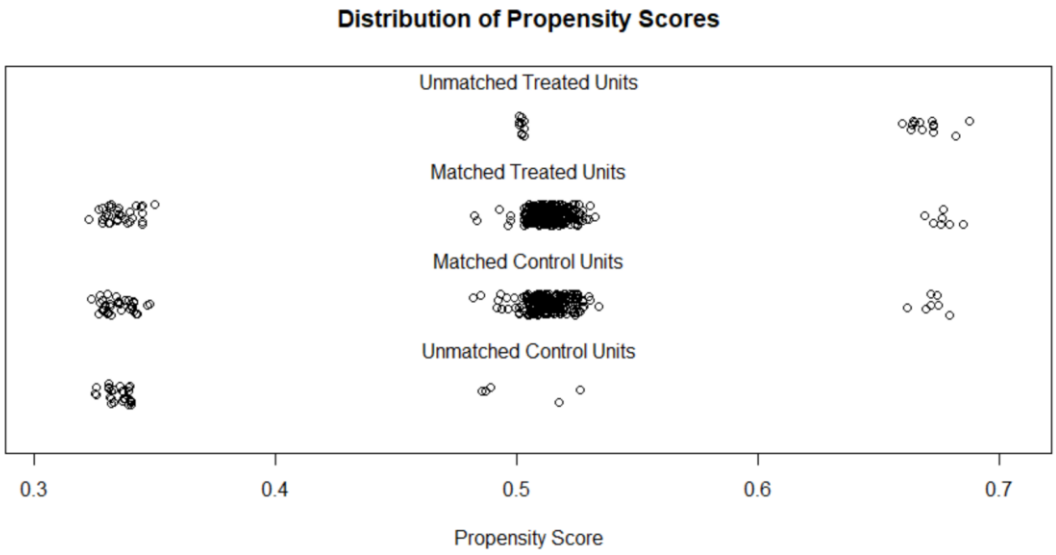

Figure 7.2: Pilot Study Analysis A propensity score distribution after version 1 of matching

Figures 7.3 and 7.4 support the balance improvement statistics above in showing that prior attainment and gender became more imbalanced after matching. A treatment of 0 is the non-TfM group and the treatment of 1 is the TfM group. The 'unadjusted sample' portrays the balance in the specific covariate before matching, whilst the 'adjusted sample' is after matching.

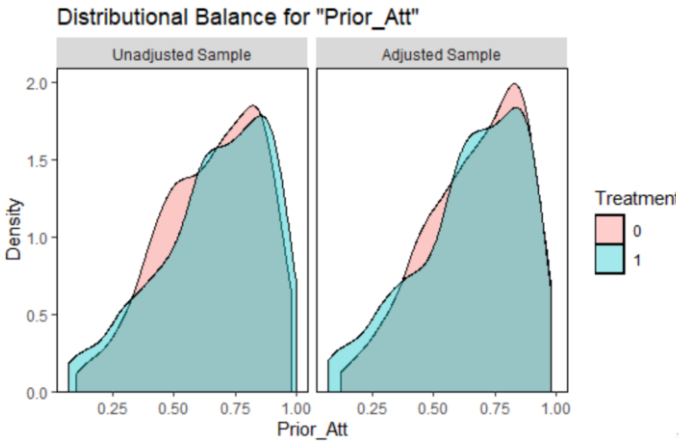

Figure 7.3: prior attainment balance before and after matching

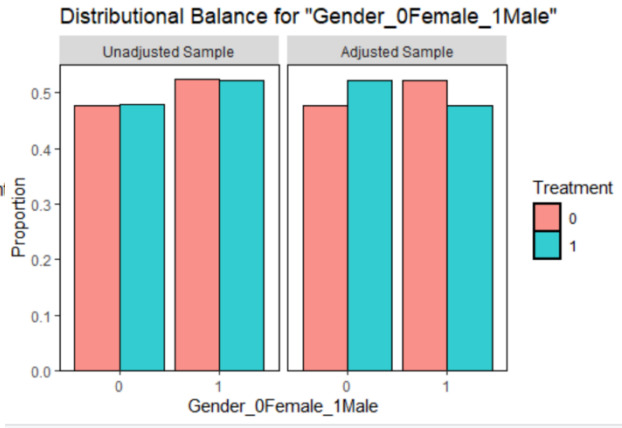

Figure 7.4: gender balance before and after matching

Figures 7.5 and 7.6 show the balance for pupil premium and SEND before and after matching and confirm that this version of matching helped to establish more balance for these two covariates.

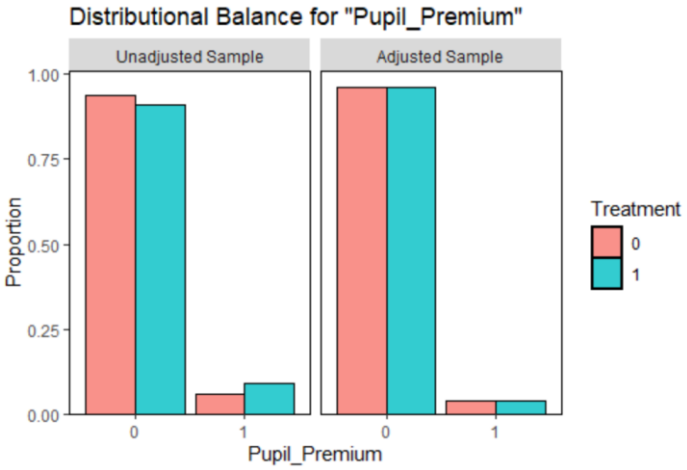

Figure 7.5: pupil premium balance before and after matching

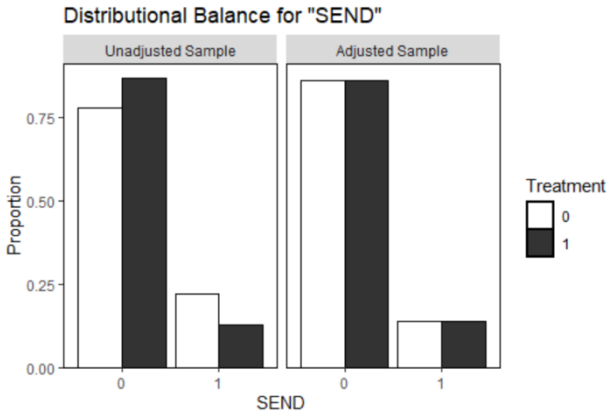

Figure 7.6: SEND balance before and after matching

Whilst the ideal outcome of matching would be for all covariates to become more balanced, gender and prior attainment were not the main drivers of the propensity score before matching and therefore, the further imbalance was not a huge concern. That said, further versions of matching were carried out nevertheless to seek the best overall balance.

7.6.2 Version 2

In the second version of matching, the same 1:1 nearest neighbour without replacement algorithm was used, but this time with a tighter caliper of 0.05, and chosen to address the issue of one continuous covariate (prior attainment) being seemingly overwhelmed by the binary covariates. Again, the propensity score was used as the distance measure, resulting in 588 observations being matched. In version 1, 596 observations were matched, so a tighter caliper predictably removed more observations from the sample.

Table 7.3 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 2 (nearest neighbour caliper, 0.05)	294	294	67.5	65.8	-48.1	0.52	0.53	-171.8	0.05	0.05	88.5	0.14	0.13	96.3

Table 7.3: Pilot Study Analysis A version 2 matching results

It is apparent that prior attainment and gender were still more imbalanced than the unmatched data. Prior attainment was not as imbalanced as it was in version 1, but gender was more imbalanced. The balance between the two treatment groups for pupil premium is very similar in the two versions and remains almost identical for SEND.

The graphic below illustrates the balance (or imbalance) after matching with version 2. Figure 7.7 shows the propensity score distribution for unmatched and matched units, which shows there is a region of 'common support' for the matched units and also shows that, as with version 1, the propensity scores are 'lumpy' due to the lack of continuous covariates.

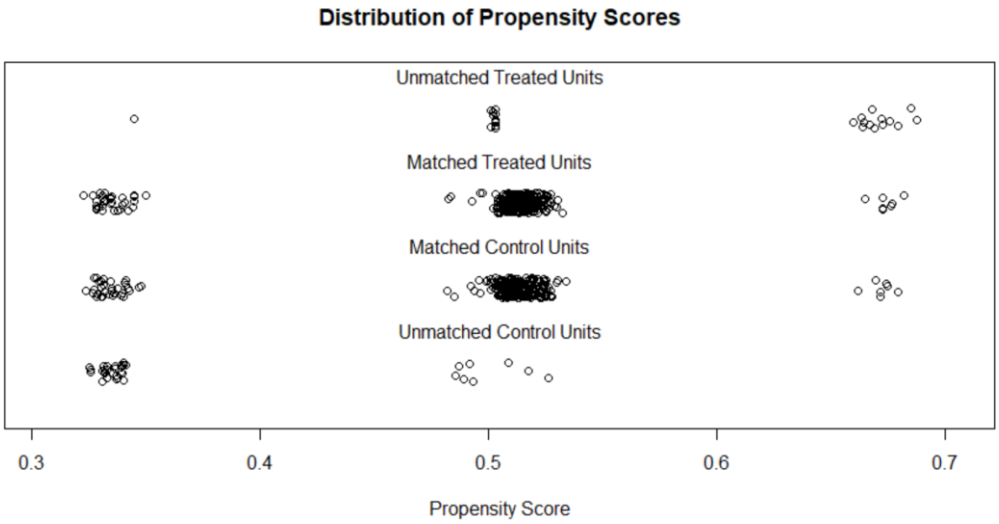

Figure 7.7: Pilot Study Analysis A propensity score distribution after version 2 of matching

Figures 7.8 and 7.9 support the balance improvement statistics above in showing that the prior attainment and gender covariates have become more imbalanced after matching. The adjusted sample graphs show a greater difference for the two treatment groups.

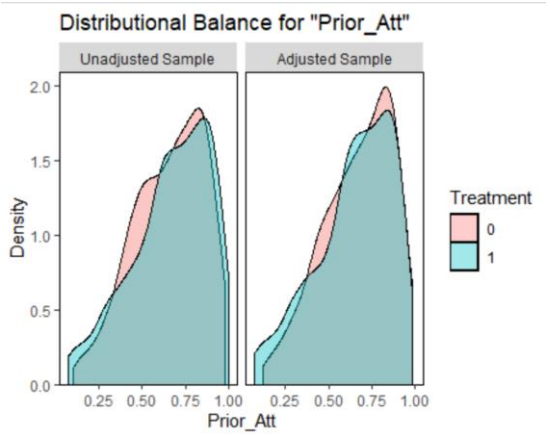

Figure 7.8: prior attainment balance before and after matching

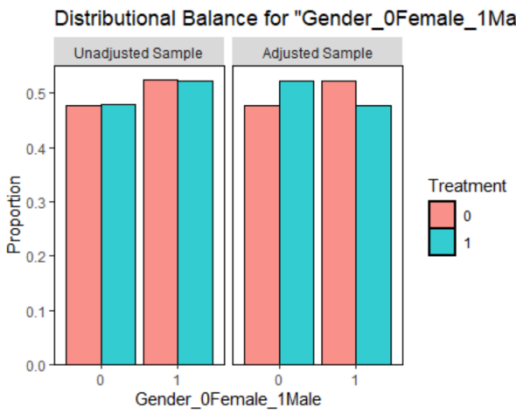

Figure 7.9: gender balance before and after matching

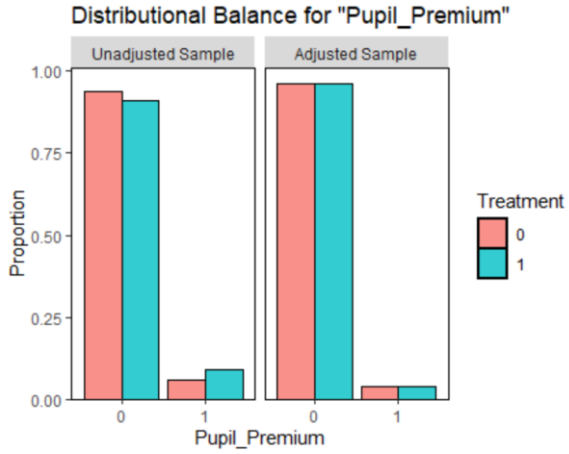

Figure 7.10: pupil premium balance before and after matching

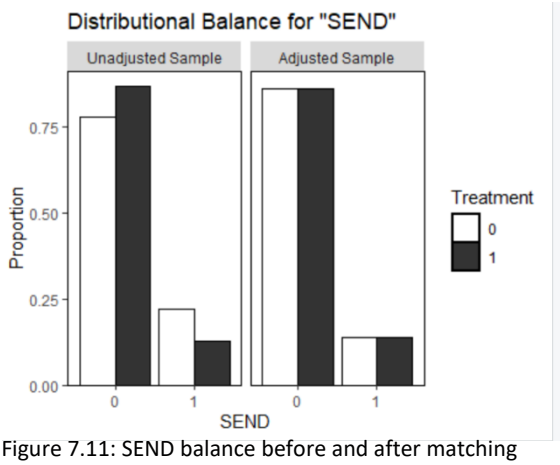

Figure 7.11: SEND balance before and after matching

Figures 7.10 and 7.11 show the balance for pupil premium and SEND before and after matching and confirm that this version of matching has helped to establish more balance for these two covariates.

As mentioned, the ideal outcome of matching is for all covariates to become more balanced but in this case, gender and prior attainment have become more imbalanced. Thus, a tighter caliper of 0.05 has not helped to create more balance across the two treatment groups for these covariates. Nevertheless, as gender and prior attainment were not the main drivers of the propensity score before matching, the imbalance is not a huge concern. Yet, further matches methods were still carried out to see if balance across all covariates could be improved.

7.6.3 Version 3

A further version of matching was carried out to see if an even tighter caliper could improve the balance between gender and prior attainment across the two treatment groups. 1:1 nearest neighbour matching without replacement was used again, but with the caliper tightened to 0.03. This time, 578 observations were matched and therefore a tighter caliper removed more people from the sample.

Table 7.4 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 3 (nearest neighbour, caliper 0.03)	289	289	67.8	65.8	-115.5	0.52	0.53	-314.7	0.05	0.05	76.5	0.12	0.13	92.4

Table 7.4: Pilot Study Analysis A version 3 matching results

The balance improvement statistics show that, with a tighter caliper, better balance was not achieved across all four covariates. Prior attainment and gender were yet again more imbalanced than before matching, whilst pupil premium and SEND were more balanced than before matching but more imbalanced than in versions 1 and 2. Because of the imbalance, visual plots of these results were not produced.

7.6.4 Version 4

It became clear that a tighter caliper was not improving balance for the covariates. Therefore, the fourth version trialled 1:1 nearest neighbour matching without a caliper. The purpose of this round of matching was to see if the use of the initial caliper in version 1 achieved better balance than if no caliper was used at all. This time, 644 observations were matched and thus many more non-TfM observations were retained without a caliper in place as matches were found.

Table 7.5 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 4 (nearest neighbour, no caliper)	322	322	65.3	66.3	-6.5	0.52	0.52	-24.1	0.06	0.09	5.1	0.21	0.07	21.3

Table 7.5: Pilot Study Analysis A version 4 matching results

Of the four versions, this round of matching found the best match for prior attainment and gender, but still slightly more imbalanced than before matching. For pupil premium and SEND, matching without a caliper resulted in poorer balance than the previous versions. Up to now, versions 1 and 2 had optimised the balance for these two covariates and therefore, visual plots of version 4 were not produced.

7.6.5 Version 5

The first four versions showed that matching with a caliper generated more balanced groups for the key drivers of the propensity score (pupil premium and SEND). The fifth version trialled a different matching method: genetic matching. Genetic matching is a form of nearest neighbour matching and was used in a 1:1 ratio without replacement, again with the propensity score used as the distance measure. This version resulted in 644 matched observations, which used all observations in the TfM group.

Table 7.6 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 5 (genetic matching)	322	322	66	66.3	71.6	0.52	0.52	-148.1	0.06	0.09	5.1	0.2	0.13	21.3

Table 7.6: Pilot Study Analysis A version 5 matching results

The balance improvement column above shows that genetic matching improved balance for prior attainment and was the first version to do so. For gender, pupil premium and SEND, however, versions 1 and 2 were still superior in optimising balance. As such, version 5 overall did not optimise overall balance and therefore visual plots were not produced.

7.6.6 Version 6

The final version of matching for the pilot study trialled the optimal full matching method. This matching method led to all 652 observations being matched, meaning that no observation was disregarded from the sample. Table 7.7 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 6 (full matching)	330	322	67.2	66.3	1.1	0.52	0.52	54.4	0.08	0.09	78.9	0.12	0.13	93.2

Table 7.7: Pilot Study Analysis A version 6 matching results

It is clear from the table above that the optimal full matching method improved balance for prior attainment and gender compared to the unmatched balance. All of the balance improvement statistics are between 0 and 100, thus suggesting that balance was improved on all covariates.

The graphic below illustrates the balance achieved after optimal full matching in version 6. Figure 7.12 shows the distribution of propensity scores for unmatched and matched units, showing there is a region of 'common support' despite the distribution being 'lumpy' due to the lack of continuous covariates.

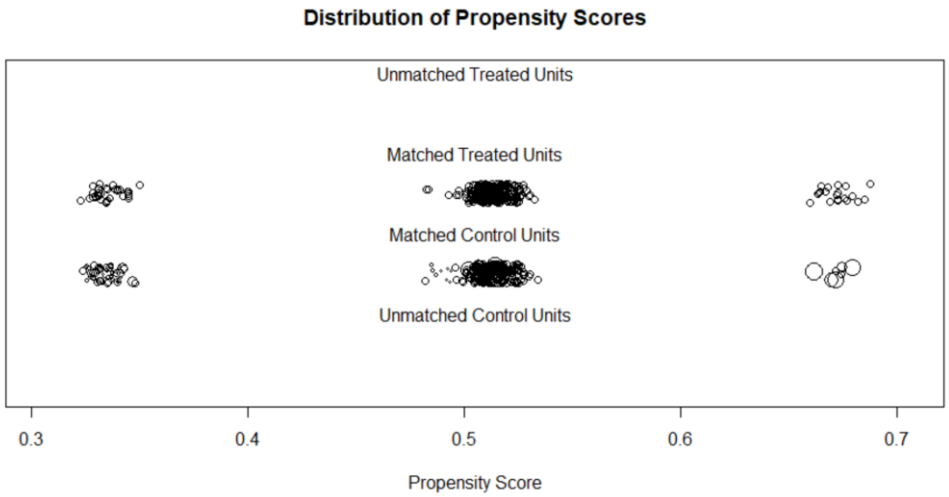

Figure 7.12: Pilot Study Analysis A propensity score distribution after version 6 of matching

Figures 7.13-7.16 show the balance achieved after optimal full matching for each of the four covariates.

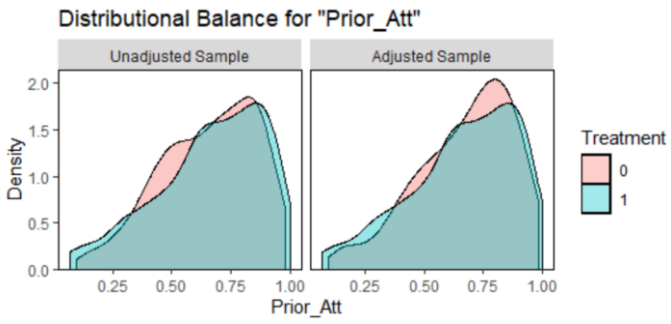

Figure 7.13: prior attainment balance before and after matching

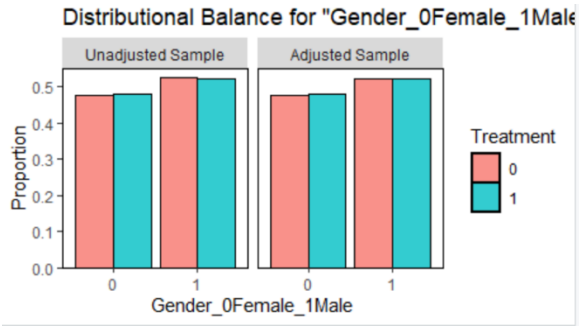

Figure 7.14: gender balance before and after matching

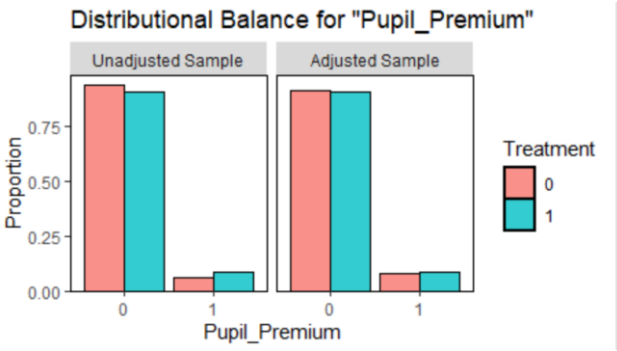

Figure 7.15: pupil premium balance before and after matching

Figure 7.16: SEND balance before and after matching

From the above versions of matching, versions 1 and 2 created greater balance improvement on pupil premium and SEND. Since these covariates were the main drivers of propensity score in the unmatched sample, it was decided that version 1 was to be used to estimate the treatment effect, with versions 2 and 6 as the sensitivity analyses.

7.7 Estimation of Treatment Effect

Although it was not possible to achieve perfect balance across all four covariates, the figures for versions 1, 2 and 6 of matching show that there was no substantial imbalance for the covariates that were the main drivers of the propensity score. Greifer (2022) stated that visual depictions of distributional balance complement numerical summaries and therefore it is important to use these in the main study.

Estimating treatment effect is the fifth step in Leite's (2016) steps to propensity score matching. For this analysis of the pilot study, version 1 was used to estimate the treatment effect, with versions 2 and 6 used as the sensitivity analyses.

Treatment effects were estimated through the use of paired t-tests on the outcome measure (year 7 assessment percentage). It was found that the non-TfM group had a slightly higher mean outcome measure by 5%, with a small p-value (= 0.0089) which indicated statistical significance with a 95% confidence interval for the range [1.3%, 8.8%] (t(297)=2.6,p=0.0089, 95% CI[1.3,8.8]). The confidence interval is a range of values that one can be 95% confident contains the true value: in this case one might reasonably expect the effect of the difference in curricula to lie between 1.3% and 8.8%. With this confidence interval, there is still a 1-in-20 chance that the identified interval does not include the actual value. The effect found is the average treatment effect

on the 298 treated (ATT) and for the topics which formed the assessment: negative number, fractions and place value. The results cannot be generalised further than this.

However, these results are not meaningful as an evaluation of TfM. As discussed at the start of this chapter, the effects of Covid cannot be disaggregated and therefore, this chapter is only used as a test of PSM methods. Furthermore, outcome 1 was an assessment taken by students at the end of one term of different curricula and therefore the analysis only looks at a very short time period.

7.8 Sensitivity Analysis

Leite's (2016) sixth and final step for propensity score matching is to conduct a sensitivity analysis in order to validate the found treatment effect.

Using version 2, it was also found that the non-TfM group had a slightly higher mean outcome measure by 4.4% on the percentage score. A small p-value (= 0.019) indicated statistical significance and the 95% confidence interval predicts the actual effect of the difference in curricula to lie between 0.73% and 8.12% (t(293)=2.4,p=0.019, 95% CI[0.73,8.12]).

Since version 6, which used full optimal matching, did not result in the same number of subjects in the non-TfM and TfM groups, the sensitivity analysis was conducted differently. A linear regression model was estimated and found that TfM had an average effect of -0.024 (-2.4%) on the percentage score, indicating that the non-TfM group did better and a p-value=0.132.

Thus, after propensity scores based on gender, pupil premium, prior attainment and SEND, the matched analysis for all three versions suggests the effect of the TfM curriculum compared to the previous curriculum was negative. This result must be treated with caution as a huge limitation of PSM is the uncertainty of knowing that all confounders are controlled for, and it was highlighted earlier that the effects of Covid cannot be disaggregated from this analysis. Indeed, it is reasonable to assume Covid played a bigger role in learning and student outcomes than any change in curriculum. Nevertheless, in a sensitivity analysis, one would hope that the results would concur with those found in the initial treatment effect and in this case, concurrence can be seen.

7.9 Analysis B - Characteristics before Matching

Analysis B will use outcome measure 2, which was a further assessment sat by all students but with students sitting only one tier: foundation or higher. The assessment was on number, fractions, area and perimeter and was taken after nearly a full year of the two alternative curricula so, Covid aside, is a better assessment of the effect of TfM. Analysis B will look at the higher tier outcome measure 2 and will use outcome measure 1 as a continuous covariate with the aim of seeking less clustered propensity scores as were seen in analysis A.

As with analysis A, the first step was to analyse the distribution of baseline characteristics for the two treatment groups using unpaired t-tests, z-tests, and difference-in-means. The results are below:

	Non-TfM group	TfM group	
Number of students	224	263	
Covariates	Mean	Mean	p-value
Prior Attainment % (mean, standard deviation)	75.3%, 0.14	74.5%, 0.15	0.53
Gender	0.54	0.56	0.68
Pupil Premium	0.027	0.065	0.04
SEND	0.12	0.07	0.07
Assessment 1 % (mean, standard deviation)	70.7%, 0.14	62.5%, 0.19	<0.00001
Outcome Measure			Unpaired t-test result
Mean outcome % (mean, standard deviation)	68.7%, 0.17	55.3%, 0.19	t(483) = 8.1, p<0.00001, 95% CI [10.1%, 16.6%]

Table 7.8: Pilot Study Analysis B pre-matching covariate distribution

Table 7.8 indicates that, on average, before any matching or adjustment, the non-TfM did better on average by 13.4%. Differences in the covariates between the two treatment groups highlighted a need for matching before the treatment effect was estimated.

For the covariates in table 7.8, two-sample z-tests revealed small p-values for pupil premium (= 0.04) and SEND (= 0.07), whilst an unpaired t-test gave a small p-value for assessment 1 (<0.00001); indicating that the same difference would unlikely be seen if random allocation to the two treatment groups was the only adjustment made.

Using an unpaired t-test, the difference in the mean outcome measure yielded a small p-value, suggesting that we would not have seen the same difference in means had we randomly allocated participants to the two treatment groups (t(483)=8.1, p<0.00001, 95% CI [10.1%,16.6%]).

7.10 Analysis B - Propensity Score & Matching

In the same way as analysis A, Leite's (2016) six steps for propensity score matching were followed for this analysis. When the propensity score for each student was estimated, it was prior attainment, SEND and assessment 1 that were found to be the biggest predictors of group allocation. All five covariates were used in the subsequent matching process, but particular attention was paid to the balance of these three driving covariates. The histogram below shows the distribution of propensity scores for both treatment groups. It is clear that the inclusion of another continuous covariate (assessment 1) helped to overcome the seemingly 'clustered' or 'lumpy' propensity score distribution that was seen in analysis A.

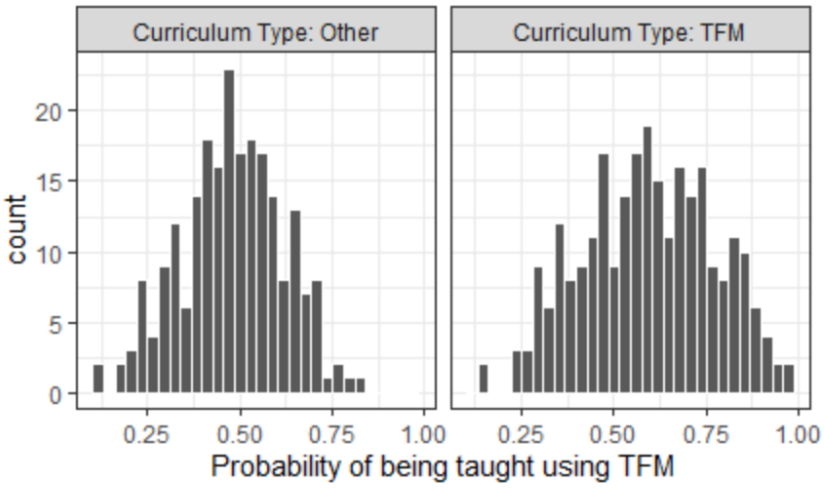

Figure 7.17: Pilot Study Analysis B Propensity Scores

7.10.1 Version 1

The first version of matching in analysis B mirrored version 1 matching in analysis A. The aim was to be able to pre-register the steps that were taken for the main study. It is important to have a systematic approach to the pilot study. As such, 1:1 nearest neighbour matching without replacement was used for version 1, with a caliper of 0.1. Propensity score was used as the distance measure, resulting in 360 observations being matched.

Table 7.9 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	180	180	75.4	75.6	58.1	0.52	0.53	40.8	0.02	0.03	85.3	0.11	0.08	54	0.69	0.69	96.2

Table 7.9: Pilot Study Analysis B version 1 matching results

For all five covariates, more balance has been achieved through this version of matching. The figures below illustrate this:

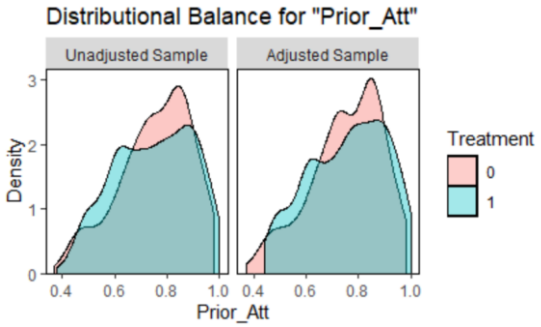

Figure 7.18: prior attainment balance before and after matching

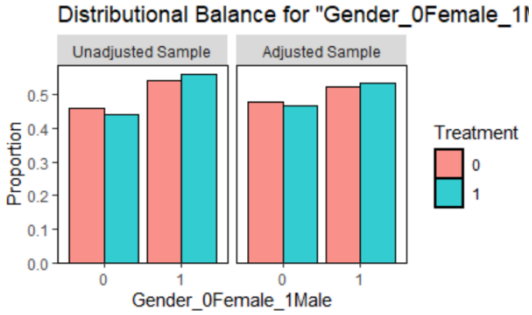

Figure 7.19: gender balance before and after matching

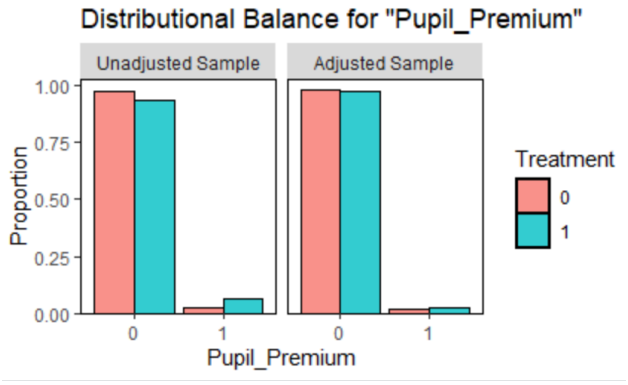

Figure 7.20: pupil premium balance before and after matching

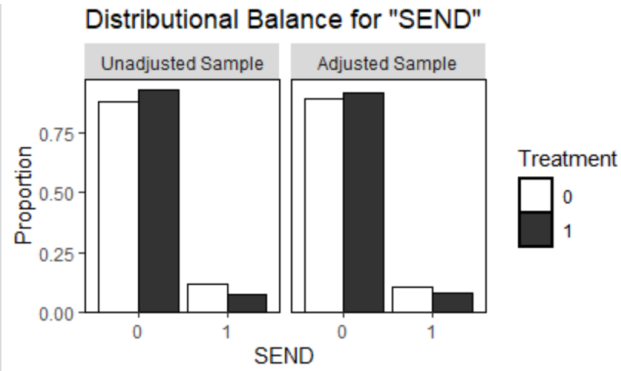

Figure 7.21: SEND balance before and after matching

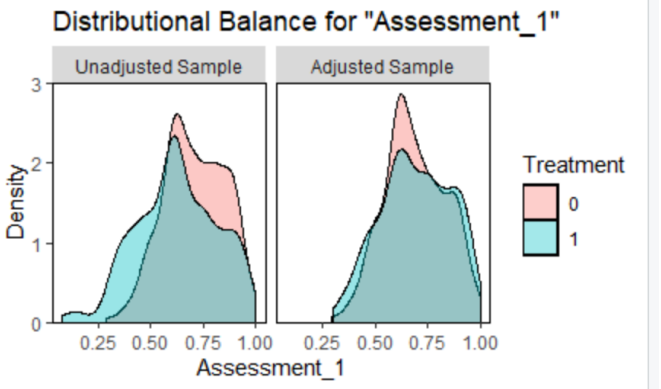

Figure 7.22: assessment 1 balance before and after matching

Since this matching method achieved better balance on all five covariates, rather than altering the matching method used, the focus of subsequent versions was the width of the caliper used.

7.10.2 Version 2

Version 2 used the same matching method as above but sought to see what the implication of tightening the caliper to 0.05, half of its original value. The results are below:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 2 (nearest neighbour, caliper 0.05)	170	170	75.7	76.0	70.2	0.54	0.52	5.9	0.02	0.02	84.5	0.1	0.06	26.9	0.69	0.70	93.9

Table 7.10: Pilot Study Analysis B version 2 matching results

Other than prior attainment, the balance for the rest of the covariates was not as good in version 2 compared with version 1. Therefore, it was decided to try to widen the caliper to assess the implications rather than tighten it anymore.

7.10.3 Version 3

For version 3, the same matching method as version 1 and 2 was used, but this time with a caliper of 0.2 which was double that used in version 1. The results of the balance are below:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 3 (nearest neighbour, caliper 0.2)	189	189	75.5	75.5	97.4	0.53	0.52	15.4	0.02	0.02	86	0.1	0.08	67.1	0.69	0.68	88.7

Table 7.11: Pilot Study Analysis B version 3 matching results

For prior attainment, pupil premium and SEND, more balance was achieved than with version 1. For gender and assessment 1, the balance was better than before matching although not as well balanced as version 1. Widening the caliper overall helped to achieve balance across all covariates in version 3. The graphics below illustrate this:

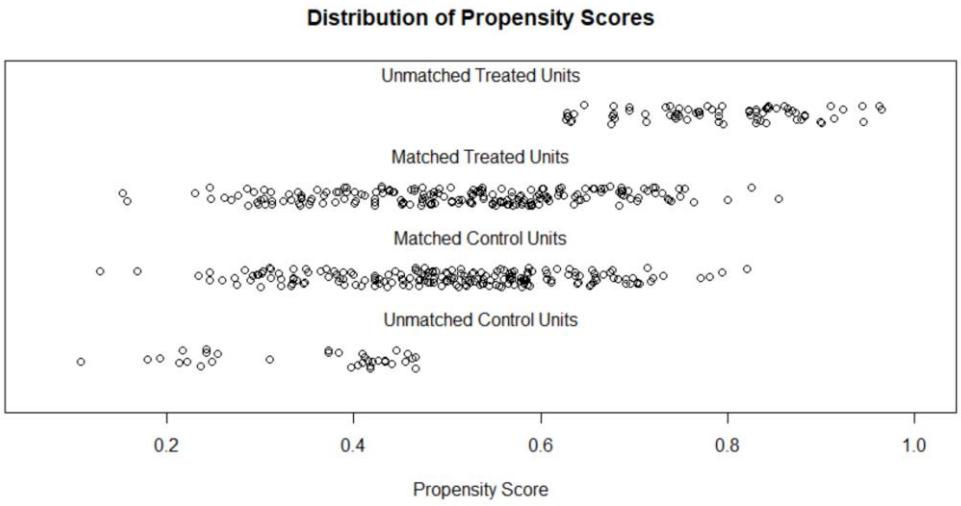

Figure 7.23: Pilot Study Analysis B propensity score distribution after version 3 of matching

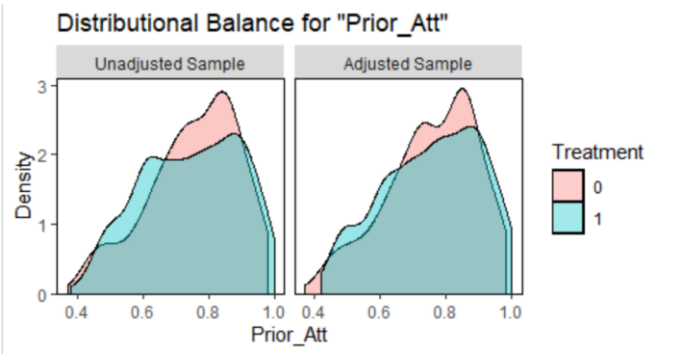

Figure 7.24: prior attainment balance before and after matching

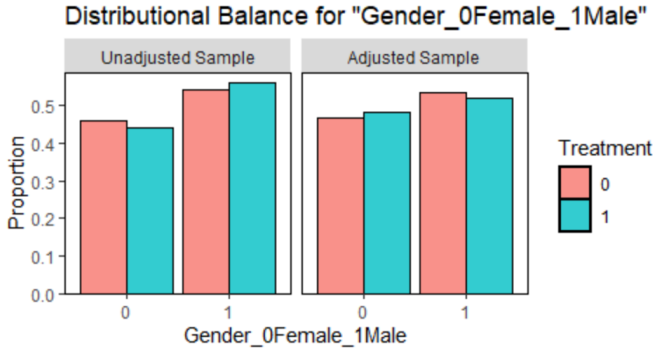

Figure 7.25: gender balance before and after matching

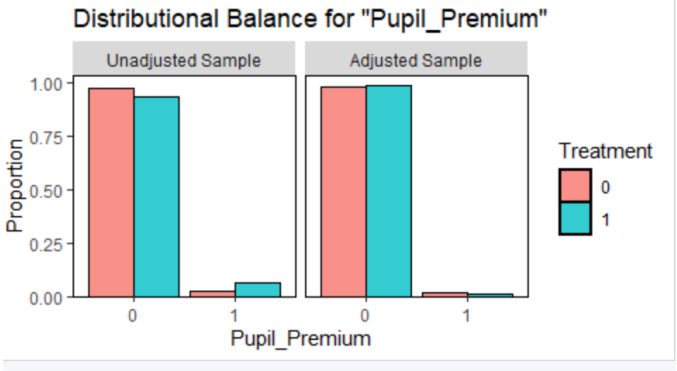

Figure 7.26: pupil premium balance before and after matching

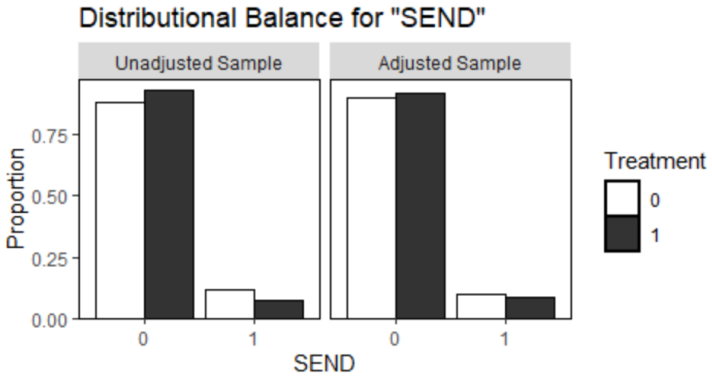

Figure 7.27: SEND balance before and after matching

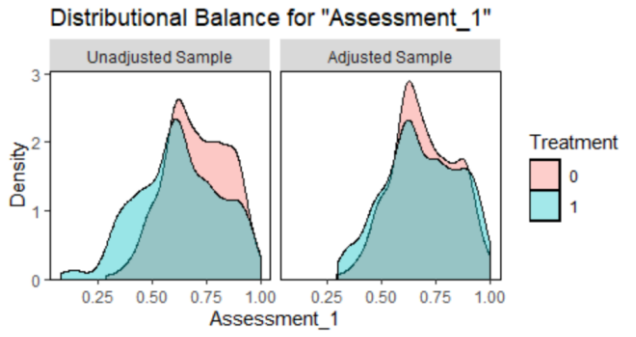

Figure 7.28: assessment 1 balance before and after matching

7.10.4 Version 4

Version 4 investigates the impact of widening the caliper to 0.3. Table 7.12 shows the results:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 4 (nearest neighbour, caliper 0.3)	200	200	75.4	75.6	70.9	0.55	0.52	-33.3	0.03	0.02	86.8	0.10	0.08	68.9	0.70	0.68	78.9

Table 7.12: Pilot Study Analysis B version 4 matching results

A wider caliper retained more observations but worsened the balance of covariates. Although four out of the five covariates were still more balanced than the unmatched sample, they were not as well balanced as previous versions. Gender became more imbalanced than the unmatched sample. Thus, it was concluded that 0.2 was the optimal caliper to use for this method of matching.

7.10.5 Version 5

Before estimating the treatment effect, it was important to try matching without a caliper since the previous versions focused on varying the width of the caliper. Nearest neighbour matching was still used, again, without replacement. The results are below:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 5 (nearest neighbour, no caliper)	224	224	75.4	73.9	-75.6	0.54	0.59	-161.8	0.03	0.08	-29.7	0.12	0.06	-20.2	0.71	0.59	-44.5

Table 7.13: Pilot Study Analysis B version 5 matching results

Without a caliper, all five covariates became more imbalanced than they were before matching and therefore it was concluded that a caliper was needed for this analysis to generate balance across the covariates.

7.11 Estimation of Treatment Effect

Of the versions of matching above, version 3 was used to estimate the treatment effect since a caliper of 0.2 helped to achieve more balance on prior attainment, SEND and assessment 1 than version 1. Since prior attainment and SEND were two of the key drivers of the propensity score, it was deemed appropriate to use version 3.

Paired t-tests were used to estimate the treatment effect on the outcome measure (year 7 higher outcome 2). It was found that the non-TfM group had a slightly higher mean percentage score by 1.05% with a small p-value (< 0.00001), thus indicating significance. The confidence interval was 95% for the range [6.9%, 14.1%] which indicates that one can reasonably expect the effect of the difference in curricula to lie between 6.9% and 14.1% (t(188)=5.7,p=<0.00001), 95% CI[6.9,14.1]). This is the average treatment effect on the treated (ATT): in this case on the 189 treated students for which matches were found. The treatment effect found only holds true for the specific outcome measure used: an assessment on number skills, fractions, area and perimeter.

7.12 Sensitivity Analysis

Since version 1 also achieved adequate balance across all covariates, the results were used as one part of the sensitivity analysis. The results from matching version 1 also indicated that the non-TfM group did better on the specific outcome measure, achieving on average 9.65% higher than the TfM group. A very small p-value (< 0.00001) also indicated significance and a confidence interval of 95% for the range [6.2%, 13.1%] suggests that one can expect the effect of the difference in curricula to lie between 6.2% and 13.1% (t(179)=5.5,p=<0.00001, CI[6.2,13.1]).

The second part of the sensitivity analysis used genetic matching. Following implementation of the method, the treatment effect was estimated to align with version 1 and version 3 in showing that the non-TfM group did better. In this case, the mean of the differences was 0.1568, suggesting that the TfM group did 15.68% better on average. A very small p-value (< 0.00001) shows statistical significance and the 95% confidence interval for the range [12.4%, 18.9%] suggests that one can expect the effect of the difference in curricula to lie between 12.4% and 18.9%.

Thus, after propensity scores based on gender, pupil premium, prior attainment, SEND and assessment 1, the matched analysis suggests the effect of the TfM curriculum compared to the previous curriculum was negative when using the assessment 2 higher as the outcome measure. As

with analysis A, this result must be treated with caution due to the difficulty of not knowing that all confounders are controlled for when using propensity score matching. Given that the TfM group were the group impacted by Covid, it may not be surprising to find that these analyses revealed that the non-TfM group performed better.

7.13 Analysis C - Characteristics before Matching

Analysis A considered the effect of TfM using outcome measure 1 which all students in both cohorts sat. Analysis B used the higher tier outcome measure 2, which was only sat by a selection of the sample. Analysis C uses the foundation tier outcome measure 2. As with analysis B, outcome measure 1 was used as a continuous variable in this analysis with the aim of seeking a less clustered distribution of propensity scores as seen in analysis A.

As with the previous two analyses, the first step was to analyse the distribution of baseline characteristics for the two treatment groups using z-tests, unpaired t-tests and difference-in-means. The results are below:

	Non-TfM group	TfM group	
Number of students	110	79	
Covariates	Mean	Mean	p-value
Prior Attainment % (mean, standard deviation)	44.5%, 0.14	35.5%, 0.16	<0.00001
Gender	0.49	0.41	0.24
Pupil Premium	0.13	0.18	0.35
SEND	0.43	0.32	0.12
Assessment 1	38.8%, 0.12	40.2%, 0.17	0.55
Outcome Measure			Unpaired t-test result
Mean outcome % (mean, standard deviation)	52.2%, 0.15	39.8%, 0.16	t(165) = 5.3, p<0.00001, 95% CI [7.8%, 17.0%]

Table 7.14: Pilot Study Analysis C pre-matching covariate distribution

Table 7.14 indicates that, before any matching or adjustment, the non-TfM did better on average by 12.4%. Differences in the covariates between the two treatment groups highlighted a need for matching before the treatment effect was estimated.

For the above covariates, an unpaired t-test revealed a small p-value for prior attainment (< 0.0001), indicating that the same difference would be unlikely seen if random allocation to the two treatment groups was the only adjustment made. The z-tests for the other four covariates did not reveal significance. An unpaired t-test revealed that there was significant difference between the two groups on the outcome measure (t(165)=5.3, p<0.00001, 95% CI [7.8%, 17.0%]).

7.14 Analysis C - Propensity Score & Matching

When the propensity score was estimated, the three covariates that were the biggest predictors of treatment allocation were prior attainment, SEND and assessment 1, which were also the three main drivers of propensity score in analysis B. All five covariates were used in the matching process, and careful attention was paid to ensure that balance of these three drivers was achieved. Figure 7.29 shows the distribution of propensity scores for both treatment groups:

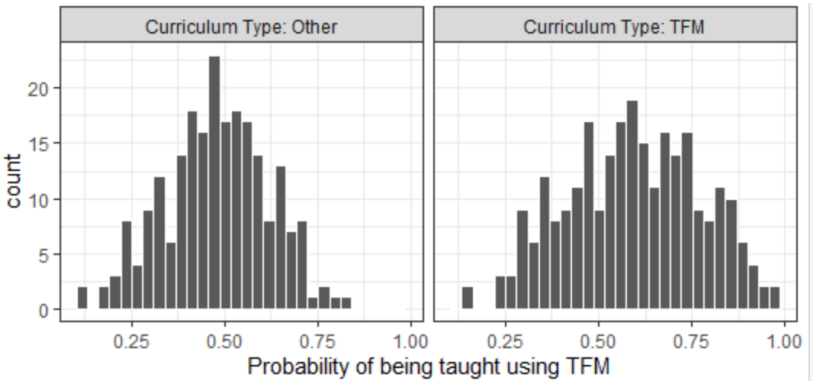

Figure 7.29: Pilot Study Analysis C Propensity Scores

7.14.1 Version 1

The first version of matching in analysis C mirrored version 1 in the previous two analyses; 1:1 nearest neighbour matching without replacement, with a caliper of 0.1. Propensity score was used as the distance measure, resulting in 104 observations being matched. Table 7.15 shows the covariate distributions in the original and matched sample, as well as the balance improvement:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	52	52	41.3	38.6	70.5	0.46	0.42	55.2	0.17	0.13	23	0.34	0.38	47.9	0.40	0.37	-91.4

Table 7.15: Pilot Study Analysis C version 1 matching results

The balance for prior attainment, gender, pupil premium and SEND improved in this version of matching, but assessment 1 became more imbalanced when compared to the unmatched data.

7.14.2 Version 2

The next version of matching used the same method as above but with a wider caliper of 0.2 to see if balance for assessment 1 could be achieved. The results are below:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 2 (nearest neighbour, caliper 0.2)	56	56	41.4	39.4	60.9	0.46	0.46	100	0.16	0.14	64.2	0.34	0.38	67.8	0.40	0.38	-20.7

Table 7.16: Pilot Study Analysis C version 2 matching results

Version 2 of matching improved the balance for assessment 1 when compared to version 1, but the covariate was still more imbalanced when compared to the unmatched data. This version of matching improved the balance for prior attainment, gender, pupil premium and SEND in comparison to version 1, suggesting that a wider caliper was helping to improve the overall balance. Since it was thought that version 2 may be used for either treatment effect estimates or sensitivity check, the plots below were produced to compare the balance pre- and post- matching. Figure 6.30 shows the distribution of propensity scores, showing there is a region of common support. Figures 7.31-7.35 show that the balance has been improved after matching for all covariates, other than assessment 1.

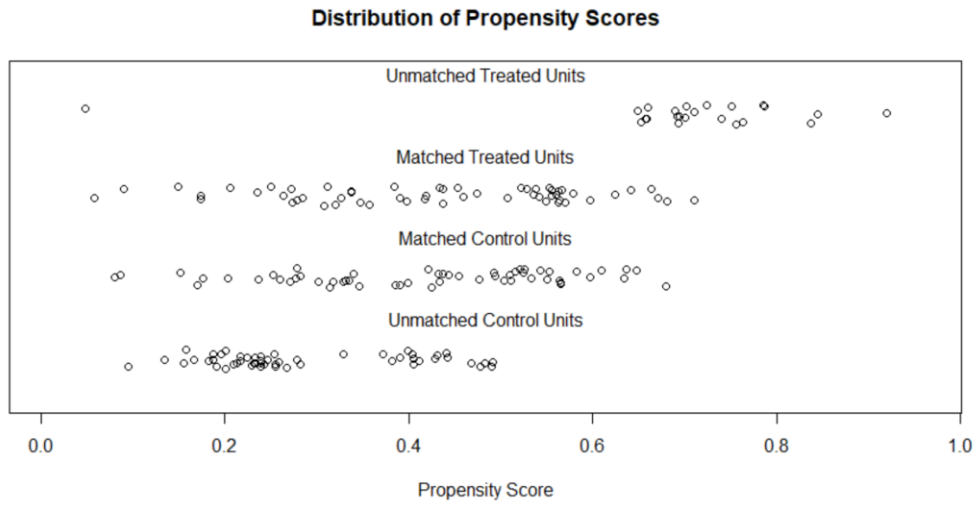

Figure 7.30: Pilot Study Analysis C propensity score distribution after version 2 of matching

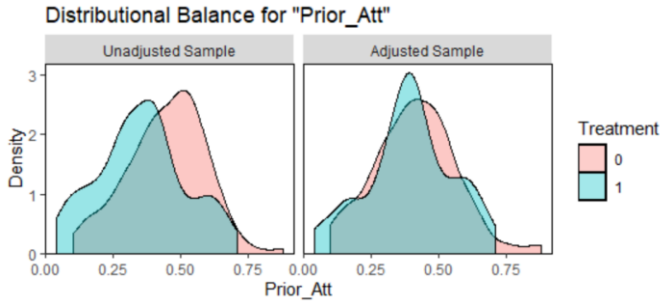

Figure 7.31: prior attainment balance before and after matching

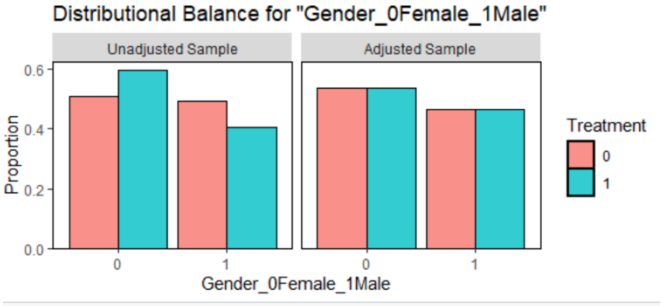

Figure 7.32: gender balance before and after matching

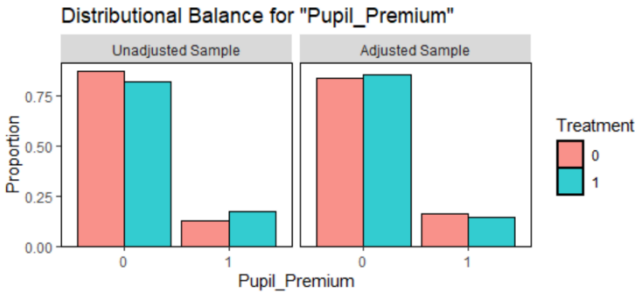

Figure 7.33: pupil premium balance before and after matching

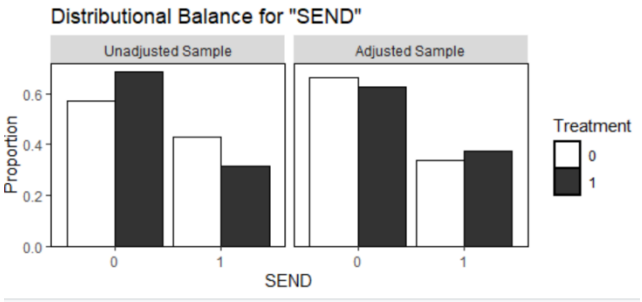

Figure 7.34: SEND balance before and after matching

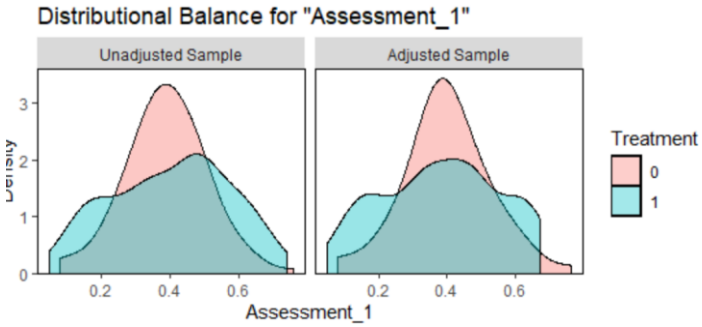

Figure 7.35: assessment 1 balance before and after matching

7.14.3 Version 3

This version used the same matching method as version 2, but widened the caliper to 0.3. The results are below:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 3 (nearest neighbour, caliper 0.3)	58	58	41.7	39.5	74.7	0.47	0.43	59.8	0.16	0.12	31	0.34	0.36	84.4	0.40	0.38	-54.9

Table 7.17: Pilot Study Analysis C version 3 matching results

The covariate balance for prior attainment, gender, pupil premium and assessment 1 achieved in version 3 was not as good as that realised in version 2 so it was concluded that, if a

caliper was to be used, version 2 provided the best balance (despite assessment 1 being more imbalanced than the unmatched data). Since assessment 1 was found to be a key driver of the propensity score, further matching methods were trialled to try and achieve balance across all covariates, as follows.

7.14.4 Version 4

In this version, the same matching method was used but without a caliper. The results are below, showing that more observations were retained.

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 4 (nearest neighbour, no caliper)	79	79	40.7	35.5	41.9	0.43	0.41	70.5	0.15	0.18	49.3	0.39	0.32	31.5	0.39	0.4	-9.1

Table 7.18: covariate balance after version 4 of matching for pilot study analysis C

The balance for all covariates other than assessment 1 improved from the unmatched data, but the -9.1 balance improvement score for assessment 1 indicated that it was not as imbalanced as with previous versions of matching. It was therefore used to estimate treatment effects. The results from version 2 did yield better balance for prior attainment, gender, pupil premium and SEND, but it is important to try and achieve balance across all five covariates and this result has produced balance for assessment 1. Figures 7.36-7.41 show that the best balance so far has been achieved across all five covariates, as well as the distribution of propensity scores for the matched units:

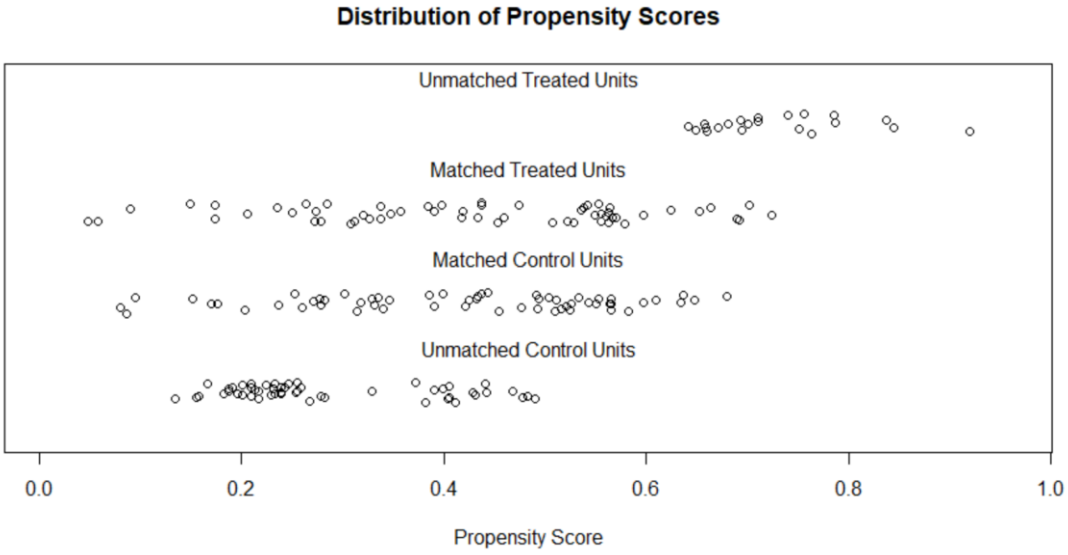

Figure 7.36: Pilot Study Analysis C propensity score distribution after version 4 of matching

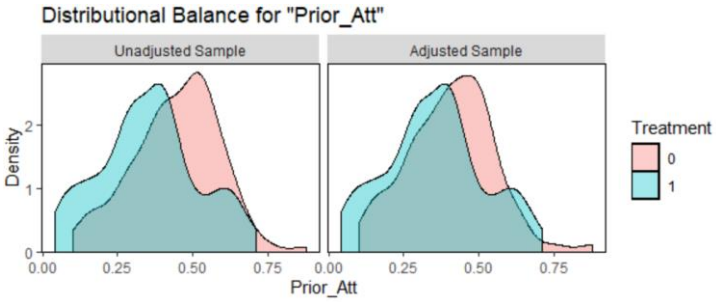

Figure 7.37: prior attainment balance before and after matching

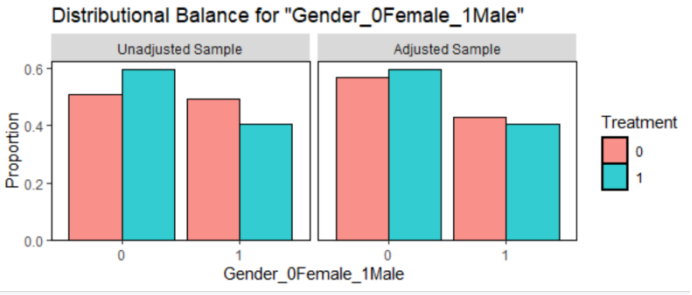

Figure 7.38: gender balance before and after matching

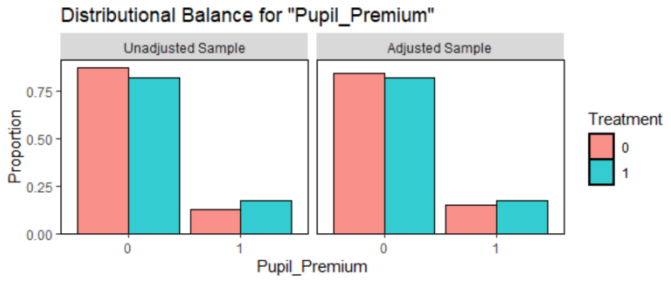

Figure 7.39: pupil premium balance before and after matching

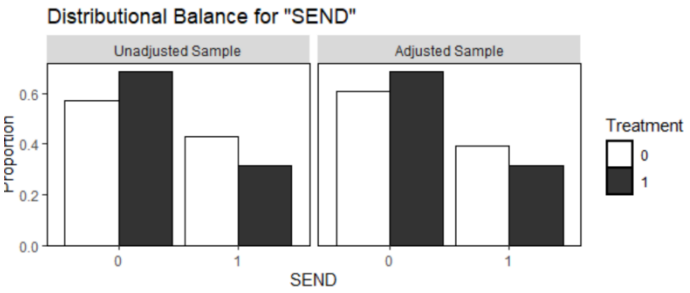

Figure 7.40: SEND balance before and after matching

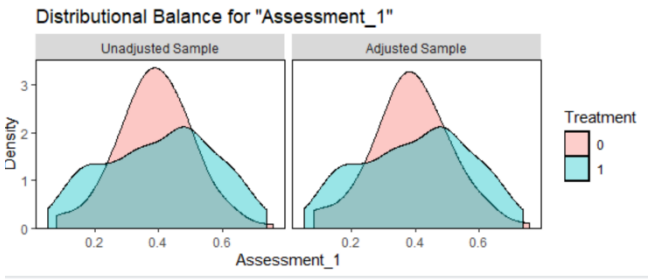

Figure 7.41: assessment 1 balance before and after matching

7.14.5 Version 5

The final version of matching for this analysis trialled a different method of matching: genetic matching. The results are below:

	No. of students		Prior Attainment %			Gender			Pupil Premium			SEND			Assessment 1 %		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 5 (genetic matching)	79	79	42.0	35.5	28.4	0.37	0.41	55.8	0.18	0.18	100	0.38	0.32	42.9	0.39	0.4	-13.8

Table 7.19: Pilot Study Analysis C version 5 matching results

The table shows that, compared to before matching, four of the five covariates had improved balance but, assessment 1 still became more imbalanced. That said, as the version used a different matching method, it was used as a sensitivity analysis. Figures 7.42-7.47 show the balance produced:

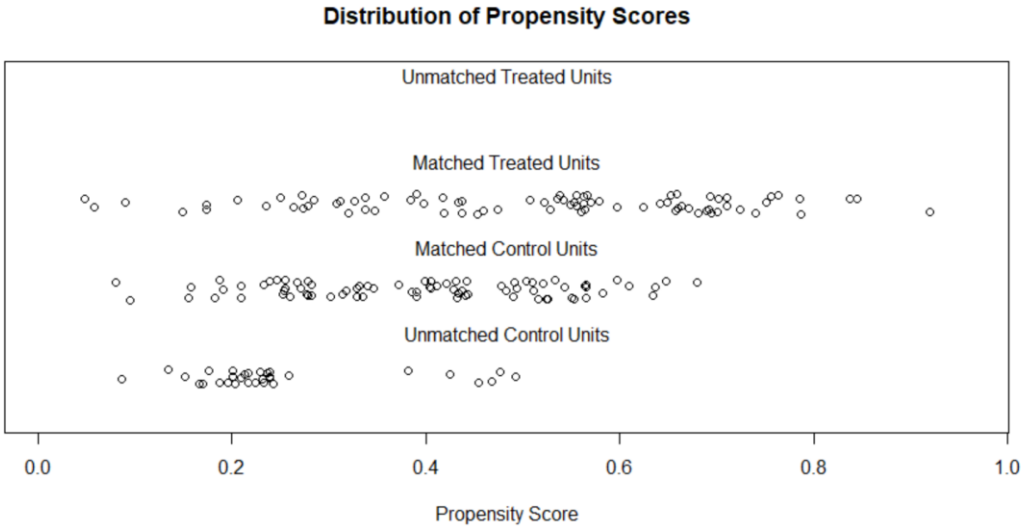

Figure 7.42: Pilot Study Analysis C propensity score distribution after version 5 of matching

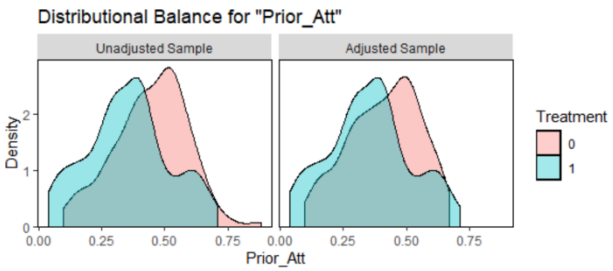

Figure 7.43: prior attainment balance before and after matching

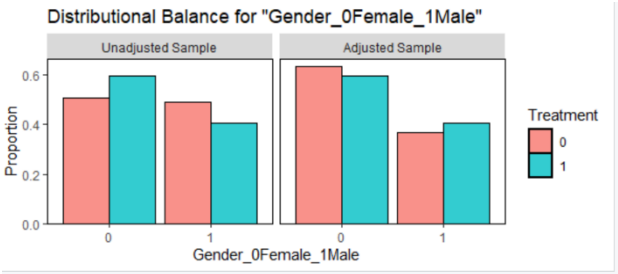

Figure 7.44: gender balance before and after matching

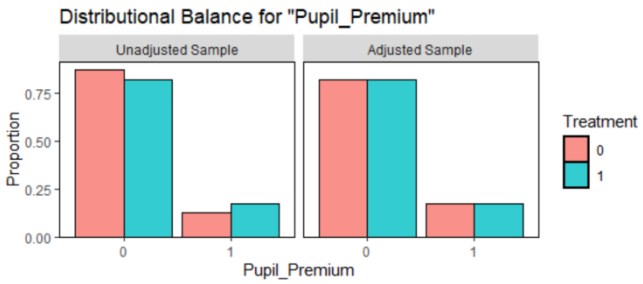

Figure 7.45: pupil premium balance before and after matching

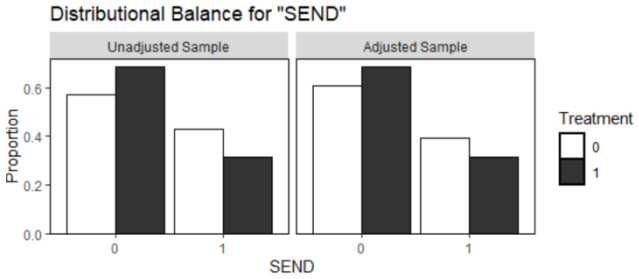

Figure 7.46: SEND balance before and after matching

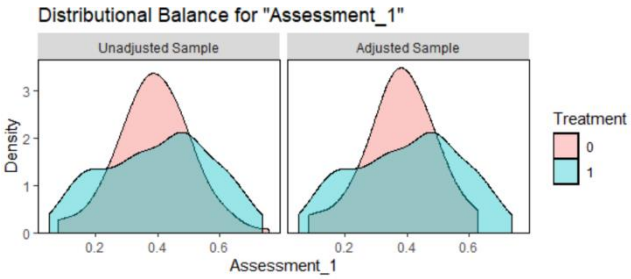

Figure 7.47: assessment 1 balance before and after matching

7.15 Estimation of Treatment Effect

Of the five versions of matching above, version 4 yielded the best overall balance, with the assessment 1 covariate the least imbalanced when comparing to the other versions. A paired t-test on the outcome measure (year 7 foundation outcome 2) found that the non-TfM group has a higher mean percentage score by 11.4% with a small p-value (<0.05), thus indicating significance. The confidence interval was 95% for the range [6.5%, 16.2%] which indicated that one can reasonably expect the effect of the difference in curricula to lie between 6.5% and 16.2%. Again, this is the average treatment effect on the 79 treated students (ATT) for which matches were found on the specific outcome measure, which was an assessment on number skills including place value, negatives, fractions, area and perimeter.

7.16 Sensitivity Analysis

Since versions 2 and 5 yielded very good balance for all covariates other than assessment 1, they were used for the sensitivity analysis check.

Using version 2, it was found that the non-TfM group did better on the foundation outcome measure 2, achieving on average, 12.3% higher than the TfM group. A small p-value (<0.00001) indicated significance and a confidence interval of 95% for the range [6.8%, 17.7%] suggested that one can expect the effect of the difference in curricula to lie between 6.8% and 17.7% on the percentage score (t(55)=4.5,p=<0.00001, 95% CI [6.8,17.7]).

Using version 5, it was also found that the non-TfM group did better on the foundation outcome measure 2, achieving on average, 9.84% higher than the TfM group. A small p-value (<0.05) indicated significance and a confidence level of 95% for the range [4.9, 14.8] suggesting that one can expect the effect of the difference in curricula to lie between 4.9% and 14.8% on the percentage score (t(78)=3.9,p=0.0001, 95% CI[4.9,14.8]).

As with analysis A and B for the pilot study, these results must be viewed cautiously since the effects of Covid cannot be disaggregated.

7.17 Lessons for the Main Study

As discussed at the beginning of this chapter, the purpose of the pilot study was to trial propensity score matching methods on one set of data to help pre-determine the steps to take forward to the main study, since the effects of Covid could not be disaggregated from the data provided. The pilot study helped to highlight some key issues with using propensity score matching on the available data. The main issue is that, for analysis A, most of the covariates used were binary and therefore the propensity scores for each observation were very clustered. Analysis B and C showed that, when using more than one continuous covariate (prior attainment and assessment 1), the propensity scores become less 'lumpy'. Therefore, for the main study, schools were asked to provide as many continuous variables as possible. Another issue highlighted was the difficulty in achieving balance across all included covariates. As the balance improved for some covariates in the pilot study, it become more imbalanced for others. Therefore, it is important to look at the key drivers of the propensity score before matching to help decide on which matching result is the best to use to estimate treatment effect.

7.18 Objectives Realisation

Each analysis within the pilot study involved applying different versions of matching before estimating the treatment effects. Each analysis started with the same version of matching: 1:1 nearest neighbour matching without replacement and with a caliper of 0.1. Following the first version, matching was trialled with wider and narrow calipers, as well as without a caliper for the same type of matching. For the subsequent versions, different types of matching were trialled, such as genetic matching and optimal full matching. It is important to test a variety of matching methods before estimating treatment effects to ensure that optimal balance across the two treatment groups is achieved on the given covariates, which is why each analysis assessed at least 5 different versions of matching.

As stated, the purpose of the pilot study was to confirm the efficacy of the steps to use for the main study and to determine a direction where there is enormous flexibility surrounding choice of matching method, caliper and included covariates. The chapter has achieved these objectives and the subsequent section seeks to set out the steps for the main study through pre-registration.

7.19 Steps for the Main Study: Pre-Registration

It was important to pre-register the process that was used ahead of a main study and therefore, following the lessons learnt from the pilot study, the flow chart below sets out the steps that were taken in the matching process for the main study. After propensity score estimation, the first matching version was the nearest neighbour matching with a caliper of 0.1. Following this, different calipers were used (wider and narrower). Greedy and full matching were also be used so that a range of matching algorithms were assessed before estimating treatment effects.

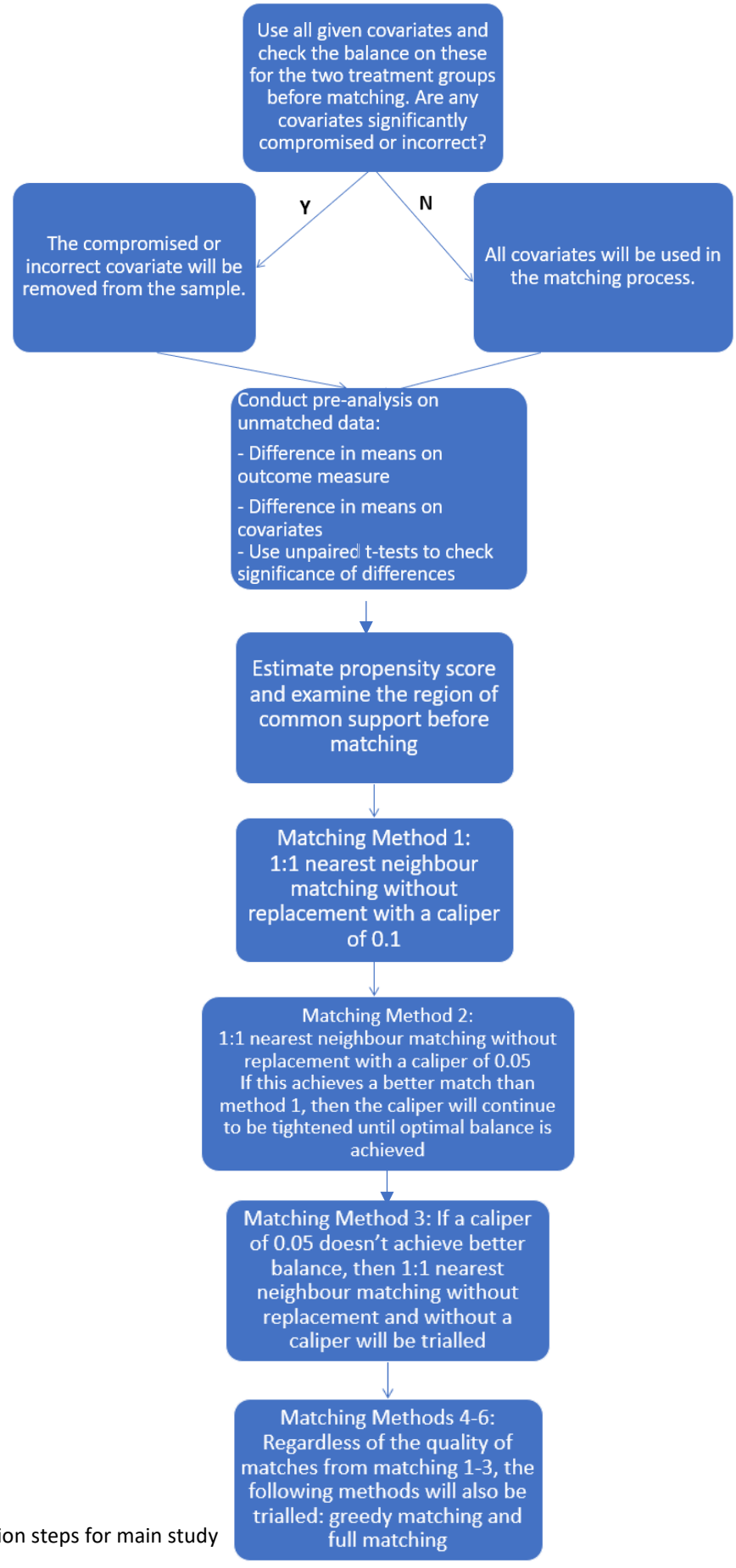

Figure 7.48: pre-registration steps for main study

After all matching methods outlined above were trialled, the results which produced the best matches were taken into the next stage of estimating the treatment effect. The criteria for 'best matches' are where covariates that were initially imbalanced become balanced without unbalancing ones that were initially balanced. The treatment effect was then estimated using paired t-tests. Alternative matching methods to the one taken to estimate the treatment effect were used as the sensitivity analyses.

These steps were pre-registered on osf.io and can be found in the OSF registries via the link: https://osf.io/rfvmz

Chapter Eight: Implementation of Methods – Main Study (Part 1)

8.1 Introduction

The previous chapter detailed the methods used and the results of the pilot study, focusing on school X. At the end of the pilot study chapter, the importance of 'pre-registering' the steps that plan to be used for the main study was highlighted, and a flow-chart of propensity score matching steps was shown to help overcome the danger of 'the garden of forking paths' which was outlined by Gelman, warning of bias where the researcher may land on the approach that influences their viewpoint (Gelman and Loken, 2013).

The overall steps that the main study follow align with Leite's (2016) six steps to propensity score matching and the intricate decision-making steps of the crucial matching process follow the flow-chart shown in the previous chapter, pre-registered on https://osf.io/rfvmz.

The main study is split into two parts. Each of the two parts takes one participating school and seeks to evaluate the impact of TfM on assessment outcomes, compared to previous pedagogical approaches, using the same non-experimental methods as adopted in the pilot study.

For school Y, three analyses are attempted. Two of these are included for methodological interest alone (Analysis A and B), whilst analysis C is the one suitable analysis for addressing any causal effect of the difference in curricula.

8.2 Interview and School Context

School Y is a comprehensive, co-educational secondary school and sixth form, catering for students from age 11 to 18 for boys and girls. There are approximately 1200 students on roll, and the most recent Ofsted inspection of 2013 graded the school as 'Outstanding' in all areas. DfE performance tables show an overall progress 8 score for the school as being "well above average".

The interview for School Y was conducted with Teacher Y who has worked at the school for 13 years and is Head of Maths, having held this position for 3 years. Teacher Y helped to lead the change to TfM in the 2017-18 academic year. The semi-structured interview used the same open questions as the pilot study, which can be found in Appendix B, and the outcome of the interview fits two main themes: available data for analysis and the school's transition to TfM.

When asked about their understanding of TfM and the main changes to their curricula since the transition, teacher Y commented that their KS3 curriculum had been stripped back hugely to ensure that longer is spent on topics in years 7 and 8, which can be considered the 'threshold' years ensuring students have deep conceptual understanding of the maths they have studied before embarking on the KS4 curriculum, which school Y starts with year 9 students. Teacher Y also remarked that another big change following the transition to TfM has been the interactive nature of the lessons, using frequent AfL and diagnostic questions. Additionally, each lesson across all three key stages seeks to embed fluency, reasoning and problem-solving so that students are challenged to think and reason mathematically, as well as answer non-standard problems. School Y's transition to TfM was spearheaded through their involvement in the regional Maths Hub which provided the school with lesson resources and time to explore schemes of work such as White Rose, before formalising their own unique curriculum.

Following the interview, it was considered that school Y's perception of TfM largely aligned with the criteria set out in the chapter 2 of this thesis. A stripped back curriculum to ensure depth of understanding through small steps fits with the following two defining characteristics:

- *Where topics are taught for a longer period to ensure depth of understanding*
- *Where there is an emphasis on teaching for conceptual understanding and forming mathematical connections.*

By seeking to ensure that each lesson builds in fluency, reasoning and problem-solving, school Y's perception of TfM aligns with the following three defining characteristics:

- *Where variation is embedded in chosen examples and exercises*
- *Where there is insistence on key mathematical language*
- *Where problem-solving is considered a key aim of the curriculum.*

The final characteristic of TfM that was set out in the literature review was *'Maths for All' – where all students are exposed to the same mathematical content and some content is not preserved for higher attaining students*. Whilst school Y has not changed to 'mixed attainment' groupings, the year group are all exposed to the same topics but at different paces. Any classes that grasp content quickly are challenged through depth rather than moving onto new content.

School Y first introduced TfM in the 2017-18 academic year, but as many of the topic assessments have been revised since the transition to TfM, only two possible assessments could be used for comparison. Details of these assessments are defined in the next section of this chapter. Teacher Y confirmed that there were no other contemporaneous events that occurred alongside the

transition to TfM that could have caused a change in assessment outcomes, other than the Covid-19 crisis which has had a knock-on effect globally. Dates that cross the pandemic years (March 2020 onwards) have been highlighted in the analysis.

The sample size provided by School Y was 602 students, made up of the following:

- 209 students that were in year 10 during the 2020-21 academic year and had been taught using TfM
- 208 students that were in year 10 during the 2019-20 academic year and had not been taught using TfM but instead with a previous pedagogical approach
- 185 students that were in year 10 during the 2018-19 academic year and had not been taught using TfM but instead with a previous pedagogical approach.

It is important to acknowledge that some of the data provided overlaps with the Covid pandemic and therefore any treatment effect detected could not be disaggregated from the effects of such. Where the outcome measure overlaps with the pandemic, the analyses will only be kept in this chapter for methodological purposes to allow for some of the interesting issues which arise.

8.3 Sample and Data Preparation

For this sample, two pieces of comparable assessment data were provided. The 2020-21 and 2019-20 year 10 had both sat the same assessment in year 10, and the 2020-21 year 10 and 2018-19 year 10 had sat the same assessment when in year 9. For the year 9 assessment, students either sat the 'higher' or the 'foundation' tier which were different assessments. This was the same for the year 10 assessment, with students entered for either the 'higher' or the 'foundation' paper. As such, it was initially thought that school Y's data would be split into four mini-studies for analysis, titled by the assessment outcome of interest:

- Analysis A: Year 10 Higher
- Analysis B: Year 10 Foundation
- Analysis C: Year 9 Higher
- Analysis D: Year 9 Foundation

Whilst this was the planned structure for analysis based on the given data, the sample size for analysis D was too small (21 non-TfM students and 36 TfM students) and therefore, the decision was made to remove the analysis from this study.

The covariates and data provided by school Y for each cohort differed slightly depending on the data records for each year. For the analysis of year 10 assessment outcomes (analysis A and B), the following covariates and data were provided:

- Gender (male or female)
- Pupil Premium status
- EAL status (English as an additional language)
- FSM status
- SEND status
- Midyis vocabulary score
- Midyis maths score
- Midyis non-verbal score
- Midyis skills score
- Midyis overall score
- KS2 grammar scaled score
- KS2 maths scaled score
- KS2 reading scaled score
- Outcome measure: year 10 assessment percentage (higher or foundation), which was a full GCSE past paper.

The pilot study used only the KS2 Maths scaled score as its baseline, whereas analysis A and B of this part of the main study used all available KS2 scores. It was shown in the pilot study that better matching is achieved when there are more continuous variables present in the analysis. The Midyis data was collected by the school from the Centre for Evaluation and Monitoring (CEM) which over 3000 secondary schools nationwide use to assess students on vocabulary, maths, non-verbal and literacy skills. It was decided that the Midyis overall score would be used as the chosen covariate since it is calculated using the component scores. In total, 78 subjects across analysis A and B were removed because of missing data. This is a substantial proportion of subjects to remove due to missing data than observed in the pilot study but recognising analysis A and B as having methodological value, it was decided to continue with the exploration.

For the analysis of year 9 assessment outcome (analysis C), the following covariates and data were provided:

- Gender (male or female)

- EAL status
- FSM status
- Pupil Premium status
- SEND status
- KS2 Maths scaled score
- Midyis overall score (note: whilst for the current year 10 cohort, the separate midyis components were available, only the overall score was used since this was the only component available for the current year 12 cohort).
- Outcome measure: year 9 assessment percentage (higher or foundation), which was split into three units. Unit 1 assessment was on Number; Unit 2 assessment was on Algebra; and Unit 3 was Data.

For students that sat the higher tier of the year 9 assessment, which made up analysis C, the pre-matching difference in means analysis revealed that the non-TfM group of students had an average of 0 for FSM which is deemed infeasible. Therefore, the FSM covariate was removed from the sample, and not used in any further analysis. For analysis C, 35 subjects were removed because of missing data, absences during assessments, and errors in the data cells. Whilst this number is larger than one would like, proportionally it equated to 13% which can be deemed a moderate level of 'missingness', as set out in the methodology chapter, and therefore complete case analysis is regarded an effective approach for dealing with the missing values. Following the unit removal, a sample size of 125 non-TfM and 117 TfM students remained, which is an adequate sample size.

The aim of analysis A, B and C was to conduct a propensity score matching analysis of the (ATT) impact of changing from non-TfM (coded 0) to TfM (coded 1) on the outcome assessment percentage, matching on propensity based on the following, as applicable for each analysis: gender (coded 0 for female, 1 for male), pupil premium (coded 0 for non-PP, 1 for PP), EAL (coded 0 for non-EAL, 1 for EAL), SEND (coded 0 for no additional need, 1 for any additional need), FSM (coded 0 for non-FSM, 1 for FSM), Midyis overall score (taken as scaled score), KS2 grammar score (taken as scaled score), KS2 maths score (taken as scaled score), and KS2 reading score (taken as scaled score). Once the propensity score matching process was complete, for analysis C, the outcome measure (taken as percentage scores) was considered for the two matched groups to look at the average treatment effect on the treated (ATT).

8.4 Analysis A: Characteristics before Matching

As discussed, analysis A and B are important to include for methodological purposes. Since this thesis is exploring alternative methods for assessing the impact of TfM, it is important to show what can happen in different versions of propensity score matching.

Analysis A used the data provided by school Y on the year 10 assessment percentage of the students that sat the higher tier. Like the pilot study, the distribution of baseline characteristics for the non-TfM and TfM groups were analysed to determine the difference between the two groups for each covariate and the outcome measure. Using z-tests, unpaired t-tests and difference-in-means analysis, the following results were obtained:

	Non-TfM group	TfM group	
Number of students	94	168[21]	
Covariates	Mean	Mean	p-value
KS2 Maths Score (mean, standard deviation)	108.79, 4.45	106.94, 5.76	0.004
KS2 Grammar Score (mean, standard deviation)	112.67, 5.45	110.61, 6.77	0.008
KS2 Reading Score (mean, standard deviation)	105.64, 2.45	109.36, 6.48	<0.00001
Gender	0.36	0.5	0.03
Pupil Premium	0.021	0.06	0.11
SEND	0.14	0.14	0.92
FSM	0	0.04	0.008
EAL	0.096	0.196	0.02
Midyis overall (mean, standard deviation)	117.62, 9.73	113.29, 13.54	0.003
Outcome Measure			Unpaired t-test result
Outcome percentage (mean, standard deviation)	33.9%, 15.02	38.1%, 18.24	t(225) = -2.03, p=0.04, 95% CI [0.12%, 8.37%]

Table 8.1: Main Study 1 Analysis A characteristics before matching

Several comments can be made from table 8.1. Without any matching or adjustment, the TfM group did slightly better (by 4.2%) on average on the outcome measure. However, differences in the covariates between the two treatment groups highlighted a need for matching before outcome measures could be considered.

For the above covariates, two-sample z-tests and unpaired t-tests revealed small p-values (< 0.05) for KS2 reading, KS2 maths, KS2 grammar, Midyis, FSM, EAL and gender, indicating that the same difference would unlikely be seen from random allocation to the two treatment groups alone. Pupil premium had a reasonably small p-value, but not at the conventional level of significance (0.05). SEND had a very large p-value of 0.92, suggesting that the difference is not significant for this

[21] The numerical imbalance in non-TfM and TfM students in analysis A is because of the large number of students that were removed from the sample due to missing data.

covariate. Using an unpaired t-test, the difference in the mean outcome measure yielded a p-value which suggested the difference is not compatible with a model in which random allocation to groups alone was a causal factor (t(225)=-2.03, p=0.04, 95% CI [0.12%, 8.37%]).

8.5 Analysis A: Propensity Score & Matching

Following Leite's (2016) six steps to propensity score matching, and the intricate steps set out in the pre-registration process for matching, six versions of matching were conducted to ensure that optimal balance was found before treatment effects were estimated. Unlike the pilot study which walked through all the decision-making steps for each version of matching, all six versions in the main study are presented together. The type of matching used for each version is shown in brackets by each version number Table 8.2, below, along with the mean for each covariate after matching and balance improvement.

The pre-matching analysis of the propensity score revealed that KS2 maths, gender, KS2 grammar and KS2 reading were the main drivers of the propensity score, based on the significant codes. The propensity score was calculated in the same way for School Y as School X using a logit model. Consequently, careful consideration was given to these covariates when seeking to optimise balance. The small p-values that these covariates generated indicates that the imbalance between the two treatment groups is most unlikely to just come from random allocation to the two groups and, therefore, may be perceived as the 'real differences' between the two groups.

	No. of students		KS2 Maths			Gender			EAL			Pupil Premium			SEND			FSM			MidYis			KS2 Grammar			KS2 Reading		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	50	50	106.58	107.222	65.3	0.4	0.52	13.2	0.14	0.18	60.3	0.04	0.04	100	0.18	0.16	-338.7	0.0	0.02	52	114.5	113.7	83.4	110.6	110.4	91.3	106.2	106.4	94.6
Version 2 (nearest neighbour, caliper 0.2)	55	55	106.98	107.07	95.1	0.44	0.49	60.6	0.15	0.2	45.8	0.04	0.05	52.5	0.16	0.16	100	0	0.04	12.7	114.9	115.8	80.2	110.6	110.6	96.5	106.1	106.6	86.3
Version 3 (nearest neighbour, no caliper)	94	94	108.79	106.73	-11.2	0.36	0.57	-53.8	0.096	0.26	-58.5	0.02	0.09	-66.9	0.14	0.14	100	0	0.07	-78.7	117.6	112.2	-26.2	112.7	110.7	5.1	105.6	111.9	-67.7
Version 4 (genetic matching)	94	94	108.79	106.73	-11.2	0.36	0.57	-53.8	0.096	0.26	-58.5	0.02	0.09	-66.9	0.14	0.14	100	0	0.07	-78.7	117.6	112.2	-26.2	112.7	110.7	5.1	105.6	111.9	-67.7
Version 5 (full matching)	94	168	106.23	106.94	61.4	0.63	0.5	6.2	0.35	0.196	-50.1	0.09	0.09	22.2	0.17	0.14	-596.3	0	0.04	0	112.9	113.3	90	107.9	110.6	-30.9	107.2	109.4	40.6
Version 6 (optimal matching)	94	94	108.79	107.65	38.4	0.36	0.48	15.4	0.096	0.16	36.6	0.02	0.03	72.2	0.14	0.15	-133.3	0	0	100	117.6	113.4	1.9	112.7	110.7	5.1	105.6	107.1	61.7

Table 8.2: Main Study 1 Analysis A propensity score variants and result balance

After consideration of the overall balance improvement for each version of matching (using the same methods that were used for school X), it was concluded that nearest neighbour matching with a caliper of 0.2 (version 2) yielded the best overall set of matches since the balance improvement for each covariate was positive. Versions 1 and 6 were used as sensitivity analyses following the treatment effect estimate since they were the second-best overall matches.

Figures 8.1 and 8.2 below show the propensity score distribution for the non-TfM group (curriculum type: other) and the TfM group (curriculum type: TfM) before matching and after matching.

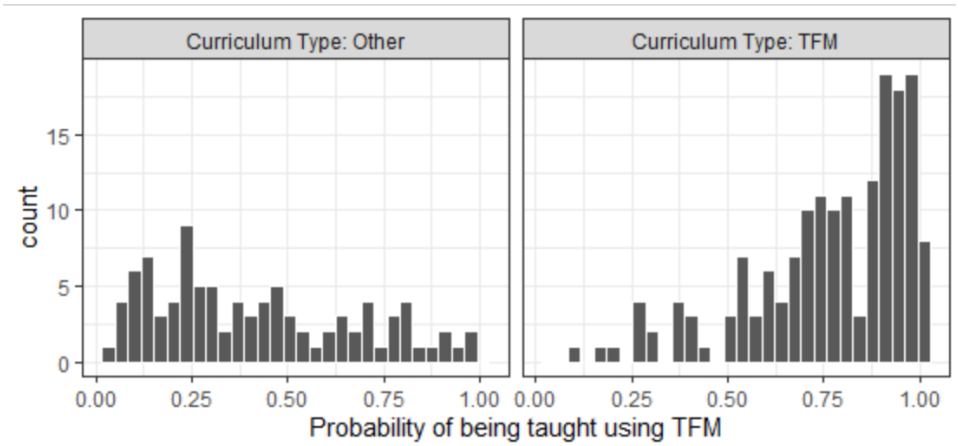

Figure 8.1: Main Study 1 Analysis A Propensity Scores

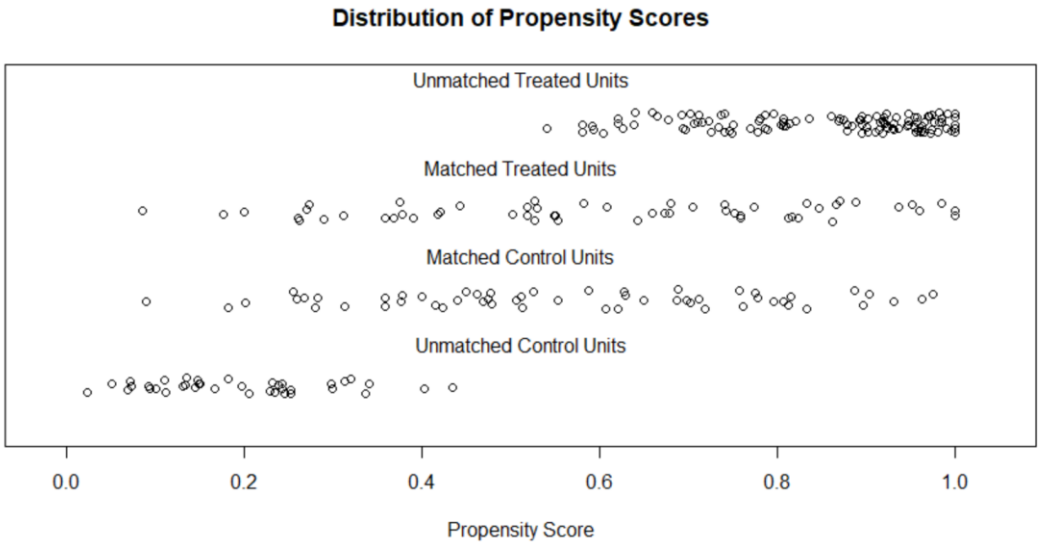

Figure 8.2: Main Study 1 Analysis A propensity score distribution after version 2 of matching

Figures 8.3-8.11 show the improved balance for each covariate for version 2 of matching, which was used to estimate treatment effects.

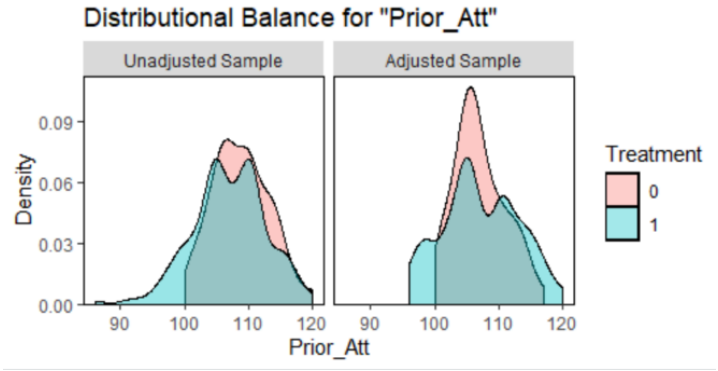

Figure 8.3: prior attainment balance before and after matching

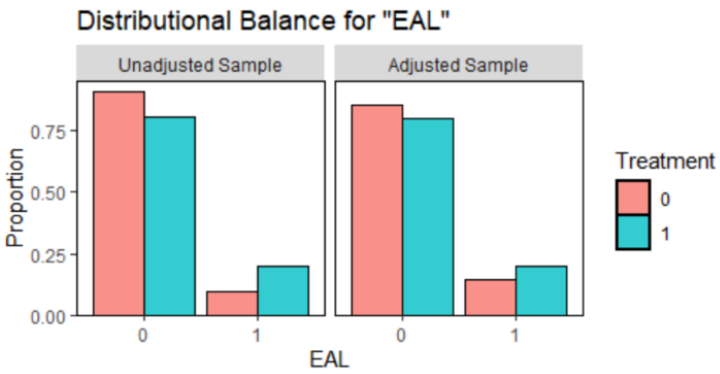

Figure 8.4: EAL balance before and after matching

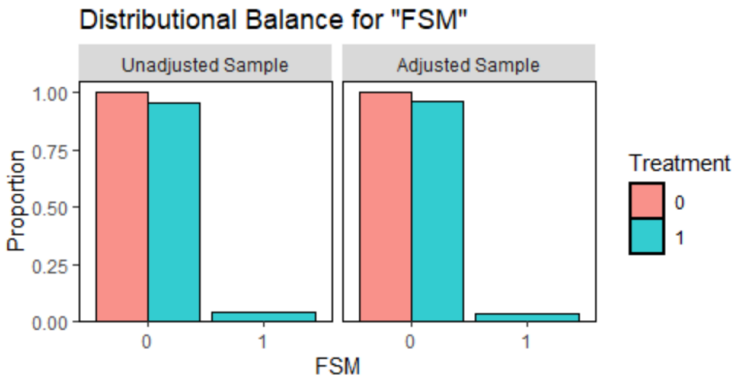

Figure 8.5: FSM balance before and after matching

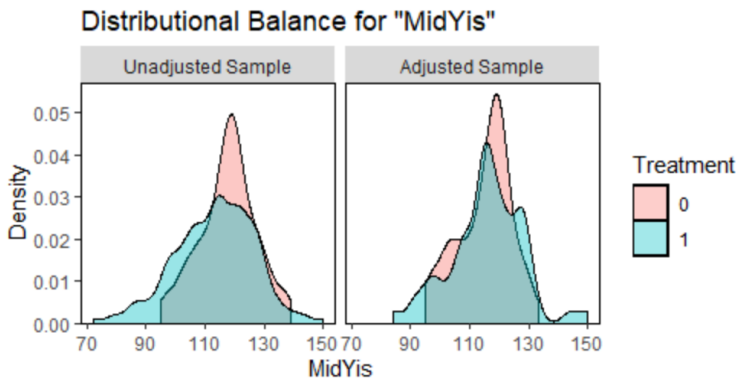

Figure 8.6: MidYis balance before and after matching

Figure 8.7: SEND balance before and after matching

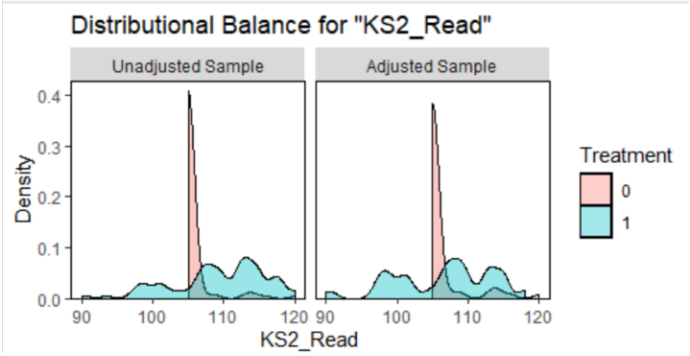

Figure 8.8: KS2 Reading balance before and after matching

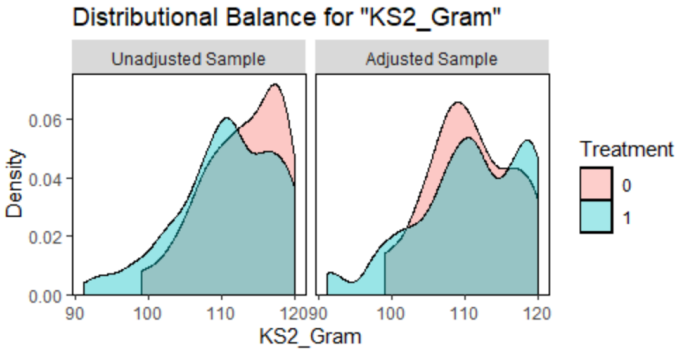

Figure 8.9: KS2 Grammar balance before and after matching

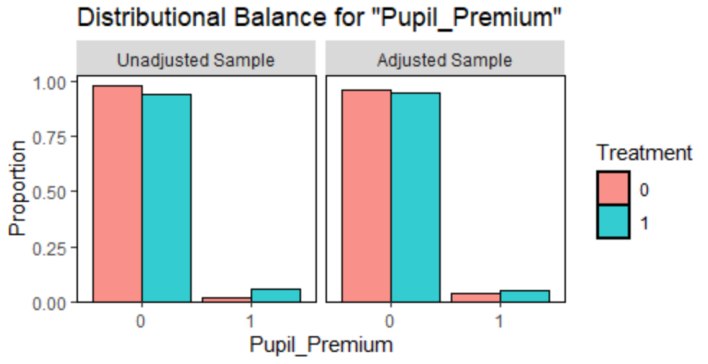

Figure 8.10: pupil premium balance before and after matching

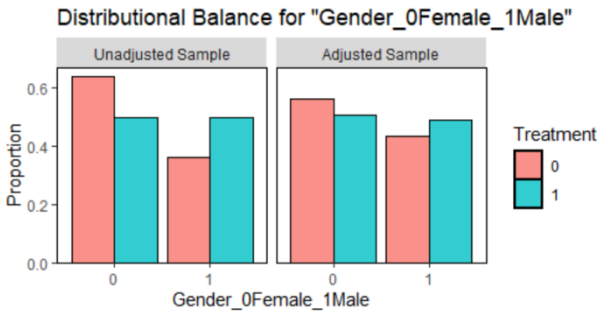

Figure 8.11: gender balance before and after matching

8.6 Analysis A: Treatment Effects and Sensitivity Analysis

Since version 2 created the most balanced groups for comparison, the matched groups were used to estimate the treatment effect. As with the pilot study, treatment effects were estimated through the use of paired t-tests on the outcome measure (year 10 higher assessment percentage). Using a t-test, it was found that the TfM group had a higher mean outcome by 5.3% on the assessment percentage, with a small p-value (= 0.035) which indicated statistical significance with a 95% confidence interval for the range [0.38%, 10.16%]. The confidence interval suggests that one can be 95% confident that the effect of the TfM curricula lies between 0.38% and 10.2% (t(93)=2.1,p=0.035, 95% CI [0.38,10.16]).

The sensitivity analysis that used version 6 of matching concurred with the results from version 2. Using a further t-test, it was estimated that the TfM group had a higher mean outcome by 4.02% on the assessment percentage, with a reasonably small p-value (= 0.12) but not at the conventional 0.05 threshold of significance (t(93)=1.6,p=0.12), 95% CI [1.04,9.8]). Interestingly, the sensitivity analysis which used version 1 of matching estimated that the TfM group had a higher mean outcome by 10.96% on the assessment percentage, with a small p-value (= 0.00054) for the range [5.02%, 16.9%] (t(49)=3.7,p=0.0005, 95% CI [5.02,16.9]). The estimate of 10.96% is substantially higher than the estimate of 5.3% and 4.02% in version 2 and 6 respectively. Thus, the sensitivity analysis suggests that the result found is not robust to changes in the propensity score specification and may suggest that, in this specific case, propensity score matching is not a very powerful way of getting to the causal effect of curriculum change. Of course, it is also worth pointing out again that data provided for analysis A overlapped with Covid, too and therefore causal conclusions were be made from this section.

8.7 Analysis B: Characteristics before Matching

The second analysis of the data provided by school Y reviews year 10 assessment percentage of the students that sat the foundation tier. Like analysis A, the distribution of baseline characteristics for the non-TfM and TfM groups were analysed to determine the difference between the two groups for each covariate and the outcome measure. Using two-sample z-tests, unpaired t-tests and difference-in-means analysis, the following results were obtained:

	Non-TfM group	TfM group	
Number of students	43	17	
Covariates	Mean	Mean	p-value
KS2 Maths Score (mean, standard deviation)	101.19, 3.59	94.82, 4.7	< 0.00001
KS2 Grammar Score (mean, standard deviation)	105.12, 5.57	96.59, 6.97	< 0.00001
KS2 Reading Score (mean, standard deviation)	105.23, 1.73	96.29, 6.02	0.0002
Gender	0.488	0.65	0.27
Pupil Premium	0.07	0.24	0.16
SEND	0.40	0.65	0.09
FSM	0.02	0.12	0.27
EAL	0.16	0.59	0.005
Midyis overall (mean, standard deviation)	102.47, 8.7	112.76, 11.39	0.01
Outcome Measure			Unpaired t-test result
Outcome percentage (mean, standard deviation)	50.5%, 12.84	30.4%, 14.59	t(49) = 7.4, p<0.00001, 95% CI [14.7%,25.5%]

Table 8.3: Main Study 1 Analysis B characteristics before matching

Before discussing the z-test and t-test results, it is important to highlight that the sample group size is very small for meaningful analysis. It has only been included as it highlights key methodological lessons learnt about the common support region when using propensity score matching with a small population, which follows.

Table 8.3 shows that, without any matching, the average outcome percentage is very different for the two treatment groups, with the non-TfM group going around 20% higher on average. The differences in the covariates between the two groups highlight the need for matching.

For the covariates listed in table 8.3, unpaired t-tests and two-sample z-tests revealed small p-values (< 0.05) for KS2 maths, KS2 reading, KS2 grammar, EAL, and Midyis, indicating that the same differences would be unlikely seen if random allocation to the two treatment groups was the only

adjustment made. Gender, pupil premium, SEND and FSM has reasonably small p-values (0.27, 0.16, 0.085 and 0.27 respectively), but not at the conventional level of significance (0.05).

Using an unpaired t-test, the difference in the mean outcome measure yielded a small p-value of which suggested the difference is not compatible with a model in which random allocation to groups alone was a causal factor ($t(49)=7.4, p<0.00001$, 95% CI [14.7%, 25.5%]).

8.8 Analysis B: Propensity Score & Matching

The estimate of the propensity score for analysis B revealed that none of the included covariates were considered main drivers of the propensity score since none of them yielded a 'significant' p-value, though that may be influenced by the small sample size. As such, it was important within analysis B to try to gain balance for all covariates. In total, five versions of matching were conducted until the researcher was satisfied that the best balance had been obtained to estimate the treatment effects. The type of matching used for each version is defined in brackets in Table 8.4.

	No. of students		KS2 Maths			Gender			EAL			Pupil Premium			SEND			FSM			MidYis			KS2 Grammar			KS2 Reading		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	2	2	96.5	97	92.1	1	0.5	-215.1	0	0.5	-17.5	0	0	100	0	0.5	-98.6	0	0	100	104.5	100.5	61.2	106	103	64.8	105	103.5	83.2
Version 2 (nearest neighbour, no caliper)	17	17	98.6	94.8	40.8	0.47	0.65	-11.2	0.29	0.59	30.9	0.18	0.24	64.5	0.35	0.65	-16.8	0.06	0.12	37.7	102.8	112.8	2.9	102.7	96.6	28.3	105.6	96.3	-4.6
Version 3 (genetic matching)	17	17	99.2	94.8	31.6	0.53	0.65	25.9	0.29	0.59	30.9	0.12	0.24	28.9	0.65	0.65	100	0.06	0.12	37.7	102.5	112.8	0.1	102.2	86.6	33.8	104.8	96.3	4.6
Version 4 (full matching)	43	17	96.9	94.8	66.9	0.79	0.65	6.9	0.07	0.59	-22.3	0.005	0.24	-39.4	0.14	0.65	-101.3	0.06	0.12	37.7	106.1	112.8	35.4	102.4	96.6	31.3	104.96	96.3	3.1
Version 5 (optimal matching)	17	17	98.8	94.8	38.1	0.41	0.65	-48.3	0.24	0.59	17	0.18	0.24	64.5	0.29	0.65	-40.2	0.06	0.12	37.7	102	112.8	-4.5	103.4	96.6	20	105.7	96.3	-5.3

Table 8.4: Main Study 1 Analysis B propensity score variants and result balance

On the surface, it appears that genetic matching (version 3) yielded the best overall balance since the balance improvement statistic for each covariate was positive. However, exploration of the propensity score distribution in Figures 8.12 and 8.13 for both treatment groups reveals that there is virtually no common support before matching and therefore there are no subjects alike:

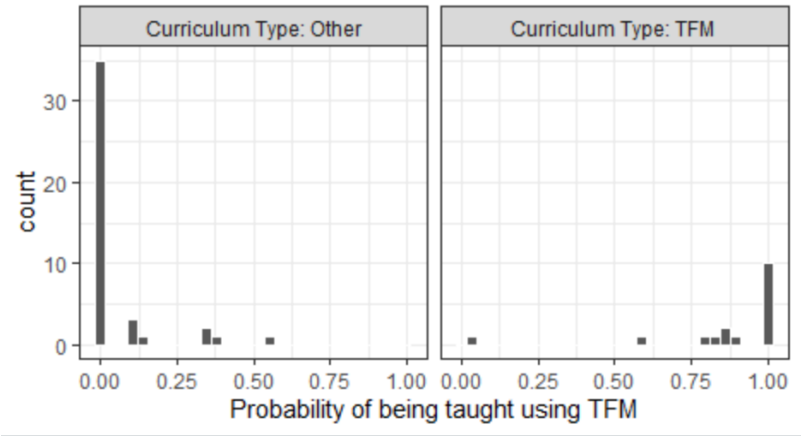

Figure 8.12: Main Study 1 Analysis B Propensity Scores

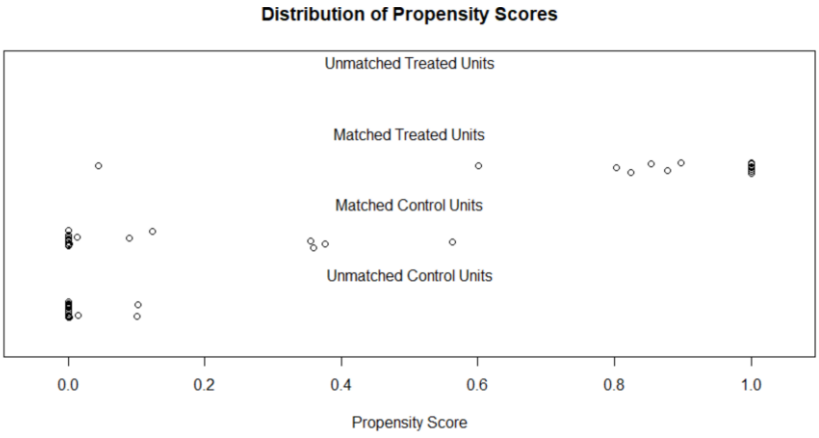

Figure 8.13: Main Study 1 Analysis B propensity score distribution after version 3 of matching

The visual depiction of the lack of common support exhibited in Figures 8.12 and 8.13 helps explain why, in version 1 of matching, only two subjects from each treatment group were matched. Nearest neighbour matching without a caliper, genetic, full and optimal matching resulted in more matches made but for subjects that are not alike on propensity scores. A selection of the covariate distribution plots below from version 3 (genetic matching) illustrates this:

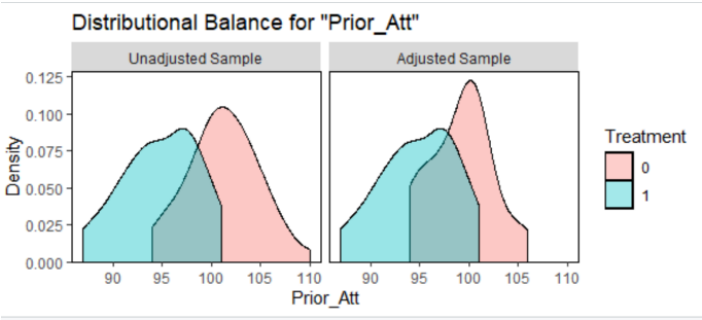

Figure 8.14: prior attainment balance before and after matching

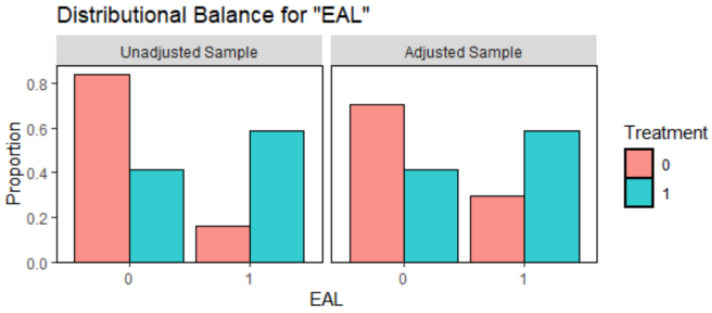

Figure 8.15: EAL balance before and after matching

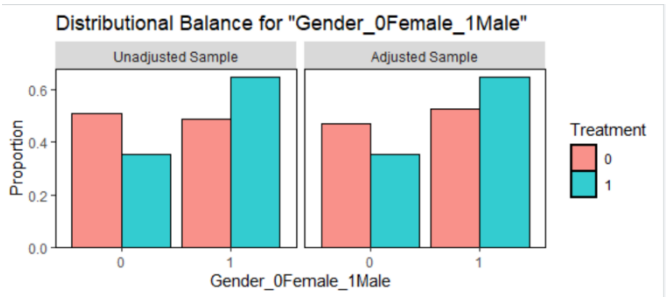

Figure 8.16: gender balance before and after matching

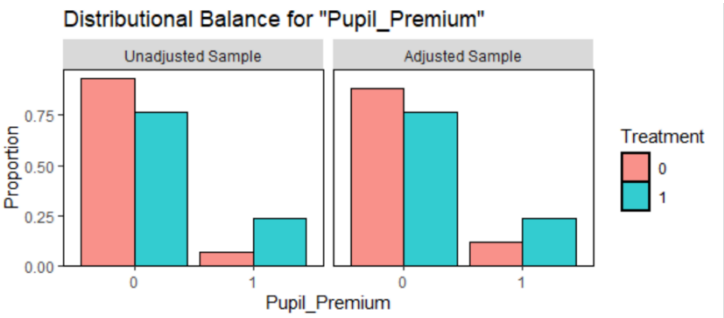

Figure 8.17: pupil premium balance before and after matching

Thus, whilst the balance improvement statistics in Table 8.4 suggest that version 3 of matching resulted in well-matched groups, it is clear that the two treatment groups actually remained very imbalanced after matching. The 'adjusted sample' charts in Figures 8.14-8.17 show very little balance improvement and therefore when matches were forced (version 3 and later in Table 8.4), students who were very different on their propensity scores were the ones who were matched.

It was decided, therefore, that analysis B was not to be taken any further in terms of treatment estimation. However, it is important to include it within this chapter as a discussion point as an example of the importance of looking beyond the surface when conducting propensity score matching. If there is no region of common support before matching, then the matches that will result are of unlike subjects that are not comparable on the given covariates.

8.9 Analysis C: Characteristics before Matching

The third analysis of the data provided by school Y uses the year 9 outcome percentage for the students that sat the higher tier. The distribution of the baseline characteristics for the non-TfM and TfM groups can be seen below:

	Non-TfM group	TfM group	
Number of students	125	117	
Covariates	Mean	Mean	p-value
KS2 Maths score (mean, standard deviation)	107.49, 4.4	109.29, 4.74	0.003
Gender	0.52	0.462	0.37
Pupil Premium	0.045	0.06	0.69
SEND	0.12	0.0513	0.06
EAL	0.09	0.21	0.01
Midyis overall (mean, standard deviation)	115.26, 9.72	122.03, 11.24	<0.000001
Outcome Measure			Unpaired t-test result
Outcome Percentage: Y9 unit 1 (mean, standard deviation)	56.7%, 19.3	56.6%, 14.99	t(232)=0.007, p=0.99, 95% CI [-4.4%,4.4%]

Table 8.5: Main Study 1 Analysis C characteristics before matching

Without matching, the non-TfM group did 0.1% better on the assessment compared to the TfM group. As before, the differences in covariates between the two treatment groups indicate the need for matching. Unpaired t-tests and two-sample z-tests found small p-values (< 0.05) for KS2 maths, EAL and Midyis, indicating that the same difference would unlikely be seen if random allocation to the two treatment groups was the only adjustment made. Gender, pupil premium and SEND had p-values not at the conventional level of significance (0.37, 0.69, and 0.055 respectively).

Using an unpaired t-test, the difference in the mean outcome measure yielded a small p-value which suggested the difference is not compatible with a model in which random allocation to groups alone was a causal factor (t(232)=0.007,p=0.99, 95% CI [-4.4%,4.4%]).

8.10 Analysis C: Propensity Score & Matching

The estimate of the propensity score revealed the main drivers to be Midyis and EAL, suggesting that the imbalance of these covariates is unlikely to come from random allocation ($p < 0.05$). Six versions of matching were conducted until the best matches had been defined to estimate the treatment effects. The type of matching used for each version is in brackets in Table 8.6. In this

analysis, many different calipers were trialled in comparison to the previous analyses, providing evidence of more matches being found with a tighter caliper.

	No. of students		KS2 Maths			Gender			EAL			Pupil Premium			SEND			FSM		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	71	71	108.9	108.5	78.9	0.54	0.44	-68.6	0.13	0.11	88	0.06	0.06	100	0.04	0.07	59	118.6	118.9	95.6
Version 2 (nearest neighbour, caliper 0.05)	65	65	108.5	108.2	86.3	0.54	0.45	-57.9	0.12	0.12	100	0.06	0.06	100	0.05	0.08	55.2	118.2	118.1	98.9
Version 3 (nearest neighbour, caliper 0.03)	59	59	108.3	108.8	74.6	0.56	0.51	13	0.1	0.05	56.6	0.05	0.07	-43.4	0.05	0.08	50.7	118.9	118.9	89.2
Version 4 (nearest neighbour, caliper 0.035)	60	60	108.4	108.7	82.4	0.55	0.48	-14	0.1	0.05	57.3	0.05	0.05	100	0.05	0.08	51.5	118.4	119.1	89.4
Version 5 (genetic matching)	117	117	107.8	109.3	16.5	0.5	0.46	26.9	0.09	0.21	5.1	0.05	0.06	27.7	0.11	0.05	12.9	115.7	122	6.1
Version 6 (full matching)	125	117	110.1	109.3	54.8	0.399	0.46	-7.5	0.17	0.21	68.6	0.06	0.06	27.7	0.1	0.05	22.2	112.4	122	94.9

Table 8.6: Main Study 1 Analysis C propensity score variants and result balance

After consideration of the overall balance improvement for each version of matching, it was concluded that nearest neighbour matching with a caliper of 0.035 (version 4) yielded the most balance and the full matching results (version 6) were used within the sensitivity analysis.

Figures 8.18-8.19 show the propensity score distribution for the non-TfM group (curriculum type: other) and the TfM group (curriculum type: TfM) before and after matching:

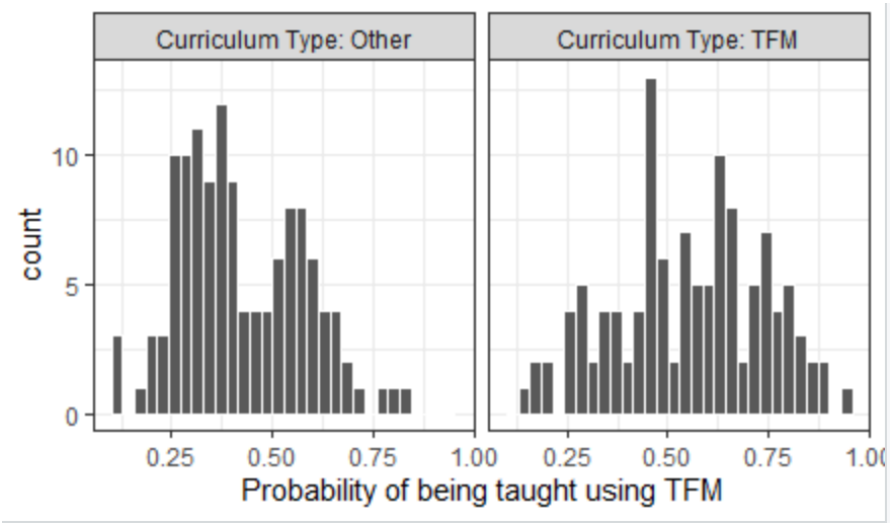

Figure 8.18: Main Study 1 Analysis C Propensity Scores

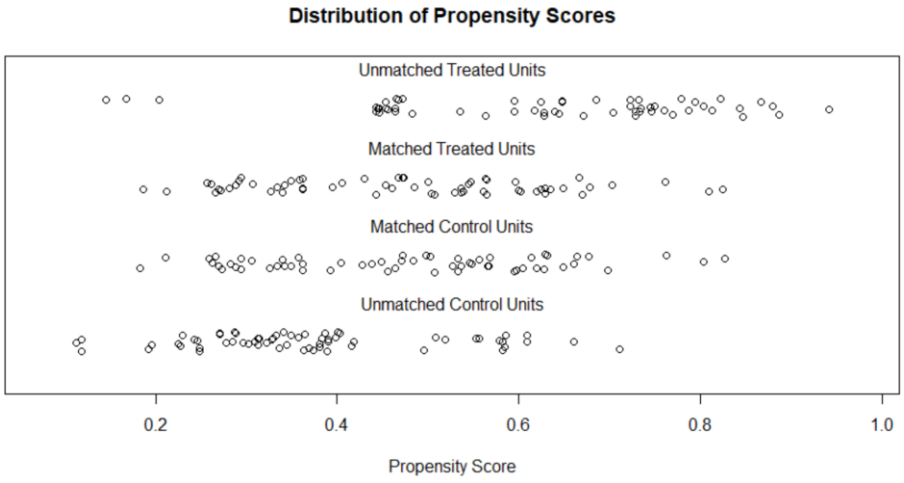

Figure 8.19: Main Study 1 Analysis C propensity score distribution after version 4 of matching

Figures 8.20-8.25 show the improved balance for each covariate after version 4 of matching, which was used to estimate treatment effects.

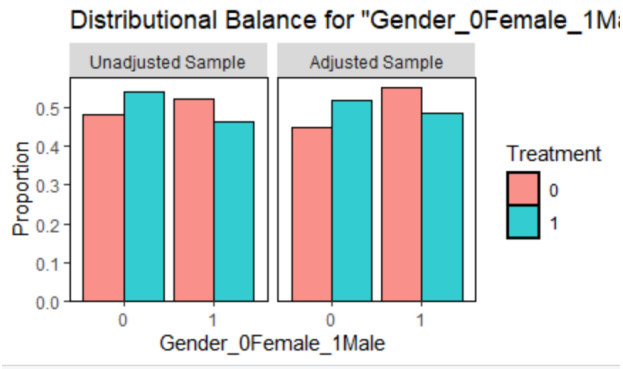

Figure 8.20: gender balance before and after matching

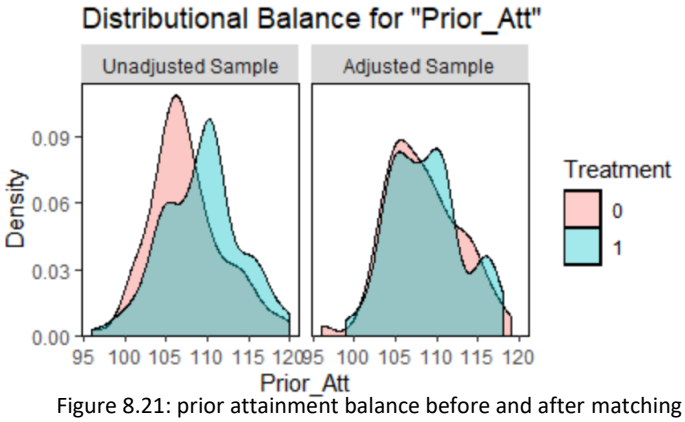

Figure 8.21: prior attainment balance before and after matching

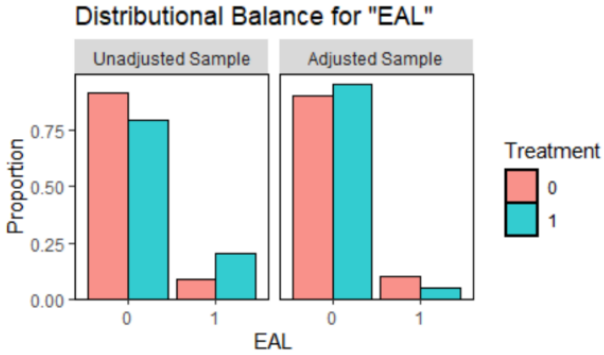

Figure 8.22: EAL balance before and after matching

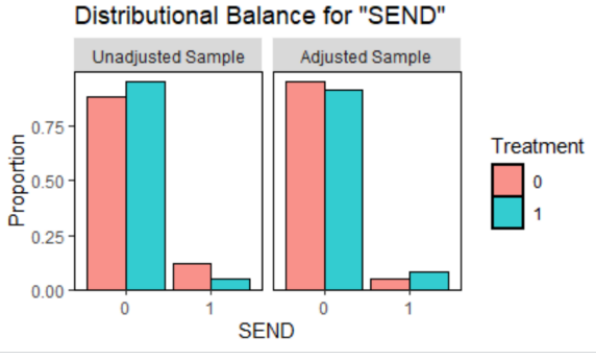

Figure 8.23: SEND balance before and after matching

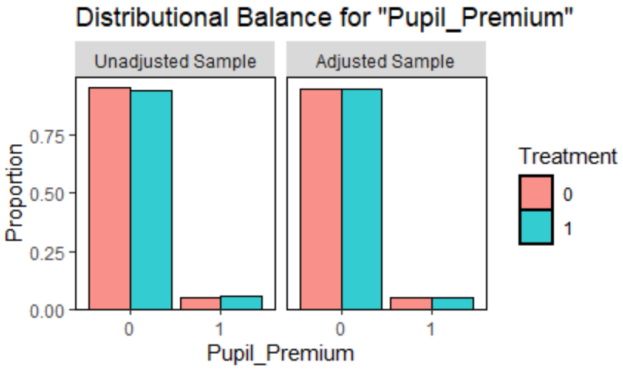

Figure 8.24: pupil premium balance before and after matching

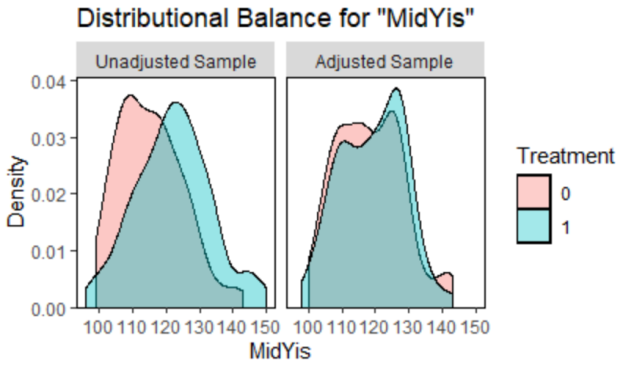

Figure 8.25: MidYis balance before and after matching

8.11 Analysis C: Treatment Effects and Sensitivity Analysis

Since version 4 created the most balanced groups for comparison, matched groups were used to estimate the treatment effect. Through the use of paired t-tests on the outcome measure (year 9 higher assessment percentage), it was found that the non-TfM group had a higher mean outcome by 3.65%. The 95% confidence interval was for the range [-1.06%, 8.36%] with a p-value of 0.13. This is not at the conventional level of significance (t(59)=1.55,p=0.13, 95% CI[-1.06,8.36]). A p-value which is not <0.05 suggests that it cannot be confidently argued that an effect of the difference in curricula on the outcome measure has been detective. It is possible that the difference found is reasonably compatible with a model in which there is no effect and where students were just randomly assigned to two groups.

A non-significant p-value is to be expected if a 95% confidence interval crosses zero, as it does in this result. When the interval crosses zero, the difference in outcomes observed is reasonably compatible with a model in which there is no effect for a difference in curricula, but when individuals are randomly allocated to groups. It is possible that there could be no difference at all on the outcome measure since zero lies in the interval. In this case, the interval of [-1.06%, 8.36%] suggests that it is unlikely that TfM causes more than an average of 1.06% improvement in the outcome or that non-TfM causes more than an average of 8.36% improvement in outcomes over the alternative. the change to a TfM curricula has a difference in either direction outside of the range.

The sensitivity analysis which used version 6 of matching was conducted using a linear regression model. A paired t-test could not be used following the method of full matching since the number of subjects in each of the two treatment groups was different. The linear regression model suggested that the students that were taught using TfM did worse than the non-TfM group, with an estimated effect -1.44% (p=0.995) for the TfM curricula, suggesting that the TfM curricula has a negative effect on student attainment for this specific assessment.

Overall, on the surface, it appears that, after reviewing propensity scores based on gender, pupil premium, EAL, SEND, MidYis and KS2 maths, the non-TfM group did better on the year 9 higher assessment as found by version 4 and version 6 of matching. However, it is worth noting that the pre-matching averages on the outcome measure for the two treatment groups were minimally different (56.6% for the non-TfM group and 56.7% for the TfM group) and therefore, before matching, the groups looked similar in terms of attainment levels.

However, beyond the surface, when interpreting a confidence interval which passes zero and a non-significant p-value, it cannot be confidently argued that an effect has been detected since it is possible that the difference in outcomes is reasonably compatible with a model in which there is no effect. On top of this, as always with PSM, any results must be treated with caution in case of unknown confounders that have not been considered that may have causal effect.

8.12 Conclusion

In conclusion, there are several key points to make. First, initial estimation of the treatment effect in analysis A showed that, following propensity score matching on the included covariates, the TfM group did have a higher mean outcome by 4% compared to the non-TfM group. However, a sensitivity analysis showed that the result was not robust to different PSM specifications, suggesting that firm conclusions on the impact of TfM cannot be made. On top of this, an awareness of the dates from which the data was provided reveals that the effects of the Covid-19 pandemic cannot be disaggregated from this analysis.

Analysis B serves as a warning to researchers using PSM. Whilst balance improvement statistics showed better balance in some of the versions of matching, a visual exploration of the region of common support showed that there were no alike subjects in the TfM and non-TfM group and therefore it was not possible to estimate any treatment effects. It is crucially important when conducting PSM to ensure there is a region of common support before estimating treatment effects.

Finally, with analysis C, the data provided by school Y preceded the Covid-19 pandemic. On the surface, it showed that the non-TfM group achieved more highly on the year 9 higher assessment. However, because the confidence interval crossed zero, and the p-value was non-significant, one cannot confidently distinguish the effects of the curricula.

Overall, the data from school Y has provided some interesting methodological considerations: the need to look at the region of common support, the need to conduct sensitivity analyses, and the need to consider the confidence interval and p-value when estimating treatment effects.

Chapter Nine: Implementation of Methods – Main Study (Part 2)

9.1 Introduction

Part two of the main study will also follow Leite's (2016) steps to propensity score matching, and the decision-making process that was set out in the pre-registration following the pilot study, helping to overcome the danger of 'the garden of forking paths' (Gelman, 2013).

Like part one of the main study, part two will focus on one participating school (school Z); seeking to evaluate the impact of TfM on assessment outcomes, compared to previous pedagogical approaches, using propensity score matching methods.

9.2 Interview and School Context

School Z[22] is a co-educational Catholic secondary school and sixth form, catering for approximately 1375 students from age 11 to 18. The most recent Ofsted inspection of 2014 graded the school as 'Outstanding' in all areas. The school converted to academy status in 2016 under the same headteacher. DfE performance tables show an overall progress 8 score for the school as being "well above average" and recent school league tables rank the school as the highest comprehensive in its county, based on the percentage of students achieving grade 5 or above in English and Maths GCSEs.

The interview for School Z was conducted with Teacher Z who has worked at the school for 12 years and is Assistant Headteacher, having held this position for 3 years. Teacher Z, as a trained Mastery specialist, helped to lead the change to TfM in the 2015-16 academic year after being involved in the England-Shanghai teacher exchange. The semi-structured interview used the same open questions as the pilot study, and the outcome of the interview fits two main themes: available data for analysis and the school's transition to TfM.

When asked about their understanding of TfM and the main changes to their curricula since the transition, teacher Z explained how the department transitioned to TfM, outlining that the change was implemented in year 7 first, and then rolled out to an additional year group in each

[22] For the purpose of this study, the school used in part 2 of the main study is referred to as 'School Z'

subsequent academic year. Teacher Z stated that the main change to the curriculum was that a significant amount of time was now spent on fewer topics in the curriculum, particularly in years 7-9. Previously, topics were taught in two or three-week cycles and each year group covered approximately twelve topics each academic year. With the mastery curriculum, only six topics are taught in each academic year at Key Stage 3, but to a much greater depth than before. Due to teacher Z's involvement in the Maths Hub, changes to pedagogy following the shift to TfM largely incorporated the NCETM's 'Five Big Ideas'. Teacher Z commented that the main changes in each classroom was the amount of time dedicated to allowing students to 'think deeply' (mathematical thinking) and carefully thought-out examples and exercises as part of variation (fluency). It was also noted that every lesson at Key Stage 3 is planned to include fluency, reasoning and problem-solving as well as diagnostic AfL to ensure that intervention can take place in the classroom, thus helping to ensure that the class can stay and progress together.

Following the interview, it was considered that school Z's perception of TfM largely aligned with the criteria set out in the literature review of this thesis. A curriculum where fewer topics are taught for longer periods of time to ensure depth of understanding through small steps fits with the following two defining characteristics:

- *Where topics are taught for a longer period to ensure depth of understanding*
- *Where there is an emphasis on teaching for conceptual understanding and forming mathematical connections.*

By seeking to ensure that each lesson builds in fluency, reasoning and problem-solving, school Z's perception of TfM aligns with the following three defining characteristics:

- *Where variation is embedded in chosen examples and exercises*
- *Where there is insistence on key mathematical language*
- *Where problem-solving is considered a key aim of the curriculum.*

The final characteristic of TfM that was set out in the literature review was *'Maths for All' – where all students are exposed to the same mathematical content and some content is not preserved for higher attaining students*. School Z moved to 'banded' groups, where students are more mixed in their ability and not set. There are some 'top band', some 'middle band' and some 'support band' classes in each year group where all classes are taught the same topics, but the level of challenge may differ between a 'top band' and 'support band' group. In Key Stage 4, school Z still sets the students despite having transition to a mastery curriculum since the GCSE examinations remain 'tiered'.

School Z first introduced TfM in the 2015-16 academic year. During the subsequent years, assessments were re-written and refined year-on-year to ensure a balance of fluency, reasoning and problem-solving. Therefore, there are only two possible assessments that can be used for comparison. Details of these assessments are given in the next section of this chapter. Teacher Z confirmed that there were no other contemporaneous events that occurred alongside the transition to TfM that could have caused a change in assessment outcomes.

The sample size provided by School Z was 1144 students across different year groups. Data was provided on the following cohorts:

- 187 students that were in year 7 during the 2013-14 academic year (pre-TfM)
- 192 students that were in year 7 during the 2014-15 academic year (pre-TfM)
- 189 students that were in year 7 during the 2015-16 academic year (TfM)
- 187 students that were in year 7 during the 2016-17 academic year (TfM)
- 192 students that were in year 10 during the 2018-19 academic year (pre-TfM)
- 197 students that were in year 10 during the 2019-20 academic year (TfM).

It is important to note that the final sample of 197 students that were in year 10 during the 2019-20 academic year are particularly interesting in this research since they substantially overlap with the 2016-17 year 7 cohort and therefore had experienced a 4-year period of TfM by the time they finished year 10. The sample sizes, though, are not identical (197 in year 10 and 187 in year 7), since it is common for students to join a school at any point during a cohort's five-year journey through years 7 to 11.

It was outlined earlier in this thesis that this research seeks to explore the impact of longer-term embedded mastery approaches, and therefore part of this chapter's analysis will consider the year 7 and 10 cohort, of whom the students are the same, as one. It may be noted that in the same way, the 2015-16 year 7 cohort will also be made up of the same students as the 2018-19 year 10 cohort. This cohort of students were taught using TfM in year 7 but were taught using previous pedagogical approaches in years 8-10 as the school steadily transitioned, refining curricula and pedagogy for two academic years before rolling it out into subsequent years. As such, the 2015-16 year 7 cohort experienced a one year dose of TfM, whilst the 2016-17 year 8 cohort experienced a four year dose of treatment.

9.3 Sample and Data Preparation

9.3.1 Year 7 Data

For the four cohorts of students on which year 7 data was provided, the comparable outcome measure was a test sat at the end of the Autumn term on number skills and some algebra. This included: factors, multiples, types of number, BIDMAS, fractions and simplifying expressions. It is important to highlight that this outcome measure was an assessment sat after just 13 weeks of different curricula. Therefore, it is possible that a difference in outcome may not be identified due to such a short period of time elapsing. Since it is one of only two consistent outcome measures that can be provided by school Z, the data will still be analysed as planned. As well as the assessment percentage for each student, the following covariate data was provided:

- Gender
- SEND status
- FSM status
- EAL status
- Pupil Premium status
- Y7 Entry assessment (this was done in-house in September upon entry and was focused on student's written mathematical ability of calculations).

9.3.2 Year 10 Data

For the two cohorts of students on which year 10 data was given, the comparable outcome measure was a half-term assessment that was sat at the end of the first half-term in October. It is important to note, therefore, that for the 2019-20 year 10 cohort, the data was not impacted by Covid since schools were closed in March 2020, but the assessment was sat in October 2019. For both cohorts, the assessment included questions on the following topics: multiplying and dividing with decimals, simplifying algebra, standard form, angles, surds, laws of indices, and circle theorems.

For this outcome measure, the focus of the analysis is the impact that the difference in curricula has had over almost four years, and therefore provides an opportunity to look at the long-term impact of TfM. As well as the assessment percentage for each student, the following covariate data was provided:

- Gender

- SEND status
- FSM status
- EAL status
- Pupil Premium status
- Y9 assessment percentages (this was done in-house at the end of year 9 to help set the students and was focused on a mixture of number, algebra, shape and geometry)
- Y7 Entry assessment (this was done in-house in September upon entry and was focused on student's written mathematical ability of calculations).

Although the given Y9 assessment percentages were a continuous covariate, which are needed for PSM to work well, it was decided to omit the variable from the analysis since it was a result gathered after treatment allocation had already taken place. Treatment allocation took place in year 7 when the group either began their secondary education with TfM or a previous pedagogical approach. If the year 9 assessment percentages were included, the results would be skewed since the data was gathered after 'treatment' had begun. All covariates should be measures of group differences pre-allocation.

9.3.3 Analyses

Based on the outcome measures available, school Z's data is split into two mini-studies for analysis, one of which will consider the aforementioned cohort of students who were taught using the mastery approach in years 7 through to year 10. Each analysis is titled by the assessment outcome of interest below:

- Analysis A: Year 7 December – this analysis will use the four cohorts of students for which year 7 data has been provided.
- Analysis B: Year 10 Half Term 1 – this analysis will use the 2015-17-year 7 cohort which became the 2019-20-year 10 cohort who had TfM for almost four academic years and will compare it to the 2015-16-year 7 cohort which became the 2018-19-year-10 cohort who only had a one year does of TfM in year 7 before continuing their secondary journey with previous pedagogical approaches.

Analysis B will include the year 7 December percentage as a continuous covariate and the year 10 Half Term 1 percentage as the outcome measure. By looking at the data this way, this analysis is investigating the impact of around 3 years of difference in curricula.

The aim was to conduct two separate propensity score matching analyses of the (ATT) impact of changing from non-TfM (coded 0) to TfM (coded 1) on the outcome assessment percentage, matching on propensity based on the following: gender (coded 0 for female, 1 for male), SEND (coded 0 for no additional need, 1 for any additional need), FSM (coded 0 for non-FSM, 1 for FSM), EAL (coded 0 for non-EAL, 1 for EAL), Pupil Premium (coded 0 for non-PP, 1 for PP), Y7 entry test (taken as percentage given).

When preparing the data for analysis, 32 students were removed from the sample for analysis A (4% of the sample for analysis A) due to missing data and 40 students were removed from analysis B (5% of the sample for analysis B). These numbers are larger than one might hope; the missing data in each instant was either the Y7 entry test, the Y7 outcome percentage, or the Y10 assessment percentage. However, as was outlined in the methodology, the proportion of 'missingness' is deemed small and therefore complete case analysis is the most effective method for dealing with the missing covariate data. Moreover, one of the methodological lessons following the pilot study was the need for continuous covariates, and therefore, for the main studies, it was deemed important to ensure that each included student in the sample had these pieces of data attributed. Once the PSM process was complete, the outcome measures (taken as percentage scores) were considered for the two matched groups to look at the average treatment effect on the treated (ATT).

9.4 Analysis A: Characteristics before Matching

The first analysis of the data provided by school Z based on year 7 assessment percentages. As with the pilot study and part one of the main study, the distribution of baseline characteristics for the non-TfM and TfM groups were analysed to determine the differences between the two groups on each given covariate and the outcome measure. Using unpaired t-tests, z-tests, and difference-in-means analysis, the following results were obtained:

	Non-TfM group	TfM group	
Number of students	333	367	
Covariates	Mean	Mean	p-value
Y7 Entry % (mean, standard deviation)	32.9, 7.1	31.1, 7.66	0.001
Gender	0.55	0.49	0.14
SEND	0.09	0.08	0.6
FSM	0.05	0.02	0.06
EAL	0.003	0.11	< 0.0001
Pupil Premium	0.09	0.11	0.29
Outcome Measure			Unpaired t-test result
Y7 Dec Test % (mean, standard deviation)	65.8%, 16.12	71.2%, 18.64	t(696)=-4.12, p = <0.0001, 95% CI [2.82%, 7.97%]

Table 9.1: Main Study 2 Analysis A characteristics before matching

Table 9.1 shows that without any matching, the TfM group did better (by 5.4% with a 95% confidence interval for the range [2.82%, 7.97%] on average on the outcome measure. An unpaired t-test revealed a small p-value for this, suggesting the difference is not compatible with a model in which random allocation to groups alone was a causal factor (t(696)=-4.12,p<0.00001, 95% CI [2.82%, 7.97%]). That said, the two groups were imbalanced on the given covariates and therefore PSM was needed to create matched groups before making any conclusions on the outcome measure. For the given covariates listed above, unpaired t-tests and two-sample z-tests revealed a p-value at the conventional level of significance (< 0.05) for Y7 entry, FSM, EAL and pupil premium, indicating that the same difference would be unlikely seen if random allocation to the two treatment groups was the only adjustment made.

9.5 Analysis A: Propensity Score & Matching

Seven versions of matching were conducted to ensure that optimal balance was found before treatment effects were estimated. The type of matching used for each version is in brackets by each version number in Table 9.2, along with the mean value for each covariate after matching and the balance improvement. The pre-matching analysis of the propensity score revealed that Y7 entry, FSM, EAL and PP were the main drivers of the propensity score based on the significance codes and so careful consideration was given to these covariates to try to reach the best balance. The propensity score calculation remained in line with that from School X and School Y. The small p-values that these covariates generated indicated that the imbalance between the two treatment groups is most unlikely to just come from random allocation to the two groups and therefore may be perceived as the 'real differences' between the two groups.

The PSM started with nearest neighbour matching, using a caliper of 0.1. In the first version of matching, better balance across all covariates was achieved, and therefore different calipers were trialled to achieve an optimal balance. Two subsequent versions of matching trialled tighter calipers, followed by matching without a caliper as well as three different versions.

	No. of students		Y7 Entry test (baseline)			Gender			SEND			FSM			EAL			PP		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	286	286	31.97	32.27	84.5	0.5	0.49	75.1	0.08	0.08	38.4	0.02	0.02	73.4	0.004	0.004	100	0.07	0.08	28.2
Version 2 (nearest neighbour, caliper 0.05)	271	271	32.5	32.2	84.9	0.51	0.52	86.8	0.08	0.07	67.5	0.01	0.02	71.9	0.004	0.004	100	0.06	0.07	54.5
Version 3 (nearest neighbour, caliper 0.03)	255	255	32.5	32.6	94.6	0.52	0.48	23.1	0.07	0.08	-3.7	0.02	0.02	70.1	0.004	0.004	100	0.06	0.08	19.5
Version 4 (nearest neighbour, no caliper)	333	333	32.93	30.45	-37.1	0.55	0.45	-66	0.09	0.07	-111.7	0.05	0.01	-60.2	0.003	0.12	-10.5	0.09	0.11	26
Version 5 (genetic matching)	333	333	32.93	30.45	-37.1	0.55	0.45	-66	0.09	0.07	-111.7	0.05	0.01	-60.2	0.003	0.12	-10.5	0.09	0.11	26
Version 6 (full matching)	333	367	31.96	31.12	53.6	0.51	0.49	59.4	0.07	0.08	3.4	0.02	0.02	98.3	0.11	0.11	100	0.1	0.11	37
Version 7 (optimal matching)	333	333	32.93	31.16	2	0.55	0.49	-1.7	0.09	0.08	20.6	0.05	0.02	8.5	0.003	0.02	85.8	0.09	0.12	-23.3

Table 9.2: Main Study 2 Analysis A propensity score variants and result balance

When looking at the balance improvement (which was assessed in the same way as school X and Y) for each version of matching on each covariate, it was concluded that nearest neighbour matching with a caliper of 0.05 (version 2) was to be used to estimate the treatment effect since it yielded the best overall set of matches and balance improvement at its greatest. Version 1 yielded the second-best set of matches and therefore was subsequently used as a sensitivity analysis on the treatment estimation.

Figure 9.1 shows the propensity score distribution for the non-TfM group (curriculum type: other) and the TfM group (curriculum type: TfM) before matching, showing there is an area of common support.

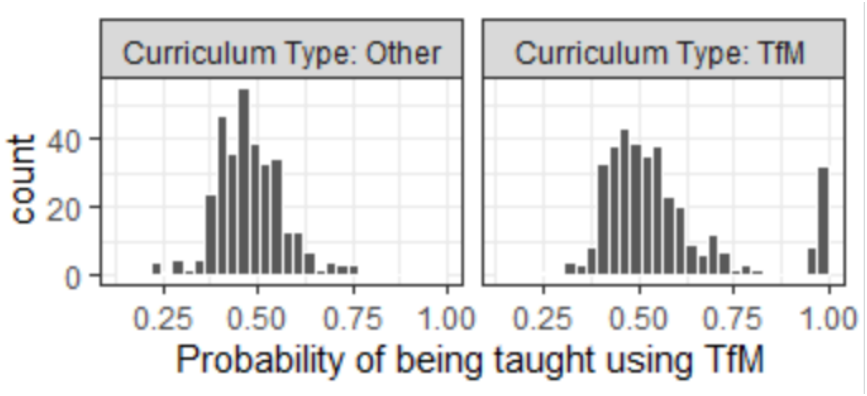

Figure 9.1: Main Study 2 Analysis A Propensity Scores

Figures 9.2-9.8 show the improved balance for each covariate for version 2 of matching, which was used to estimate treatment effects.

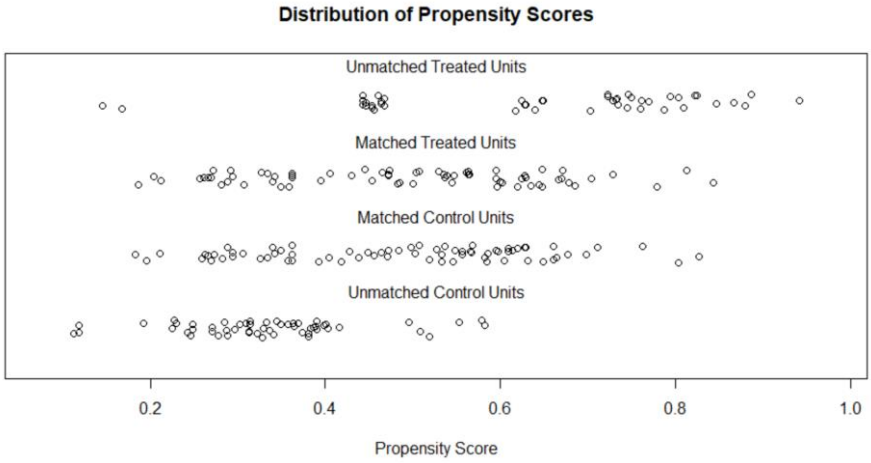

Figure 9.2: Main Study 2 Analysis A propensity score distribution after version 2 of matching

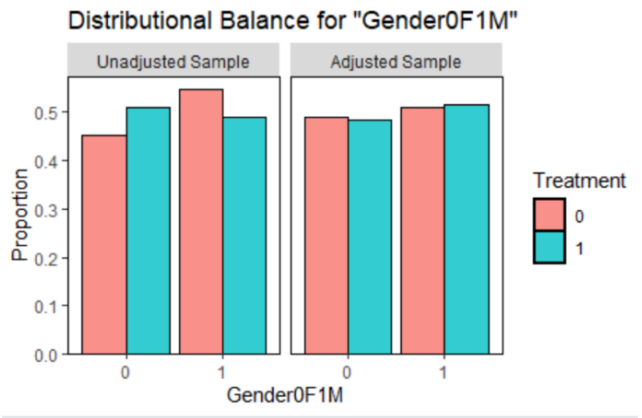

Figure 9.3: gender balance before and after matching

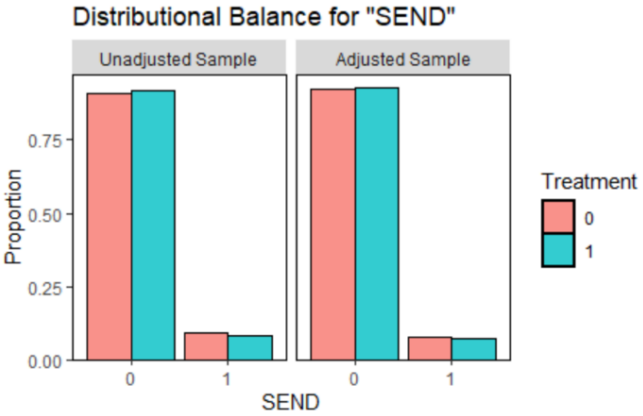

Figure 9.4: SEND balance before and after matching

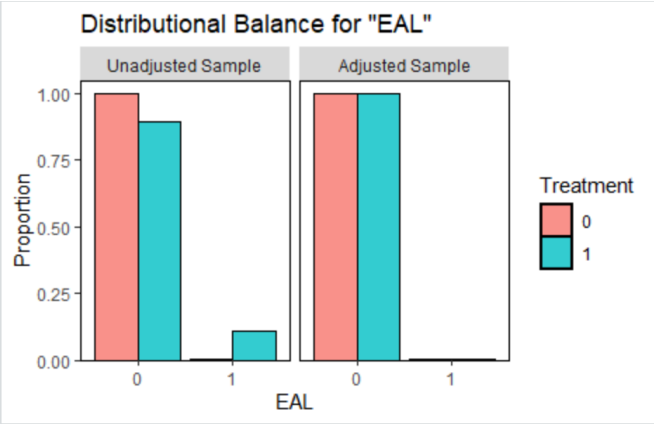

Figure 9.5: EAL balance before and after matching

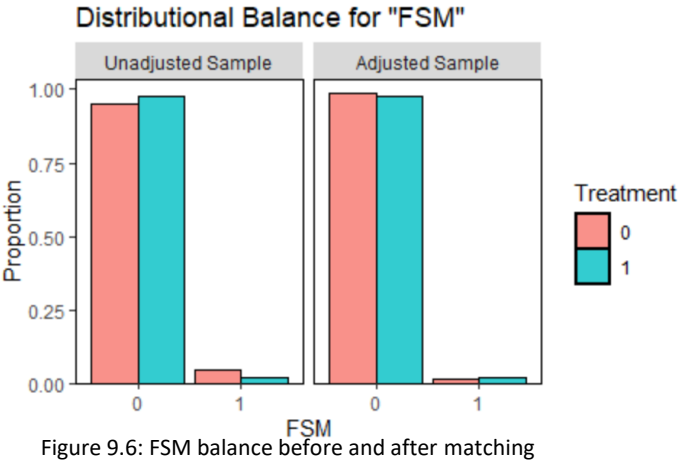

Figure 9.6: FSM balance before and after matching

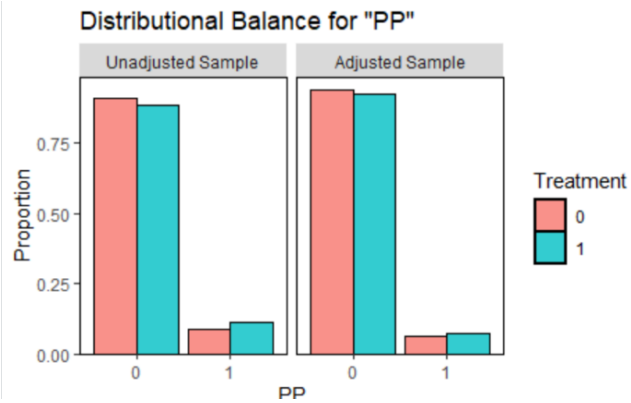

Figure 9.7: pupil premium balance before and after matching

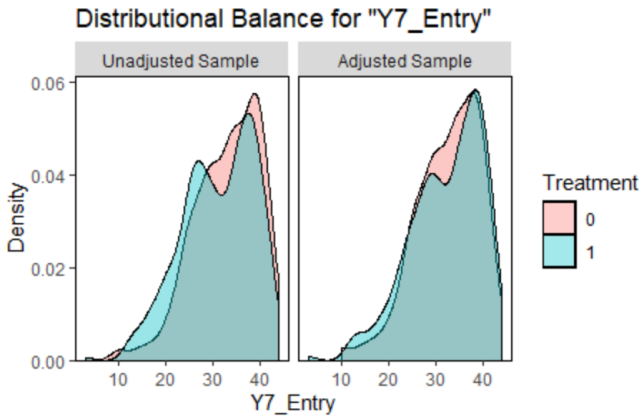

Figure 9.8: Y7 Entry balance before and after matching

Similarly, figures 9.9-9.15 show the balance improvement for each covariate for version 1 of matching, showing that it was a sensible choice to also use this version for sensitivity analysis.

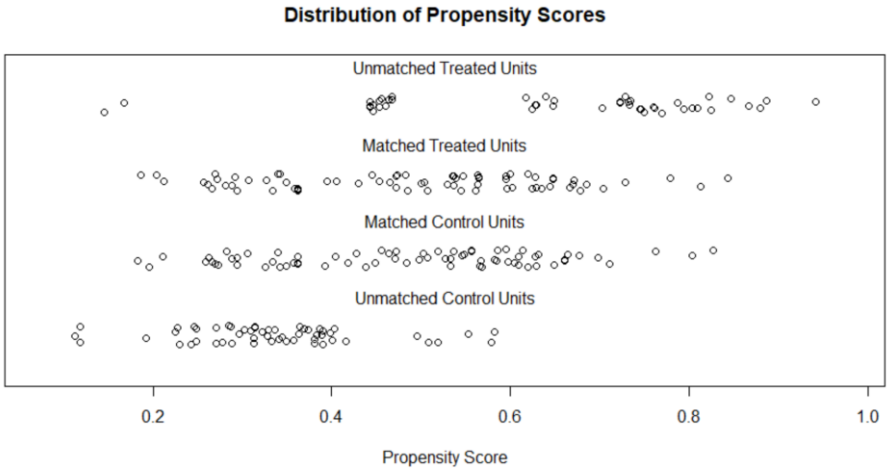

Figure 9.9: Main Study 2 Analysis A propensity score distribution after version 1 of matching

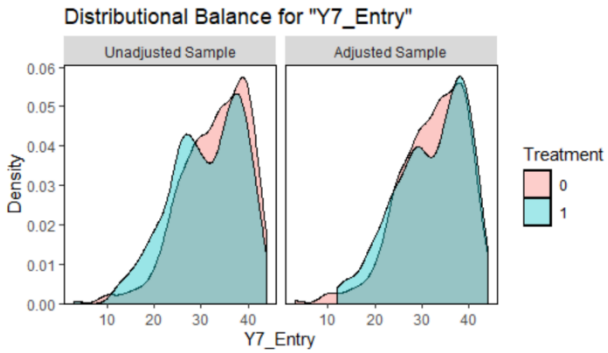

Figure 9.10: Y7 Entry balance before and after matching

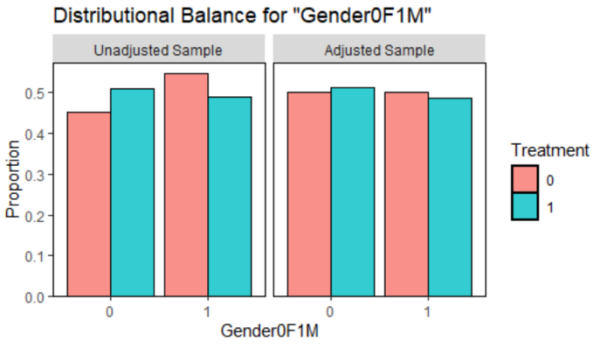

Figure 9.11: Gender balance before and after matching

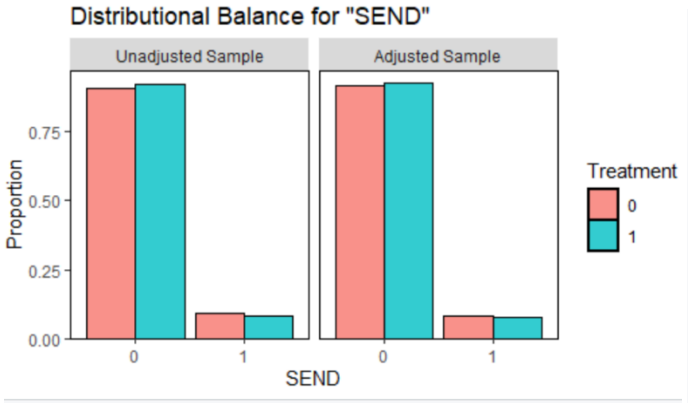

Figure 9.12: SEND balance before and after matching

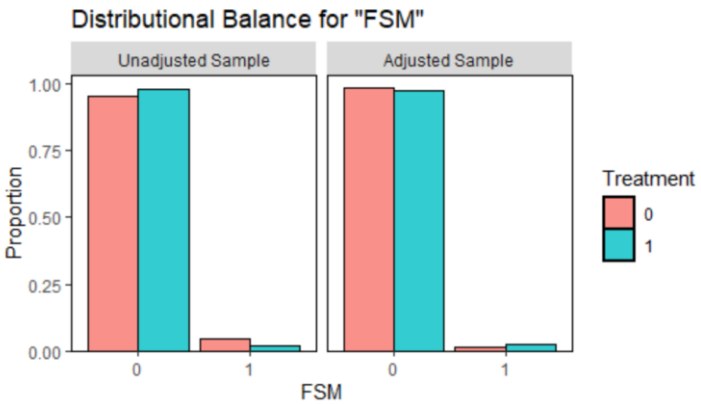

Figure 9.13: FSM balance before and after matching

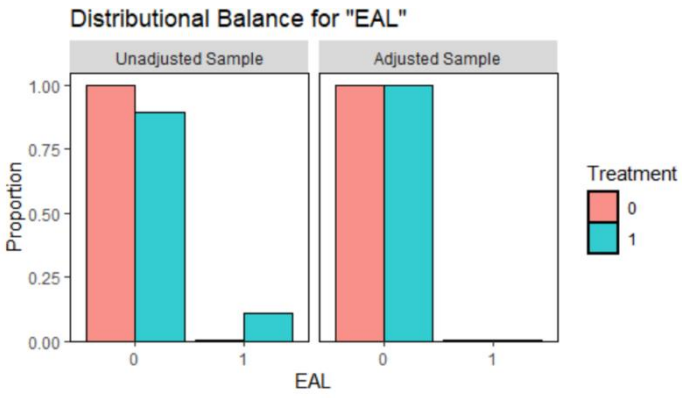

Figure 9.14: EAL balance before and after matching

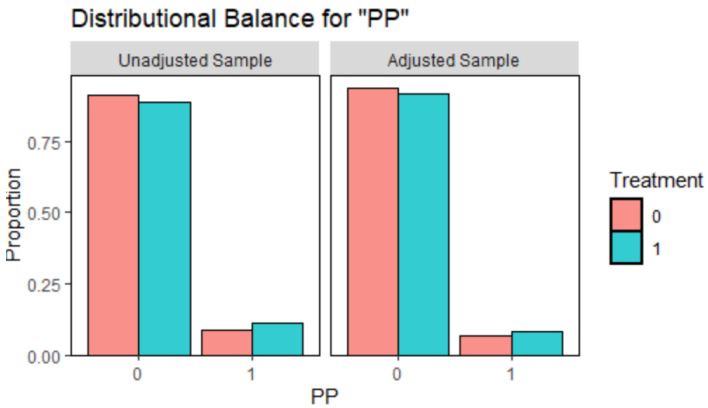

Figure 9.15: Pupil Premium balance before and after matching

9.6 Analysis A: Treatment Effects and Sensitivity Analysis

Since version 2 created the most balanced groups for comparison, the matched groups were used to estimate the treatment effect. As with the pilot study, treatment effects were estimated through the use of paired t-tests on the outcome measure (year 7 December assessment percentage). A paired t-test found that the TfM group had a higher mean outcome by 6.5% on the assessment percentage, with a small p-value (<0.00001) which indicated statistical significance with a 95% confidence interval for the range [3.7%, 9.4%] (t(270)=4.5,p=<0.00001, 95% CI [3.7,9.4]). The confidence interval suggests that one can be 95% confident that the effect of the difference in curricula lies between 3.7% and 9.4%.

The sensitivity analysis which used version 1 of matching estimated that the TfM group had a higher mean outcome also by 6.5%, with a small p-value (<0.0001) for the range [3.7%, 9.3%] (t(285)=4.6,p=<0.00001, 95% CI[3.7,9.3]). Since the sensitivity analysis concurred with the results of the main treatment effect estimation, one can be more confident that the result is robust to alternative specifications of the propensity score.

Following the estimate of the treatment effect and the subsequent sensitivity analysis, one can conclude that, after propensity scores based on gender, SEND, FSM, EAL, pupil premium and Y7 entry test, the matched analysis suggests positive effect of the TfM curriculum compared to the previous curriculum when using the year 7 December assessment percentage as the outcome measure. As ever with propensity score matching, this result must be treated with caution due to the difficulty of not knowing that all confounders are controlled for.

9.7 Analysis B: Characteristics before Matching

The second analysis of the data provided by school Z addresses year 10 assessment percentages. Like analysis A and the previous analyses in this thesis, the distribution of baseline characteristics for the non-TfM and TfM groups were analysed to determine the differences between the two groups on each given covariate and the outcome measure. Using unpaired t-tests, z-tests, and difference-in-means analysis, the following results were obtained:

	Non-TfM group	TfM group	
Number of students	135	140	
Covariates	Mean	Mean	p-value
Y7 Entry Raw Score (mean, standard deviation)	35.3, 5.3	32.5, 6.1	<0.001
Y7 Dec Assessment % (mean, standard deviation)	75.7, 13	76.2, 17	0.78
Gender	0.46	0.47	0.84
SEND	0.04	0.03	0.7
FSM	0.01	0.01	0.95
EAL	0.06	0.17	0.003
Pupil Premium	0.07	0.06	0.94
Outcome Measure			Unpaired t-test result
Half Term 1 Assessment % (mean, standard deviation)	67.0%, 19.2	67.2%, 16.2	t(262)=-0.069, p = 0.94, 95% CI [-4.37%, 4.08%]

Table 9.3: Main Study 2 Analysis B characteristics before matching

Table 9.3 shows that without any matching, the TfM group did marginally better (by 0.2%) on average on the year 10 outcome measure. Because the difference is only marginal, the unpaired t-test revealed a large p-value, suggesting that the difference is not statistically significant (t(262)=-0.069,p=0.94, 95% CI [-4.37%,4.08%]). It is clear from table 9.3 that the two groups were imbalanced on some of the given covariates (most notably Y7 Entry score and EAL) and therefore PSM was needed to create matched groups before making any conclusions on the outcome measure. Unpaired t-tests and two-sample z-tests revealed small p-values for the Y7 Entry score and EAL covariates, indicating that the same difference would be unlikely seen if random allocation to the two treatment groups was the only adjustment made.

9.8 Analysis B: Propensity Score & Matching

Table 9.3 highlights that the FSM covariate was already very well balanced before matching and therefore any small changes in the balance following propensity score matching would flag up as causing large imbalances between the two treatment groups. After calculating the propensity score for each student in the sample based on the given covariates, it was clear that FSM was not a key driver of the propensity score. Therefore, the decision was made to remove FSM as a covariate. The propensity score estimation revealed that they key-drivers of the propensity score were the Y7 Entry score, Y7 Dec percentage and EAL. Therefore, in the versions of matching, particular attention was paid to the balance of these covariates.

For the analysis, eight versions of matching were conducted to ensure that optimal balance across the two treatment groups was achieved before treatment effects were estimated. The type of matching used for each version is in brackets by each version number in Table 9.4, along with the mean for each covariate after matching and the balance improvement.

The PSM started with nearest neighbour matching, using a caliper of 0.1. In the first version of matching, greater balance was achieved on the key drivers of the propensity score, but gender, SEND and pupil premium become more imbalanced. Therefore, different versions of matching were tried with tighter and wider calipers, and these can be seen below.

	No. of students		Y9 Exam Percentage (Baseline)			Gender			SEND			FSM			EAL			PP		
	Non-TfM	TfM	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv	Non-TfM	TfM	Balance Imprv
Version 1 (nearest neighbour, caliper 0.1)	130	130	65.0	63.9	64.9	0.49	0.5	63.7	0.04	0.04	100	0.015	0.008	-741.5	0.08	0.08	100	0.07	0.05	-274
Version 2 (nearest neighbour, caliper 0.3)	138	138	64.8	61.6	0.2	0.49	0.51	31.7	0.04	0.04	100	0.014	0.007	-692.7	0.08	0.08	100	0.07	0.06	-76.2
Version 3 (nearest neighbour, caliper 0.5)	138	138	64.8	61.1	-16.3	0.49	0.5	65.8	0.04	0.04	100	0.01	0.01	100	0.08	0.08	100	0.07	0.07	100
Version 4 (nearest neighbour, no caliper)	143	143	65.9	60.8	-56	0.48	0.52	-97.8	0.04	0.02	-126.1	0.01	0.007	-665	0.08	0.18	-12.8	0.06	0.06	-70
Version 5 (nearest neighbour, caliper 0.4)	138	138	64.8	61.2	-12.8	0.49	0.5	65.8	0.04	0.04	100	0.01	0.01	100	0.08	0.08	100	0.07	0.07	100
Version 6 (genetic matching)	143	143	65.9	60.8	-56	0.48	0.52	-97.8	0.04	0.02	-126.1	0.01	0.007	-665	0.08	0.18	-12.8	0.063	0.056	-70
Version 7 (full matching)	143	153	62.6	62.6	98.7	0.54	0.49	-125	0.05	0.03	-40.9	0.02	0.01	-972.5	0.17	0.17	100	0.07	0.06	-249.6
Version 8 (optimal matching)	143	143	65.9	63.1	15.7	0.47	0.5	-31.9	0.04	0.04	24.6	0.01	0.007	-665	0.07	0.11	62.4	0.06	0.056	-70

Table 9.4: Main Study 2 Analysis B propensity score variants and result balance

Table 9.4 shows that none of the eight versions of matching have balanced the gender distribution any more than they were before matching. In fact, the negative balance improvement scores show that the two treatment groups have become more imbalanced on gender. It is because the two treatment groups were relatively well balanced before matching that the balance improvement scores have some large negative magnitudes. Thus, it was important to choose the version of matching regardless of the gender distribution. Instead, close attention was paid to Y7 Entry and EAL as these were found to be the key drivers of the propensity score in Table 9.3. Based on this, it was concluded that nearest neighbour matching with a caliper of 0.1 (version 1) be used to estimate the treatment effect. Version 2 (nearest neighbour with a caliper of 0.2) yielded the second-best set of matches where the key propensity score drivers were more balanced than before matching whilst maintaining some balance across the other covariates.

Figure 9.16 below shows the propensity score distribution for the non-TfM group (curriculum type: other) and the TfM group (curriculum type: TfM) before matching, showing there is an area of common support.

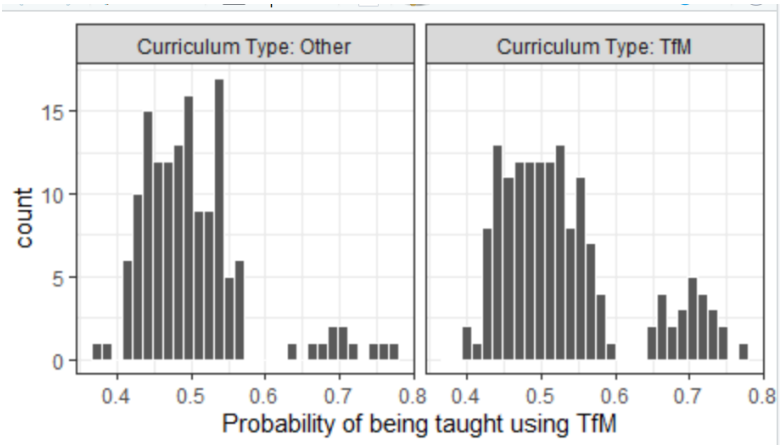

Figure 9.16: Main Study 2 Analysis B Propensity Scores

Figure 9.17 shows the propensity score distribution for version 1 of matching, which was the version used to estimate treatment effects.

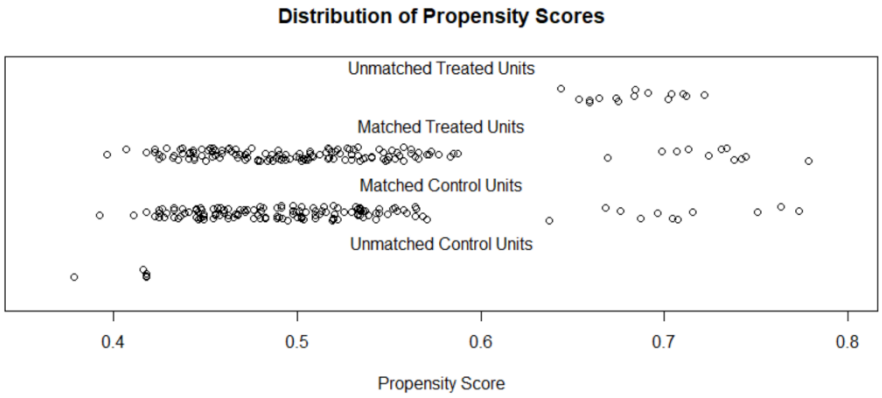

Figure 9.17: Main Study 2 Analysis B Propensity Score Distribution after version 5 of matching

Figures 9.18-9.21 show the balance change for each covariate in version 1 of matching. These figures show an improved balance after matching for the key drivers of the propensity score (Y7 entry, Y7 Dec percentage, and EAL).

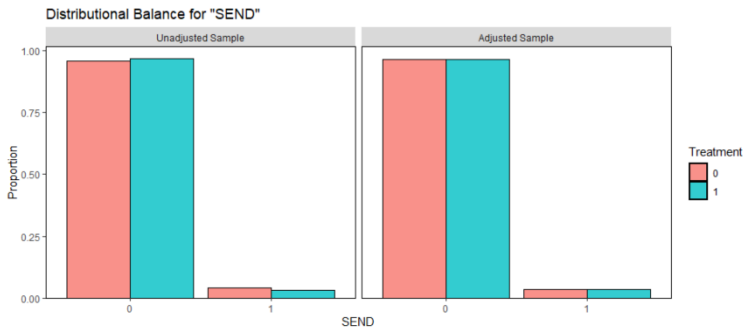

Figure 9.18: SEND balance before and after matching

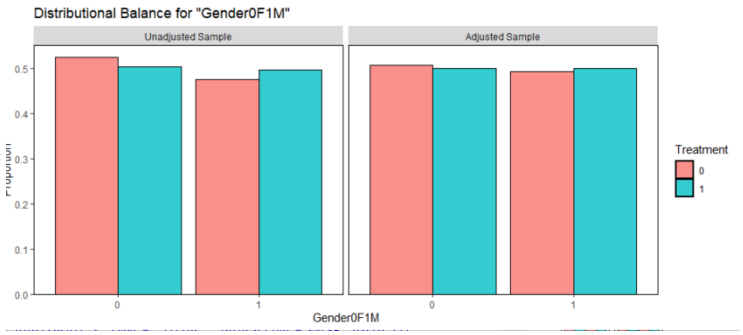

Figure 9.19: Gender balance before and after matching

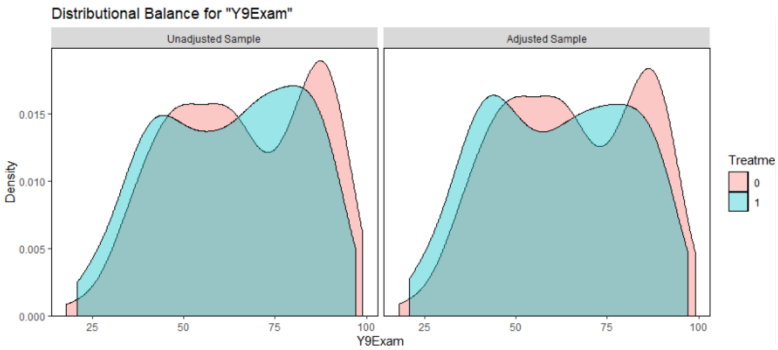

Figure 9.20: Y9 Exam balance before and after matching

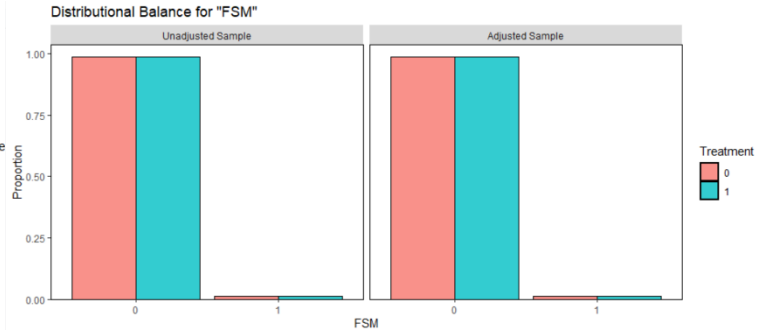

Figure 9.21: FSM balance before and after matching

Similarly, Figures 9.22- 9.28 below show the balance improvement for each covariate as well as the propensity score distribution for version 2 of matching, showing that it was a sensible choice to also use this version for sensitivity analysis.

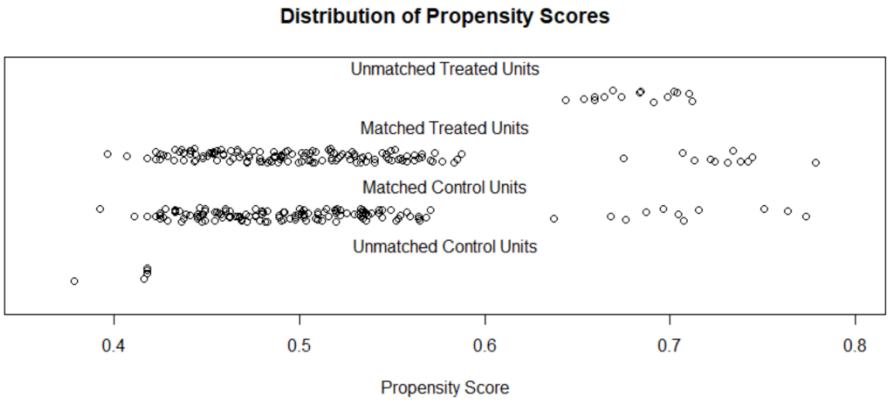

Figure 9.22: Main Study 2 Analysis B Propensity Score Distribution after version 3 of matching

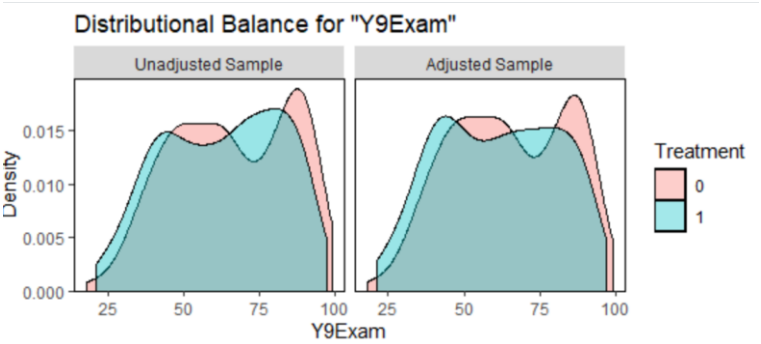

Figure 9.23: Y9 Exam balance before and after matching

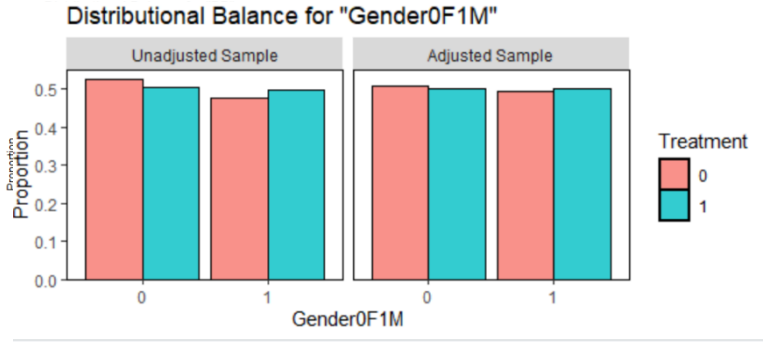

Figure 9.24: Gender balance before and after matching

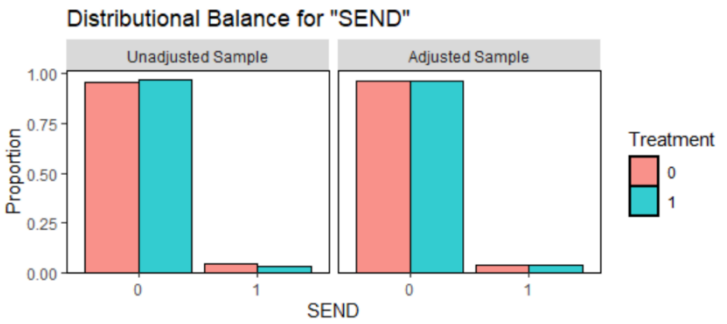

Figure 9.25: SEND balance before and after matching

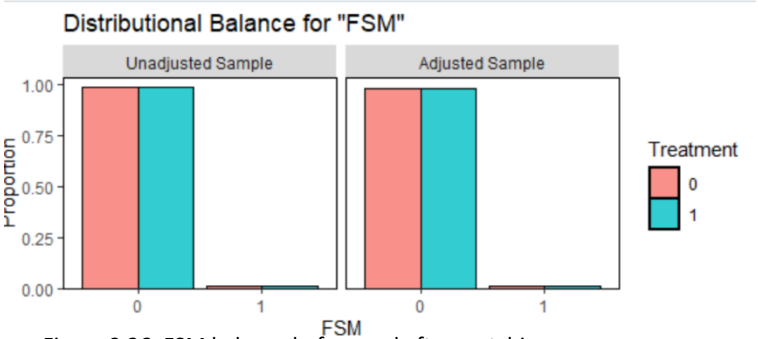

Figure 9.26: FSM balance before and after matching

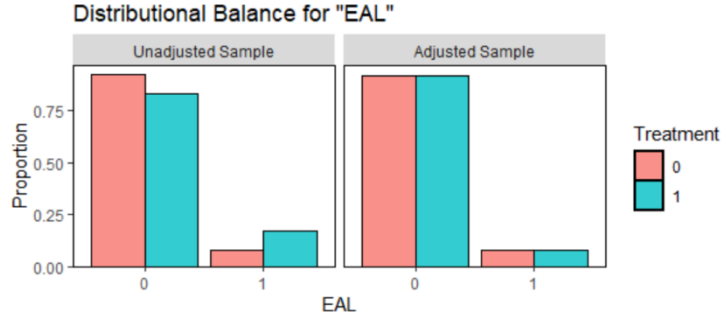

Figure 9.27: EAL balance before and after matching

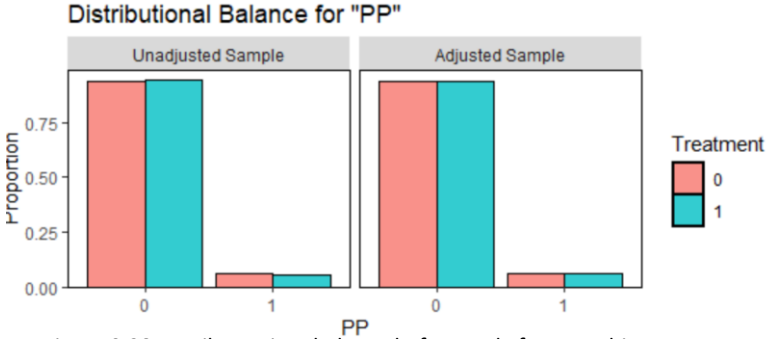

Figure 9.28: Pupil Premium balance before and after matching

9.9 Analysis B: Treatment Effects and Sensitivity Analysis

Since version 1 created the most balanced groups for comparison, the matched groups were used to estimate the treatment effect. As with the pilot study, treatment effects were estimated through the use of paired t-tests on the outcome measure (year 10 half-term 1 assessment percentage). A paired t-test found that the TfM group had a higher mean outcome by 4% on the assessment percentage. The 95% confidence interval was for the range [-0.54%, 8.92%] with a p-value of 0.08 which is not at the conventional level of significance (t(103)=1.7,p=0.08, 95% CI[-0.54,8.92]). A non-significant p-value is to be expected if a 95% confidence interval crosses zero. If a confidence interval of a difference crosses zero, it means that the effect of difference in treatments could be zero.

The sensitivity analysis which used version 2 of matching estimated that the TfM group had a higher mean outcome also by 3.97% but again, with a non-small p-value (0.08) for the range [-0.6%, 8.54%] (t(108)=1.7,p=0.08, 95% CI[-0.6,8.54]). Again, the confidence interval crossed zero, meaning that there could be no difference at all between the TfM and the non-TfM group in this case.

Overall, the treatment effect estimation and sensitivity analysis concur in their suggestion that the TfM group achieved around 4% higher than the non-TfM group on the year 10 assessment measure following propensity scores based on Y7 entry, Y7 Dec test, gender, SEND, EAL and pupil premium. However, since the 95% confidence interval for this result passes zero, there is some uncertainty as to whether the TfM curriculum has indeed added a positive effect, particularly with a non-significant p-value. Therefore, the results from this analysis must be interpreted cautiously.

That said, before any matching took place, the mean outcome percentage for both treatment groups were almost identical (67% for non-TfM and 67.2% for TfM). Propensity score matching has allowed a greater difference to be detected (of around 4%), despite the need to interpret these results cautiously due to the confidence intervals ([-0.54%, 8.92%], [-0.6%, 8.54%]).

9.10 Conclusion

In conclusion, analysis A of the year 7 assessment data showed that following a propensity score based on the included covariates, TfM had a positive effect on student attainment. A

sensitivity analysis of the results showed concurred with the initial treatment effect estimation, and therefore the results are robust to changes in the propensity score matching specification. The treatment effect estimation and the sensitivity analysis also yielded significant p-values for the 95% confidence interval. As such, it can be concluded that compared to the previous curriculum, TfM has improved student outcomes. It is important to acknowledge that this result cannot be generalised beyond this specific sample with this specific assessment and included covariates.

Whilst analysis B suggests that TfM has a 4% positive effect on the year 10 assessment, it is important to view this analysis as another methodological lesson. A non-significant p-value and a confidence interval that passes zero suggests we cannot be certain of this positive effect. However, the data provided by school Z all preceded Covid and therefore there is no risk of the pandemic impeding the reliability of the results.

Chapter Ten: Discussion

The previous results chapters took each individual participating school and explored whether propensity score matching methods can be used as a tool to assess if TfM is associated with a change in student assessment outcomes. The first result section used school X as the pilot study to act as a learning case for the main study, which featured school Y and school Z. Each individual case offered key methodological lessons since this research is the first of its kind; trialling propensity score matching methods for evaluating mastery at the individual level.

This discussion chapter is structured as follows: a summary of the key findings from each individual case; a discussion taking an overview of the three cases, and the methodological value of them; a consideration of the main research question and whether this has been answered; as well as key implications emerging from this research.

10.1 Summary of Main Findings

10.1.1 Results Analysis

School X

For school X, the pilot school, three analyses were conducted. The first analysis took the assessment for which the entire 2017-18 and 2020-21 year 7 cohorts took. Following propensity scores based on gender, pupil premium, SEND, and KS2 maths scores, it was found that the non-TfM outperformed the TfM group by 5% on the raw score.

The second and third analyses of school X examined the second year 7 assessment, which split the same students into either the higher or foundation tier. Analysis B investigated the results of the higher tier assessment. Following propensity scores based on gender, pupil premium, KS2 maths scores, SEND and the first year 7 assessment, it was found that the non-TfM group outperformed the TfM group by 10.5%. The third analysis for school X, which looked at the foundation assessment and used the same covariates for the propensity score, also concluded that the non-TfM group outperformed the TfM group, this time by 11%. However, as outlined in the results chapter, these findings cannot be considered reliable since the 2020-21-year 7 cohort experienced disrupted schooling due to the COVID-19 pandemic. Therefore, PSM methods cannot

account for all known differences since one would expect the pandemic to be a highly influential and negative factor on student outcome.

As well as not being able to disaggregate the effects of the COVID-19 pandemic, the pilot study highlighted some key methodological warnings. The first of those was the importance of ensuring that continuous covariates are used in the propensity score. In the first analysis, the propensity scores appeared 'lumpy' as the majority of covariates were binary. When including the first year 7 assessment in the second and third analyses, the propensity scores were much more well distributed between 0 and 1. The other key methodological learning following the pilot study was to look closely at the key drivers of the propensity score, out of the included covariates, and to ensure that the balance of these is improved as it is unlikely to ever achieve perfect (100%) balance for all covariates.

School Y

For school Y, three analyses were conducted. The first of these looked at the year 10 higher assessment outcomes. Following propensity scores based on KS2 Maths scores, KS2 Grammar scores, KS2 Reading scores, gender, pupil premium, SEND, FSM, EAL and Midyis, it was found that the TfM group outperformed the non-TfM group by 4%. These results must be viewed with caution, however, as an awareness of the dates from which the data was provided reveals an overlap with the COVID-19 pandemic, and therefore the effects of disrupted education cannot be disaggregated. That said, to see an improvement in assessment outcomes for the TfM group despite the cohort experiencing disruption to their education is worthy of acknowledgement and may imply that TfM is associated with improved assessment outcomes.

The second analysis from school Y was only included in the results chapter for methodological purposes since it highlighted the importance of visually examining the region of common support. The PSM specifications removed a number of students rendering the remaining group too small to make any comparisons. For those that were left, their propensity scores were too dissimilar to provide meaningful assessment. Note that, when there is little common support, the matching algorithms (especially nearest neighbour) will still run in the statistical programme and provide an 'answer'. However, the resulting matches will actually have very different propensity scores, and therefore it is flawed to try to argue that the difference in outcomes can be associated to the difference in treatments. It became clear, therefore, during this analysis that it is important to

take care over common support and to avoid running an algorithm and assuming the answer is meaningful.

The third analysis from school Y investigated the effects of the TfM curriculum on the year 9 higher assessment outcome with the data provided preceding the COVID-19 pandemic. It was found that the non-TfM group had a higher mean outcome by 3.65%, but the 95% confidence crossed zero. If the confidence interval crosses zero, it indicates that the possibility that there was no effect of the difference in treatments should not be discounted.

Overall, the data from school Y provided some interesting methodological lessons: the need to look at a visual depiction of the region of common support and the need to conduct sensitivity analysis, as well as thinking carefully about the confidence interval to check the results are robust.

School Z

For school Z, two analyses were conducted. The first of these used the year 7 assessment on the four cohorts of students, all of whom sat the assessment before the COVID-19 pandemic. It was found that, following propensity scores based on the entry assessment percentage, gender, SEND, FSM, EAL and pupil premium, the TfM students outperformed the non-TfM students by 6.5% on the assessment percentage. Though it is noteworthy that this was a small 'dose' of TfM since the assessment was taken after one term of different curricula.

The second analysis investigated the effects of the mastery approach on the year 10 assessment scores, which looked at the longer-term impact since one of the cohorts used for comparison had been taught using TfM approaches since they were in year 7. In this analysis, it was found that the TfM group outperformed the non-TfM group by 4%. However, the results of the second analysis were not found to be statistically significant and the confidence interval crossed zero, meaning that the possibility that there was no effect of the difference in treatments should not be discounted. That said, it was highlighted at the end of the results section for school Z that, before any matching took place, the average outcome percentage for the TfM and non-TfM groups were almost identical (67% for non-TfM and 67.2% for TfM) and therefore PSM methods suggest that there might be a greater difference (a difference of 4%).

10.1.2 TfM Impact Assessment

Now that the main findings for the three schools have been outlined, it is important to acknowledge that, whether a positive or negative effect of TfM has been found, each result is only applicable to that particular assessment, on the particular students at that particular school. The effect is the average treatment effect for the treated (ATT) and cannot be generalised beyond the sample and the used assessment, i.e., the effect of the treatment actually applied and is distinct from ATE. Average treatment effect (ATE) is the measure used to compare treatments in randomised experiments and measures the difference in outcomes between treatment and control units for the population under consideration. The ATT does not look at the entire population; merely those that have received 'treatment'.

In chapter four, different methods for evaluating interventions were discussed, leading to the rationale for using PSM methods. RCTs are inherent to post-allocation trial effects and by using pre-existing data, this research has avoided the risk of such biases. It was discussed the multiple regression models are based on the assumption that there is no correlation between the independent variables and that the relationship between the independent and dependent variable is linear. This method was not appropriate to answer the research question that framed this thesis since there is often interaction between different independent variables in the education context. An ITS model was also not suitable since they rely on longitudinal data which was not appropriate for this research. Many schools refine their assessments over time, and it was crucially important to have a consistent and comparable outcome measure for the cohorts of students under analysis in this research. Longitudinal data would have led to more issues in finding schools that could provide comparable outcome measures for pre- and post- TfM groups of students. Along with the interviews which gave contextual insight, PSM methods ensured there were no systematic differences on key observed covariates and therefore was the best approach for the research question.

10.2 Methodological Lessons and Limitations

This research sought to explore methods that are not intrinsically linked to post-allocation differences in the way that experimental methods such as RCTs are. It also sought to explore the longer-term impact of embedded TfM approaches which the existing evidence base to date had not addressed. In doing so, this research has exposed many methodological considerations and concerns when using observational research methods to assess whether a transition to TfM approaches can

be associated with changes in student assessment outcomes. As such, this next section takes an overview of the three cases and discuss these methodological lessons. This is particularly important since this research sought to trial alternative methods to assess the impact of TfM.

10.2.1 Lessons and Limitations

From two of the school cases, the need to analyse the region of common support carefully and visually following the matching process emerged. Similar to the discussion around common support earlier in this chapter, there is a danger of looking at the balance improvement statistics and seeing a balance improvement across the covariates, but not being aware of the lack of common propensity scores across the two treatment groups. Thus, when conducting PSM, it is important to use a visual plot to assess the region of common support before deciding on a matching specification and estimating the treatment effect and avoid 'push-button' thinking. It was found that including more than one continuous covariate in the propensity score specification helped to overcome 'lumpy' propensity scores and therefore ensure that there was more overlap between the two groups. Critically, PSM does not work unless there is a substantial region of common support.

The need to conduct a sensitivity analysis on the results was an important lesson. It is important to check that the results found are robust to changes in the propensity score specification by conducting the treatment effect estimate using a different matching version. When the results concur, then one can be reasonably confident that the results are robust, but when they do not, they must be interpreted cautiously.

The key methodological lessons outlined above suggest that there should be a consideration of the limitations of using propensity score matching for the outlined research question. As with any research method, it is important to recognise and acknowledge these limitations. Perhaps the most glaring limitation is that, when using propensity score matching methods for a set of given covariates, it is impossible to be certain that all confounders have been controlled for. This limitation is similar to that highlighted in the critique of RCTs where it is only up to the point of allocation where differences between the two treatment groups are controlled for.

A recurring theme throughout this research has been the issue with data overlapping the COVID-19 pandemic since it is impossible to disaggregate the effects of this from a change in curriculum on student attainment. Finding schools that can provide adequate data was challenging on two levels. Firstly, it was difficult to get schools to engage due to the intensity of working in

education, perhaps accentuated by Covid. Secondly, the nature of the data gathered was not always ideal. For PSM to work effectively, this research found that more than one continuous covariate was needed. It was clear that many schools predominantly have discrete or binary data available on demographics, which led to clustered or 'lumpy' propensity scores. Although costly, an RCT brings all the infrastructure needed to gather the desired data, whereas in the instance of an observational study, there is a reliance on the data already gathered.

A key concern with the data already gathered was whether it gave information on all known differences between the two treatment groups. Semi-structured interviews were conducted to try to ensure that the change to a TfM curriculum was the only substantial change that happened at that time for each participating school. For example, that there were no changes in leadership or whole school policy for each school around the time that each assessment was taken. That said, there are many other influential factors that can affect student performance, where schools do not necessarily have relevant data for. In particular, family income and parental qualification were cited as two important factors in the methodology chapter and much of the literature references students' attitude towards Maths, parent involvement, quality of teachers and peer influence as others (Wang, 2007). It is important to highlight that where a model may not include all influential factors, the results may be affected by hidden (i.e., omitted variable) bias (Morgan et al., 2008).

10.2.2 PSM Lessons and Limitations

Relevant to PSM methods in general is the need to agree a set direction for the propensity score matching process. The risk of the 'garden of forking paths' was discussed in the methodology chapter. Gelman (2013) warned that where a researcher relies on probabilities in decision making there is often a situation where there are many different routes to arrive at a decision. If chance processes are involved, some of the routes may give a result which is spurious and a researcher not alert to this issue might consciously, or unconsciously, choose the paths that lead to the apparently strongest results. There were many different PSM specifications open for this research and therefore, there was a need to set a direction and to pre-register that direction as was done following the pilot study in this case.

It is also important to ensure that the choice of propensity score variant is driven by its key drivers. The three cases collectively showed that it was unlikely to ever achieve one hundred percent balance across all included covariates. When choosing a matching specification as the one to take

forward to estimate the treatment effect, it is important to ensure that the covariates which drove the propensity score have become more balanced across the two groups. On the other hand, if any covariates are well-balanced before matching, small changes are likely to be flagged up as causing large imbalances in the balance improvement statistics, and therefore, there may be the need to remove covariates that are already adequately balanced. For example, in analysis B for school Z, had the researcher not looked closely at the key drivers of the propensity score before any matching method was implemented, a version of matching may have been chosen that had not optimised the balance of the key drivers.

The central research question of this thesis was as follows:

"How might non-experimental methods be adapted to address whether Teaching for Mastery is associated with improved assessment scores, compared to previous pedagogical approaches?"

It is appropriate to return to first principles and the arguments set out in the literature review to consider if this research has answered the intended research question and indeed, if PSM can ever answer a research question which looks to attribute cause. The principle aim of this research was to find a non-experimental way of ascribing cause so that it can be concluded that the difference in student outcomes is caused by the difference in treatments. The seminal piece of research into the effects of mastery is an RCT, which was explored and critiqued in the earlier chapters. Usually, RCTs are short-term due to their expensive nature and therefore do not tell us anything about any long-term effects of an intervention. Further to this, experimental approaches cannot account for post-allocation differences which are not the difference in treatments, particularly trial effects such as a group knowing that they are getting the new treatment or knowing they are under scrutiny. An RCT is considered the 'gold-standard' of education research since it accounts for causal factors associated with pre-allocation group differences, but it was argued early in this thesis that there is room for observational research in the field to help overcome the issue with post-allocation differences. PSM is an attempt to match students on factors which impact on the treatment they receive and outcome to minimise the impact of pre-allocation group differences, whilst also avoiding the post-allocation trial effects of RCTs, addressing issues of blinding to group allocation whilst assessing long-term effects.

That said, none of the methods explored in this research have been found to deal with other post-allocation differences. If something other than the difference in treatments happens to one of the two groups after allocation, that something can never be discounted as the cause of any difference in outcomes. At best, one can use their theoretical understanding of the context to try to avoid this. In this research, semi-structured interviews were used as a means to account for any

other explanation for a difference in outcomes, such as changes in teachers or school policy. However, there are instances in this research where a huge post-allocation difference occurred, e.g., a global pandemic. This was the main challenge in conducting this research. The aim was to collect data from schools that did not overlap with the pandemic or use data pre-pandemic.

10.3 Key Implications

There are 40 Maths Hub across the country with thousands of schools engaging with TfM programmes offered by the Hubs that are coordinated by the NCETM. Many schools are continuing to transition to a TfM approach to teaching mathematics. The central aim of the Maths Hubs programme is to improve mathematics education and student outcomes nationally and, therefore, schools engaging with the professional development offered to them through the Hubs should consider taking an evidence-informed approach. Where possible, institutions should seek to keep assessments constant so that they are in a position to establish if a change in pedagogy, practice and curriculum is associated with improved assessment outcomes; both in the short and long- term. The analyses outlined in this research have shown that propensity score matching, when it does work, can detect deeper change than a surface-level difference-in-means analysis might reveal. True difference can only be detected when the groups of students that are compared are matched across a set of given covariates.

It has become standard, according to Cartwright and Hardie (2012), that policy makers base their recommendations on evidence. However, they have raised questions as to whether the methods that policy makers rely on, predominantly ones that imitate practices in medicine such as RCTs, collect the best evidence. Experimental methods, like RCTs, do not help policy makers to predict if policies would be effective and therefore, according to Cartwright and Hardie (2012), there is a need to consider alternative approaches for evidence-based policy that exclude trial biases and effects. Therefore, because the observational methods trialled in this research have allowed the exploration of the longer-term impact of a policy (in this case, TfM approach), then propensity score matching methods could be a useful tool for education policy makers allowing individual schools to determine if initiatives have or have not impacted student outcomes. When doing this, it is important to consider and be aware of 'confirmation bias' (Wason, 1960). Confirmation bias has been referred to as a 'cognitive error' made by people when they are only willing to accept new information when it confirms what they already believe (Nickerson, 1998). Those that demonstrate 'confirmation bias' are likely to intentionally seek out evidence that supports already solidified

opinions and purposefully refuse, albeit subconsciously, to accept any evidence that goes against those beliefs (Nickerson, 1998). A big question in educational research is, "*does one subject results that they agree with to the same scrutiny that they subject results to that they disagree with*" (Ritchie et al., 2012)?

10.4 Conclusion

Whilst it has not been possible to ascertain if TfM is associated with improved assessment scores compared to previous pedagogical approaches, this research has contributed to the field. It is the first of its type to explore using observational methods as a means to investigate the longer-term impact of embedded TfM at the individual level. It has shown, to an extent, that observational methods can be used to ascribe cause when the researcher accommodates the key methodological lessons outlined in the previous chapter.

PSM methods have enabled the overcoming of the challenge of dealing with some of the post-allocation differences that are inherent to experimental methods, such as trial effects, thereby showing advantages of the method over RCTs. It has also been shown that pre-allocation group differences can be minimised through matching on given covariates. To complement this, semi-structured interviews provided a way to account for other events such as change in leadership and school ethos which could impact student attainment.

It has not been possible, however, to definitively determine if a change to TfM can be associated with a change in student assessment scores because of the huge post-allocation difference which is the Covid pandemic. There have been three instances in this research where the data preceded Covid, however the results from these threw up other methodological considerations such as non-significant p-values and confidence intervals which crossed zero.

10.5 Future Research

The ideal analysis would have been to explore GCSE results pre- and post- TfM but, because of the COVID-19 pandemic, this was not possible and therefore the research focused on topic and end-of-year assessments. There is scope, therefore, to take this research further when the effects of

the pandemic have subsided[23] so that schools are able to compare student attainment year-on-year again and to review GCSE performance. A larger sample size of schools will lead to more clarity in whether the observational methods trialled can answer the central research question. An exploration of primary schools and the effect of mastery on pupil attainment at Key Stages 1 and 2 is also viable.

Access to a more substantial database, such as the National Pupil Database, would also inform future research where data on student attainment at various levels, demographics, absence and attendance can be found. There are huge advantages to using the National Pupil Database as it would give the researcher access to numerous critical covariates, many of them continuous, for a large student cohort. It would also provide access to standardised attainment data which could be used as the outcome measure. However, it must be used in conjunction with data defining which schools transitioned to TfM and when. Students that have had which teaching approach need identifying so that it is clear which data to extract from the national pupil database. Moreover, it is important to ensure the data taken from the database does not overlap with the global pandemic in the same way that some of the data has in this research.

Overall, this research has paved the way for observational approaches to be used as a tool to investigate if TfM approaches can be associated with improved student outcomes. Propensity score matching methods can overcome the trial biases and effects issue inherent within experimental studies, whilst also minimising pre-allocation group differences. Anything which happens directly to the two treatment groups after allocation, however, holds the potential for playing a causal role and therefore it is important to conduct semi-structured interviews to complement the quantitative method to give the researcher an insight into the school context so they can be sure that the transition to TfM was the only real change at the time.

[23] Of course, there is difficulty in describing such a time when schools will be clear of the “Covid effects”.

References

Abaasa, A., Mayanja, Y., Asiki, G., Price, M. A., Fast, P. E., Ruzagira, E., Kaleebu, P., & Todd, J. (2021). Use of propensity score matching to create counterfactual group to assess potential HIV prevention interventions. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86539-x

Adams, A., Karunakaran, M. S., Klosterman, P., Knott, L., & Ely, R. (2016). Using Precise Mathematics Language to Engage Students in Mathematics Practices. *Psychology of Mathematics & Education of North America*, *38*, 1158–1165. https://eric.ed.gov/?id=ED583792

Akobeng, A. (2005). Understanding randomised controlled trials. *Archives of Disease in Childhood*, *90*, 840–844. https://adc.bmj.com/content/90/8/840.citation-tools

Almond, N. (2022). *Fluency, Reasoning and Problem Solving: What This Looks Like In Every Maths Lesson*. Third Space Learning. https://thirdspacelearning.com/blog/fluency-reasoning-problem-solving/

Anderson, L. (1975). Major Assumptions of Mastery Learning. *Annual Meeting of the Southeast Psychological Association.* https://eric.ed.gov/?id=ED150172

Ark Academy Plus. (n.d.). *Our principles.* Ark Academy. https://www.arkcurriculumplus.org.uk/curriculum-intent/our-principles

Askew, M. et al., (2015). *Teaching for Mastery Questions, tasks and activities to support assessment*. NCETM. https://www.exeterconsortium.com/uploads/1/1/5/9/115936395/mastery_assessment_y6.pdf

Austin, J. L., & Howson, A. G. (1979). Language and mathematical education. *Educational Studies in Mathematics, 10*(2), 161-197.

Austin, P. (2012). A comparison of 12 algorithms for matching on the propensity score. *Statistics In Medicine*, *33*(6), 1057-1069. doi: 10.1002/sim.6004

Austin, PC. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat.,* 10(2), 150–161.

Axbey, H. (2020). 'Fusion or Friction? UK Teachers' Experiences of Cross-Cultural Teaching in China.' *Imagining Better Education: Conference Proceedings*, 15-26.

BERA (2011). Ethical Guidelines for Educational Research. https://www.bera.ac.uk/publication/bera-ethical-guidelines-for-educational-research-2011

Ballantine, N. (2018). *'What's the same? What's different?' Variation theory in practice*. https://nicolaballantine.wordpress.com/2018/04/15/whats-the-same-whats-different-variation-theory-in-practice/

Bao, J. (2002). Comparative study on composite difficulty of Chinese and British School mathematics curricula. (Unpublished Doctoral Dissertation), East China Normal University, Shanghai, China.

Bao, J., Huang, R., Yi, L., & Gu, L. (2003a). Study in bianshi teaching [In Chinese]. *Mathematics Teaching [Shuxue Jiaoxue], 1*, 11–12.

Bao, J. (2004). A comparative study on composite difficulty between new and old Chinese mathematics textbooks. *How Chinese learn mathematics: Perspectives from insiders*, 208-227. https://doi.org/10.1142/9789812562241_0008

Barton, C. (2021). *Variation Theory*. https://variationtheory.com/

Becker, J. P., & Miwa, T. (1986). Proceedings of the U. S. – Japan seminar on mathematical problem solving. Retrieved from ERIC Document Reproduction Service No. ED 304 315.

Becker, S. O., & Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *Stata Journal*, *2*, 358–377.

Bhide, A., Shah, P. S., & Acharya, G. (2018). A simplified guide to randomized controlled trials. *Acta Obstetricia et Gynecologica Scandinavica*, *97*(4), 380–387. https://doi.org/10.1111/aogs.13309

Blausten, H., Gyngell, C., Aichmayr, H., & Spengler, N. (2020). Empowering Teachers to Build a Better World. *SpringerBriefs in Education*. doi:10.1007/978-981-15-2137-9

Bloom, B. (1968). All Our Children Learning – A Primer for Parents, Teachers, and other Educators. *McGraw-Hill.* ISBN 9780070061187.

Bloom, B. (1968). Learning for Mastery. Instruction and Curriculum. *Regional Education Laboratory for the Carolinas and Virginia, Topical Papers and Reprints*, *1*, 1–12. https://eric.ed.gov/?id=ED053419

Boaler, J., & Staples, M. (2008). Creating mathematical futures through an equitable teaching approach: the case of Railside school. *Teachers College Record*, *110*(3), 608–645.

Boylan, M. (2018). *Where did maths mastery come from?* Schools Week. https://schoolsweek.co.uk/where-did-maths-mastery-come-from/

Boylan, M., Maxwell, B., Wolstenholme, C., Jay, T., Demack, S. (2018). The Mathematics Teacher Exchange and 'Mastery' in England: The Evidence for the Efficacy of Component Practices. *Educ. Sci.*, *8*, 202. https://doi.org/10.3390/educsci8040202

Boylan, M., Wolstenholme, C., Maxwell, B., Jay, T. (2017). Longitudinal evaluation of the Mathematics Teacher Exchange: China-England – Third Interim Report. Department for Education.

Boylan, M., Wolstenholme, C., Demack, S., Maxwell, B., Jay, T., Adams, G., & Reaney, S. (2019). Longitudinal evaluation of the Mathematics Teacher Exchange: China-England - Final Report. Department for Education.

Burns, R. B. (2000). Introduction to research methods (4th ed.). Frenchs Forest, NSW, Australia: Pearson Education.

Cai, J., & Hwang, S. (2002). Generalized and generative thinking in US and Chinese students' mathematical problem solving and problem posing. *The Journal of mathematical behavior, 21*(4), 401-421.

Cai, J., Nie, B. (2007). Problem solving in Chinese mathematics education: research and practice. *ZDM Mathematics Education, 39*, 459–473. https://doi.org/10.1007/s11858-007-0042-3

Cai, J., & Lester, F. K. (2005). Solution representations and pedagogical representations in Chinese and U. S. classrooms. *Journal of Mathematical Behavior, 24*, pp. 221-237.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal Of Economic Surveys*, *22*(1), 31-72. doi: 10.1111/j.1467-6419.2007.00527.x

Cambridge Maths Hub. (2016). *Variation: developing mathematical thinking through intelligent practice*. https://cambridgemathshub.org/wp-content/uploads/2016/08/Variation-Mastery-Conference-KS1.pdf

Cartwright, N., & Hardie, J. (2012). Evidence-Based Policy: A Practical Guide to Doing It Better (Illustrated ed.). Oxford University Press.

Cepeda, M. (2003). Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. *American Journal Of Epidemiology*, *158*(3), 280-287. doi: 10.1093/aje/kwg115

Charlie's Angels. (2016). *England-China teacher exchange 2016/17: Seeing Shanghai teaching at first hand*. NCETM. https://www.ncetm.org.uk/features/england-china-teacher-exchange-2016-17-seeing-shanghai-teaching-at-first-hand/

Chen, C. (2019). *UK teachers exchanging math insights in Shanghai*. Chinadaily.Com.Cn. https://global.chinadaily.com.cn/a/201911/13/WS5dcb09a0a310cf3e35576eb8.html

Choi, J., Dekkers, O. M., & le Cessie, S. (2018). A comparison of different methods to handle missing data in the context of propensity score analysis. *European Journal of Epidemiology*, *34*(1), 23–36. https://doi.org/10.1007/s10654-018-0447-z

Cifarelli, V & Cai, J. (2005). The evolution of mathematical explorations in open-ended problem solving situations. *Journal of Mathematical Behavior, 24*, 302-324.

Cimbricz, S. (2013). *Academic language*. Retrieved from https://www.brockport.edu/oat/docs/ALanguage_Cimbricz%20FINAL1.pdf

Clapham, A., & Vickers, R. (2018). Neither a borrower nor a lender be: Exploring 'teaching for mastery' policy borrowing. *Oxford Review of Education, 44*(6), 787-805. doi:10.1080/03054985.2018.1450745

Clarke, D., Keitel, C., & Yoshinori, S. (2006). Mathematics classrooms in twelve countries: the insider's perspective. Rotterdam: Sense Publishers.

Cochran WG & Rubin DB. (1973). Controlling bias in observational studies: a review. *Sankhyā: Indian J Stat, Ser A., 35*(4), 417–446.

Cornfield J, Haenszel W, and Hammond EC et al. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J Natl Cancer Inst., 22*(1), 173 - 203.

Cummings, Kelsey. (2015). How Does Tutoring to Develop Conceptual Understanding Impact Student Understanding?. *BSU Honors Program Theses and Projects*. Item 96. http://vc.bridgew.edu/honors_proj/96

Department for Education. (2013). Mathematics programmes of study: key stage 3. National curriculum in England. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/239058/SECONDARY_national_curriculum_-_Mathematics.pdf

Department for Education (2014). The National Curriculum in England: Key stages 3 and 4 framework document. Retrieved April 05, 2021, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/840002/Secondary_national_curriculum_corrected_PDF.pdf

Department for Education. (2014). National curriculum: secondary curriculum. https://www.gov.uk/government/publications/national-curriculum-in-england-secondary-curriculum

Department for Education (2016). South Asian method of teaching maths to be rolled out in schools. Retrieved August 05, 2020, from https://www.gov.uk/government/news/south-asian-method-of-teaching-maths-to-be-rolled-out-in-schools

Department for Education. (2018). National Statistics: Revised GCSE and equivalent results in England: 2016 to 2017. https://www.gov.uk/government/statistics/revised-gcse-and-equivalent-results-in-england-2016-to-2017

Department for Education. (2019). Shanghai maths exchange shows power of international partnership [Press release]. https://www.gov.uk/government/news/shanghai-maths-exchange-shows-power-of-international-partnership

Department for Education and Employment (DfEE) (1999). The National Numeracy Strategy: Framework for Teaching Mathematics from Reception to Year 6. London: DfEE.

Desimone, L.M., T.M. Smith, S.A. Hayes and D. Frisvold. (2005). Beyond accountability and average mathematics scores: Relating state education policy attributes to cognitive achievement domains. *Educational Measurement: Issues and Practice*, 24(4), 5-18. Available at: https://doi.org/10.1111/j.1745-3992.2005.00019.x.

Ding, L., Jones, K., & Sikko, S. A. (2017). An Expert Teacher's Use of Teaching with Variation to Support a Junior Mathematics Teacher's Professional Learning. *Teaching and Learning Mathematics through Variation: Confucian Heritage meets Western Theories*, 241-266.

Dresher, R. (1934). Training in mathematics vocabulary. *Educational Research Bulletin*, *13*(8), 201-204.

Drury, H., & Mathematics Mastery. (2014). Mastering mathematics: teaching to transform achievement. Oxford: Oxford University Press.

Drury, H. (2018). How To Teach Mathematics For Mastery. Oxford University Press.

Education Endowment Foundation. (2015). Mathematics Mastery: Secondary Evaluation. London: Institute of Education.

Ejdemyr, S. (2020). *R Tutorial 8: Propensity Score Matching*. R. https://sejdemyr.github.io/r-tutorials/statistics/tutorial8.html

Ekanayake, Athula. (2015). Validity and Reliability in Case Study Research in Accounting: A Review and Experience. *Modern Sri Lanka Studies*, 161-185.

English, L.D. (1997). The Development of Fifth-Grade Children's Problem-Posing Abilities. *Educational Studies in Mathematics, 34,* 183–217. https://doi.org/10.1023/A:1002963618035

English L., Sriraman B. (2010). Problem Solving for the 21st Century. *Theories of Mathematics Education*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00742-2_27

Fetzer, M., & Tiedemann, K. (2015). The interplay of language and objects in the mathematics classroom. *CERME9,* 1387-1392.

Freitas, B. (2021). *Using Propensity Score Matching to Uncover Shopify Capital's Effect on Business Growth*. Shopify Engineering. https://shopify.engineering/propensity-score-matching-shopify-capital

Gant, T., & Crowland, K. (2017). A Practical Guide to Getting Started with Propensity Scores. *Data & Information Management Enhancement (DIME)*. https://pdf4pro.com/view/a-practical-guide-to-getting-started-with-propensity-scores-433aba.html

Garrido, M. M., Kelley, A. S., Paris, J., Roza, K., Meier, D. E., Morrison, R. S., & Aldridge, M. D. (2014). Methods for Constructing and Assessing Propensity Scores. *Health Services Research, 49*(5), 1701-1720. https://doi.org/10.1111/1475-6773.12182

Geary, D.C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*(1), 4-15.

Gelman, A. and Loken, E. (2013). "The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time." http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Ghazali, N. H. C., & Zakaria, E. (2011). Students' procedural and conceptual understanding of mathematics. *Australian Journal of Basic and Applied Sciences*, *5*(7), 684–691. Retrieved from http://www.ajbasweb.com/old/ajbas/2011/July-2011/684-691.pdf

Greifer, N. (2022). *Matching Methods*. https://cran.r-project.org/web/packages/MatchIt/vignettes/matching-methods.html

Greifer, N. (2022). *Assessing Balance*. R-Project. https://cran.r-project.org/web/packages/MatchIt/vignettes/assessing-balance.html#recommendations-for-balance-assessment

Gu, L. (1981). The visual effect and psychological implication of transformation of figures in geometry. Paper presented at annual conference of Shanghai Mathematics Association. Shanghai, China.

Gu, L. Y. (1991). *Xuehui jiaoxue* [Learning to teach]. Beijing: People's Educational Press.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2, 405–420.

Gu, L., & Huang, R. Marton, F. (2004). Teaching with variation: a Chinese way of promoting effective mathematics learning. *How Chinese Learn Mathematics: Perspectives from Insiders*, *309-347.*

Gu, L., Yang, Y., & He, Z. (2015). Qingpu mathematics teaching reform and its impact on student learning. *How Chinese teach mathematics: Perspectives from insiders,* 435-454.

Gu, F., Huang, R., & Gu, L. (2017). Theory and development of teaching through variation in mathematics in China. *Teaching and Learning Mathematics through Variation*, 13–21. https://doi.org/10.1016/j.lindif.2017.11.017

Gurganus, S. P., & Wallace, A. H. (2005). Teaching for Mastery of Multiplication. *Teaching Children Mathematics, 12*(1), 26-33. National Council of Teachers of Mathematics.

Hamilton, L. (2011) Case studies in educational research. British Educational Research Association.

Hattie, J. (2017). *Hattie's 2017 Updated List of Factors Influencing Student Achievement.* Available at: http://www.evidencebasedteaching.org.au/hatties-2017-updated-list/

Hayes, A. (2022). *Multicollinearity*. Investopedia. https://www.investopedia.com/terms/m/multicollinearity.asp

Heckman, J. J., Ichimura, H., & Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, *64*, 605–654.

Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture toward a stronger collective quantitative proficiency. *Educational Researcher*, *39*(3), 229–240.

Hiebert, J., J.W. Stigler, J.K. Jacobs, K.B. Givvin, H. Garnier, M. Smith, H. Hollingsworth, A. Manaster, D. Wearne and R. Gallimore. (2005). Mathematics teaching in the United States today (and tomorrow): Results from the TIMSS 1999 video study. *Educational Evaluation and Policy Analysis*, *27*(2), 111-132. Available at: https://doi.org/10.3102/01623737027002111.

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, *25*, 2230–2256.

Ho, D., Imai, K., King, G., & Stuart, E. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*(3), 199-236.

Holm, E. (2021). *15 Multiple Regression | Introduction to Research Methods*. https://bookdown.org/ejvanholm/Textbook/multiple-regression.html

Huang, R., & Leung, F. K. S. (2005). Deconstructing teacher-centrerdness and student-centeredness dichotomy: A case study of a Shanghai mathematics lesson. *The Mathematics Educators,* 15(2), 35-41.

Huang, H. M. E., & Witz, K. G. (2011). Developing children's conceptual understanding of area measurement: A curriculum and teaching experiment. *Learning and Instruction*, *21*(1), 1-13. https://doi.org/10.1016/j.learninstruc.2009.09.002

Intaros, P., Inprasitha, M., & Srisawadi, N. (2014). Students' problem solving strategies in problem solving-mathematics classroom. *Procedia - Social and Behavioral Sciences*, *116*, 4119–123. https://doi.org/10.1016/j.sbspro.2014.01.901.

Jackson, R. (2020). *Mastery in primary mathematics | STEM*. Stem Learning. https://www.stem.org.uk/news-and-views/opinions/mastery-primary-mathematics

Jacques, L. (2018). What is teaching with variation and is it relevant to teaching and learning mathematics in England? *UCL, Institute of Education*.

Jerrim and Vignoles. (2015). Mathematics Mastery: Overarching Summary Report. *Education Endowment Foundation*. https://files.eric.ed.gov/fulltext/ED581180.pdf

Jerrim, J. & Vignoles, A. (2016). The link between East Asian 'mastery' teaching methods and English children's mathematics skills. *Economics of Education Review, 50,* 29-44.

Johnson, H. C. (1944). The effect of instruction in mathematical vocabulary upon problem solving in arithmetic. *Journal of Educational Research, 38,* 97-110.

Kangaroo Maths. (n.d.). *Home.* https://kangaroomaths.co.uk/

Kazemi, E., & Stipek, D. (2001). Promoting conceptual thinking in four upper-elementary mathematics classrooms. *The Elementary School Journal*, *102*(1), 59–80.

Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2006). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology, 42*(1), 11–26.

Kenton, W. (2021). *What Is a Monte Carlo Simulation?* Investopedia. https://www.investopedia.com/terms/m/montecarlosimulation.asp

Kontopantelis, E., Doran, T., Springate, D. A., Buchan, I., & Reeves, D. (2015). Regression based quasi-experimental approach when randomisation is not an option: interrupted time series analysis. *BMJ*, *350*(jun09 5), h2750. https://doi.org/10.1136/bmj.h2750

Kranda, J. (2008). Precise mathematical language: Exploring the relationship between student vocabulary understanding and achievement. Unpublished MA thesis, Department of teaching, learning, and teacher Education, University of Nebraska-Lincoln.

Krummheuer, G. (1999). The narrative character of argumentative mathematics classroom interaction in primary education. *CERME1,* 331-341.

Kullberg et al. (2016). Teaching one thing at a time or several things together? –Teachers changing their way of handling the object of learning by being engaged in theory-based professional learning community in Mathematics and Science. *Teachers and teaching. Theory and Practice*, 22(6), 1–15.

Kullberg, A., Runesson, U., & Marton, F. (2017). What is made possible to learn when using the variation theory of learning in teaching mathematics? *ZDM Mathematics Education*, *49*(4). doi:10.1007/s11858-017-0858-4

Leighton, D. (2020). *Maths Mastery Toolkit: A Practical Guide To Mastery Teaching And Learning*. Third Space Learning. https://thirdspacelearning.com/blog/maths-mastery-techniques-primary-free-resource/

Leite, W. L. (2016). Practical Propensity Score Methods Using R (1st ed.). SAGE Publications, Inc.

Lesh, R. & Zawojewski, J. (2007). Problem solving and modeling. *Second handbook of research on mathematics teaching and learning.* Charlotte, NC: Information Age Publishers, 763-804.

Lester, F. K. (1994). Musings about mathematical problem-solving research: 1970–1994. *Journal for Research in Mathematics Education*, *25*(6), 660–675. Retrieved from http://www.jstor .org/stabl e/74957 8

Lester, Frank K. Jr. (2013). "Thoughts About Research On Mathematical Problem- Solving Instruction". *The Mathematics Enthusiast*, *10*(1). Available at: https://scholarworks.umt.edu/tme/vol10/iss1/12

Lester, F. K., & Kehle, P. E. (2003). From problem-solving to modeling: The evolution of thinking about research on complex mathematical activity. *Beyond constructivism: Models and modeling perspectives on mathematics problem solving, learning, and teaching,* 501-518.

Levine, D. (1985). Improving student achievement through mastery learning programs. *Jossey-Bass.* ISBN 9780875896458.

Liljedahl, P. (2008). The AHA! experience: Mathematical contexts, pedagogical implications. Saarbrücken, Germany: VDM Verlag.

Liljedahl, P., & Santos-Trigo, M. (2019). *Mathematical Problem Solving: Current Themes, Trends, and Research (ICME-13 Monographs)* (1st ed. 2019 ed.). Springer.

Ling, L. M., Chik, P., & Pang, M. F. (2006). Patterns of Variation in Teaching the Colour of Light to Primary 3 Students. *Instructional Science*, *34*(1), 1–19. https://doi.org/10.1007/s11251-005-3348-y

Lim, C. S. (2007). Characteristics of Mathematics Teaching in Shanghai, China: Through the Lens of a Malaysian. In *Mathematics Education Research Journal,* 19(1), 77-89.

Lunt, M. (2013). Selecting an Appropriate Caliper Can Be Essential for Achieving Good Balance With Propensity Score Matching. *American Journal Of Epidemiology*, *179*(2), 226-235. doi: 10.1093/aje/kwt212

Mason, J., Burton, L.,& Stacey, K. (1982). Thinking mathematically. Harlow: Pearson Prentice Hall.

Mathematics Mastery (2019). *Secondary Programme*. https://www.mathematicsmastery.org/wp-content/uploads/2019/10/MM_Secondary_Brochure_2019.pdf

Mathematics Mastery (n.d.). *Ofsted and Curriculum - The Structure*. https://www.mathematicsmastery.org/ofsted-and-curriculum-the-structure/

Mattock, P. (2023). *Welcome to Goal Free Problems!* http://goalfreeproblems.blogspot.com/

McCourt, M. (2019). *Teaching for Mastery*. John Catt Educational.

Mcginn, K.M.; Booth, J.L. (2018). Precise mathematics communication: The use of formal and informal language. *Bordón Rev. Pedagog., 70*, 165–184. doi:10.13042/Bordon.2018.62138

Merriam, S. B. (1985). The Case Study in Educational Research: A Review of Selected Literature. *Journal of Educational Thought*, *19*(3). https://eric.ed.gov/?id=EJ330243

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*, 118–124.

Ministry of Education. (2012). Mathematics Syllabus *Secondary* One to Four, Ministry of Education, Singapore.

Ministry of Education, Singapore. (n.d.). TIMSS & PIRLS International Study Centre. http://timssandpirls.bc.edu/timss2015/encyclopedia/countries/singapore/the-mathematics-curriculum-in-primary-and-lower-secondary-grades/

Moore, A. W., Anderson, B., Das, K., & Wong, W. K. (2006). Combining Multiple Signals for Biosurveillance. *Handbook of Biosurveillance*, 235–242. https://doi.org/10.1016/B978-012369378-5/50017-X

Morgan, P., Frisco, M., Farkas, G., & Hibel, J. (2008). A Propensity Score Matching Analysis of the Effects of Special Education Services. *The Journal Of Special Education*, *43*(4), 236-254. doi: 10.1177/0022466908323007

Morgan, C.J. (2018). Reducing bias using propensity score matching. *J. Nucl. Cardiol. 25*, 404–406. https://doi.org/10.1007/s12350-017-1012-y

Mutawah, M., Thomas, R., Eid, A., Mahmoud, E., & Fateel, M. (2019). Conceptual Understanding, Procedural Knowledge and Problem-Solving Skills in Mathematics: High School Graduates Work Analysis and Standpoints. *International Journal of Education and Practice,* 7(3), 258-273. https://doi.org/10.18488/journal.61.2019.73.258.273

NCETM. (n.d). *Five Big Ideas in Teaching for Mastery*. https://www.ncetm.org.uk/teaching-for-mastery/mastery-explained/five-big-ideas-in-teaching-for-mastery/

NCETM. (2014). *Mastery approaches to mathematics and the new national curriculum*. https://www.ncetm.org.uk/media/2tlkwtz5/developing_mastery_in_mathematics_october_2014-pd.pdf

NCETM. (2016). *The Essence of Maths Teaching for Mastery. Mastery Explained, 1.* https://www.ncetm.org.uk/media/uhjhtxy1/the-essence-of-maths-teaching-for-mastery-june-2016.pdf

NCETM. (2019). *Teaching for Mastery: What is happening in primary maths, and what next?* https://www.ncetm.org.uk/media/2ljdu4kh/ncetm_primary_teachingformastery_report_july2019.pdf

NCETM. (2021). *Supporting Research, Evidence and Argument.* https://www.ncetm.org.uk/teaching-for-mastery/mastery-explained/supporting-research-evidence-and-argument/

NCSL. (2013). Report on research into maths and science teaching in the Shanghai region. Nottingham: NCSL.

NCSL. (2014). Report on the international Maths Research Programme, China 2014. Nottingham: NCSL.

NRICH. (2015). *Mastering Mathematics and Problem Solving.* https://nrich.maths.org/11796

Nahdi, D. S., & Jatisunda, M. G. (2020). Conceptual understanding and procedural knowledge: a case study on learning mathematics of fractional material in elementary school. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/1477/4/042037

National Council of Supervisors of Mathematics, U.S.A. (1977). Position paper on basic mathematical skills. Washington, DC: National Institute of Education

National Council of Teachers of Mathematics. (1991). Professional standards for teaching mathematics. Reston, VA: Author.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Nilsson, P. (2014). When teaching makes a difference: developing science teachers' pedagogical content knowledge through learning study. *International Journal of Science Education, 36*(11), 1794–1814.

Nosek, B. A. (2018). *The preregistration revolution*. PNAS. https://www.pnas.org/content/115/11/2600

OECD. (2012). PISA 2012 Results in Focus What 15-year-olds know and what they can do with what they know. Available from: http://www.oecd.org/pisa/keyfindings/pisa-2012- results-overview.pdf

Ofqual. (n.d). *GCSE outcomes in England.* https://analytics.ofqual.gov.uk/apps/GCSE/Outcomes/

Pate, K., & Norman, N. (2019). Maths Progress Core Textbook 1: Second Edition (Maths Progress Second Edition) (2nd ed.). Pearson Education.

Pittard, V. (2018). *Mastery: solving the problem*. Oxford Education Blog. https://educationblog.oup.com/primary/mastery-solving-the-problem

Planas, N., Morgan, C., & Schütte, M. (2018). Mathematics education and language. Lessons from two decades of research. *Developing research in mathematics education. Twenty years of communication, cooperation and collaboration in Europe*, 196–210. London: Routledge.

Pólya, G. (1965). Mathematical discovery: On understanding, learning and teaching problem solving (Vol. 2). New York, NY: Wiley.

Rahman, M. (2019). 21st Century Skill "Problem Solving": Defining the Concept. *Asian Journal of Interdisciplinary Research*, *2*(1), 64–74. https://doi.org/10.34256/ajir1917

Raiker, A. (2002). Spoken language and mathematics. *Cambridge Journal of Education, 32*(1), 45-60.

Raynor WJ. (19830. Caliper pair-matching on a continuous variable in case-control studies. *Commun Stat Theory Methods, 12*(13), 1499–1509.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect. *PLoS ONE*, *7*(3). https://doi.org/10.1371/journal.pone.0033423

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362. https://doi.org/10.1037/0022-0663.93.2.346

Robinson, J.T., Fischer, L., Wiley, D., & Hilton, J. (2014). The Impact of Open Textbooks on Secondary Science Learning Outcomes. *Educational Researcher 43*(7), 341-351.

Rosenbaum P.R. and Rubin D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*. *70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Rosenbaum P.R. (1989). "Optimal Matching for Observational Studies," *Journal of the American Statistical Association, 84*, 1024-1032.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, 127, 757–763.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*(3). https://doi.org/10.1214/08-aoas187

SSDD Problems. (2021). *Same Surface, Difference Deep Structure maths problems from Craig Barton @mrbarton maths.* https://ssddproblems.com/

Sainani, K. L. (2012). *Propensity Scores: Uses and Limitations*. Wiley Online Library. https://onlinelibrary.wiley.com/doi/full/10.1016/j.pmrj.2012.07.002

Salkind, N. (2010). Covariate. *Encyclopaedia of Research Design*.

Sammons, P., Sylva, K., Melhuish, E., Siraj, I., Taggart, B., Toth, K., & Smees, R. (2014). Influences on students' GCSE attainment and progress at age 16 Effective Pre-School, Primary & Secondary Education Project (EPPSE). Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/373286/RR352_-_Influences_on_Students_GCSE_Attainment_and_Progress_at_Age_16.pdf

Saretsky, G. (1975). The John Henry Effect: Potential Confounder of Experimental vs. Control Group Approaches to the Evaluation of Educational Innovations. *Education Resources Information Centre*.

Saritas, T. & Akdemir, O. (2009). Identifying Factors Affecting the Mathematics Achievement of Students for Better Instructional Design. http://www.itdl.org/journal/dec_09/article03.htm

Schoenfeld, A. H. (1992). Learning to Think Mathematically: Problem Solving, Metacognition, and Sense-Making in Mathematics. *Handbook for Research on Mathematics Teaching and Learning,* 334-370, New York: MacMillan.

Sedgwick, P., & Greenwood, N. (2015). Understanding the Hawthorne effect. *BMJ*, h4672. https://doi.org/10.1136/bmj.h4672

Sekaran, U. (2003). Research Methods for Business, 4th (Ed.). *John Wiley & Sons, Inc*. New York.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental design for generalized causal inference.* Boston: Houghton-Mifflin.

Shawa, L. (2017). Ethics in Educational Research. *Education Studies for Initial Teacher* Development, 432–443. Retrieved from https://www.researchgate.net/publication/312069857_Ethics_in_educational_research

Siu, M. K. (2004). Official curriculum in mathematics in ancient China: How did candidates study for the examination? *Chinese learn mathematics: Perspectives from insiders,* 157-188, Mahwah, NJ: World Scientific.

Skemp, R.R. (1976). Relational Understanding and Instrumental Understanding. *Mathematics Teaching, 77*, 20-26.

Solent Maths Hub. (2019). *Variation.* https://solentmathshub.org.uk/variation

Stanic, G. M. A., & Kilpatrick, J. (1988). Historical perspectives on problem solving in mathematics curriculum. *Research agenda for mathematics education: The teaching and assessing of mathematical problem solving,* 1-22.

Steinbring, H. (2005). The Construction of New Mathematical Knowledge in Classroom Interaction—An Epistemological Perspective. *Mathematics Education Library, 38*. Berlin, New York: Springer.

Stripp, C. (2014). *A Teacher's Guide to Using the Mastery Approach.* Retrieved August 05, 2020, from https://www.trueeducationpartnerships.com/schools/a-teachers-guide-to-using-the-mastery-approach-in-the-classroom/

Stuart, E. A. (2007). Estimating Causal Effects Using School-Level Data Sets. *Educational Researcher*, *36*(4), 187–198.

Stuart, E. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, *25*(1). doi: 10.1214/09-sts313

Sun, X. H. (2011). An insider's perspective: "Variation problems" and their cultural grounds in Chinese curriculum practice. *Journal of Mathematics Education, 4*(1).

Sun, X. H., Teresa B, N., & Loudes E, O. (2013). What different features of task design are associated with goals and pedagogies in Chinese and Portuguese textbooks: the case of addition and subtraction. *Task Design in Mathematics Education. Proceedings of ICMI Study, Oxford, 22*(6), 411-419.

Thoemmes, Felix & Kim, Eun. (2011). A Systematic Review of Propensity Score Methods in the Social Sciences. *Multivariate Behavioral Research*. 46. pp. 90-118. 10.1080/00273171.2011.540475.

UK Health Security Agency. (2020). *Interrupted time series study.* Retrieved 2 January 2022, from https://www.gov.uk/guidance/interrupted-time-series-study

Vander Linde, L. F. (1964). Does the study of quantitative vocabulary improve problem solving? *Elementary School Journal*, *65*, 143-15.

Vikstrom, A. (2014). What makes the difference? Teachers explore what must be taught and what must be learned in order to understand the particular character of matter. *Journal of Science Teacher Education.* doi:10.1007/ s10972-014-9397-9.

Wang, J. (2007). A trend study of self-concept and mathematics achievement in a cross-cultural context. *Mathematics Education Research Journal, 19*(3), 33–4.

Wang, Y. and Barmby, P. and Bolden, D. (2017). 'Understanding linear function: a comparison of selected textbooks from England and Shanghai.', *International Journal of Science and Mathematics Education*. *15*(1), 131-153.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology,* 12, 129-140.

Watson, A., Geest, E. D., & Prestage, S. (2003). Deep Progress in Mathematics - The Improving Attainment in Mathematics Project. https://www.researchgate.net/publication/46166585_Deep_Progress_in_Mathematics_-_The_Improving_Attainment_in_Mathematics_Project/citations

Watson, A., & Mason, J. (2006). Variation and Mathematical Structure. *Mathematics Teaching Incorporating Micromath*, *194*, 3–5. http://www.pmtheta.com/uploads/4/7/7/8/47787337/variation_2007.pdf

Weidmann, B., & Miratrix, L. (2020). Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *Journal of Policy Analysis and Management*, *40*(3), 964–986. https://doi.org/10.1002/pam.22236

White Rose Maths (2018). *Mathematical Misconceptions.* https://whiterosemaths.com/latest-news/mathematical-misconceptions/

White Rose Maths (2019). *Conceptual vs Fluency.* https://whiterosemaths.com/latest-news/conceptual-vs-fluency/

White Rose Maths (2020). *Order, Order! The Importance of Sequencing*. https://whiterosemaths.com/latest-news/order-order-the-importance-of-sequencing/

White Rose Maths (2020). *Secondary SOL | White Rose Maths | FREE Maths Teaching Resources*. https://whiterosemaths.com/resources/secondary-resources/secondary-sols/

White Rose Maths (2021). *About White Rose Maths | White Rose Maths | Confidence in Maths*. (2021, March 15). White Rose Maths. https://whiterosemaths.com/who-we-are/about-white-rose-maths/

White Rose Maths (2021). *Thinking Through Variation* – White Rose Maths. https://resources.whiterosemaths.com/courses/thinking-through-variation/

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology*, *18*, 327–350.

Yin, R. K. (2003). Case study research, design and methods (3rd ed.). California, USA: SAGE Publications.

Zazkis, R. (2000). Using code-switching as a tool for language mathematical language. *For the Learning of Mathematics*, *20(*3*),* 38-43.

Zanutto, E. (2021). A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data. *Journal Of Data Science*, *4*(1), 67-91. doi: 10.6339/jds.2006.04(1).233

Appendices

Appendix A

Dear XXXX,

I hope you don't mind me contacting you and apologies for the lengthy email, but I have a really exciting opportunity that you may be interested in.

This academic year, I embarked upon a Masters' Research Degree as I'm really interested in exploring the impact of Teaching for Mastery in Maths. So many schools have worked so hard implementing TfM over the last few years, and I know that from speaking to many teachers, there have been so many benefits to the approach such as student motivation and engagement in lessons. I'm now really interested in the quantitative impact of TfM in terms of student achievement through data analysis and have decided to pursue this at Masters' level so that I can conduct an in-depth analysis.

I was hoping you might be interested in being involved in my research as I know you've being a driving force in the implementation of TfM at your school.

The first part of my research will consist of interviews where I try to gleam your/the school's motivation for implementing Teaching for Mastery and how you've rolled it out across the Maths department. I'd be interested in finding out how you define TfM and what they key changes have been to your curriculum during and since the implementation. Depending on Covid restrictions, these interviews could be done via Zoom or in person and I am currently aiming to have these done by the end of this academic year. As well as speaking to yourself, it would be great to interview some other Maths teachers at your school so I can get a well-rounded overview of how TfM looks at your school, which will feed into my research and thesis.

The main part of the research will then take place next academic year where I would need to get hold of as much data as possible from you. I'm currently planning an observational style methodology where I examine the data that you have available pre- and post- TfM implementation. I'm looking to compare data of cohorts (such as test results) that have being taught using TfM approaches with cohorts that haven't. For example, if you have rolled out TfM to years 7, 8, and 9 so far, then I'd want to compare their data with data of your current year 10 cohort (who haven't had TfM) when they were in years 7, 8 or 9. Therefore, I do require up to date data as well as data from a few years ago before TfM was implemented.

To ensure that the comparison of cohorts is as valid as possible, I will be using a matching method (such as propensity score matching) to ensure that the cohorts I compare are statistically similar and therefore matched on important covariates such as: gender, pupil premium status, prior attainment scores, and SEND status to name a few. Therefore, I will need this information alongside student test scores. For ethical considerations, I will ask you to anonymise all student names.

At the minute, I'm just putting feelers out to schools in the Maths Hub that I know have adopted Teaching for Mastery and it would be great to know if this sounds like something that you might be interested in being involved in? I'll try to make it as little work for you as possible – the main job would be anonymising the data and collating it from your current tracking systems and sending it to me.

Appendix B

Introduction/Purpose of Interviews and Research Project

1. Project: to see if quantitative observational methods can attribute changes in student outcomes to the transition to Teaching for Mastery
2. Interview purpose: to gain contextual insight into the school and your department, and trying to understand your transition to Teaching for Mastery and how that looks in your school
3. Semi-structured interview so that the interviewee can determine the direction the interview goes for each question

Areas of Interviews:

1. What is your understanding of Teaching for Mastery? What would be the defining features of the approach to you?
2. Tell me about your school's transition to Teaching for Mastery – how, why and when did you implement it?
3. What made you transition to Teaching for Mastery?
4. Have you used any particular mastery programmes, or have you developed your own resources?
5. What makes the resources 'mastery'? How are they different to 'non-mastery'?
6. What would you say have been the biggest changes to your curriculum? What did your previous curriculum look like?
7. What would you say have been the biggest changes to the departmental pedagogy since the change? How did this differ before the transition?
8. Do you have your mastery schemes of work available for me to view? If you have pre-mastery schemes of work, it'd be really useful to compare and contrast them.

Quantitative Methods

1. My main methodology is going to be quantitatively seeking to compare student performance pre- and post- Teaching for Mastery and therefore, I will need a consistent outcome variable. Do you have assessment data available for students before the curriculum change and since the curriculum change, where tests are comparable, such as year 10 exams or year 11 mocks?
2. I will also need information on confounding variables and demographics. Will you have data available on student gender, prior attainment level (KS2 score), pupil premium status, SEND status, EAL status.
3. 'Perfect confounds' are an issue I will need to overcome to be able to attribute any changes in assessment scores to the transition to TfM. Are you able to give any insight into events that happened in the school at the same time as the curriculum change which can't be disaggregated, such as change of HOD or school ethos?

Appendix C

A sample of the code used:

```
> library(readr)
> WSchoolDataforRV2HFINAL <- read_csv("~/Masters/Weydon School Data/WSchoolDataforRV2HFINAL.csv")
> View(WSchoolDataforRV2HFINAL)
> View(WSchoolDataforR)
> library(MatchIt)
> library(dplyr)
> library(ggplot2)
> library(readr)
> #Pre-analysis using unmatched data: difference-in-means on outcome measure
> WSchoolDataforRV2HFINAL %>%
+   group_by(TFM) %>%
+   summarise(n_students = n(),
+             mean_math = mean(OutcomeH),
+             std_error = sd(OutcomeH) / sqrt(n_students))
> #Generate standardised test scores for both the TfM and non-TfM group (z-scores)
> WSchoolDataforRV2HFINAL %>%
+ mutate(test = (OutcomeH - mean(OutcomeH)) / sd(OutcomeH)) %>%
+ group_by(TFM) %>%

 + summarise(mean_math = mean(test))
 > #Check if the difference in means is statistically significant using an unpaired t-test
 > with(WSchoolDataforRV2HFINAL,t.test(OutcomeH,TFM,paired=F,conf.level=0.95))
 > #Difference-in-means: pre-treatment covariates
 > WSchoolDataforRV2HFINAL_cov <- c('Prior_Att','Gender_0Female_1Male','Pupil_Premium','SEND','Assessment_1')
 > WSchoolDataforRV2HFINAL %>%
 + group_by(TFM) %>%
 + select(one_of(WSchoolDataforRV2HFINAL_cov)) %>%
 + summarise_all(funs(mean(., na.rm = T)))
 > #Evaluate if the means are statistically distinguishable using t-test
 > with(WSchoolDataforRV2HFINAL, t.test(Prior_Att ~ TFM))
 > with(WSchoolDataforRV2HFINAL, t.test(Gender_0Female_1Male ~ TFM))
 > with(WSchoolDataforRV2HFINAL, t.test(Pupil_Premium ~ TFM))
 > with(WSchoolDataforRV2HFINAL, t.test(SEND ~ TFM))
 > with(WSchoolDataforRV2HFINAL, t.test(Assessment_1 ~ TFM))
```

```
> #Propensity Score Estimation
> m_ps <- glm(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
+ family = binomial(), data = WSchoolDataforRV2HFINAL)
> summary(m_ps)
> #Calculate propensity score for each student
> prs_df <- data.frame(pr_score = predict(m_ps, type = "response"), TFM = m_ps$model$TFM)
> head(prs_df)
> #Examine the region of common support using ggplot (histogram)
> labs <- paste("Curriculum Type:", c("TFM", "Other"))
> prs_df %>%
+ mutate(TFM = ifelse(TFM == 1, labs[1], labs[2])) %>%
+ ggplot(aes(x = pr_score)) +
+ geom_histogram(color = "white", bins=30) +
+ facet_wrap(~TFM) +
+ xlab("Probability of being taught using TFM") +
+ theme_bw()
> install.packages("rgenoud")
> install.packages("Matching")
> #Execute a Matching Algorithm
> WSchoolDataforRV2HFINAL_nomiss <- WSchoolDataforRV2HFINAL %>% # MatchIt does not allow missing values
> result_1 <- matchit(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
WSchoolDataforRV2HFINAL_nomiss, method = "nearest", distance = "glm", caliper = 0.1)
> #Summarise matching method used
> result_1
> summary(result_1)
> #Store matched data as a data frame
> dta_m <- match.data(result_1)
> dim(dta_m)
> plot(result_1, type = 'jitter', interactive = FALSE)
> library(cobalt)
> bal.plot(result_1, "Prior_Att", which = "both")
> plot(result_1, type = "qq", interactive = FALSE, which.xs = c("Prior_Att"))
> bal.plot(result_1, "Gender_0Female_1Male", which = "both")
> plot(result_1, type = "qq", interactive = FALSE, which.xs = c("Gender_0Female_1Male"))
> bal.plot(result_1, "Pupil_Premium", which = "both", type = "ecdf")
> plot(result_1, type = "qq", interactive = FALSE, which.xs = c("Pupil_Premium"))
> bal.plot(result_1, "SEND", which = "both", mirror = TRUE, type = "histogram", colors = c("white", "black"))
```

```
> plot(result_1, type = "qq", interactive = FALSE, which.xs = c("SEND"))

> bal.plot(result_1, "Assessment_1", which = "both")

> plot(result_1, type = "qq", interactive = FALSE, which.xs = c("Assessment_1"))

> #Examine covariate balance in matched sample (by t-tests of difference in means)

> dta_m %>%

+ group_by(TFM) %>%

+ select(one_of(WSchoolDataforRV2HFINAL_cov)) %>%

+ summarise_all(funs(mean))

> with(dta_m, t.test(Prior_Att ~ TFM))

> with(dta_m, t.test(Gender_0Female_1Male ~ TFM))

> with(dta_m, t.test(Pupil_Premium ~ TFM))

> with(dta_m, t.test(SEND ~ TFM))

> result_2 <- matchit(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
WSchoolDataforRV2HFINAL_nomiss, method = "nearest", distance = "glm", caliper = 0.05)

> result_2

> summary(result_2)

> result_3 <- matchit(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
WSchoolDataforRV2HFINAL_nomiss, method = "nearest", distance = "glm", caliper = 0.2)


> result_3

> summary(result_3)

> result_4 <- matchit(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
WSchoolDataforRV2HFINAL_nomiss, method = "nearest", distance = "glm", caliper = 0.3)

> result_4

> summary(result_4)

> result_5 <- matchit(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
WSchoolDataforRV2HFINAL_nomiss, method = "nearest", distance = "glm")

> result_5

> summary(result_5)

#Back to result_3 to estimate treatment effect

> #Store matched data as a data frame

> dta_m3 <- match.data(result_3)

> dim(dta_m3)

> plot(result_3, type = 'jitter', interactive = FALSE)

> library(cobalt)

> bal.plot(result_3, "Prior_Att", which = "both")

> plot(result_3, type = "qq", interactive = FALSE, which.xs = c("Prior_Att"))

> bal.plot(result_3, "Gender_0Female_1Male", which = "both")
```

```
> result_3
> plot(result_3, type = "qq", interactive = FALSE, which.xs = c("Gender_0Female_1Male"))
> bal.plot(result_3, "Pupil_Premium", which = "both", type = "ecdf")
> plot(result_3, type = "qq", interactive = FALSE, which.xs = c("Pupil_Premium"))
> bal.plot(result_3, "SEND", which = "both", mirror = TRUE, type = "histogram", colors = c("white", "black"))
> plot(result_3, type = "qq", interactive = FALSE, which.xs = c("SEND"))
> bal.plot(result_3, "Assessment_1", which = "both")
> plot(result_3, type = "qq", interactive = FALSE, which.xs = c("Assessment_1"))
> #Examine covariate balance in matched sample (by t-tests of difference in means)
> dta_m3 %>%
+ group_by(TFM) %>%
+ select(one_of(WSchoolDataforRV2HFINAL_cov)) %>%
+ summarise_all(funs(mean))
> with(dta_m3, t.test(Prior_Att ~ TFM))
> with(dta_m3, t.test(Gender_0Female_1Male ~ TFM))
> with(dta_m3, t.test(Pupil_Premium ~ TFM))
> #Estimate Treatment Effects
> with(dta_m3, t.test(OutcomeH ~ TFM,paired=T,conf.level=0.95))
#Sensitivity Check using Result_1
> #Estimate Treatment Effects
> with(dta_m, t.test(OutcomeH ~ TFM,paired=T,conf.level=0.95))
#Sensitivity Check using genetic matching
> result_5 <- matchit(TFM ~ Prior_Att + Gender_0Female_1Male + Pupil_Premium + SEND + Assessment_1,
+ method = "genetic", data = WSchoolDataforRV2HFINAL_nomiss)
> result_5
```