Maynooth University
National University of Ireland Maynooth

Extensions of Bayesian Non-Parametric Causal Inference Machine Learning Methods with Applications to Large Scale Educational Studies

A thesis submitted in fulfillment of the requirements for the Ph.D. degree in Statistics

By:

Nathan McJames

Under the supervision of:

Professor Andrew Parnell

Professor Ann O'Shea

Hamilton Institute
Maynooth University
Maynooth, Co. Kildare, Ireland

June 13, 2025

Declaration

I hereby declare that I have produced this manuscript without the prohibited assistance of any third parties and without making use of aids other than those specified.

The thesis work was conducted from September 2020 to September 2024 under the supervision of Professor Andrew Parnell and Professor Ann O'Shea in the Hamilton Institute, Maynooth University.

In particular, chapters 3 to 6 of this thesis are based on research conducted collaboratively with my supervisors and co-authors, and parts of the text reflect this through the use of collective terms such as "we" and "our". Where such language appears, it is intended to acknowledge the collaborative nature of the research, or to include the reader in the discussion and development of the work.

Nathan McJames,

Maynooth, Ireland,

June, 2025.

Sponsor

This work was supported by a Science Foundation Ireland grant number 18/CRT/6049.

Collaborations

Andrew Parnell: As my supervisor, Professor Parnell (Maynooth University) supervised and collaborated on the work of all chapters. This includes reviewing and editing all chapters.

Ann O'Shea: As my supervisor, Professor O'Shea (Maynooth University) supervised and collaborated on the work of all chapters. This includes reviewing and editing all chapters.

Yong Chen Goh: Yong Chen Goh collaborated on the work of Chapter 4 by assisting with the derivation of the log-likelihood formulas used in the multivariate Bayesian Causal Forests extension. She also suggested a number of helpful R packages, which were useful in the data analysis stage.

Publications

Chapters 3, 4, 5, and 6 of this thesis have been published in, or submitted to, peer-reviewed journals. Chapter 3 has been published in *Educational Review*, Chapter 4 has been published in the *Journal of the Royal Statistical Society Series A: Statistics in Society*, and Chapter 5 has been published in *Learning and Instruction*. Chapter 6 has recently been submitted.

Peer-reviewed journal articles:

- McJames N., Parnell A., & O'Shea A. (2023). Factors affecting teacher job satisfaction: A causal inference machine learning approach using data from TALIS 2018. *Educational Review*, 1–25. https://doi.org/10.1080/00131911.2023.2200594.
- McJames, N., O'Shea, A., Goh, Y. C. & Parnell, A. (2024). Bayesian causal forests for multivariate outcomes: Application to Irish data from an international large scale education assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 1-23. https://doi.org/10.1093/jrsssa/qnae049.
- McJames N., Parnell A., & O'Shea A. (2024). Little and often: Causal inference machine learning demonstrates the benefits of homework for improving achievement in mathematics and science. *Learning and Instruction*, 1–11. https://doi.org/10.1016/j.learninstruc.2024.101968.

Currently submitted:

- McJames, N., O'Shea, A., & Parnell, A. (2024). Bayesian causal forests for longitudinal data: Assessing the impact of part-time work on growth in high school mathematics achievement. *arXiv pre-print*. https://arxiv.org/abs/2407.11927.

List of Acronyms

ATE Average Treatment Effect

BART Bayesian Additive Regression Trees

BCF Bayesian Causal Forests

CART Classification and Regression Tree

CI Confidence Interval/Credible Interval

CPD Continual Professional Development

DGP Data Generating Process

DiD Difference-in-Differences

GRF Generalised Random Forests

HSLS High School Longitudinal Study

ICATE Individual Conditional Average Treatment Effect

ICE Individual Conditional Expectation

IEA International Association for the Evaluation of Educational Achievement

ILSA International Large Scale Assessment

IPW Inverse Probability Weighting

IRT Item Response Theory

LTMLE Longitudinal Targeted Maximum Likelihood Estimation

MCMC Markov Chain Monte Carlo

MVN Multivariate Normal

NCES National Center for Education Statistics

OECD Organisation for Economic Cooperation and Development

PEHE Precision Estimating Heterogeneous Effects

RCT Randomised Controlled Trial

RMSE Root Mean Squared Error

SES Socioeconomic Status

STEM Science Technology Engineering and Mathematics

SUTVA Stable Unit Treatment Value Assumption

TALIS Teaching and Learning International Survey

TIMSS Trends in International Mathematics and Science Study

Acknowledgements

I want to start by thanking my supervisors, Professor Andrew Parnell and Professor Ann O'Shea, for everything you have done over the past four years. I have had the greatest PhD experience I could have asked for. Thank you for the many ideas, suggestions, comments, and meetings, without which I would have been completely lost. I'm also grateful for the many conferences and workshops you found for me to attend, and for letting me know about the different tutoring and lecturing opportunities that I enjoyed and learned so much from. Your kindness and encouragement has meant a lot. I will miss our regular meetings. I hope I have been able to repay even just a fraction of the trust and confidence that you both placed in me four years ago.

I also want to thank Professor James Gleeson, Professor Claire Gormley, Professor Ken Duffy, Professor David Malone, Janet Clifford, Patsy Finn, and all involved in creating and managing the CRT. I have benefitted and learned so much from the many events and training activities I have attended as part of this programme. For organising these CRT events and always being on hand to help with bookings, travel plans, and everything else, a special thank you goes to Joanna O'Grady, Rosemary Hunt, and Kate Moriarty.

Thank you Professor Ioanna Manolopoulou and all of the Department of Statistical Science at UCL for allowing me to visit and being so welcoming during my research visit last summer. I had a great time in London and learned so much from hearing about the research projects you are working on.

Thank you to all of my friends, especially my many new friends from the Hamilton Institute that I have worked alongside for the past four years. I want to single out my office mates Akash, Amit, Bill, Chang, Darshana, Fergal, Jonny, Pramit,

Shauna, and Tzirath for making all of the CRT events so fun and memorable. I couldn't have asked for better friends to share this journey with.

Claire, I will never be able to express how much your support and encouragement has meant to me over the past four years. Thank you for everything.

Finally, Mum, Dad, and Josh, I never could have started let alone finished any of this without you. Mum and Dad, I hope you know how grateful I am for all of the opportunities you've worked so hard to give me.

Summary

When exploring how a unique individual's characteristics can lead to variations in their response to treatment, Bayesian non-parametric causal inference machine learning methods based on Bayesian Additive Regression Trees (BART) and Bayesian Causal Forests (BCF) have emerged as leading approaches. This thesis presents a series of studies focused on extending and applying these methods to large scale educational studies.

We begin by demonstrating the broad potential for these methods in educational studies by applying BART to English data from the Teaching and Learning International Survey (TALIS 2018). By estimating the effect of multiple treatments on teacher job satisfaction, we identify positive factors such as continual professional development and induction activities that may be used to improve job satisfaction, thus encouraging teachers to stay in their jobs and new entrants to join the profession.

Our second contribution is a multivariate extension of Bayesian Causal Forests, designed to estimate the effect of an intervention on multiple outcome variables simultaneously. By allowing the tree structure of BCF to benefit from the shared information across all outcome variables, we demonstrate the performance gains made possible with this approach. Applying this method to Irish data from the Trends in International Mathematics and Science Study (TIMSS 2019), we also investigate the effect of a number of home-related factors on student achievement such as having access to a study desk at home, often being absent, or often feeling hungry when arriving at school.

Later, we augment this multivariate model in order to investigate the separate effects of homework frequency and homework duration on student achievement in

mathematics and science, again using data from TIMSS 2019. We find that while increasing homework frequency can lead to greater homework benefits, increasing homework duration beyond 15 minutes has no additional effect.

Our final contribution is a longitudinal extension of BCF, designed to estimate treatment effects from multiple waves of data, using a structure similar to that of the difference-in-differences approach. With the help of simulation studies, we demonstrate the performance gains made possible with our new method. Applying this model to data from the High School Longitudinal Study of 2009 (HSLS), we also reveal the negative effects of participation in intensive part-time work by high school students.

Contents

List of Figures

List of Tables

1 Introduction

This thesis will focus on developing and applying new extensions of Bayesian non-parametric causal inference machine learning methods to large-scale educational studies. Our aim is to broaden the applicability of these methods, improve their performance, and allow them to be applied in situations that they would otherwise not be well suited to. In doing so, we also aim to tackle several important research questions from the world of education, which are of importance because of their implications for education policy both in Ireland and internationally. In this first chapter, we provide an introduction to the datasets we will work with, the challenges they pose, and some of the existing methods that have been used to study them. Finally, we outline the key contributions made within each chapter of this thesis.

1.1 Large Scale Educational Studies

1.1.1 Overview of Datasets Used in This Thesis

Large scale educational studies is a broad term used to describe a number of research initiatives, often designed to measure and compare the outcomes and features of educational systems across different countries (Rutkowski et al., 2010). These studies provide valuable information about educational systems, learning outcomes, and factors that influence student performance. They are organised by large national or international cooperative institutions, and typically involve

representative samples of students or teachers from participating countries in the grade level of interest. Often repeated at regular intervals, they offer up-to-date information and allow for the analysis of trends in national achievement outcomes over time. Due to their strengths, datasets from large scale educational studies have become an increasingly popular area of research, and have been used in a diverse range of applications (Hernández-Torrano and Courtney, 2021).

The first dataset that we will work with in this thesis is called the Teaching and Learning International Survey (TALIS, OECD, 2019a). Unlike most educational studies which focus on students, TALIS is the world's largest international study of teachers and school principals. It is widely regarded as a highly important educational study because of the very valuable insights it offers into the working and learning environments of school systems in participating countries. Organised by the Organisation for Economic Cooperation and Development (OECD), TALIS first took place in 2008. Further cycles were conducted in 2013 and 2018, with work on a fourth cycle currently in progress. TALIS 2018 is the largest of the studies to date and took place in 48 countries. The focus is on teachers of lower secondary school students, but participating countries also have the option of involving primary school teachers and upper secondary school teachers.

Ireland did not take part in TALIS 2018, so our analysis in Chapter 3 is based on the English subset of the data. This subset of the data contains 2009 primary and 2376 lower secondary school teachers. Principals of 162 primary and 157 lower secondary schools are also included. As part of the study, teachers are asked to complete a questionnaire on a wide variety of topics such as personal background, current teaching duties, their perception of the school climate, and job satisfaction. Principals, meanwhile, are given a questionnaire on aspects related to the school characteristics, leadership, staffing, and policies that are in place in the school. Being the only large scale international dataset involving representative samples of teachers, TALIS is uniquely placed as a resource for researchers investigating factors related to school working and learning environments.

One of the longest running large scale assessments in education is the Trends in International Mathematics and Science Study (TIMSS, Broer et al., 2019). TIMSS first took place in 1995 and has taken place every four years since then, with TIMSS 2019 being the most recent study for which data is publicly available. Data from

TIMSS 2023 is expected to be released at the end of 2024. It is one of a number of studies regularly conducted by the International Association for the Evaluation of Educational Achievement (IEA). TIMSS 2019 took place in 64 different countries making it the largest of the TIMSS studies conducted to date (Mullis et al., 2020).

As part of the study, students from the fourth and eighth grades of participating countries (aged approximately 10.5 and 14.5 years on average) are given a short assessment in mathematics and science to measure their achievement in these subjects. Surveys are also given to the students (or their parents in the case of the fourth grade students), their teachers, and their school principals. Specifically, the Irish eighth grade version of data from TIMSS 2019 that we will explore later in Chapters 4 and 5 involved a representative sample of 4118 students, and more than 500 teachers from almost 150 schools. These surveys collect important contextual information such as the socioeconomic backgrounds of the students, the education level of their parents, teaching practices within the classrooms, and the availability of school resources. This makes TIMSS an excellent source of information for researchers investigating factors associated with student achievement in, or attitudes towards studying mathematics and science.

The final dataset explored in this thesis is the High School Longitudinal Study of 2009 (HSLS, Ingels et al., 2011). HSLS is the most recent of a series of five longitudinal studies organised by the National Center for Education Statistics (NCES). An important feature of HSLS is that unlike many educational studies which employ a cross sectional design, it uses a longitudinal design which follows the same cohort of students over time. This allows for an analysis of changes in student achievement, or indeed other important outcomes over time.

The first wave of data collection for HSLS took place in the fall of 2009 when the target population were in the ninth grade of high school. More than 20,000 students from nearly 1000 high schools took part and were given a short mathematics assessment designed to measure their algebraic reasoning abilities. Similarly to TIMSS 2019, questionnaires were also completed by the students, their parents, their teachers, and school administrators. These questionnaires facilitated the collection of important background information such as socioeconomic status, sense of belonging at school, parental education, school type, and location. A follow up of these students then took place in the spring of 2012 when the students were in

the eleventh grade. The mathematics ability of the students was re-assessed and updated questionnaires were completed to track any changes in student, family, or school circumstances. Further follow ups have also taken place, and HSLS is in fact an ongoing longitudinal study. However, Waves 1 and 2 are the only waves to administer standardised mathematics assessments, so Chapter 6 will focus on these waves only.

1.1.2 Critical Perspectives on Large Scale Educational Studies

Before continuing, we should acknowledge that despite the many strengths of the datasets discussed above, they have also occasionally been the subject of criticism, and their results sometimes brought into question. Concerns have been raised in the past, for example, regarding the potential of countries, motivated by a desire to achieve strong standings in international league tables, to "fudge" sampling designs (Berliner, 2011), while others fear that the low-response rates within some countries could limit the generalisability of what should be fully representative samples (Eivers, 2010). Language translation challenges (Upsing and Hayatli, 2021) and cultural differences (He et al., 2022) also present difficulties. Lastly, aside from the potential for technical challenges, some researchers also fear that competitive international comparisons can lead to a detrimental homogenisation of education systems, and stifle innovation (Zhao, 2020).

One of the key criticisms faced by the organisers of large scale education studies such as the OECD, is that they can create what is viewed as a narrow, one size fits all view of education (Volante et al., 2017). Their prioritisation of standardised test scores and cross country comparisons (Niemann et al., 2017) can shift focus to international benchmarks and away from local region specific needs as may be the case in developing countries (Murphy, 2014). Due to the established nature of large scale assessments in many countries, organisations such as the OECD have developed a strong position in relation to educational policy (Rutkowski, 2007; Sellar and Lingard, 2014). Some fear this position of authority held by such organisations can prompt individual countries to adopt policies based on improving standings in international rankings, rather than what may be best for their own students (Zhao, 2020).

Another common area of criticism in relation to large scale educational studies is the standards and values promoted by these assessments (Engel et al., 2019), which may not be reflective of or helpful for advancing the needs of all participating countries. Notably, studies such as TIMSS and related studies such as PISA or even national versions such as HSLS predominantly focus on what many consider to be a narrow view of what counts as educational success, such as high test scores in STEM oriented subjects, considered essential for the modern economy and skills for the 21st century (Delahunty, 2024). Consequently, some fear there is a risk that education systems from across the world can be increasingly shaped by a single vision of achievement that may not be appropriate for local contexts (d'Agnese, 2015).

Despite these concerns and criticisms, however, the data collected from large scale educational studies such as TALIS, TIMSS, or HSLS can certainly offer significant value. These datasets can help identify gaps in and barriers to equity in areas like teacher job satisfaction and student achievement. For example, the findings from Chapter 3, which explore factors influencing teacher job satisfaction, need not be seen solely as a means to ensure adequate staffing. Instead, they can be understood in light of evidence showing that more satisfied teachers contribute to more motivated and happier students as well (Toropova et al., 2021), highlighting broader benefits. Similarly, Chapter 4's findings, which point to the negative impact of arriving at school feeling hungry on student achievement, can be viewed through the wider lens of advocating for free school meal programmes, not only to boost academic outcomes but also to promote student well-being (McKelvie-Sebileau et al., 2023). In this way, large scale educational studies can inform meaningful educational improvements that extend beyond narrow, standardised metrics, and this is an aim this thesis also seeks to support.

1.2 Challenges Posed by Data From Large Scale Educational Studies

There are some specific challenges that are met when analysing data from large scale educational studies that are worth discussing briefly. The first is that ow-

ing to the nature of large scale educational data, which often involves students nested within classrooms and schools, there is a clear hierarchical aspect to the data. This is a feature that is true of all of the datasets explored in this thesis. Accounting for this multi-level structure of the data is therefore very important in order to correctly capture the uncertainty in model estimates created by the sampling design of the studies, and to adjust for between group differences. In the case of the TALIS data in Chapter 3, the recommended strategy by the study organisers is to use Fay's Balanced Repeated Replication method, for which replication weights are provided in the dataset (OECD, 2019b). For the TIMSS data in Chapters 4 and 5, meanwhile, the models used are equipped with a random intercept at the classroom level. However, due to data confidentiality concerns, the public use version of the HSLS data used in Chapter 6 does not contain school or classroom identifiers, precluding a hierarchical modelling strategy. This limitation is discussed in more detail in Chapter 6.

Large scale educational studies often employ complex sampling designs to ensure that their samples are representative of the target populations. These designs can involve methods such as stratification and probability sampling, often based on features such as school size. As a result, sampling weights are required in order to adjust for the varying probabilities of selection. TIMSS 2019, for example, adopted a stratified two-stage cluster sampling design. In the first stage, schools were sampled with a probability proportional to their size. In the second stage, classes were sampled from the participating schools. To address the issue of some selected schools declining to participate, non-participation adjustments were also applied to the sampling weights. Accounting for these adjustments and sampling weights is necessary to ensure that the parameter estimates resulting from an analysis are reflective of the entire population.

One challenge specific to working with large scale education data is the use of plausible values to estimate student achievement. Studies such as TIMSS and HSLS typically employ Item Response Theory (IRT, Cai et al., 2016) to create estimates of student achievement based on test answers. This process is complex, as it aims to accurately measure student achievement using relatively short tests, which are necessary to minimise the burden on participating schools and students. Further complicating matters is the fact that not all students answer the same

questions. TIMSS, for example, used a rotated booklet design, while HSLS employed an adaptive computerised assessment. The result is a degree of uncertainty in the achievement estimates of students. To address this uncertainty, it is common for studies like TIMSS and HSLS to report five plausible values for each student's achievement rather than a single point estimate. These values are random draws from the posterior distribution of the student's achievement estimate, reflecting the uncertainty that is present (Wu, 2005; Khorramdel et al., 2020). Plausible values are treated analogously to multiple imputations in a statistical analysis. In the Bayesian framework, this involves running separate chains for each of the five plausible values and then pooling the results together after burn-in. This ensures that the uncertainty in parameter estimates is not understated, and was performed appropriately in Chapters 4, 5 and 6.

Finally, an important challenge that is encountered when working with all observational studies is the ever present issue of confounding. Confounding variables have the potential to create an illusion of causality, resulting in misleading or incorrect findings. This challenge is perhaps especially relevant in the field of education, where policy decisions based on secondary analyses of large scale education datasets have the potential to impact tens of thousands of students, teachers, and entire education systems, often requiring many millions of euros in funding to implement. This challenge, which provides a motivation for the methods extended and employed throughout this thesis, is what we will discuss next.

1.3 Causal Inference and Educational Studies

Causal inference is the field of study concerned with cause and effect. Within this domain, practitioners are often interested in investigating questions of the form "Does A cause B?", and if so, "What effect does A have on B"? Answering these questions can be a seemingly simple but deceptively difficult task. A common "cause" of this difficulty is the presence of confounding variables. A confounding variable is one which simultaneously influences both the likelihood of receiving treatment, and the outcome variable of interest (Greenland et al., 1999). Students from advantaged socioeconomic backgrounds, for example, may be more likely to wear designer shoes. These same students may also have higher achievement on

average due to increased exposure to beneficial factors such as increased physical and human educational resources at home. The confounding variable, socioeconomic status, creates an illusion of causality by making it appear as if there is a direct relationship between designer shoes and student achievement when in fact this may not be the case.

The best way to avoid the unwanted effects of confounding variables is to perform an experiment via a Randomised Controlled Trial (RCT). By randomly assigning students to control and treatment groups, RCTs break the link between the confounding variables and the decision to receive treatment. In this ideal scenario, the causal effect τ of an intervention can be estimated by simply looking at the difference in the average outcomes of the control and treatment groups. Using y_i to denote the realised outcome associated with observation i, and Z_i to indicate if observation i was assigned to a control or treatment group ($Z_i = 1$ for treatment, $Z_i = 0$ for control), of which we will say there are n_C and n_T members respectively, an estimate for the average effect of treatment Z on y can be obtained with

$$\hat{\tau}_{RCT} = \sum_{Z_i=1} \frac{1}{n_T} y_i - \sum_{Z_i=0} \frac{1}{n_C} y_i.$$

However, RCTs are often not feasible due to factors such as cost, complexity, or ethical considerations (West et al., 2008). Fortunately, several causal methods have been developed for working with observational data.

Many of these approaches are founded on the Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Sekhon, 2008), which postulates that for every observation i, there are two potential outcomes that may be observed: One that would be observed under treatment, $y_i(Z_i = 1)$, and one that would be observed under control, $y_i(Z_i = 0)$. The challenge lies in the fact that we only ever observe one of these potential outcomes. As a result, direct calculation of the treatment effects for any observation, $\tau_i = y_i(Z_i = 1) - y_i(Z_i = 0)$, is not possible. This is known as the fundamental problem of causal inference. Many causal inference methods therefore rely on finding clever ways to determine what the unobserved potential outcomes should be, even though we never observe them directly. For these approaches to work, three key assumptions must be met (Angrist et al., 1996; Kurz, 2022):

Assumption 1: The Stable Unit Treatment Value Assumption. This assumption requires that the potential outcomes of any observation i must be independent of the treatment status Z_j of any other observation j. This assumption also requires that there should be "no multiple versions" of the same treatment. This ensures there are not multiple potential outcomes corresponding to different versions or types of the same treatment.

Assumption 2: The Ignorability Assumption. The potential outcomes of observation i must be independent of whether or not observation i received treatment. In other words, there must be no residual confounding that we can not control for with the available covariates x_i. In notation, this means that we assume $y_i(Z_i = 0), y_i(Z_i = 1) \perp Z_i|x_i$.

Assumption 3: The Overlap Assumption. For every observed set of covariates, there must be a non-zero probability of receiving, or not receiving treatment: $0 < P(Z_i = 1|x_i) < 1$.

Two popular methods that rely on these assumptions involve the use of propensity scores (Pan and Bai, 2018). The propensity score π_i of student i represents the probability of them receiving a treatment Z conditional on their characteristics x_i. Using $Z_i = 1$ to indicate that student i did receive treatment, and $Z_i = 0$ to indicate that student i did not receive treatment, the propensity score π_i is defined as $\pi_i = P(Z_i = 1|x_i)$. The propensity score, by capturing important information related to the non-random selection of individuals into treatment, can be used to recover treatment effects in two different ways. The first, known as inverse probability weighting (Kurz, 2022), replaces the simple difference in mean outcomes of the control and treatment groups with the following:

$$\hat{\tau}_{IPW} = \sum_{Z_i=1} \frac{1}{n_T} \frac{y_i}{\hat{\pi}_i} - \sum_{Z_i=0} \frac{1}{n_C} \frac{y_i}{1 - \hat{\pi}_i}.$$

The result is that the treated units who were unlikely to receive treatment, and

the untreated units who were likely to receive treatment are up-weighted. This accounts for the non-random selection mechanisms which may be present owing to the confounding variables, and leads to a more "balanced" estimate of the mean outcomes under both treatment conditions. Henderson (2019), for example, used this approach to investigate the effect of "dual enrolment" courses on the high school achievement of students in the US. Meanwhile, McNealis et al. (2024) and Bowman et al. (2023) have used similar approaches to investigate the effects of supplemental instruction and maternal education on academic outcomes of students.

A second strategy involving the use of propensity scores is called propensity score matching (Stuart, 2010). Many variations exist in the specifics of how this method is used, but the key idea and the underlying principles remain the same. In a common version of this approach, each member of the treated group is paired with a member of the control group with a similar propensity score. The result is a new dataset that mimics the conditions of an RCT by ensuring that members of the control and treatment groups do not differ substantially in selection probabilities. Estimating the causal effect of treatment can then proceed by looking at the difference in the mean outcomes of the treatment group, and the group of paired observations with similar propensity scores. This was the approach adopted by Kretschmann et al. (2014) to investigate the effect of skipping a grade on the achievement of German elementary school students. Other examples of this method in use include McCormick et al. (2013), who examined the impact of teacher-student relationships on student achievement, and Ponzo (2013), who studied the effect of bullying on student achievement.

Other causal methods, such as the difference-in-differences (DiD, Donald and Lang, 2007) approach are applicable to specific types of data, namely longitudinal data. This approach is useful when data is available on a sample of individuals at an initial time 1 and a later time 2. Interest is in determining the effect of some treatment or intervention that was applied to a subset of the sample at a point between time 1 and time 2. The key assumption underpinning this approach is called the parallel trend assumption. It requires that the outcome variable of the treated units would have followed a similar trajectory to that of the untreated units if they had not received treatment. This means the untreated units can be used as an

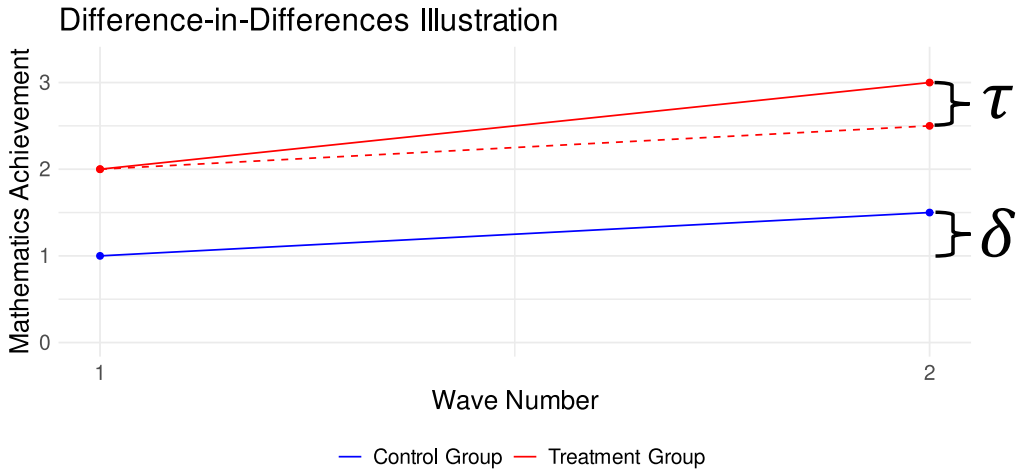

Figure 1.1: Illustration of the difference-in-differences approach where the solid blue and red lines indicate the observed trend for the control and treatment groups respectively. The dashed red line indicates the counterfactual trend for the treated group, had it not received treatment. δ shows the difference in achievement experienced by the control group, while τ represents the difference in this difference experienced by the treated group.

indication of what the difference in outcomes for the treated units would have been had they not received treatment, allowing us to recover the counterfactual trend, and outcome at time 2 for the treated group. See Figure 1.1 for an illustration of this in action. For a review of how the difference-in-differences design has been used in education research in the past, the reader is referred to Corral and Yang (2024).

Other popular methods include instrumental variables (Newhouse and McClellan, 1998), two way fixed effects (Imai and Kim, 2021), and Bayesian causal networks (Pearl, 1995). A key limitation of these extra methods and the ones discussed above is that they are often restricted to providing an estimate of the average effect of treatment. Often, however, interest is in understanding how unique individuals or subsets of individuals with specific characteristics respond to treatment. This is referred to as estimating heterogeneous treatment effects,

and has important applications in tailored medicine, and careful policy design. Doctors, for example, are often interested in knowing how their specific patient may respond to a drug, and policy makers are often interested in understanding how an education intervention may affect a specific group of students.

When understanding heterogeneity in treatment effects is important, a family of methods based on causal inference machine learning has emerged as a leading approach (Caron et al., 2022a). This family of methods relies on using the advanced predictive capabilities of machine learning models to learn from patterns in the available data in order to predict the counterfactual outcomes of the observations. Any sufficiently accurate model can be used for this task, but the best candidates are models that are flexible, allowing them to adapt to complicated features in the data such as non-linear relationships and interaction terms, while also providing reliable uncertainty quantification. For these reasons, approaches based on flexible Bayesian tree-based regression models such as Bayesian Additive Regression Trees (BART, Chipman et al., 2010; Hill, 2011; Carnegie et al., 2019; Dorie et al., 2022) and Bayesian Causal Forests (BCF, Hahn et al., 2020) have become very popular. These are the methods that we will work with and extend throughout this thesis, and they deserve a more thorough introduction and discussion than is possible here, so we will defer a detailed introduction of these methods and our extensions of them to Chapter 2. But first, we provide an outline for the remaining chapters of this thesis, and highlight the key contributions contained in each.

1.4 Thesis Outline and Contribution

After introducing BART, BCF, and their applications for causal inference in Chapter 2, Chapter 3 uses English data from TALIS 2018 to demonstrate the use of BART for causal inference in a study investigating factors affecting teacher job satisfaction. Teacher shortages and attrition are issues of international concern (UNESCO, 2015). A common reason for teachers leaving the profession is a lack of job satisfaction (Madigan and Kim, 2021; Klassen and Chiu, 2010; Wang et al., 2015). Therefore, identifying factors that may be positively or negatively impacting teacher job satisfaction can be a valuable method for reducing turnover and encouraging new teachers to join the profession. Given the widespread concern

about teacher shortages, many studies have been conducted in this area. However, our study makes an important contribution in areas that have been relatively under explored in much of the existing literature.

First, our use of the flexible non-parametric BART model is the first time, to our knowledge, that an advanced statistical model of this type has been used in this area. Our use of a flexible BART model is likely to be very valuable in this setting, as job satisfaction is known to have a non-linear U-shaped relationship with covariates such as age (Guarino et al., 2006; Boe et al., 1997), for example. Secondly, our investigation focuses on specific and implementable factors, adding practical value to our findings. This is in contrast to many prior studies which have examined general aspects of school environments, such as overall collaboration. These existing findings are useful and important, of course, but it is not always immediately clear how to improve collaboration, and consequently job satisfaction. Our investigation of factors such as induction schemes, however, does offer a direct pathway to new or updated policy decisions that may improve job satisfaction. Finally, given the well-documented teacher shortages currently facing schools in England, our focus on a representative sample of teachers from a country experiencing these shortages makes our results highly relevant and timely.

In Chapter 4 we develop and introduce a new multivariate extension of Bayesian Causal Forests and apply it to Irish data from TIMSS 2019. Through simulation studies, we demonstrate the improved accuracy, precision, and uncertainty quantification of the approach. This is made possible by allowing the tree structures of the underlying BART model to benefit from the shared information across all outcome variables. We also demonstrate the robustness of the proposed approach to violations of the model assumptions affecting only one of the outcome variables.

In our application to the TIMSS 2019 data, we investigate the effect of three home-related factors on student achievement in both mathematics and science: often being hungry when arriving at school, often being absent, and having access to a study desk. We find that often arriving at school feeling hungry has a clear negative effect on student achievement (Vik et al., 2022). Interestingly, this effect is found to be less pronounced in schools with fewer resources - these are schools which are more likely to benefit from free school meal schemes in Ireland (Department of Social Protection, 2023). Our finding that having access to a study

desk at home improves student achievement in mathematics, meanwhile, points towards an opportunity to educate parents about the importance of students having dedicated study spaces at home. Finally, our finding that often being absent can negatively affect student achievement highlights the importance of schools investigating the reasons for these absences, and offering extra supports to affected students where necessary (Vesić et al., 2021).

Chapter 5 focuses on an application of our newly developed multivariate BCF model to examine the impact of homework frequency and duration on student achievement in mathematics and science, using data from TIMSS 2019. Homework, like teacher shortages, is a topic of interest to many and has been the subject of much debate in the literature (Kohn, 2006; Buell, 2008). However, despite its important role in the daily lives of students, significant gaps remain in the literature on homework. First, existing studies have typically focused on the cumulative weekly time students spend on homework, overlooking how homework is distributed throughout the week (Trautwein et al., 2002). Chapter 5 seeks to address this limitation by untangling the effects of homework frequency and duration, and determine the optimal distribution of homework throughout the week. Secondly, much of the existing literature has focused on individual subjects in isolation, with few exceptions (e.g., Eren and Henderson, 2011; Fernández-Alonso et al., 2017). With the multivariate BCF model from Chapter 4, however, we are able to identify subject specific differences in the effects of homework, allowing teachers to tailor homework plans in mathematics and science separately. Finally, a widely held belief among many researchers is that homework has a greater effect for students from advantaged socioeconomic backgrounds, who may have better home resources and parental support (Patall et al., 2008; Tan et al., 2020). However, through examining the heterogeneous treatment effect estimates from the BCF model, Chapter 5 finds no evidence to support this, at least at the eighth grade. This suggests that homework can remain a valuable tool for improving student achievement across all socioeconomic backgrounds.

Motivated by two waves of data on high school mathematics achievement in the US, Chapter 6 develops and introduces a new longitudinal extension of BCF. The proposed model combines the structure and interpretability of the traditional difference-in-differences approach with the accuracy and flexibility offered by mod-

ern Bayesian non-parametric methods. It relaxes the parallel trend assumption of the difference-in-differences model, enables the estimation of individual level growth in student achievement, and also provides estimates for the heterogeneous effects of treatment on this growth. In a simulation study we benchmark the proposed model against competing methods, and demonstrate the improved predictive performance and other advantages offered by the new approach.

Our application of this model to the HSLS data investigates the effect of intensive part-time work on high school mathematics achievement, where intensive part-time work is defined as upwards of 20 hours of work per week during the school year (Lee and Staff, 2007). For many high school students, part-time jobs have become an integral part of their daily routine (Singh and Ozturk, 2000). Prior research shows that this part-time work can have a significant impact on their educational journeys (Bachman and Schulenberg, 2014). Opinions differ, however, on whether this impact can be given a causal attribution, and gaps remain in our understanding of how student backgrounds and motivations may moderate the effects of part-time work. Our findings from Chapter 6 point towards a widening achievement gap between those with initially high and low mathematics achievement. Our results show that on average, intensive part-time work has a small but negative impact on student achievement. For a subset of students with a low sense of school belonging, however, we find that part-time work may actually have a positive effect. These findings have implications for policies aimed at ensuring young students are not adversely affected by very long working hours. They also highlight the importance of early monitoring of achievement gaps, before they are allowed to become established (Morgan et al., 2016).

Finally, we conclude our work in Chapter 7 by discussing the contributions, limitations, and promising areas of future work for each of the chapters of this thesis.

2

Bayesian Additive Regression Trees and Bayesian Causal Forests for Causal Inference

2.1 Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART, Chipman et al., 2010) is a Bayesian non-parametric modelling tool for regression and classification tasks. It has become very popular among researchers because of its excellent predictive performance and uncertainty quantification. Building upon the earlier Bayesian Classification and Regression Tree (Bayesian CART) model of Chipman et al. (1998), it can be used to approximate any unknown function f with an ensemble of decision trees. Individually, each decision tree contributes only a small amount to explaining the variance of the unknown function f. Together, however, their combined contributions allow the model to effectively capture complicated patterns such as non-linear relationships or interaction terms.

The most fundamental unit of a BART model is the decision tree, which consists of three key parts: 1) A root, 2) A set of decision rules, and 3) A set of terminal nodes. An example diagram of a BART model with two decision trees is provided in Figure 2.1. The roots of the trees, in the shape of a diamond, represent the starting point of the decision process. Here, assuming the set of decision rules is not empty, in which case the root of the tree is also the sole terminal node of

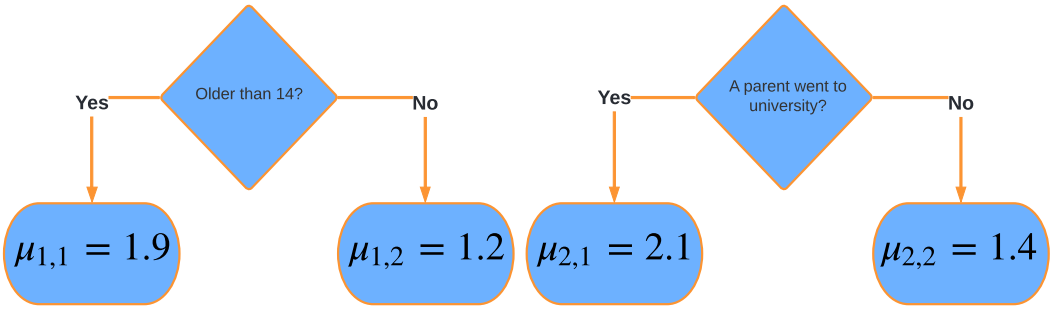

Figure 2.1: Diagram of two simple decision trees. By following the decision rules from the root to the terminal nodes, the contributions from individual trees can be combined into one final prediction for each observation.

the tree, the data is partitioned into two subsets which themselves can be further partitioned into even smaller subsets with additional decision rules. Following the set of decision rules leads to the terminal nodes of the tree, where a unique result or prediction is assigned to each subset of the data.

Figure 2.1 shows that for an observation with an age of 15, the contribution provided to the model by the first tree is 1.9. For the same student, with at least one parent who went to university, the contribution provided by the second decision tree is 2.1. Adding these two results together yields a final prediction of 4.0. By combining predictions from more decision trees, the final prediction from the model can be informed by more covariates, more splitting values on those covariates, and gradually capture more and more complex relationships.

Mathematically, the BART model can be written as follows:

$$y_i = \sum_{j=1}^{J} g(T_j, M_j, x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Here, y_i represents the outcome variable associated with each observation i. The function $g()$ returns the correct prediction from a given tree T_j of J total trees, where M_j denotes its set of terminal nodes, and x_i is the vector of covariates for

observation i. The error term ϵ_i is assumed to follow a normal distribution with mean 0 and variance σ^2.

The model is fitted to the data using a process called Markov Chain Monte Carlo, which allows the decision rules, the structure of the trees, and the terminal node parameters providing the final predictions to learn from the data in a gradual step by step manner. The process starts with all trees as stumps - trees where the root of the tree is also the sole terminal node of the tree. Next, one of four possible operations is selected at random to apply to the first tree: Grow, Prune, Change, or Swap. If grow is selected, then a terminal node of the tree is selected at random along with a variable to split on and a splitting value is also chosen from the set of available valid options. If prune is selected then a parent node of two terminal nodes is selected, and transformed into a terminal node by removing its children nodes from the tree. The change operation takes a randomly selected non-terminal node, and assigns to it a new decision rule. Finally, the swap operation selects a parent-child pair which are both internal nodes, and swaps their splitting rules with each other.

Once the chosen operation has been applied to the tree, the resulting structure of the updated tree is either accepted or rejected using a Metropolis-Hastings step. In order to prevent the trees from growing too large, a prior is placed on the structure of each tree, which says that the probability of any node at depth d being non terminal is given by $\alpha(1+d)^{-\beta}$, where α and β can be adjusted to modify the strength of the prior. If we apply this to a tree with terminal nodes labelled as $h_{j,1}...h_{j,K}$, and non-terminal nodes labelled as $b_{j,1}...b_{j,L}$, we have that:

$$P(T_j) = \prod_{k=1}^{K} \alpha(1+d(h_{j,k}))^{-\beta} \prod_{l=1}^{L} [1-\alpha(1+d(b_{j,l}))^{-\beta}]$$

where $d()$ is a function for returning the depth of any given node.

With the proposed structure of the tree accepted or rejected, the next step is to update the terminal node parameters of the tree. This step is performed conditional on the partial residuals $R_{i,j}$, given by y less the contributions from all trees except tree j. In other words, the update is performed so that the new terminal node parameters attempt to explain the leftover variation in y that is not explained by the other trees in the ensemble:

$$R_{i,j} = y_i - \sum_{k \neq j} g(T_k, M_k, x_i)$$

While completing this step, to ensure that each tree contributes an approximately equal amount to the overall predictions from the model, a normal prior is placed on each of the K terminal node parameters of tree j:

$$\mu_{j,k} \sim N(\mu_\mu, \sigma_\mu^2)$$

After scaling the response variable y to follow a normal distribution with mean 0 and standard deviation 1 during data pre-processing, a sensible choice for μ_μ and σ_μ^2 is 0 and $\frac{1}{J}$ respectively. This ensures that the prior on the combined contribution of the terminal node parameters from all trees, $E[Y|X] \sim N(J\mu_\mu = 0, J\sigma_\mu^2 = 1)$ is appropriate for covering the range of all observed y values. This also helps to limit the influence of each individual tree, encouraging all trees to contribute a small amount to the final predictions, acting as a form of regularisation.

Now with the structure of the first tree updated, along with its terminal node parameters, analogous updates are applied to each of the remaining trees in the ensemble. Then, with the final tree updated, it is possible to calculate the combined contribution of all trees, and the final residuals:

$$R_i = y_i - \hat{y}_i = y_i - \sum_{j=1}^{J} g(T_j, M_j, x_i)$$

The final step is to update the residual variance term σ^2 which has the conjugate prior:

$$\sigma^2 \sim \text{Inv-Gamma}\left(\nu/2, \nu\lambda/2\right)$$

A good strategy for selecting ν and λ is to note that σ is likely to be less than the standard deviation of the raw y values. Given the flexibility of BART, it is also reasonable to expect that σ is likely to be less than a residual standard deviation $\hat{\sigma}$ obtained with a less flexible model such as ordinary least squares regression. Sensible choices for ν and λ therefore place a high prior probability on σ being less than $\hat{\sigma}$. Chipman et al. (2010), for example, suggest setting $\nu = 3$ and choosing

λ such that $P(\sigma < \hat{\sigma}) = 0.9$. After repeating the above steps for a pre-specified number of iterations, the result is a posterior of trees, terminal node parameters, and sampled σ^2 values.

Because of the impressive performance of BART, it has seen applications in many different areas such as medicine, economics, and education (Pierdzioch et al., 2016; Sparapani et al., 2016; McJames et al., 2023b, Chapter 3). This has also motivated researchers to extend the model in various ways, allowing it to be applied in different scenarios that are not well suited to the standard BART implementation described above. Some important BART extensions include a multinomial logistic regression variant by Murray (2021), a version designed for situations involving local linearities by Prado et al. (2021a), and a BART model well suited to high dimensional data by Linero (2018). Additional BART extensions and applications are discussed in Hill et al. (2020), but now we turn our attention to an extra important use for BART: the identification of heterogeneous treatment effects from observational data.

2.2 BART for Causal Inference

One of the many areas that BART has found success in is the estimation of heterogeneous treatment effects from observational data. An important contribution was made to this area by Hill (2011), who showed that by using BART to flexibly model the response surface of the outcome of interest, accurate estimation of causal effects with reliable uncertainty quantification is possible. Hill's paper considered the situation in which we have data on n observations, each accompanied by a response variable y_i, a set of covariates x_i, and a treatment variable Z_i where $Z_i = 1$ indicates that observation i did receive treatment, and $Z_i = 0$ indicates that observation i did not receive treatment. The goal is to understand the impact that treatment may have on y, and how this effect may vary depending on the specific covariates of the treated unit.

The approach is based on the previously described Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Sekhon, 2008), also known as the potential outcomes framework. Hill's contribution was to show that although we only ever observe one of the potential outcomes, by using BART to estimate them with $E\left[y_i(Z_i = 1)|x_i\right]$

and $E\left[y_i(Z_i = 0)|x_i\right]$, we can estimate τ_i with $\hat{\tau}(x_i) = E\left[y_i(Z_i = 1)|x_i\right] - E\left[y_i(Z_i = 0)|x_i\right]$. The BART based approach to heterogeneous treatment effect estimation proceeds by fitting a BART model to the whole of the data, using the x_i covariates and Z_i indicators to establish the decision rules within the trees. The trained BART model can then assign to any observation two predictions, one for each potential outcome: $\hat{y}_i(Z_i = 1) = \hat{f}(x_i, Z_i = 1)$, and $\hat{y}_i(Z_i = 0) = \hat{f}(x_i, Z_i = 0)$. From here, estimates of individual conditional average treatment effects, $\hat{\tau}(x_i)$ (ICATEs), can be obtained by calculating the difference between each individual's predicted potential outcomes: $\hat{\tau}(x_i) = \hat{f}(x_i, Z_i = 1) - \hat{f}(x_i, Z_i = 0)$. By averaging these individual conditional average treatment effects over the population, the mixed average treatment effect (MATE, Li et al., 2023), $MATE = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}(x_i)$, can also be estimated.

To provide a visual description of the approach, Figure 2.2 reproduces an example provided in Hill (2011) which shows how BART can be used to estimate the response surfaces corresponding to the two potential outcomes. The ICATEs are given by the difference in height between the red and the green curves, which approximate the true response surfaces for the control and treatment groups coloured in black and blue respectively.

This BART based approach has become very popular because of its strong performance and the desirable ability to provide Bayesian credible intervals for the estimated treatment effects. Studies adopting this approach include Hu and Gu (2021); Gill et al. (2023); Blette et al. (2023); Carnegie et al. (2019); Dorie et al. (2022). Recent work has categorised this approach as belonging to a family of machine learning based approaches called an S-Learner, or Single-Learner, as a single model is fitted to the dataset and then used to derive treatment effect estimates (e.g. Caron et al., 2022a). Finally, owing to the success of this method, others have been inspired to build upon the approach in various ways. One such development is the Bayesian Causal Forest model of Hahn et al. (2020), which we will discuss next.

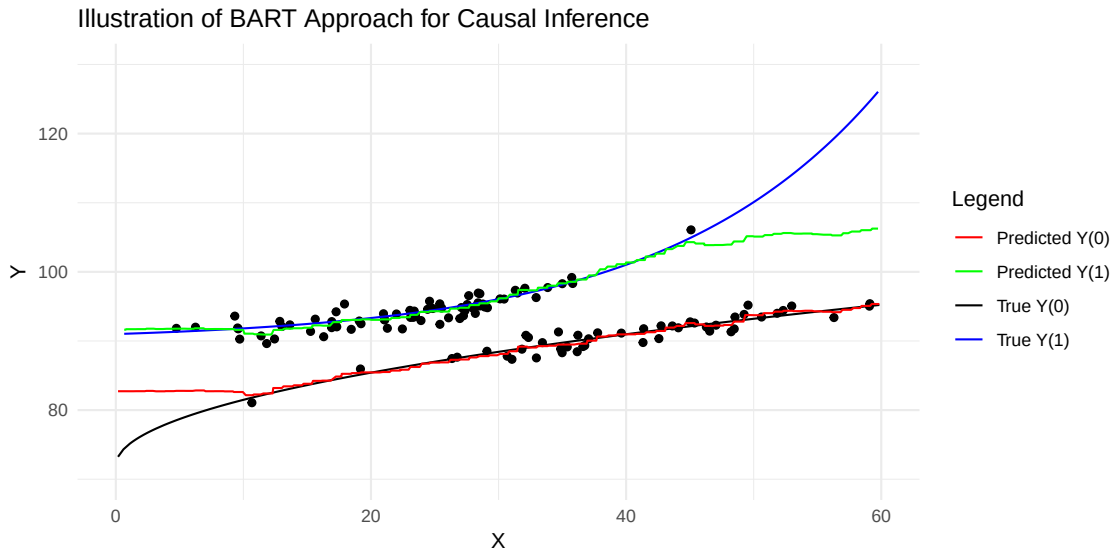

Figure 2.2: Illustration of the BART based approach for causal inference, reproducing the simulated example from Hill (2011). By flexibly modelling the response surface for the control and treatment groups, ICATEs can be estimated based on the difference between the two potential outcomes for each observation.

2.3 Bayesian Causal Forests

Bayesian Causal Forests (BCF, Hahn et al., 2020) is a Bayesian non-parametric causal inference machine learning algorithm based on Bayesian Additive Regression Trees. It uses the Robinson (1988) treatment effect parameterisation, with BART as a foundation, to split the estimation of y into two separate parts: A prognostic effect μ, and a treatment effect τ:

$$y_i = \mu(x_i, \hat{\pi}_i) + \tau(x_i)Z_i + \epsilon_i.$$

With Z_i coded as 1 for treatment and 0 for control, the model takes on a very intuitive interpretation, where $\mu(x_i, \hat{\pi}_i)$ represents the expected outcome for observation i in the absence of treatment, and $\tau(x_i)$ represents the effect of the intervention or treatment under investigation. Within the $\mu()$ and $\tau()$ parts of the model, x_i represents the covariates associated with observation i, used for establishing the decision rules in the trees. The additional covariate included in the $\mu()$

part of the model, $\hat{\pi}_i$, is the previously described propensity score. It represents the estimated probability of observation i, receiving treatment, conditional on the observed covariates: $\pi_i = P(Z_i = 1|x_i)$. The inclusion of this 'clever covariate' was recommended by Hahn et al. (2020) in order to avoid a phenomenon called regularisation induced confounding, and has been found to help improve the estimation of treatment effects not only in the BCF model, but also when used as a covariate as part of the BART based approach described above. Finally, ϵ_i is the familiar random error term, which as before, is assumed to follow a normal distribution with mean 0 and variance σ^2.

The key advantage of this model over the BART based approach is that because the treatment effect is estimated by a separate part of the model, it is possible to apply separate priors and amounts of regularisation to $\mu()$ and $\tau()$. For example, given that it is reasonable to expect less heterogeneity in the treatment effects than in the prognostic effect of the control variables, it is common to apply greater regularisation to the τ part of the model. This can be accomplished with the α and β priors, which are responsible for deciding the preferred tree depths. In the μ part of the model, a choice of $\alpha = 0.95$, $\beta = 2$, is common, while in the τ part of the model, a stricter setting of $\alpha = 0.25$, $\beta = 3$ is often preferred. Similarly, the scale of the terminal node parameters σ^2_μ and σ^2_τ can also be adjusted if desired.

An extra important advantage is that separate sets of covariates can also be used in the μ and τ parts of the model. This can be especially useful in situations where there may be a strong a-priori belief that the confounding variables may be different to the covariates responsible for moderating the effect of treatment. Additionally, as the estimated treatment effects are now captured by a separate part of the model, it is possible to make inference on the treatment effects directly, without the need to perform the post-processing steps associated with the BART based approach. Finally, with the treatment effect estimates disentangled from the prognostic effect of the covariates on $\mu()$, it is possible to study aspects such as variable importance with model explainability tools developed to work with the original BART model (Inglis et al., 2022a,b).

The BCF model, which shares the same impressive features of BART such as strong predictive performance, and uncertainty quantification, has gained considerable traction in both applied and methodological research areas. Studies adopting

the BCF model to investigate different research questions include Schwartz et al. (2021), O'Neill et al. (2024), and McJames et al. (2023a). On the methodological side, researchers have extended the BCF model in different directions to improve its applicability to specific research questions. Notable examples include Starling et al. (2021) and Caron et al. (2022b). The BCF model also provides a foundation for three different extensions of the model which have been developed in the course of writing this thesis. An outline of the variations of BART and BCF used in the remainder of this thesis is what we will discuss next.

2.4 Variations of BART and BCF Used in This Thesis

2.4.1 Chapter 3

In Chapter 3, rather than introducing a novel extension of BART or BCF, we apply the BART-based approach to causal inference outlined in Chapter 2.2 to a significant issue in education. This chapter investigates the impact of various factors on teacher job satisfaction. For each factor Z under investigation, a separate BART model $y = f(X, Z) + \epsilon$ is fitted to the data to predict teacher job satisfaction based on the observed covariates and treatment status. Individual conditional average treatment effects are then averaged across the sample using a weighted mean to obtain an estimate of the mixed average treatment effect (MATE, Li et al., 2023) for each factor:

$$MATE = \sum_{i=1}^{n} w_i \tau_i = \sum_{i=1}^{n} w_i \left[\hat{f}(x_i, Z_i = 1) - \hat{f}(x_i, Z_i = 0) \right]$$

2.4.2 Chapter 4

Motivated by data from the Trends in International Mathematics and Science Study (TIMSS 2019, Mullis et al., 2020), which includes data on student achievement in both mathematics and science, Chapter 4 introduces a multivariate extension of BCF, allowing BCF to be applied in situations where we are interested in estimating the effect of a single intervention on multiple outcome variables si-

multaneously. The BCF model structure remains the same:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}(x_i, \hat{\pi}_i) + \boldsymbol{\tau}(x_i)Z_i + \boldsymbol{\epsilon}_i$$

but $\boldsymbol{Y}_i$ is now a length p vector of outcome variables, predicted by $\boldsymbol{\mu}()$ and $\boldsymbol{\tau}()$ which are ensembles of multivariate Bayesian Additive Regression Trees, with length p error term $\boldsymbol{\epsilon}_i$.

Accordingly, multivariate normal priors are now used for the terminal node parameters of the $\boldsymbol{\mu}()$ and $\boldsymbol{\tau}()$ trees, while an Inverse-Wishart prior is placed over the residual covariance matrix Σ:

$$\boldsymbol{\mu}_{j,k} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_\mu = \sigma_\mu^2 \boldsymbol{I}\right), \quad \boldsymbol{\tau}_{j,k} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_\tau = \sigma_\tau^2 \boldsymbol{I}\right)$$
$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}\left(\nu, \boldsymbol{\Sigma_0}\right).$$

Importantly, one prior that remains the same is the tree prior $P(T_j)$. This is because the same tree structure is used for all of the p outcome variables. The effect of this is to encourage the model to carefully choose predictor variables and tree structures appropriate for all outcome variables. A potential drawback of this decision, which is tested and demonstrated in Chapter 4, is that this choice of prior may not be well suited to situations where different predictors or tree structures are required for separate outcome variables. However, given the focus of Chapter 4 on two highly correlated outcomes, and on home related factors that are likely to influence student achievement in similar ways, this decision is justified.

A further extension used in Chapter 4 is the inclusion of a random intercept term, $\boldsymbol{\alpha}_{class,i}$, into the BCF model:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}(x_i) + \boldsymbol{\tau}(x_i)Z_i + \boldsymbol{\alpha}_{class,i} + \boldsymbol{\epsilon}_i,$$

where we assume that each $\boldsymbol{\alpha}_{class}$ comes from a multivariate normal distribution with a population mean $\boldsymbol{\mu}_\alpha$ and population covariance matrix $\boldsymbol{\Sigma}_\alpha$: $\boldsymbol{\alpha}_{class} \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$, where the prior on $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Sigma}_\alpha$ is: $\boldsymbol{\mu}_\alpha \sim N(\boldsymbol{m} = \mathbf{0}, \boldsymbol{s} = 0.01\boldsymbol{I})$, and $\boldsymbol{\Sigma}_\alpha \sim \mathcal{W}^{-1}(a = 1, \boldsymbol{\Omega_0} = 0.1\boldsymbol{I})$.

Yeager et al. (2019) were the first to introduce this important feature in the

single outcome context, which we adapt to our multivariate setting. While not the primary focus of the chapter, this is an important feature to include, necessitated by the hierarchical structure of the TIMSS dataset, whereby students are nested within classrooms. This random intercept feature allows the model to account for classroom specific variations in achievement that are not captured by the available covariates in $\boldsymbol{\mu}()$.

Finally, we note that while the treatments of interest in Chapter 4 (having a study desk at home, often arriving to school feeling hungry, or often being absent) are not subject specific treatments, meaning that an indication of $Z_i = 1$ implies both outcome variables are affected by the treatment, there is nothing to prevent the study of subject specific treatments with this model. Such a situation may arise, for example, when studying the effect of paid tuition on student achievement, as it is possible for students or parents to pay for extra tuition in neither, one and not the other, or both subjects. This is made possible by replacing the multivariate BCF parameterisation above with the following:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}(x_i) + \boldsymbol{\tau}(x_i) \circ \boldsymbol{Z}_i + \boldsymbol{\alpha}_{class,i} + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{Z}_i$ is now a length p treatment status indicator, connected to $\boldsymbol{\tau}()$ through a Hadamard product, $\circ$, allowing the treatment status of individual outcomes within each student to be different.

2.4.3 Chapter 5

Chapter 5 is based primarily on the application of the newly developed multivariate BCF model to investigate the effect of homework frequency and duration on student achievement in mathematics and science, using the same TIMSS 2019 data as Chapter 4. While not emphasised in the chapter, an important extension is made to BCF in order to make this possible. The default implementation of BCF is applicable only in situations where there is interest in the effect of a single binary treatment on the outcome, or in the case of multivariate BCF, outcomes of interest. This is an important limitation in the context of homework, because in order to build a nuanced understanding of how students respond to homework, and understand the optimal way in which homework can be distributed, it is im-

portant to account for the differing frequencies (number of days per week), and durations (number of minutes each evening) that homework can be assigned with.

In order to account for the differing levels of homework frequency and duration, the multivariate BCF model applied to the TIMSS data in Chapter 5 uses the following structure:

$$y_{i,j} = \mu_j(x_i) + \tau_{j,1}(x_i)Z_{i,j,1} + \tau_{j,2}(x_i)Z_{i,j,2} + \tau_{j,3}(x_i)Z_{i,j,3} + \tau_{j,4}(x_i)Z_{i,j,4} + \alpha_{class,i,j} + \epsilon_{i,j}$$

where $y_{i,j}$ is the achievement of student i in subject j ($j = 1$ for mathematics or $j = 2$ for science). For a student who receives homework up to one or two times per week with a duration of up to fifteen minutes, their achievement in subject j is given by $\mu_j(x_i)$, where x_i denotes the characteristics associated with student i. Students who receive homework with a greater frequency or duration belong to the treatment groups $Z_{i,j,1} \dots Z_{i,j,4}$:

- Frequency of three or four times per week $\rightarrow Z_{i,j,1} = 1$,
- Frequency of every day $\rightarrow Z_{i,j,2} = 1$,
- Duration of fifteen to thirty minutes $\rightarrow Z_{i,j,3} = 1$,
- Duration of greater than thirty minutes $\rightarrow Z_{i,j,4} = 1$.

We estimate the causal effect of belonging to these groups with $\tau_{j,k}(x_i)$, where $k = 1 \dots 4$ as above. Note that this is an example of a situation where the treatment (homework) may apply differently to mathematics and science, as it is possible a student may receive mathematics homework with a different frequency and or duration to science homework.

This extension is closely related to the "no multiple versions of treatment" aspect of the stable unit treatment value assumption. In the case of homework, this extension was necessary because variations in homework frequency and duration may constitute multiple versions of the same treatment. Therefore, this simple yet powerful extension is also likely to be very useful in other domains such as medicine, where multiple drugs or variations of the same treatment are commonly tested in parallel with each other. As an example, a model with the following (univariate) structure based on the homework model could potentially be useful

in a medical context:

$$y_i = \mu(x_i, \hat{\pi}_i) + \tau_1(x_i)Z_{i,\text{Medicine-1}} + \tau_2(x_i)Z_{i,\text{Medicine-2}} + \epsilon_i$$

where two separate $\tau()$ ensembles are used for different medicines being compared in a clinical trial.

Chapter 5 itself does not discuss this model in very technical terms, so this is a good opportunity to note two additional advantages of this model that are not discussed in Chapter 5 which focuses primarily on the application to the TIMSS data. First, given the multiple treatment effect estimating parts of the model, there is the interesting possibility to incorporate separate moderating variables into each of the $\tau()$ parts of the model. This may be useful in cases where there is a strong a-priori belief that separate effect moderators may be appropriate for the treatments under investigation.

Second, given the model jointly estimates the effects of multiple treatments simultaneously, the model may provide a promising approach for tackling the issue of multiple comparisons that arises when multiple hypotheses are tested at once. A full exploration of this promising possibility is beyond the scope of this thesis, but we suggest that further exploration of this possibility may include experimentation with different strengths of priors over the $\tau()$ parts of the model in order to shrink the false discovery rate to a pre-defined threshold. Such an approach could be further validated with large scale simulation studies designed to test how often false discoveries are made in a controlled setting, relative to the false discovery rate obtained with a simpler BCF approach involving the use of many separate analyses.

2.4.4 Chapter 6

Finally, motivated by longitudinal data on student achievement from the High School Longitudinal Study of 2009 (HSLS, Ingels et al., 2013), Chapter 6 introduces a longitudinal extension of BART and BCF for modelling individual level growth in mathematics achievement, and the heterogeneous impact of part-time work on this growth. The idea is to model the trajectory of student achievement as the cumulative sum of separate BART ensembles, each of which pertain to a specific

period of time after an initial base year. Using $y_{i,t}$ to denote the mathematics achievement of student i at time t, and $x_{i,t}$ to denote the available covariates collected on student i up to time t, the model for student achievement becomes:

$$y_{i,t} = \mu(x_{i,1}) + \sum_{w=1}^{T-1} G_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}, \hat{\pi}_{i,w+1}) I(t > w) + \epsilon_{i,t}$$

where

$$G_{w+1}() = \delta_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}, \hat{\pi}_{i,w+1}) + \tau_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}) Z_{i,w+1}$$

This means that student achievement at Wave 1 is given by $\mu(x_{i,1})$, which represents the initial starting achievement of the students. Moving forward one step in time, student achievement at Wave 2 is given by $\mu(x_{i,1}) + G_2(x_{i,2}, y_{i,1}, \hat{\pi}_{i,2})$, where $G_2(x_{i,2}, y_{i,1}, \hat{\pi}_{i,2})$ represents the growth in achievement experienced by student i in the period between Waves 1 and 2. Achievement in later waves up to Wave T is given by adding this to $G_3()$, then $G_4()$, as far as $G_T()$.

Within each $G_w()$, there are two separate BART/BCF ensembles, one which estimates the growth that would have been experienced in the absence of treatment $\delta_w()$, and one which estimates the impact of the intervention of interest on this growth, $\tau_w()$. This structure is analogous to the difference-in-differences method, where $\delta_w()$ represents the difference in achievement from Wave $w-1$ to w for the control group, and $\tau_w()$ represents the difference in this difference experienced by the treatment group. Our model, which extends BART and BCF to the longitudinal setting, however, has a number of important advantages.

First, by using BART and BCF as a foundation, our model can readily detect individual variations in mathematics achievement growth. Similarly, the same flexibility also allows us to investigate heterogeneity in the treatment effects, thus providing a richer and more nuanced understanding of how individual characteristics can affect the quantities of interest, than more traditional methods which are often limited to providing an average treatment effect only. An additional strength of the model is that by controlling for the many confounding variables that may jointly influence the likelihood of receiving treatment and the outcome of interest, we are able to relax the easily violated parallel trends assumption that is necessary

with traditional difference-in-differences based approaches (Roth et al., 2023).

Being a Bayesian model, we can also specify separate priors for each of the $\mu()$, $\delta_w()$, and $\tau_w()$ parts of the model. This is useful, as if we expect the growth between time periods to only be a fraction of the initial achievement level from Wave 1, we can build this information into the prior, σ_δ^2 of the terminal node parameters of the growth trees. Moreover, by devoting a separate ensemble of decision trees to each of the quantities of interest, we can make use of existing model explainability tools developed specifically for BART in order to extract variable importance and interaction metrics (Inglis et al., 2022a,b).

Finally, two extra features are introduced to the longitudinal BCF model of Chapter 6 in order to tackle challenges posed by missing data. The first is to address missing data in the covariates used by the model. For this challenge, we borrow the method introduced by Kapelner and Bleich (2015), which treats missingness as an inherent and useful feature of the data. The procedure involves directing observations with missing data to left or right children of nodes which are being split on, thus allowing the model to learn from any patterns present in the missingness of the data. The second new feature is to address missing data in the treatment variable itself. Here, we make the assumption that the treatment indicators are missing at random, and introduce an additional Gibbs sampling step to impute the missing Z_i values in each iteration of the MCMC algorithm, conditional on the rest of the available data.

2.5 Chapter Summary

This chapter introduced Bayesian Additive Regression Trees and Bayesian Causal Forests, describing how they fit into the framework of the Neyman-Rubin causal model, and how they estimate heterogeneous treatment effects from observational data. We also outlined the variations of BART and BCF used in the following chapters, describing the challenges that motivated the extensions, and the key idea behind the proposed solutions. Full details on all models and added features can be found in the relevant chapters, with detailed derivations of log-likelihoods used in the Metropolis-Hastings steps, and posterior distributions used in Gibbs sampling steps deferred to the appendices of each chapter.

Now, with our understanding of BART and BCF in place, we move on to the first of the research chapters of this thesis, which shows how BART can be applied to an important issue in the world of education - teacher shortages and teacher job satisfaction.

3

Factors Affecting Teacher Job Satisfaction: A Causal Inference Machine Learning Approach Using Data From TALIS 2018

3.1 Introduction

3.1.1 Background

Teacher supply and demand is an important challenge faced by many countries around the world (UNESCO, 2015). The scale of this problem is partly reflected in the teacher shortages currently facing many countries including England (Hilton, 2017), Ireland (O'Doherty and Harford, 2018), the United States (Wiggan et al., 2021), and many others. The scale of the challenge currently facing England is made clear by a recent House of Commons report which reveals that the 2019 five-year retention rate was at its lowest level since 1997, with 32.6% of teachers entering the profession in 2014 no longer teaching in classrooms five years later (Long and Danechi, 2021). These sustained high levels of attrition have led to a situation where the total number of all qualified teachers in England working outside of the state funded sector in 2019 (350,000) was nearly as high as the number of teachers working inside it (454,000). This comes at a time when secondary school pupil numbers are expected to rise by 7% in England between 2020 and 2026, thus placing increasing pressure on already difficult recruitment and retention targets.

Teacher shortages are often more pronounced in Science, Technology, Engineering and Mathematics (STEM) subjects (Han and Hur, 2021). This is also noted in the House of Commons report, which showed the subject specific vacancy rate of unfilled teaching posts for these subjects was higher than the average (Science 1.4%, Technology 1.7%, Maths 1.4%) (Long and Danechi, 2021).

Teacher shortages may arise as a result of insufficient numbers of new entrants to the profession or high levels of qualified teachers leaving their posts. In addition to the serious challenges associated with not having enough teachers, higher levels of teacher turnover have been shown to negatively affect student learning and also incur large economic costs (Levy et al., 2012; Sorensen and Ladd, 2020). Encouraging teachers to stay in their posts is therefore very important. Research has shown that job satisfaction is one of the key predictors of a teacher's intention to remain in the profession (Madigan and Kim, 2021; Klassen and Chiu, 2010; Wang et al., 2015). Consequently, it is vital to identify factors that can improve job satisfaction in order to boost retention rates of qualified teachers and to attract new entrants to start their career. In addition to the economic and staffing implications of job satisfaction which are the primary focus of this study, Toropova et al. (2021) point out that happier teachers tend to have happier students, and more satisfied teachers provide higher quality teaching to their students as well (Spilt et al., 2011; Klusmann et al., 2008). Job satisfaction has also been shown to predict teacher self-efficacy which is another significant area of study within the literature (Burić and Kim, 2021). These reasons combine to make teacher job satisfaction a crucially important outcome of interest.

We decided therefore to investigate the effect that a number of selected factors may have on teacher job satisfaction. Job satisfaction is a term for which no single definition exists, but a widely accepted version describes job satisfaction as "the pleasurable emotional state resulting from the appraisal of one's job as achieving or facilitating the achievement of one's job values" (Locke, 1969, p. 316). Informally, job satisfaction can be thought of as an overall sense of contentment with one's career.

Our approach in this study can be best described as an application of a causal inference machine learning method. We employ these cutting-edge statistical models in order to identify specific, implementable steps that may be taken to enhance

the job satisfaction of qualified teachers. This is a key advantage of our approach, because it allows school principals and other policy makers to determine specific steps that may be taken as part of a school strategy for improving job satisfaction. Our method makes use of Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), a cutting-edge modelling tool which enables us to detect non-linear relationships and interactions which would not normally be found in a standard linear model. Additionally, this strategy allows us to control for a much larger number of background (confounding) variables than would normally be possible when using a linear model. Furthermore, we demonstrate how this approach can be used to identify subgroups of teachers who are most (or least) likely to benefit from the positive effects of a given treatment.

Our study uses data from the third cycle of the Teaching and Learning International Survey (TALIS) which took place in 2018 (OECD, 2019a). TALIS is the world's largest survey of teachers and principals and has taken place every five years since 2008. A fourth cycle is due to take place in 2024. Participating teachers and principals are asked to complete questionnaires on a wide variety of topics such as: personal background; current teaching duties; their perception of the school climate; and job satisfaction. TALIS 2018 is the largest of the surveys to-date with 48 countries participating, and includes data on approximately 260,000 teachers from 15,000 primary, and lower and upper secondary schools. For the purpose of this study, however, we will limit our investigation to the data from England. This subset of the entire dataset contains a representative sample of 2009 primary and 2376 lower secondary school teachers for a total sample size of 4385.

We decided to focus on the English subset of the TALIS data for a number of reasons. Firstly, the English subset of the data is able to provide us with a relatively large sample size of teachers from both primary and lower secondary schools. This is advantageous for machine learning models, as it enables BART to more easily detect relationships between variables in the data, and this is essential for reliably producing accurate results. A second important factor we considered is that England is currently facing a serious teacher recruitment and retention problem (Hilton, 2017). This important contextual factor makes England a more appropriate choice than a country not currently facing such difficulties. Furthermore, a number of initiatives such as the Early Career Framework (Daly et al.,

2021; Department for Education, 2019) have recently been introduced in England, thus making our investigation of mentorship and induction schemes particularly relevant to the English context.

With these data, and using a causal inference machine learning approach, we attempt to answer the following research question: What are the specific and implementable factors that have the most positive (or negative) impact on teacher job satisfaction? The factors we consider include: participation in induction schemes; high levels of participation in continual professional development; team teaching; observing other teachers; mentorship schemes; teaching in a public school; class size; out-of-field teaching; and having a part-time contract. Our decision to include these factors in our investigation has been informed by previous studies which show they have a strong association with both teacher job satisfaction and retention. We now discuss these findings in more detail in a literature review, focusing on key aspects relevant to our research.

3.1.2 Literature Review

Induction and Mentoring Programmes

Induction is a broad term used to describe different activities or supports put in place for teachers to assist them in adapting to the ethos or practices of a new school (Allen, 2005). Induction programmes are frequently designed with newly-qualified teachers in mind, but we will use the slightly more inclusive definition from the TALIS questionnaire, which broadens the scope of induction activities to include supports for experienced teachers who have recently begun teaching in a different school (OECD, 2018).

Mentoring describes the arrangement whereby a newly-qualified teacher is assigned a more experienced member of staff at their school, who will advise and assist them as they begin their career (Allen, 2005). Roles of a mentor can vary, as can frequencies of meetings between a mentor and their mentee. For consistency, we will once again use the more general definition provided in the TALIS questionnaire which allows mentoring to encompass any situation where a more experienced teacher supports a less experienced one, who need not be newly qualified (OECD, 2018).

New teachers are commonly faced with many challenges in the classroom after they qualify (Guarino et al., 2006). To mitigate the risk of newly qualified teachers encountering difficulties, induction and mentoring schemes are often provided to support them during this formative stage of their career. In fact, induction for new teachers is statutory in many countries, such as in England where this is the case in state schools, and a new scheme for early career teachers has recently been introduced called the Early Career Framework (Daly et al., 2021; Department for Education, 2019).

International evidence often points towards induction and mentoring schemes as having a positive effect on the job satisfaction of participating teachers. Regression analyses of teachers in the US subset of TALIS 2018, for example, have found a strong link between the presence of a mentor and considerably higher levels of job satisfaction (Renbarger and Davis, 2019). This finding is backed by a review of ten studies on the effect of mentoring, which reveals consistent evidence in support of the positive effects of mentoring on teacher retention (Ingersoll and Kralik, 2004). Other studies based on the US subset of TALIS 2018 have also identified induction activities as having a positive effect on job satisfaction (Reeves et al., 2022). Additionally, the provision of induction supports for newly qualified teachers in their first year of teaching has also been linked to lower levels of attrition (Ronfeldt and McQueen, 2017).

Despite their widespread use, the evidence supporting the use of induction and mentoring schemes for all teachers has sometimes been brought into question. A large scale review of over 90 studies by Allen (2005) found only limited evidence that participation in induction and mentoring schemes leads to higher retention rates of qualified teachers. Indeed, survey studies of teachers undergoing statutory induction in the UK suggest that initial teacher education may be far more important for preparing new teachers for the challenges they will face in their first year of teaching (Hulme and Wood, 2022).

It is also true that while induction schemes or mentoring may be beneficial for job satisfaction and retention in the long term, not all teachers report enjoying induction or mentoring at the time. Some teachers undergoing induction in the UK report it as being a stressful experience due to the busyness of their schedule, and others report dreading meetings with mentors who provide them with criticism

(Smethem, 2007).

The effect of *being* a mentor on job satisfaction has been the focus of relatively little research in comparison to the effect of *having* a mentor. Despite this, there are still studies which show that mentoring arrangements can be mutually beneficial to both the mentor and the mentee. Lunsford et al. (2018) for example, in a study of US teachers, found that those with either a mentor or a mentee are on average more satisfied than teachers who do not.

It is important to note that most of the above findings are based on observational data, and therefore the positive correlation between induction or mentoring and job satisfaction can not be claimed to be causal in nature. A recent study with a longitudinal design which tracked a sample of newly-qualified teachers in the US over the first 5 years of their career therefore makes an important contribution (Gray and Taie, 2015). At each follow-up visit, teachers who were assigned a mentor during their first year in the classroom were more likely to still be teaching than those who did not receive this extra support, thus showing a temporal association between mentoring and retention.

Continual Professional Development

Continual professional development (CPD) can refer to a wide range of activities designed to assist teachers as they build upon and improve their professional skills (OECD, 2018). Higher levels of participation in CPD have often been linked to improved teacher job satisfaction (Wang et al., 2020; Yoon and Kim, 2022). In a joint study of English and international data from TALIS 2013, Sims (2017) was able to show that this relationship holds in both the national and international context. With two separate analyses, they demonstrated that there is a positive correlation between CPD and job satisfaction, firstly using data for England only, then again for a combined dataset of more than 50,000 teachers from 38 different countries.

CPD has also been shown to be related to higher levels of teacher retention. A survey of 500 teachers based in England, for example, who had just completed a professional development course showed that teachers who were more engaged with the CPD course were more likely to respond that the course had a positive

effect on their intention to remain teaching (Coldwell, 2017). This link was less strong for teachers who only engaged moderately or weakly with the course. Furthermore, Allen and Sims (2017), in a study of teachers at state-funded secondary schools in England, found that similar effects were still being felt two years after participation in a science subject-specific CPD course, and that participation had reduced department turnover rates by two percentage points. This finding is especially important given that STEM subject teachers are known to be at higher risk of attrition (Han and Hur, 2021).

Despite these benefits, one challenge often faced by teachers is that there may be barriers to their attendance at different CPD activities due to factors such as timetabling issues, cost of travelling to CPD events, or a lack of suitable events being organised (Zhang et al., 2020). It is unsurprising then, that the presence of barriers to attending quality CPD activities has also been linked to lower levels of job satisfaction (Renbarger and Davis, 2019).

Teacher Cooperation

Higher levels of cooperation between teachers and staff within schools has been identified as a strong correlate of job satisfaction in previous research (Lopes and Oliveira, 2020). Examples of factors contributing to high levels of cooperation within a school could include team teaching, observation of other teachers' classes, or sharing of teaching materials and resources (OECD, 2018). In fact, analysis of international data from TALIS 2013 which includes teachers from England, has shown teacher cooperation to be the most significant predictor of job satisfaction when accounting for other working conditions and teacher characteristics (Sims, 2017). Similar trends have also been found in Swedish data from TIMSS 2015, where cooperation has been identified as one of the strongest predictors of job satisfaction (Toropova et al., 2021). Although often seen as positive, teamwork can also have negative effects. Interview studies with Norwegian teachers, for example, have found that teamwork can sometimes be a source of stress, and disagreements can arise when teachers are unable to choose who they collaborate with (Skaalvik and Skaalvik, 2015).

In addition to having a positive effect on job satisfaction, teacher cooperation

has been linked to lower levels of teacher turnover in the US (Nguyen, 2021), where teachers reporting higher levels of cooperation were found to be less likely to want to leave their current school. However the same higher levels of cooperation were not associated with lower probabilities of teachers wanting to leave the teaching profession entirely.

Other Factors

In our data, a public school is defined as any school managed by a public education authority, government agency, municipality, or governing board appointed by government or elected by public franchise (OECD, 2018). Previous studies have found that job satisfaction is typically higher in private schools than in public schools. This discrepancy, however, is often attributed to the differing levels of autonomy (Lopes and Oliveira, 2020), or positive relationships with management (Sönmezer and Eryaman, 2008) which may be present in these two types of schools. Therefore, one might not expect to see a significant difference in the job satisfaction of public and private school teachers when controlling for these variables. Despite this, studies which have attempted to control for important policy, individual, and workplace level characteristics have still found significantly higher levels of job satisfaction in private schools (Small, 2020). The effect on job satisfaction of teaching in a public vs. a private school is therefore an open question.

While larger class sizes and larger student teacher ratios have often been shown not to have a large effect on student achievement (Woessmann and West, 2006; Li and Konstantopoulos, 2017), a clear connection between class size and job satisfaction has not been established. One interview study of 200 teachers in the US, for instance, found that class size was one of the top 3 reasons reported by teachers as justifications for their current levels of job satisfaction (Perrachione et al., 2008). Other studies, however, have found that class size is not a major driver of American or Japanese teacher job satisfaction when controlling for other working conditions (Reeves et al., 2017).

A second factor which is less commonly examined in relation to teacher job satisfaction is the practice of out-of-field teaching. Out-of-field teaching has been linked to lower student achievement in a number of studies (Dee and Cohodes,

2008; Hill and Dalton, 2013), but the literature available on the effects that out-of-field teaching has on job satisfaction is quite limited. Olmos (2010) and Provasnik and Dorfman (2005) found that out-of-field teachers in the US were more prone to attrition, though other studies have not found as substantial an effect (e.g. Shen, 1997).

Finally, one additional factor which has not been the subject of much research in relation to teacher job satisfaction is contract-type. Our search for studies relating factors associated with the terms of a teacher's employment and their job satisfaction returned few results. Furthermore, those studies which we did find were not focused primarily on terms of employment, but instead used it as one of a variety of control variables, and results have varied across researchers. One investigation of the effect of personal characteristics on teacher job satisfaction, for example, found teachers with permanent contracts to be less satisfied on average (Gil-Flores, 2017). In contrast, Capone and Petrillo (2020) found teachers with permanent contracts to have higher levels of job satisfaction and well-being. Other studies which have investigated the effects of part-time or full-time contracts have revealed no discernible changes in job satisfaction (e.g. Ferguson et al., 2012).

3.2 Methods

3.2.1 Data and Pre-Processing

This study uses English data from TALIS 2018 (OECD, 2019a) which provides us with a representative sample of 4385 primary and lower secondary school teachers (2009 primary, 2376 lower secondary). Each observation includes more than 30 scales describing various teacher and school characteristics such as self-efficacy, participation in CPD, and perceived cooperation among staff. The individual survey responses upon which these scales are based are also provided, as well as personal and background details for each of the teachers such as gender; school level; qualification; and years' experience. A full list of all variables used can be found in Appendix A.3. A description of how we handled missing data in these variables can be found at the end of this section.

The main variable of interest in this study is teacher job satisfaction. Teacher

job satisfaction in the TALIS data is based on the responses of teachers to eight items which gauge a teacher's overall contentment and happiness with their current working environment and profession. All eight questions share a common stem which reads "We would like to know how you generally feel about your job. How strongly do you agree or disagree with the following statements?". An example item for measuring satisfaction with the working environment is "I enjoy working at this school", and an example item for satisfaction with the profession is "The advantages of being a teacher clearly outweigh the disadvantages". Possible responses to these items lie on a 4 point Likert scale, with options ranging from strongly disagree (1) to strongly agree (4). The ordinal responses to these items have been converted into a continuous measure of job satisfaction by the organisers of the TALIS study using an approach called confirmatory factor analysis. This continuous variable is the outcome we will use in our study. Confirmatory factor analysis is a very widely used approach in the social sciences (e.g. McInerney et al., 2018; Saloviita and Pakarinen, 2021). The organisers of TALIS have also conducted a number of tests to ensure the reliability and validity of the constructed teacher job satisfaction scale (OECD, 2019b). The resulting job satisfaction scale (after combining primary and lower secondary school teachers) has a mean of 12.42, and a standard deviation of 2.28.

To ensure a representative sample during the data collection stages of TALIS, a stratified two-stage probability sampling design is employed within each country. Each teacher in the TALIS dataset is therefore assigned a set of weights to rigorously estimate population parameters of interest and their associated standard errors. These sampling weights were fully incorporated into our analysis through the Balanced Repeated Replication (BRR) procedure described in the TALIS technical report (OECD, 2019b). The resulting confidence intervals are presented in Figure 3.3. It is important to note that while our primary estimation method, Bayesian Additive Regression Trees (BART), is inherently Bayesian and provides posterior distributions for the estimated treatment effect, the application of Fay's BRR to construct confidence intervals represents a departure from a fully Bayesian framework. Specifically, Fay's BRR is a frequentist resampling method, and its use with BART estimates forgoes direct reliance on the Bayesian posterior distribution for uncertainty quantification. We adopted this approach based on

the TALIS data organiser's recommendation to appropriately reflect the complex sampling design of the TALIS data. However, an alternative method possible in future research would be to directly incorporate the hierarchical structure of the sampling design within the Bayesian framework itself, using a multilevel (hierarchical) Bayesian model. This would enable both proper accommodation of the sampling design and full retention of Bayesian uncertainty quantification.

Data from the survey can be missing for a number of reasons. Some teachers did not reach every question, and others did not answer personal questions such as those concerning their age. Of the variables we have used, 52 contained missing values, with on average 8% of the data missing. In order to maximise the data available for use, we have imputed these missing responses with the R package `missRanger` (Mayer, 2019). This method substitutes missing values with predictions based on an individual's responses to all other questions in the survey. This approach allows us to retain information that would otherwise be discarded if missing cases were excluded and offers greater accuracy than simpler methods, such as imputing with the mean value (Stekhoven and Bühlmann, 2012).

However, this method assumes that the data are missing at random (MAR) (Donders et al., 2006), an important consideration when interpreting the results. A limitation of this approach is that the uncertainty associated with imputing missing values is not captured in the main analysis. Therefore, the true 95% confidence intervals are likely to be slightly wider than reported. To address this limitation, more sophisticated techniques, such as multiple imputation (Rubin, 1996), imputation with added noise (Gold and Bentler, 2000), or treating missingness as a useful predictive feature of the data (Kapelner and Bleich, 2015), could be employed. Incorporating these methods into the analysis would be an excellent area for future work, but was outside the scope of the current study.

3.2.2 Limitations of Traditional Statistical Approaches

Commonly used approaches in international large scale assessments to investigate the relationship between a set of independent variables, X, and a dependent variable, Y, include ordinary least squares regression and other more sophisticated modelling approaches such as multilevel models. These approaches are very use-

ful but have a number of limitations. Firstly, they assume a linear relationship between each independent variable and the outcome of interest. This can lead to biased parameter estimates in some cases and can lead one to believe that there is no relationship between two variables when in fact there is. For example, the relationship between teacher attrition and age has been found to be U-shaped in a number of different studies (Guarino et al., 2006; Boe et al., 1997). Such limitations can be addressed within the framework of generalised linear models by incorporating higher order polynomial terms or interactions, but these features must be specified explicitly by the user, opening up the potential for model misspecification, especially if the functional form of the response is complex or unknown.

Thirdly, linear models can become difficult to interpret when a large number of covariates have been included as explanatory variables. This means that it can be difficult to control for a large number of factors simultaneously when investigating the association of one variable of interest with another while still maintaining the required interpretability. Consequently, researchers often limit their analysis to a smaller subset of the available data. However, not controlling for some variables may bias parameter estimates.

In Section 3.3 we have concentrated on factors which relate to measures that school principals or other policy makers could introduce immediately with the view to improving job satisfaction levels. By contrast, much of the existing literature on teacher job satisfaction uses scale scores of different psychological constructs which have been validated using approaches such as confirmatory factor analysis (e.g. McInerney et al., 2018; Saloviita and Pakarinen, 2021). Such approaches are certainly useful because, for example, they have demonstrated a link between higher levels of teacher self-efficacy and job satisfaction. Teachers' levels of self-efficacy are not easily changed however, and so these results do not provide a directly implementable process that can be used to improve job satisfaction or the outcome of interest.

3.2.3 Limitations of The Study Design

In addition to the modelling limitations discussed above, there are also inherent constraints in the nature of the study design. Large-scale assessments such as the one used in this study are typically cross-sectional and observational, meaning they capture data at a single point in time without random assignment or experimental manipulation. As a result, causal inferences cannot generally be made from such data unless specific methodological strategies are employed, such as adjusting for all relevant confounders to satisfy the assumption of no unmeasured confounding, or by using quasi-experimental techniques like instrumental variable estimation or difference-in-differences designs.

Furthermore, the directionality of observed associations is often unclear. For example, while teacher self-efficacy has often been assumed to be an antecedent of teacher job satisfaction, recent evidence suggests that the causal relationship may actually be the opposite (Burić and Kim, 2021). These ambiguities underscore the importance of caution when interpreting findings, especially in the context of making policy recommendations.

While the relatively large sample size provided by the TALIS data used in this study helps to mitigate concerns related to statistical power, it remains an important topic when analysing educational data. Some common rules of thumb, for example, recommend having at least 10 to 20 observations per predictor variable to ensure sufficient power in traditional regression analyses. Within the Bayesian framework, which includes methods such as BART and BCF, the emphasis tends to shift away from hypothesis testing and power in the frequentist sense, toward estimation and reliable uncertainty quantification, such as the use of credible intervals and posterior distributions. This distinction is discussed well in Kruschke and Liddell (2018) and Kruschke (2014).

An interesting Bayesian analogue to frequentist power analysis involves conducting simulation studies in which synthetic datasets are generated under assumed conditions. For each simulated dataset, it is assessed whether certain criteria such as obtaining a sufficiently narrow credible interval or achieving a posterior probability above a predefined threshold are met. The proportion of simulations satisfying these conditions provides a Bayesian measure of prospective study per-

formance, which can inform decisions about sample size and study design. While this type of simulation based power analysis was not performed in the present study given that we are working with publicly available, pre-collected data, it represents a valuable approach for future work, especially in the context of primary data collection and experimental design.

3.2.4 Bayesian Additive Regression Trees for Causal Analysis

With the above considerations in mind, this study aims to investigate the effect of a number of binary factors, which we call treatments, on teacher job satisfaction. Our approach will be to use the R package `bartCause` which is a causal inference machine learning package for the R programming language (Dorie and Hill, 2020; R Core Team, 2021). The `bartCause` package allows us to estimate causal effects, and has been demonstrated to be highly competitive in causal inference machine learning competitions (Dorie et al., 2019). The package owes its success to the impressive prediction capabilities of Bayesian Additive Regression Trees (BART), a Bayesian non-parametric modelling tool which is well suited to a wide variety of problems (Chipman et al., 2010). This flexibility has inspired a large number of BART extensions, such as a recent paper which shows how BART based models can be adapted for use in a mediation analysis setting (Linero and Zhang, 2022).

BART is known as a sum of trees model which can flexibly and accurately predict an outcome of interest Y using a set of covariates X. It can be seen as an extension of regression modelling that automatically identifies interactions and non-linear relationships between the variables. In the case of a single tree model, BART makes predictions by establishing a set of decision rules which when followed, assign a prediction to each observation. See Figure 3.1 for an example of a decision tree.

Bayesian methods are becoming increasingly popular in educational research (König and van de Schoot, 2018). In particular, a recent study has used BART to estimate the causal effects of private tuition on student achievement (Suk et al., 2021). BART has also been used extensively in other fields outside of education, and is a popular choice for many quantitative researchers (e.g. Prado et al., 2021b).

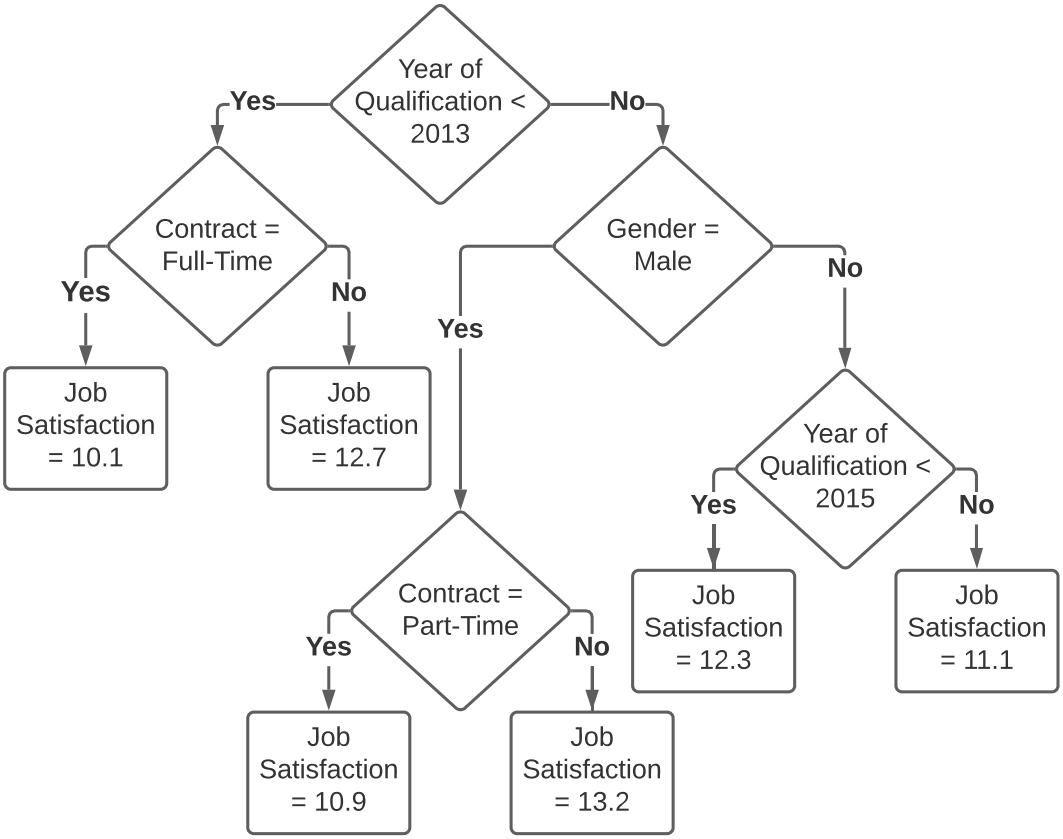

Figure 3.1: Example of a single decision tree for the TALIS data. Each teacher's information can be fed into the tree by following the decision rules. The terminal nodes provide the predictions for the job satisfaction of each teacher. In practice the BART model works by creating many different decision trees and summing the predictions together.

3.2.5 Treatment Effect Estimation

To estimate the causal effect of a given intervention on the outcome variable we adopt the Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Rubin, 1974; Sekhon, 2008). Central to the Neyman-Rubin causal model is the concept of potential outcomes which posits that there are two potential outcomes for each individual i, one that would be observed under treatment, $y_i(1)$, and one that would be observed under control, $y_i(0)$ (no treatment). The individual treatment effect would then be given by the difference between these potential outcomes: $\tau_i = y_i(1) - y_i(0)$. Observing individual i simultaneously under both treatment and control is impossible, however, and this is known as the fundamental problem of causal inference.

Estimation of τ is a difficult task, especially in the case of observational data. Challenges posed by observational data to estimating causal effects include the fact that individuals are not randomly assigned to the treatment and control groups, and that our observation of the data may not include all variables which have an influence on the outcome of interest or the non-random assignment mechanism. It is, however, possible to identify causal effects with a number of key assumptions (Kurz, 2022). These assumptions include:

(1) The stable unit treatment value assumption (SUTVA). It requires that the treatment status of any individual i should not affect the potential outcomes of any other individual j. It also requires that there should be "no multiple versions" of the same treatment, meaning that the treatment must be consistently defined for all individuals.

(2) The ignorability assumption. This requires that the potential outcomes of individual i must be independent of their treatment status conditional on their observed covariates. In other words, we require there to be no confounding variables we have not observed.

(3) The overlap assumption. This requires that every individual must have a non-zero probability of being assigned to both treatment conditions.

Assuming the above assumptions hold, the `bartCause` package estimates treatment effects with BART by predicting the two potential outcomes for each individ-

ual, using their observed characteristics and an indicator of whether or not they received treatment as predictor variables. Using these BART derived estimates of the potential outcomes, Hill (2011) showed that the conditional average treatment effect (CATE) for observation i can then be estimated as $\hat{\tau}_i = \hat{y}_i(1) - \hat{y}_i(0)$. We then compute the average of these individual-level CATEs over the sample, weighted by the sampling weights, yielding what Li et al. (2023) refer to as the mixed average treatment effect (MATE):

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i = \frac{1}{N}\sum_{i=1}^{N}\hat{y}_i(1) - \hat{y}_i(0).$$

Earlier, we identified several limitations of traditional modelling approaches, such as their reliance on linearity assumptions, difficulties in inferring causal direction from observational data, and challenges in managing complex models with many covariates. The BART approach used in this chapter can help us to tackle each of these concerns. Firstly, BART does not assume linear relationships and can flexibly capture complex, non-linear associations and interactions between variables, reducing the risk of biased estimates arising from model misspecification. Secondly, provided the three key assumptions listed above hold, the causal inference framework based on estimating potential outcomes under treatment and control conditions should allow us to estimate causal effects from the data. Finally, with automatic variable selection and the ability to handle a large number of covariates, the BART model maintains interpretability without the need to exclude important confounding factors, thus mitigating the potential for omitted variable bias.

3.2.6 Including Propensity Scores in Causal Models

Following the advice of Hahn et al. (2020), we will include an additional independent variable as a predictor in our model. This additional variable is known as the propensity score, and is defined as an individual's probability of being assigned to the treatment group. This probability can be estimated from an individual's characteristics such as their gender, year of qualification, degree type etc. Logistic regression is a common choice for this task, but we have chosen to use BART

instead to keep our approach as consistent as possible and retain the superior predictive approach.

The inclusion of the propensity score has been shown to improve the estimation of treatment effects (Hahn et al., 2020). Besides this practical advantage, it can also be interesting to look at different trends in the propensity scores for individual teachers. Analysing such trends allows us to identify, for example, which subgroups of teachers are particularly likely to belong to positive or negative treatment groups. This process can identify specific subgroups of teachers who need to be given extra support, or who would benefit from being assigned to a particular treatment group. We highlight an example of this in our next section.

In addition to serving as a control variable in the BART model, the propensity score is crucial for assessing the overlap assumption, which must hold for valid causal inference. Overlap was evaluated by visually inspecting density plots of the propensity score distributions for treated and control units to ensure no regions lacked common support. Similar visual checks were performed for all continuous covariates, while for categorical variables, contingency tables were used to confirm that no levels contained only treated or control units. It is important to note, however, that this assessment can only be applied to observed variables included in the propensity score model. Unobserved confounders, if present, may still violate the overlap assumption in ways that cannot be detected through these methods, so the ignorability assumption is still required.

3.2.7 Choice of Treatment Variables

We calculate mixed average treatment effects for each of the following treatment options (short names or abbreviations used in Figures are shown in brackets):

(1) Did the teacher take part in at least 4 CPD activities in the past year (CPD)?

(2) Did they take part in a formal/informal induction programme when they started teaching at their current school (Induction)?

(3) Do they take part in observing other teachers (Observing)?

(4) Do they take part in team teaching (Team Teaching)?

(5) Do they have a mentor (Has Mentor)?

(6) Are they a mentor to another teacher (Is Mentor)?

(7) Do they teach in a publicly managed school (Public School)? (Full definition in Appendix A.2)

(8) Do they have ≥ 30 students in their class (30+ Students)?

(9) Are they an out-of-field teacher (Out-of-field)?

(10) Do they have a part-time contract (Part-Time)?

In each of the cases above, the MATE is estimated independently of the other treatments. The set of predictor variables included in X remains unchanged, as we control for the same covariates in every assignment option (with a few exceptions). For an exact definition of each of these treatments see Appendix A.2. Appendix A.3 identifies any variables which were removed from X for a specific assignment option. For example, it would be inappropriate to control for the number of students in a class when investigating the effect of teaching a class with ≥ 30 students (Variable Code: TT3G38).

We note that by investigating multiple factors affecting teacher job satisfaction, there is an increased likelihood of introducing Type 1 errors or false positives. Common methods for controlling Type 1 errors or False Discovery Rates include the Bonferroni correction or Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). However, these method are accompanied by an increased risk of Type 2 errors. As discussed in the literature review, the decision to investigate each of the factors under consideration was founded on prior research and literature. This should help to reduce the risk of spurious discoveries, but the increased risk of false positives must still be borne in mind when interpreting the main results. As an alternative to these conventional methods, we also highlight a later extension of the Bayesian Causal Forests (BCF) model, developed in Chapter 5, which allows for the joint modeling of multiple factors. This extension may offer a more robust solution to the multiple comparisons issue, accounting for the interdependencies between the treatment variables. For a more detailed discussion of this approach, see Chapter 5.

Figure 3.2 shows the control and treatment group sizes for the different factors that we have created and are investigating. The control group size for CPD is 1618, meaning that 37% (unweighted) of teachers in the sample did not take part in 4 or more CPD events over the course of the past year. The treatment group size for this assignment option is 2767, corresponding to a 63% participation rate in at least 4 CPD events. The other segments of the plot have similar interpretations.

As can be seen from Figure 3.2, 30% of teachers met our criteria for teaching out-of-field. A more in depth analysis of these numbers reveals that 24% of secondary school teachers meet this criteria, and 37% of primary school teachers do. Further investigations also show that the subjects being taught out-of-field by teachers are different across the two school levels. We bring this point to the reader's attention to make clear that these teachers are all treated identically, and we do not make careful distinctions between reasons for teaching out-of-field. Furthermore, we do not distinguish between primary vs. secondary school teachers for this treatment effect (or indeed any of the other treatment effects).

Before continuing, we acknowledge that this decision may appear at odds with the distinct educational contexts and student age groups that primary and post-primary teachers engage with. However, this approach is based on evidence from the data preparation process undertaken by the TALIS organisers, who conducted measurement invariance testing across school levels. For the teacher job satisfaction construct used in this study, scalar invariance was established, indicating that the measure operates equivalently for both primary and post-primary teachers (OECD, 2019b). This provides empirical justification for pooling the two groups in our analysis. Nevertheless, we recognise this as an important limitation and return to this point in the discussion.

3.3 Results

This section describes the results from:

(1) Choosing a treatment assignment option to consider from the list in Section 3.2.7.

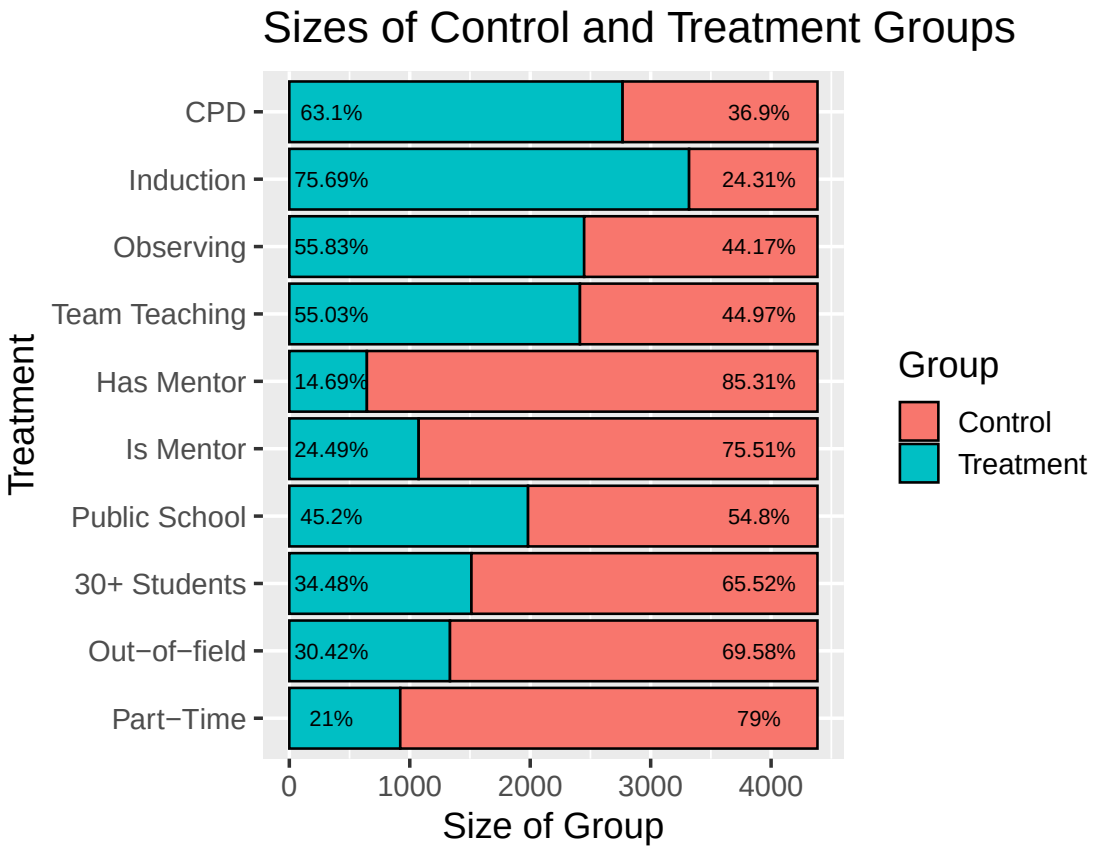

Figure 3.2: Percentage of teachers belonging to the control and treatment groups under investigation. There are different levels of balance across the groups.

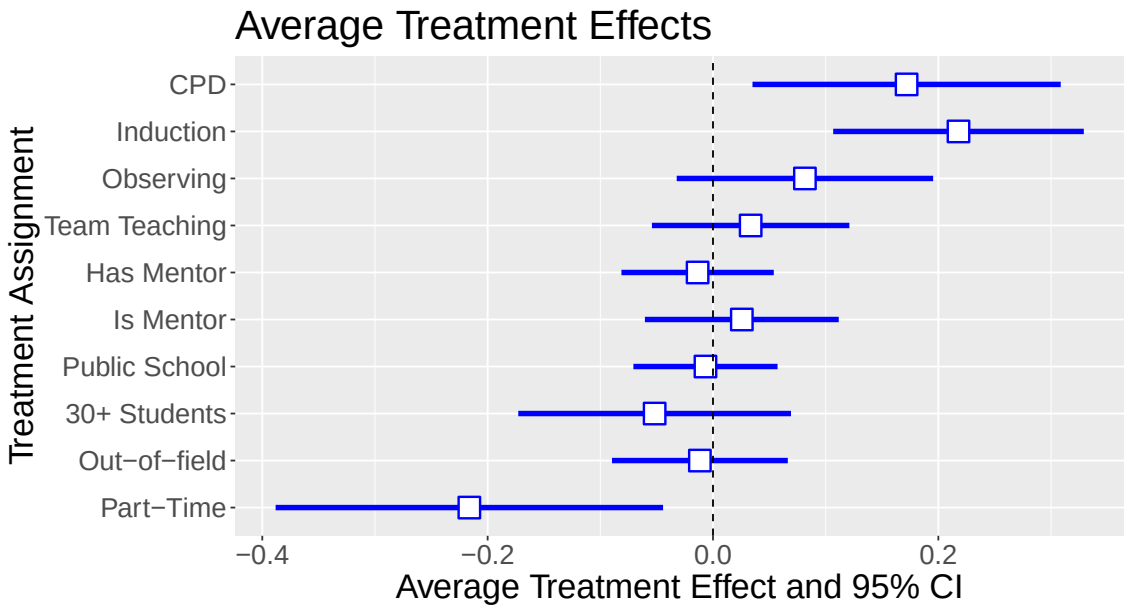

Figure 3.3: Plot of mixed average treatment effects for each treatment under investigation. The central box of each error bar represents the best available point estimate, while the full extent of the error bars represents a 95% confidence interval.

(2) Estimating the mixed average treatment effect of this assignment option on job satisfaction.

For a visual representation of these results, see Figure 3.3 which indicates the final estimate and 95% confidence interval for each of the treatment effects. Diagnostic tests were also performed for all models fitted to the data to ensure convergence had been reached, and goodness of fit statistics were calculated to ensure satisfactory predictive performance.

3.3.1 Continual Professional Development

Our results identify participation in at least 4 CPD events over the course of a year as having a positive effect on teacher job satisfaction. The 95% confidence interval for this mixed average treatment effect is [0.035, 0.309]. To give an idea of the magnitude of this treatment effect, consider that the teacher job satisfaction scale has a mean of 12.42, and a standard deviation of 2.28. Therefore, the centre-point

of this confidence interval which is at 0.172 would correspond to an increase in job satisfaction of 0.08 standard deviations, which is a small but positive improvement.

3.3.2 Induction and Mentoring Programmes

Our results show that taking part in induction when starting at a new school has a positive effect on job satisfaction. The 95% confidence interval for the mixed average treatment effect on job satisfaction is [0.107, 0.329]. Therefore, taking part in an induction scheme is associated with a mean increase in job satisfaction of 0.218 meaning that induction schemes are the most beneficial of all of the treatment assignment options we have considered.

Mentoring, however, is not identified as having a strongly positive effect. As can be seen from the 95% confidence intervals in Figure 3.3, this is true for both mentors and mentees.

3.3.3 Observation and Team Teaching

Team teaching and observing the lessons of other teachers are both identified as having a positive effect on job satisfaction. The uncertainty in these estimates is quite large however, and this is reflected in the wide 95% confidence intervals shown in Figure 3.3 which both include zero within their range. Given the large confidence intervals it may be that there are large effects of these variables, but the data here do not provide us with enough information to estimate them precisely. Alternatively there may be sub-groups for whom the estimated effect is particularly high or low. This, however, would also be difficult to ascertain with a high degree of statistical confidence.

3.3.4 Other Factors

Of the remaining factors we considered, the treatment assignment option with the largest effect in relation to job satisfaction is the possession of a part-time contract of less than 90% of a typical full-time contract's hours. This factor has the effect of reducing job satisfaction on average by 0.216, 95% confidence interval [-0.388, -0.044]. The results from analysing the propensity scores for this factor show an interesting trend. Figure 3.4 shows that experienced female teachers generally

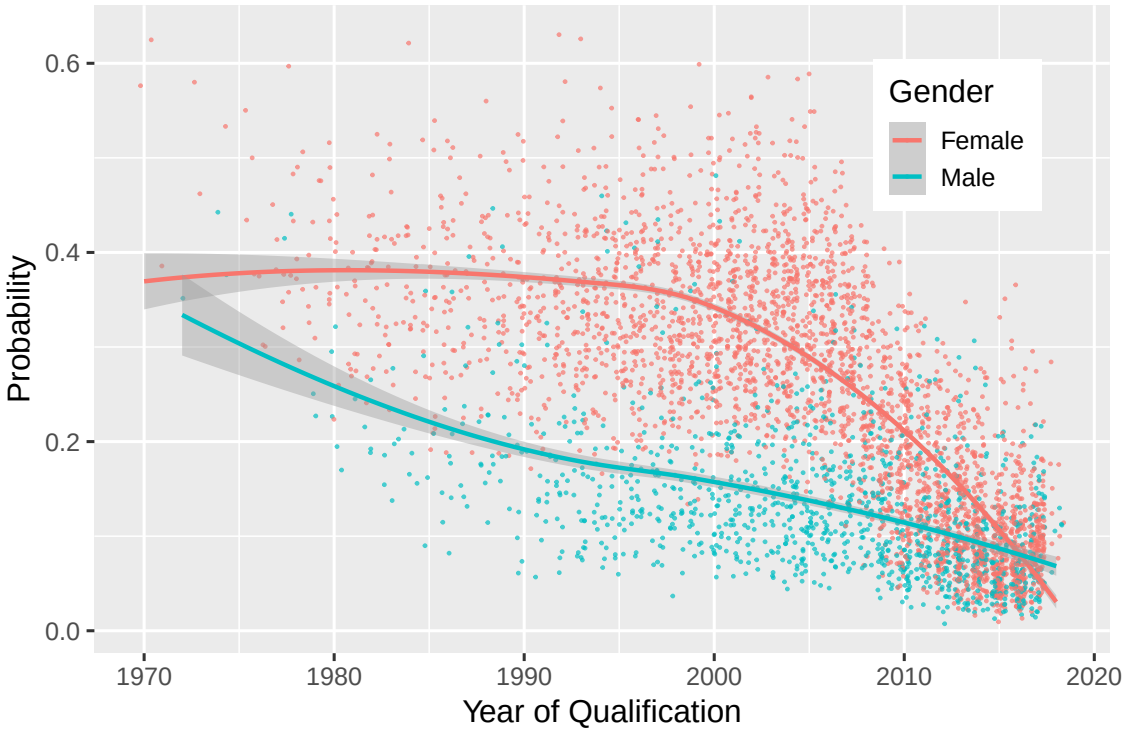

Figure 3.4: Probability of having a part-time contract. Female teachers have higher probabilities than male teachers, especially more experienced female teachers.

have much higher propensity scores (probability of being assigned to the treatment group) than their male colleagues.

The other factors we have considered are out-of-field teaching, working in a public vs. a private school, and teaching a class with 30 or more students. According to our results, these factors are not associated with a strong effect on job satisfaction. We emphasise again that these factors may indeed be very important, but the precision with which the data allows us to estimate these effects is insufficient to make such claims with a high degree of statistical confidence in this case.

3.4 Discussion

We begin by discussing our main findings in more detail, and go on to highlight some key aspects of this study which make a new and important contribution to the literature on teacher job satisfaction. We finish this section by drawing the reader's attention to some limitations of this study, and by suggesting areas for future research.

3.4.1 Main Findings

Continual Professional Development

Our results identify high levels of participation in CPD as having a positive effect on teacher job satisfaction. This is in agreement with multiple studies which have found a strong correlation between CPD and job satisfaction (e.g. Yoon and Kim, 2022; Wang et al., 2020; Sims, 2017). Crucially, our result supports these previous findings by verifying the strong positive effects of CPD using a causal inference approach, and thus we are able to infer results about causation and not just correlation. Furthermore, as job satisfaction is known to be an important predictor of teacher intentions to remain teaching (Madigan and Kim, 2021), our results also support recent findings from studies of teachers based in England which have linked CPD to higher levels of retention (e.g. Coldwell, 2017; Allen and Sims, 2017). In addition, we have ensured that our treatment effect estimates are as unbiased as possible, by removing the effect of possible confounding variables on our outcome of interest.

We noted that only 63% of teachers in the English dataset have reached this high level of CPD. Barriers to participation in CPD are known to be a key predictor of job satisfaction (Zhang et al., 2020). Our results therefore also provide strong support for this body of work by demonstrating the positive gains that can be made by removing such barriers and encouraging and enabling more teachers to engage in CPD events. For this reason, the emphasis on the importance of engagement with CPD in the Early Career Framework in England is very welcome (Department for Education, 2019). In addition, the inclusion of a 10% and 5% reduction in timetabled teaching hours for teachers in their first and second years, in order to

enable them to fully avail of the supports and training offered during this time is likely to be crucial.

We highlight the fact that our investigation has only considered a binary version of CPD. In reality, however, levels of attendance at CPD belong on a spectrum, not just high/low. Furthermore, the benefits from CPD are likely to depend on many factors such as the quality and relevance of the training to a teacher's needs. These factors warrant further investigation but were beyond the scope of this study. Despite this, we do find clear evidence in favour of recommending CPD as a measure for improving job satisfaction.

Induction and Mentoring Programmes

Our finding that induction schemes have a very positive effect on job satisfaction agrees with prior work from Ronfeldt and McQueen (2017). Contrary to the review by Allen (2005), we did not detect high levels of heterogeneity in the treatment effect estimates of this assignment option. The recent introduction of the Early Career Framework in England which includes mandatory induction for new teachers is therefore an excellent step forward, but we argue that induction schemes should also be made available more generally for all new teachers at a school, regardless of number of years qualified or experience in the classroom.

Unlike some previous studies (e.g. Ingersoll and Strong, 2011; Renbarger and Davis, 2019), our results do not identify the presence of a mentor as being beneficial for job satisfaction. There are a number of plausible reasons for this. Firstly, there may be some unobserved or unaccounted for confounding variables common to schools with mentorship schemes which bias the estimates of these analyses. Secondly, we did not consider other aspects related to mentoring, such as the subject area of the mentor. Research has shown that a mentee is more likely to benefit from a mentoring arrangement if their mentor is a teacher from the same grade level (Parker et al., 2009). Other factors such as the mentoring quality and the frequency of meetings can also be important (Richter et al., 2013). The provision of training for mentors taking part in the Early Career Framework to improve mentoring quality is therefore commended.

An indicator of whether or not a teacher is currently a mentor to another

member of staff was also included as a component in our analysis. Similarly, we did not find that this treatment was associated with an appreciable increase or decrease in job satisfaction. Again, this could be a result of our binary view of mentoring relationships, in which we only consider the presence or absence of a mentee, and fail to account for other aspects such as the quality of the mentoring relationship, which has been demonstrated to be an important predictor of job satisfaction (Lunsford et al., 2018).

Observation and Team Teaching

The fact that we have not found a clear link between team teaching or observation with job satisfaction may initially appear to be strange. The literature reviewed consistently pointed towards higher levels of teamwork and cooperation as having a positive effect on teacher job satisfaction. Therefore, we might have expected to see this reflected in our results also.

One plausible explanation for this is that higher levels of teamwork and cooperation within a school are difficult to attribute to a small number of specific practices such as team teaching and observation. Higher levels of teamwork and cooperation within a school are characterised by many different aspects such as sharing resources with colleagues and collaborating together on different projects etc. As a result, it is difficult to capture the true impact of higher levels of teamwork and cooperation as a whole by only considering two of a much larger number of indicators. Therefore, the absence of a large effect size here does not necessarily mean that team teaching and observation are not useful practices. Rather, the results indicate that only implementing one or two of these factors is unlikely to yield significant improvements in job satisfaction, and efforts should instead be focused on improving teamwork and cooperation as a whole. This is made clear by the very small treatment effect sizes that result from us considering two of these such practices in isolation.

Other Factors

We investigated whether working at a publicly owned and managed school affects job satisfaction. The results from our approach do not identify a significant

causal effect for this treatment assignment. This result is in line with work by Dahler-Larsen and Foged (2018) who attribute the difference in job satisfaction between public and private schools to differences in organisational characteristics, as opposed to the ownership of the school.

In line with research by Reeves et al. (2017), our results show that teaching a class with 30 or more students does not have a large effect on job satisfaction. We should note, however, that our finding is based on a cut-off point of 30 students. This value was chosen to ensure an approximately even split of teachers in the treatment and control groups. It is possible, however, that a different value would yield different results, and teachers at the more extreme end of the distribution with greater than 35 students may experience a more negative effect from this treatment.

Given the lack of research linking out-of-field teaching to job satisfaction we thought it was important to include this as a factor in our study. The magnitude of the treatment effect that we have obtained in our results for this factor is very small, but out-of-field teaching is a complex phenomenon (Hobbs and Törner, 2019), so it is reasonable to expect that the effects of teaching out-of-field may be dependent on a number of contextual factors such as how dissimilar the subject being taught is to one's area of expertise. A more detailed investigation of the effects of out-of-field teaching on job satisfaction is therefore warranted.

As in the study by Ferguson et al. (2012), the contract-type used in our study refers to full-time or part-time contracts. We have chosen this as it will allow us to have more evenly balanced control and treatment groups. Our results show that teachers on a part-time contract are less satisfied with their career than their full-time colleagues. Also, an analysis of the propensity scores for this treatment effect shows that experienced female teachers have much higher probabilities of being on part-time contracts than their male counterparts. Future research should investigate the reasons for this, and supports that might be put in place for teachers with childcare responsibilities.

In summary, of the factors that we have investigated, we have found that high levels of participation in CPD and induction schemes have the strongest positive influence on job satisfaction. Conversely, we have also found that possessing a part-time contract can have a negative effect on job satisfaction. In the case of

the other treatments we have studied, despite the mixed average treatment effects often pointing in the direction we had expected, there was not enough certainty in these estimates to claim the presence of a clear causal effect.

3.4.2 Contribution of This Study

We believe this study makes three main contributions to the current literature on teacher job satisfaction. The first is that we have employed a causal inference machine learning approach, bringing the power of advanced statistical modelling techniques to an important problem in the world of education. One advantage of this approach is the ability to flexibly model job satisfaction without assuming a linear relationship between the predictor variables and the outcome of interest which is a common feature of most conventional statistical models. This approach is also well suited to detecting interactions between variables and allows us to include a much wider variety of covariates than would normally be possible when using linear models. This is absolutely crucial, because it enables us to model the response surface using the propensity score along with a large number of other variables, thus accounting for many potential sources of confounding which could otherwise bias treatment effect estimates.

Second, instead of identifying important characteristics related to a teacher's working environment such as cooperation, quality of school leadership, or personal traits such as self-efficacy, we have established several specific and implementable measures that may be introduced in an attempt to improve job satisfaction. We summarise our findings with the following recommendations:

(1) Our results provide strong evidence that participation in an induction scheme when starting at a school can have a beneficial effect on teacher job satisfaction. We therefore recommend that schools not currently offering such schemes should endeavour to introduce them. We also recommend that schools currently offering induction schemes should encourage participation from all new staff, including experienced and novice teachers.

(2) We also find strong support for higher levels of participation in continual professional development having a positive effect on job satisfaction. Therefore

we suggest that school authorities should make it a priority to identify and remove any barriers to staff attendance at CPD events, whilst also ensuring a regular calendar of relevant CPD activities are available for attendance.

(3) Our finding that part-time contracts are negatively impacting on the job satisfaction of affected teachers warrants a closer examination of how concerns about job security may be affecting teachers. The fact experienced female teachers are disproportionately more likely to be on a part-time contract also requires a review into the supports that may be put in place for teachers with young families.

Specific recommendations are important because although it may be known that certain factors such as stress are negatively correlated with job satisfaction (Klassen and Chiu, 2010), it is not always obvious how best to reduce stress levels among teachers, or if a set of proposed changes will have the desired effect. This study therefore avoids this pitfall by identifying factors such as induction schemes which can be beneficial for job satisfaction, while also identifying the negative effects of factors such as part-time contracts.

Finally, the propensity scores described in Section 3.2.6, although not the primary focus of this study, provide us with an interesting insight into the types of teachers more likely to belong to the treatment and control groups we have investigated. This can help us to identify certain subgroups of teachers who have not availed of positive treatments, and we can then ensure that these activities are made available to them. This can also help us to identify subgroups of teachers who are more likely to be exposed to the negative effects of a treatment, such as experienced female teachers who we found were significantly more likely to have a part-time contract.

3.4.3 Limitations and Areas for Future Research

As discussed in the methodology, the causal inference approach that we have employed makes a number of important assumptions. Among these is the ignorability assumption, which requires that we have accounted for all potential sources of confounding when investigating a given treatment. Despite including a wide variety

of control variables in our design matrix, X, it is certainly still possible that there may be some confounding variables not collected as part of the survey. Teachers with young children for example, may be more likely to work part-time, but there is no indication in the TALIS data whether teachers have young children. Future research could include a detailed assessment of the reasonableness of these assumptions in relation to TALIS by incorporating data from external sources, and using different diagnostic methods designed to assess these assumptions.

A second limitation concerns the cross-sectional nature of the TALIS data. Because all variables were measured at the same point in time, it is not always possible to determine whether covariates included in the model are truly pre-treatment. This concern is relevant, for example, in the analysis of the effect of CPD on teacher job satisfaction, where teacher self-efficacy was included as a potential confounder. While self-efficacy may influence a teacher's decision to participate in CPD, it is also possible that the self-efficacy measure in TALIS reflects confidence gained after CPD participation. If this is the case, self-efficacy functions as a mediator rather than a confounder, meaning the estimated effect in this study should be interpreted as the direct rather than the total causal effect of CPD on job satisfaction.

A third limitation of our approach is that some aspects of the working environment such as teamwork and cooperation are very difficult to capture with binary variables. Therefore, it may be less meaningful to investigate binary factors in relation to aspects such as this, because levels of teamwork and cooperation can not be fully characterised by a simple dummy variable. Also, hours of CPD attended and the number of students in a class are both continuous variables. Therefore, their impact on job satisfaction can not be fully appreciated by artificially converting them into a binary factor. Additional studies using causal inference machine learning methods designed to handle continuous treatment variables may be better suited to this task.

A further limitation relates to the decision to analyse primary and post-primary teachers together, without disaggregating by school level. This choice was informed by measurement invariance testing conducted by the TALIS organisers, which established scalar invariance for the teacher job satisfaction construct across school levels (OECD, 2019b). This indicates that the measure functions equivalently for

both primary and post-primary teachers, allowing for meaningful comparisons and pooled analysis. Nonetheless, we acknowledge that this analytical decision may obscure some important contextual differences, and that treatment effects could function differently across teacher subgroups. Therefore, future research could focus on exploring interaction effects to examine how specific treatments impact primary and post-primary teachers in distinct ways.

Additionally, while the Bayesian Additive Regression Trees (BART) method for causal inference introduced by Hill (2011) was adopted in this study, an alternative method, Bayesian Causal Forests (BCF) (Hahn et al., 2020), could have also been considered. However, given BCF extends BART with additional complexity, BART was selected as a more accessible and appropriate starting point in this first study. Future research could also explore the application of BCF or other advanced causal inference methods to assess whether they yield different insights or improved estimation accuracy.

Finally, as our results are based on data which only includes teachers from England, we can not claim that the same treatment effects would be observed in other countries and cultures, where other factors may be more important for improving the job satisfaction of teachers. The application of a similar approach to ours, but to different countries in the TALIS data is therefore a promising area for future research.

3.5 Conclusion

Faced with increasing demand for qualified teachers in England and internationally, it is of the utmost importance to identify strategies for improving teacher job satisfaction. This can help to encourage higher retention rates of qualified teachers, and attract new entrants to start their career. Many studies which investigate factors associated with job satisfaction, however, instead of identifying specific and implementable measures for achieving this task, link higher levels of job satisfaction to positive working environments or higher levels of self efficacy. Our study has tackled this issue by employing a causal inference machine learning approach to investigate the effect of a number of treatments on job satisfaction. We encourage school management teams and educational administrations to take

note of our results which further support the provision of induction schemes for new teachers, and continual professional development for all staff. We also recommend an examination of how part-time contracts may be causing anxiety around job security and satisfaction for teachers. More generally, we advocate for further research into the specific steps that may be taken for improving job satisfaction through the use of causal inference methods.

4

Bayesian Causal Forests for Multivariate Outcomes: Application to Irish Data From an International Large Scale Education Assessment

4.1 Introduction

Estimating heterogeneous treatment effects from observational data is a complex yet essential task. The pursuit of precise and tailored interventions, informed by an understanding of how individuals uniquely respond to treatments, holds profound implications for many different fields. In this pursuit, we encounter two critical challenges. The first is the ever present issue of confounding, which arises in causal settings where Randomised Controlled Trials (RCTs) are not feasible. The second challenge lies in detecting often subtle variations in individual responses to treatment. To address these hurdles, advanced modelling strategies are required, capable of flexibly controlling for confounding variables while guarding against overfitting. As a result, there has recently been a surge of interest in applying advanced non-parametric methods to tackle these challenges (see Caron et al., 2022a, for a review).

This family of methods includes a vast assortment of techniques which exploit the predictive capabilities of advanced regression models in order to estimate het-

erogeneous treatment effects. A key strength of this family of methods is that it is very flexible, and many of the relevant techniques can be performed with virtually any suitable regression model. One of the most important contributions to this area was made by Hill (2011) who demonstrated that by using a sufficiently flexible regression model such as Bayesian Additive Regression Trees (BART, Chipman et al., 2010), it is possible to accurately estimate treatment effects. A second influential contribution was made by Hahn et al. (2020) who built on Hill's work by using Robinson's (1988) treatment effect parameterisation to separate the estimation of Y into a prognostic effect μ, and a treatment effect τ. This approach, named Bayesian Causal Forests (BCF), has a number of advantages over that of Hill (2011) as it allows separate priors to be applied to the μ and τ components of the model, and enables individual level treatment effects to be estimated directly from the data.

Due to its flexibility and impressive predictive performance, the Bayesian Causal Forest model has become one of the most popular causal inference methods available and has been the subject of multiple research papers, many of which have extended its capabilities for use in different settings. Caron et al. (2022b), for example, introduced a prior in the model which encourages BCF to focus more on important predictor variables than less influential covariates. In another recent extension, Starling et al. (2021) introduced a BCF model designed for estimating treatment effects that exhibit smooth variations across a single covariate, drawing inspiration from a BART model with a similar design outlined in Starling et al. (2020). Other extensions of the BCF model include a hierarchical version which was used to investigate the effectiveness of a growth mindset in improving student achievement in a study by Yeager et al. (2019). Our model described later will also include a hierarchical element in its application to the motivating dataset, but this will not be the primary focus of our study.

An important limitation of many causal inference methods, including Bayesian Causal Forests, is that they are only applicable to a single outcome variable subject to a binary treatment Z. Therefore, motivated by data from the Trends in International Mathematics and Science Study (Mullis et al., 2020), which includes data on both the mathematics and science achievement of eighth grade (approximately 14 - 15 year old) secondary school students, we present a multivariate

extension of BCF which is capable of estimating the causal effect of an intervention on multiple outcomes simultaneously. With our new approach, we consider the effect of a number of home-related factors on student achievement. Specifically, we attempt to answer the following three research questions: 1) "What effect does having access to a study desk at home have on student achievement in mathematics and science?", 2) "What is the impact on student achievement in mathematics and science of often arriving at school feeling hungry?", and 3) "What effect does regular absence from school have on student achievement in mathematics and science?". We investigate these factors because they have important implications for student-focused initiatives such as free school meals programmes and back to school allowances which are designed to assist students from disadvantaged backgrounds (Taras, 2005; Kennedy, 2013).

The main advantage of our multivariate approach is a potentially substantial reduction in the uncertainty associated with the causal parameters, since the model now has access to extra information through the correlation of the outcome variables and the treatment effects. Situations in which the model might exploit a strong positive correlation in the effects of an intervention could occur for many plausible reasons. To illustrate, imagine an intervention aimed at enhancing students' problem-solving skills, specifically as measured by the TIMSS cognitive domains of "Applying" and "Reasoning" (Mullis and Martin, 2017). As these skills improve, it would be reasonable to anticipate a concurrent improvement in TIMSS mathematics and science achievement scores, as both subjects require students to apply learned concepts to novel contexts and engage in logical reasoning to solve complex tasks. These cognitive processes are not entirely domain-general, of course, but share a degree of similarity such that improvements in one area could support gains in the other. Consequently, by focusing on the narrower definition of problem solving as captured by the TIMSS assessments, rather than invoking a broad or abstract notion of STEM-general problem solving, we would expect to observe a positive correlation between the magnitude of improvement in mathematics achievement and the corresponding improvement in science achievement.

Conversely, the model might also exploit a strong negative correlation in situations where one of the outcome variables serves as a mediator for another. For

instance, in an intervention focused on reducing stress levels among students, it might be observed that lowered stress levels act as a mediator by enhancing students' productivity. If the intervention significantly reduces a student's stress levels, this may be associated with a substantial increase in productivity. Therefore, this scenario demonstrates a negative correlation in treatment effects, as a larger reduction in stress levels corresponds to a more significant increase in productivity.

As an additional advantage of the proposed multivariate BCF model, we also note that by jointly modelling all outcome variables of interest simultaneously, the model provides a natural way to account for the dependence between these outcomes. By estimating the effect of each factor on all outcomes simultaneously, the model may also prove useful in situations where the multiple comparisons problem is of concern.

Whilst multivariate causal inference models are rare, our approach shares similarities with that of a recent multivariate extension of Bayesian Factor Analysis models for causal inference which demonstrates the potential for a multivariate approach to causal inference (Samartsidis et al., 2020). Also of note is a multivariate random forest based method for causal inference developed by Guo et al. (2021). Our work is different from these because the multivariate causal Factor Analysis model developed by Samartsidis et al. uses a very different structure to our BART based model. We believe our approach offers greater flexibility and may be used in a much wider variety of settings. Also, the multivariate random forest based method developed by Guo et al. (2021) requires RCT data with pre and post intervention covariates. Our model, which is designed for observational data, does not impose such requirements.

In addition to multivariate causal inference models, there is also precedent for shared tree structures to assist in the estimation of multivariate parameters in the Bayesian context. The first example of this to our knowledge is a study by Linero et al. (2020) which demonstrated the impressive performance gains made possible by this approach. Related to this work is a paper by Um et al. (2023) which develops a multivariate BART model for skewed distributions, which also highlighted the advantages of using a shared tree structure in a regression problem. The key difference which separates these previous works from our study, however, is that in contrast to the paper by Um et al. (2023) which focused on estimat-

ing a multivariate target variable Y, or the paper by Linero et al. (2020) which focuses on estimating multiple transformations of, or parameters associated with, a single outcome variable, we are interested in estimating the causal effect of an intervention on multiple outcomes.

The remainder of this chapter is organised as follows: In Section 4.2 we give some background on the Trends in International Mathematics and Science Study, the dataset motivating our multivariate approach. Section 4.3.1 describes Bayesian Additive Regression Trees, the model providing the foundation upon which Bayesian Causal Forests are built. Section 4.3.2 explains how BCF leverages the impressive predictive capabilities of BART for the purpose of estimating heterogeneous treatment effects, and Section 4.3.3 details the modifications necessary to extend BCF to the multivariate setting. In Section 4.4 we present the results of a simulation study, in which we demonstrate the substantial benefits of jointly modelling all outcome variables available. In Section 4.5 we apply our multivariate extension of BCF to the motivating dataset, TIMSS 2019. Here, we investigate the effects of a number of treatments on student mathematics and science achievement, including home study supports, being hungry at school, and absenteeism. We conclude this chapter with a discussion of our results, the limitations, and potential avenues for future research.

4.2 Trends in International Mathematics and Science Study

The Trends in International Mathematics and Science Study (TIMSS) is a large scale international assessment organised by the International Association for the Evaluation of Educational Achievement (IEA). It has taken place in many countries across the world every four years since 1995, with 64 countries participating in TIMSS 2019. As part of the study, students in the fourth and eighth grade of secondary school (typically aged approximately 10 - 11 and 14 - 15 respectively) are given a short assessment in mathematics and science, which is used to estimate their overall achievement level. The eighth grade students also complete a short background questionnaire on topics such as their home and classroom environ-

ment, and how much they like and feel confident in these subjects. The teachers and principals of these students are also given short questionnaires on their educational background, teaching practices, and school access to learning resources, thus providing us with a large number of covariates to control for as potential confounding variables. This makes TIMSS an excellent source of information for researchers investigating factors associated with student confidence and achievement in mathematics and science.

Due to its scale and comprehensive nature, TIMSS data has been the subject of many studies in the field of education since its origin in 1995. Some recent studies using data from TIMSS include Tang et al. (2022) who investigate the impact of science teacher continual professional development on student achievement in science, and Chen (2022) who considers the effect of the interaction between classroom and individual achievement levels on student confidence in mathematics. Our focus in this chapter however will be on the effect of three specific treatments on student achievement in mathematics and science. In contrast to much of the existing literature which focuses on typically just one of these outcomes, we will model achievement in both subjects jointly.

In this study, we delve into the causal factors associated with a student's household environment, specifically examining the influence of home study supports, hunger at school, and absenteeism. These factors possess the potential to exert significant impacts on a student's educational journey in various ways. Notably, they are all susceptible to the influence of confounding variables. Home study supports and access to educational resources are well-established predictors of student achievement, as evidenced by prior research (Tsai and Yang, 2015). However, given their likely correlation with socioeconomic status, it is imperative to untangle this relationship within a causal framework. The detrimental effects of students attending school on an empty stomach are widely acknowledged (Vik et al., 2022). Hunger can hinder concentration and deprive students of essential nutrients, underscoring the importance of free school meal programmes (Taras, 2005). Yet, the prevalence of hunger may also be linked to a student's socioeconomic background, necessitating careful consideration as a potential confounding factor. Lastly, absenteeism has been consistently linked to lower academic achievement (Vesić et al., 2021). However, absenteeism itself is a multifaceted issue intertwined with other

adverse factors like school bullying (Bennour, 2021; Ladd et al., 2017), making it crucial to account for these variables (and others) collected as part of the TIMSS study.

In summary, TIMSS is an excellent source of information for researchers in the field of education. TIMSS data has been used extensively to answer many important research questions over the years, but as the short discussion above highlights, there can often be multiple layers of complexity with the potential to bias the estimates of these analyses. Furthermore, much of the existing research has focused solely on achievement in one subject, employing traditional approaches such as multiple linear regression models which are not well suited to answering questions of a causal nature. For this reason we propose that a multivariate causal approach, capable of flexibly accounting for the many confounding variables that may be present, is well suited to these data.

4.3 Bayesian Non-Parametric Estimation of Heterogeneous Treatment Effects

One of the fastest growing areas of research in the causal inference literature is the application of Bayesian non-parametric machine learning methods for the estimation of heterogeneous treatment effects. Before discussing these approaches in detail, however, we must first cover some notation. In this study we will adopt the Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Rubin, 1974; Sekhon, 2008) which can be applied to situations where we are interested in the effect of a treatment Z on an outcome Y. The Neyman-Rubin causal model is based on the concept of potential outcomes, which asserts that for each observation i, there are two potential outcomes: one that would be observed under treatment $y_i(Z_i = 1)$, and one that would be observed under control, $y_i(Z_i = 0)$. Knowing both $y_i(Z_i = 0)$ and $y_i(Z_i = 1)$ would allow us to calculate the individual treatment effect for unit i, $\tau_i = y_i(Z_i = 1) - y_i(Z_i = 0)$. This is of course impossible, because we only ever observe one of the potential outcomes, and this is known as the fundamental problem of causal inference.

Although we may not observe both potential outcomes directly, we can esti-

mate them with $\hat{y}_i(Z_i = 0)$ and $\hat{y}_i(Z_i = 1)$. Then, in the presence of the correct conditions, we may estimate τ_i with $\hat{\tau}_i = \hat{y}_i(Z_i = 1) - \hat{y}_i(Z_i = 0)$. More generally, we may also estimate the conditional average treatment effect (CATE), $\hat{\tau}(x_i)$, at any covariate x_i with $\hat{y}(x_i, Z_i = 1) - \hat{y}(x_i, Z_i = 0)$. Calculating the average of these CATEs over the population of interest then yields what Li et al. (2023) refer to as the mixed average treatment effect (MATE):

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\tau}(x_i) = \frac{1}{N}\sum_{i=1}^{N}\hat{y}(x_i, Z_i = 1) - \hat{y}(x_i, Z_i = 0).$$

The conditions which are required to hold for the reliability of this approach in the univariate context are provided by Kurz (2022). However, for the sake of generality, given our focus on multivariate outcomes, we have extended and adapted these assumptions to accommodate the complexities introduced by multiple outcome variables:

(1) **The stable unit treatment value assumption (SUTVA).** The original assumption requires that the potential outcomes of any individual i must not be affected by the treatment status of any other individual j. For example, if student j is often absent from school, this must not influence the achievement level of any other student i. The SUTVA also requires that there are "no multiple versions" of the same treatment. In other words, there should not be multiple potential outcomes corresponding to different versions or types of the same treatment.

Multivariate Modification: To align with the multivariate context, we now stipulate that this must hold for all outcome variables associated with student i.

(2) **The ignorability assumption.** Also known as the unconfoundedness assumption, we require that there must be no residual confounding we have not controlled for: $y_i(Z_i = 1), y_i(Z_i = 0) \perp\!\!\!\perp Z_i | x_i$.

Multivariate Modification: In our multivariate adaptation, we extend this to each outcome variable y_k by requiring that $y_{i,k}(Z_i = 1), y_{i,k}(Z_i = 0) \perp\!\!\!\perp Z_i | x_i \ \forall i, k$.

(3) **The overlap assumption.** This requires that the propensity score for any individual i must be bounded away from zero and one: $0 < P(Z_i = 1|x_i) < 1$. For example, if it was true that students from disadvantaged backgrounds were guaranteed never to have a study desk, this would be a violation of the overlap assumption.

Multivariate Modification: This assumption remains unchanged in the multivariate adaptation as the same treatment applies to all outcome variables.

The above provides us with a very flexible approach for estimating individual treatment effects, as $y_i(Z_i = 1)$ and $y_i(Z_i = 0)$ can be estimated with any sufficiently accurate model f. A good choice for f, for a number of reasons, is Bayesian Additive Regression Trees, and this is what we will discuss next.

4.3.1 Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) is a Bayesian non-parametric machine learning algorithm that is well suited to a variety of regression and classification tasks (Chipman et al., 2010). BART can be described as a tree based ensemble method for predicting an unknown function $f(X)$ based on the contributions of many shallow trees. Individually these trees act as weak learners, each only explaining a small part of the unknown function, but when combined they are able to capture very complicated relationships and interactions between variables in the data. Owing to its impressive predictive performance, BART has become popular with researchers from many disciplines and has been used for a diverse range of applications in many fields such as medicine, economics, and education (Pierdzioch et al., 2016; Sparapani et al., 2016; McJames et al., 2023b, Chapter 3). BART is a very flexible model, which has enabled researchers to adapt or modify the underlying algorithm for various specialised use cases such as genomics, problems with local linearities and, of course, causal inference (Sarti et al., 2023; Prado et al., 2021a; Hill et al., 2020; Carnegie et al., 2019; Dorie et al., 2022).

Given an outcome variable y of length n, and a covariate matrix X consisting of n observations of d variables, the BART model can be written as follows:

$$y_i = \sum_{j=1}^{J} g(T_j, M_j, x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

where $g()$ is a function which calculates the individual contribution of each tree j of J total trees. M_j specifies the terminal node parameters associated with the j^{th} tree T_j. The residuals, ϵ_i, are assumed to be normally distributed with mean 0 and variance σ^2. Being a Bayesian model, appropriate priors are required for T_j, M_j and σ^2.

The BART model is fitted using Markov Chain Monte Carlo. Before discussing the intricacies of this procedure in detail, however, we will first consider how BART can serve as the foundational framework for a powerful causal machine learning approach: the Bayesian Causal Forest (BCF) model.

4.3.2 Bayesian Causal Forests

Bayesian Causal Forests is an advanced causal inference machine learning algorithm (Hahn et al., 2020). BCF uses BART as a foundation for estimating causal effects and shares the same desirable features such as impressive predictive performance, careful regularisation through the use of Bayesian priors, and uncertainty quantification. BCF does however have a number of advantages over BART for estimating heterogeneous treatment effects, and this is made possible by adopting the Robinson parameterisation which expresses the outcome y as:

$$y_i = \mu(x_i, \hat{\pi}_i) + \tau(x_i) Z_i + \epsilon_i$$

where $\mu()$ and $\tau()$ are both BART ensembles which work together to estimate two separate parts of the model: a prognostic effect μ, which can be thought of as the expected outcome under control when the treatment variable Z is coded as 1 for treatment, 0 for control, and a treatment effect τ, which can be interpreted as the impact on y of receiving treatment. The additional covariate $\hat{\pi}_i$ included in the $\mu()$ part of the model is the propensity score, which is simply the estimated probability of individual i receiving treatment: $\pi_i = P(Z_i = 1)$. The inclusion of the propensity score in $\mu()$ is important for avoiding a phenomenon called regularisation induced confounding, and is especially useful in situations where the

likelihood of receiving treatment is in some way related to the expected outcome under control (Hahn et al., 2020). We will therefore include an estimate of $\hat{\pi}_i$, obtained using a BART model, in all experiments in this study.

The parameterisation above has a number of important benefits associated with it. First, it allows different amounts of regularisation to be applied to the μ and τ parts of the model. It is common to apply greater regularisation to τ than to μ because we expect the degree of heterogeneity in the treatment effects to be relatively simple in comparison to y itself. This prior belief can be incorporated into the model by encouraging shallower trees in the τ ensemble. The complexity of the τ component of the model can be further reduced by using a smaller number of trees to estimate τ than for μ. Secondly, if it is known that only a subset of the variables in X are responsible for moderating the effect of Z on y, then it is possible to use a different set of covariates in $\mu()$ and $\tau()$. Finally, as τ is now an explicit part of the model, it is possible to make direct inference on the treatment effects with BCF, and this provides a more straightforward interpretation of the model.

With our understanding of Bayesian Additive Regression Trees (BART) and Bayesian Causal Forests (BCF) now in place, we can turn our attention to the main focus of this chapter: a multivariate extension of BCF which we will see has a number of important advantages over the standard univariate model.

4.3.3 Multivariate Bayesian Causal Forests

Motivated by data from TIMSS which includes information on student achievement in both mathematics and science, we now extend the BCF algorithm to the multivariate setting. This extension allows us to estimate the causal effect of a given intervention on two or more outcomes jointly, and thus we are able to improve our predictions by taking advantage of the correlation between, and the shared information across all outcome variables. With our new setup, the BCF model specification becomes:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}(x_i, \hat{\pi}_i) + \boldsymbol{\tau}(x_i) Z_i + \boldsymbol{\epsilon}_i$$

Algorithm 1 Bayesian Backfitting MCMC Algorithm for Multivariate BCF

Require: Multivariate target variable Y (n rows of p outcome variables), Feature matrix X (n rows of d covariates),
Treatment variable Z (length n; 1 for treatment; 0 for control)

Ensure: Posterior list of trees T, residual covariance matrix Σ, multivariate fitted values $\hat{\mu}$, multivariate fitted values $\hat{\tau}$

Set hyperparameter values of α_μ, β_μ, σ_μ, σ_τ, α_τ, β_τ, ν, Σ_0
Set number of μ trees N_μ, number of τ trees N_τ, number of iterations N
Set initial value $\Sigma = I$, and set all μ trees and τ trees to stumps with terminal node parameters set to 0
for iterations i from 1 to N **do**
  for μ trees j from 1 to N_μ **do**
    With τ predictions fixed at their current values: compute multivariate partial residuals $R_{\mu,j}$ from Y minus predictions of all trees except μ tree j
    Grow a new tree $T_{\mu,j}^{new}$ based on grow/prune/change/swap
    Accept/Reject tree structure with Metropolis-Hastings step using $P(T_{\mu,j}|R_{\mu,j},\Sigma) \propto P(T_{\mu,j})P(R_{\mu,j}|T_{\mu,j},\Sigma)$
    Sample μ values from multivariate normal distribution using $P(M_{\mu,j}|T_{\mu,j},R_{\mu,j},\Sigma)$
  end for
  for τ trees k from 1 to N_τ **do**
    With μ predictions fixed at their current values: Compute multivariate partial residuals $R_{\tau,k}$ from Y minus predictions of all trees except τ tree k
    Grow a new tree $T_{\tau,k}^{new}$ based on grow/prune/change/swap
    Accept/Reject tree structure with Metropolis-Hastings step using $P(T_{\tau,k}|R_{\tau,k},\Sigma) \propto P(T_{\tau,k})P(R_{\tau,k}|T_{\tau,k},\Sigma)$
    Sample τ values from multivariate normal distribution using $P(M_{\tau,k}|T_{\tau,k},R_{\tau,k},\Sigma)$
  end for
  Combine predictions from all trees to get $\hat{Y} = \hat{\mu} + \hat{\tau}Z$
  Update Σ with an Inverse-Wishart distribution using $P(\Sigma|\hat{Y})$
end for

where $\boldsymbol{Y}_i$ is a length p vector representing the i^{th} observation of the p dimensional outcome variable $\boldsymbol{Y}$, $\boldsymbol{\mu}(x_i)$ and $\boldsymbol{\tau}(x_i)$ represent the i^{th} predictions from the multivariate prognostic and treatment effect functions, $\boldsymbol{\epsilon}_i$ is the i^{th} residual, and Z_i is the familiar treatment status indicator: 1 for treatment, 0 for control.

The MCMC algorithm for fitting this model is based on that of BART, and uses a combination of Gibbs sampling and Metropolis-Hastings steps. Specifically, the contribution provided to the model by $\boldsymbol{\mu}(x_i)$ is based on an ensemble of J_μ multivariate Bayesian Additive Regression Trees. Similarly, the contribution provided to the model by $\boldsymbol{\tau}(x_i)$ is based on an ensemble of J_τ multivariate Bayesian Additive Regression Trees. The process begins by updating each of the J_μ trees belonging to the $\boldsymbol{\mu}(x_i)$ part of the model. The structure of the j^{th} such tree is updated at each iteration by choosing at random one of four possible operations to propose a new updated tree; grow, prune, change, or swap. If grow is selected, then a splitting rule is assigned to a randomly chosen terminal node which then becomes the parent of two children. If prune is selected, then a parent of two terminal nodes is chosen at random, and its children are removed from the tree. During the change operation, an internal node is chosen at random and its splitting rule is replaced with a new randomly chosen split rule. Finally, the swap operation selects a parent-child pair which are both internal nodes, and swaps their splitting rules with each other.

To prevent any member of the ensemble from growing too large, a prior $P(T_j)$ is placed on the structure of the j^{th} tree. This prior specifies that the probability of any node at depth d being non-terminal is given by $\alpha(1+d)^{-\beta}$. Therefore, for a tree T_j with terminal nodes $h_{j,1}...h_{j,K}$, and non-terminal nodes $b_{j,1}...b_{j,L}$, we have that:

$$P(T_j) = \prod_{k=1}^{K} \alpha(1 + d(h_{j,k}))^{-\beta} \prod_{l=1}^{L} [1 - \alpha(1 + d(b_{j,l}))^{-\beta}]$$

where $d()$ is a function for returning the depth of an arbitrary node, and α and β are hyper parameters which can be adjusted to place a higher probability on the preferred tree depth. The purpose of this prior is to encourage more shallow trees, thus restricting the amount of variance any one tree can explain, and helping to avoid overfitting. Note that with this prior, we allow all outcome variables to share

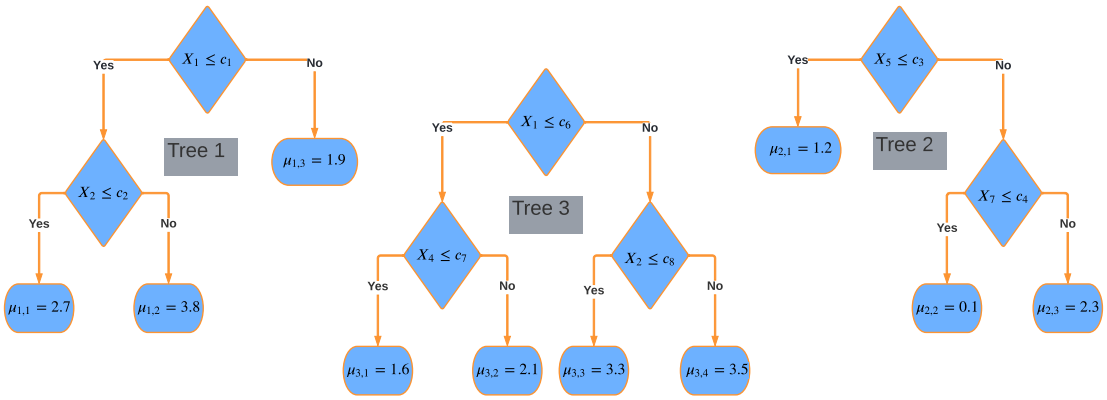

Figure 4.1: Diagram of a BART model with three decision trees as part of the ensemble. The predictions for an observation are given by following the decision rules from the root to the terminal nodes of the trees and summing the individual contributions together. For example, for an observation with $X_1 > c_1$, $X_5 < c_3$, $X_1 < c_6$ and $X_4 < c_7$, the final prediction would be given by $1.9 + 1.2 + 1.6 = 4.7$.

the same tree structure. This is made appropriate by our motivating dataset, as we expect our chosen covariates will predict both outcomes in a similar way. As a result, the algorithm is encouraged to prioritise decision rules that will contribute positively towards accurately estimating all components of $\boldsymbol{Y}$. This helps to avoid over-fitting and acts as a type of regularisation, improving predictive performance.

With the structure of the j^{th} tree defined, the decision rules at each node form a pathway directing observations to the leaves of the tree. See Figure 4.1 for an example. Terminal node parameters $\boldsymbol{\mu}_{j,k}$ are now assigned to each of the K leaves of the j^{th} tree, responsible for providing a small but important contribution to the final prediction made by the model. To safeguard against any individual trees becoming unduly influential in $\boldsymbol{\mu}(x_i)$, and to ensure that the scale of the $\boldsymbol{\mu}$ parameters is sufficient to cover the whole of the observed data, a multivariate normal prior is placed over both $\boldsymbol{\mu}$ and $\boldsymbol{\tau}$:

$$\boldsymbol{\mu}_{j,k} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mu} = \sigma_{\mu}^2 \boldsymbol{I}\right), \quad \boldsymbol{\tau}_{j,k} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_{\tau} = \sigma_{\tau}^2 \boldsymbol{I}\right)$$

With the columns of $\boldsymbol{Y}$ scaled during data pre-processing to follow a standard normal distribution, a sensible choice for the hyper parameter σ_{μ}^2 is $1/J_{\mu}$, which places a high prior probability over the range of all observed $\boldsymbol{Y}$ values. Depending

on prior beliefs regarding the scale and heterogeneity of the treatment effects, a reasonable choice for σ_τ^2 is likely to be smaller. In our experience, however, model performance is typically quite insensitive to the choice of prior when it comes from a sensible range of values - see the supplementary material for some experimentation. The combination of priors above allows the likelihood used in the Metropolis-Hastings step to be calculated in closed form as a multivariate normal distribution summed across terminal nodes. The data enter this multivariate normal distribution as partial residuals calculated from the response minus the predictions of the other trees that are not being updated.

When the terminal node parameters for tree T_j have all been sampled, a newly selected grow/prune/change/swap operation is applied to tree T_{j+1} and the process is repeated until all J_μ trees belonging to the $\boldsymbol{\mu}(x_i)$ part of the model have had their terminal node parameters updated. At this point, analogous updates are applied to each of the J_τ trees belonging to $\boldsymbol{\tau}(x_i)$. At the end of each iteration the combined contribution from all trees is subtracted from $\boldsymbol{Y}$ to calculate the residuals and the residual covariance matrix $\boldsymbol{\Sigma}$ is updated. The above process repeats for a pre-specified number of iterations and the end result is a posterior distribution of trees, terminal nodes, and $\boldsymbol{\Sigma}$ parameters.

The conjugate prior we have used for the residual covariance matrix is an Inverse-Wishart distribution:

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}\left(\nu, \boldsymbol{\Sigma_0}\right)$$

Our decision to employ the Inverse-Wishart prior in this study was motivated by practical considerations. Specifically, we utilised an RCPP (Eddelbuettel and François, 2011) implementation of the Inverse-Wishart distribution, which seamlessly integrated into our existing codebase. However, an important alternative in this context is the LKJ prior introduced by Lewandowski, Kurowicka, and Joe (Lewandowski et al., 2009). The LKJ prior has been demonstrated to offer several advantages over the Inverse-Wishart prior in certain scenarios, particularly with regard to its potential to exert less influence on the posterior distribution of the covariance matrix — a desirable trait. Despite this, our experimentation and testing with the Inverse-Wishart prior did not reveal any serious concerns. Nevertheless,

the LKJ prior presents an interesting avenue for further exploration.

For an algorithmic summary of the steps described above the reader is referred to Algorithm 1. Additionally, full derivations of the updates for each of the parameters in the model can be found in the supplementary material.

We believe our multivariate extension of Bayesian Causal Forests introduces a number of important advantages over previous approaches. Firstly, by sharing the same tree structure, the predictions made by the model are able to benefit from any correlation between the outcome variables. In this way, we expect to see greater predictive performance as information can be shared across predictions for all outcome variables. We might expect this advantage to be especially beneficial in settings where there are few observations because if there is a lack of information relating to one of the outcome variables at a specific region of the covariate space, the information available for the other outcome variables can help to guide the predictions. Finally, by jointly modelling all outcome variables we can estimate additional model parameters related to the residual covariance between different outcomes which may be of interest in certain applications.

However, it is also important to note that there may also be some disadvantages to multivariate approaches. While in the majority of applications we expect the outcome variables of interest to be strongly related in some way, this may not always be the case. As a result, if the ideal tree structure for estimating one of the outcome variables is quite different to the ideal tree structure of the other outcome variables, the model may struggle to find a structure suitable for estimating all outcome variables of interest. Though unlikely, it is clear that situations such as this may pose challenges for our model which imposes the same tree structure on all outcome variables.

An additional causal inference specific challenge that may be plausible is the situation that arises when the assumptions of the model have been met for one of the outcome variables, but not the other. For example, if a set of covariates X_1 to X_P are sufficient to control for all sources of confounding impacting an outcome Y_1, but there is an additional unobserved covariate X_U that introduces confounding in Y_2, then the unconfoundedness assumption of the model has been met for Y_1 but not for Y_2. Clearly, in this situation the treatment effect estimates for Y_2 should be biased, but perhaps due to the shared dependency of Y_1 and Y_2 on the same

tree structure, some otherwise avoidable bias might be introduced into Y_1.

While we expect situations such as this to be rare, it is important to be aware of these possible drawbacks. In the next section we explore these possibilities and their impact on model performance based on a number of simulation studies.

4.4 Simulation Studies

In this section we present evidence of the advantages and improved predictive performance of the new multivariate BCF approach. We do this via a simulation study in which we have compared the performance of our multivariate implementation of BCF with a univariate BCF model, and a univariate BART based approach from Hill (2011). To further investigate the advantages of multivariate modelling in heterogeneous treatment effect estimation we also compare our model with the multivariate BART model of Um et al. (2023), using the same approach described by Hill (2011). We compare the four different approaches above on three different data generating processes, all based on a modified version of the first Friedman dataset (Friedman, 1991), which is a commonly used benchmarking dataset as it provides a complicated non-linear pattern with complex interactions. The data generating processes are as follows:

Data Generating Process 1: With this data generating process we aim to test the performance of multivariate BCF in a setting that we believe is almost ideally suited to the model. The prognostic effect μ and the treatment effect τ use the same variables and follow a very similar functional form. We generate 10 covariates X_1 to X_{10}. Variables X_1 to X_5 are uniformly distributed random variables over the range zero to one. Variables X_6 to X_8 are random Bernoulli variables with $P(X = 0) = P(X = 1) = 0.5$. Variables X_9 and X_{10} are ordinal categorical variables with equal probability of being zero, one, two, three, or four. Our decision to incorporate a mixture of uniform, Bernoulli, and categorical variables here is to match the type of variables our model will encounter in the real life TIMSS data in the next section. The functional form for Y_1 and Y_2 is given in Table 4.1, where $\epsilon_{1,i}$ and $\epsilon_{2,i}$ come from a multivariate normal distribution with mean 0, covariance matrix $50^2 I$ and $P(Z_i = 1) = x_{4,i}$ which makes X_4 a confounding variable. The coefficients and the signal to noise ratio above have also been chosen to ensure the

scale of the data and treatment effects approximately matches what we expect in the TIMSS data. In the case of the signal to noise ratio, this was chosen by fitting standard univariate BART models to both outcome variables from the real TIMSS data, both of which resulted in a residual standard deviation of approximately 50.

Data Generating Process 2: We investigate the performance of multivariate BCF when the assumptions of the model have been met for Y_2 but not for Y_1. We make a minor modification to DGP1 by setting the coefficient of X_4 in μ_2 to zero, and we remove X_4 from the set of covariates available to the model. This means that Y_2 is not affected by any residual confounding, as X_4 has no impact on Y_2, but Y_1 is affected by unobserved confounding as we hide X_4 from the model and X_4 does have an impact on Y_1. We therefore expect that all models will perform poorly in estimating τ_1 but they should not encounter any difficulties in estimating τ_2. We anticipate that this degradation in performance will be isolated to Y_1 and that the multivariate BCF estimates for τ_2 will not be affected in any way.

Data Generating Process 3: We investigate if the multivariate BCF model can successfully adapt to scenarios that require different tree structures for both of the outcome variables Y_1 and Y_2. We make the following two changes to DGP1: 1) For μ_1, we replace $110\sin(\pi x_1 x_2)$ with $110\sin(\pi x_4 x_5)$. This ensures that Y_1 and Y_2 experience different interactions, and the tree structure will need to adapt to this. 2) For τ_1 we replace $20x_4 + 20x_5$ with $20x_4 + 20x_2$, and replace τ_2 with $10x_3 + 30x_5$. This ensures the tree structure responsible for estimating the treatment effects also will need to change. We expect the flexible tree prior $P(T_j)$ to allow the trees to grow slightly larger if necessary, and to adapt to the greater flexibility that will be required in this situation.

In the results that follow, we have generated 3000 synthetic data sets for each data generating process, 1000 with a training sample size of 1000 observations, 1000 with a training sample size of 500 observations, and 1000 with a training sample size of 100 observations. The test set contains 1000 observations in every case. For multivariate and univariate BCF we have used 50 trees in the ensemble for estimating the prognostic effect μ, and 20 trees in the ensemble for estimating the treatment effect τ. For the BART and multivariate BART approach we have used a total of 70 trees to estimate y. A total of 500 iterations were run for both the pre and post burn-in stages of model fitting. In each simulation we have

Functional Form of Y_1 and Y_2 in Data Generating Process 1
$y_{1,i} = \underbrace{(300 + 110\sin(\pi x_{1,i}x_{2,i}) + 180(x_{3,i} - 0.5)^2 + 100x_{4,i} + 120x_{6,i} + 10x_{9,i})}_{\mu_1(x_i)} + \underbrace{(20x_{4,i} + 20x_{5,i})}_{\tau_1(x_i)} Z_i + \epsilon_{1,i}$
$y_{2,i} = \underbrace{(300 + 90\sin(\pi x_{1,i}x_{2,i}) + 220(x_{3,i} - 0.5)^2 + 140x_{4,i} + 80x_{6,i} + 10x_{9,i})}_{\mu_2(x_i)} + \underbrace{(10x_{4,i} + 30x_{5,i})}_{\tau_2(x_i)} Z_i + \epsilon_{2,i}$

Table 4.1: Functional Form of Y_1 and Y_2 in Data Generating Process 1. Note that π above refers to the mathematical constant and not the propensity score described earlier.

fitted the multivariate models to both outcome variables Y_1 and Y_2, and we have fitted the univariate approaches to outcome variables Y_1 and Y_2 separately. The R package `bcf` (Hahn et al., 2020) was used for the univariate implementation of BCF and the R package `bartCause` was used for the BART approach (Dorie and Hill, 2020). The R package accompanying the paper by Um et al. (2023) was used for the multivariate BART model.

4.4.1 Results

Figure 4.2 provides a graphical illustration of how the investigated approaches compare when predicting the heterogeneous treatment effects (τ), with a training sample size of 500 and 1000 observations. The full set of results, including results with a training sample size of 100, 500, and 1000 observations can be found in the supplementary material. This figure uses the precision in estimating heterogeneous effects (PEHE; equivalent to the root mean squared error in estimating τ) to evaluate predictive performance when estimating $\tau(x_i)$: $PEHE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tau(x_i) - \hat{\tau}(x_i))^2}$.

A comprehensive overview of our findings can also be found in Table 4.2. In this table, "PEHE on τ" denotes the mean PEHE across 1000 simulations for each data generating process, while "Bias on MATE" signifies the average difference between the true and predicted mixed average treatment effect, computed over the 1000 simulations. The, "τ 95% Coverage" represents the average 95% coverage rate for estimating the heterogeneous treatment effects, also averaged across the 1000 simulations, while "MATE 95% Coverage" shows what proportion of the MATE estimates from the 1000 simulations contained the true MATE. Lastly, the "τ 95% CI Width" and "MATE 95% CI Width" rows show the average widths of the 95% credible intervals for the heterogeneous treatment effects and the MATE estimates respectively.

Looking at the PEHE results from DGP1 in Figure 4.2 which we referred to as the ideal situation, multivariate BCF clearly outperforms the other three methods when tasked with accurately predicting heterogeneity in the treatment effect τ. This is evident across both outcome variables, and across both training sample sizes. Unsurprisingly, all methods perform better with a training sample size of 1000 observations. Interestingly however, the relative performances of the tested

	Simulation Study Results, Training Sample Size = 500							
	MVBCF		BCF		BART		MVBART	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
DGP1								
PEHE on τ	**9.05±0.16**	**9.40±0.16**	9.63±0.16	9.96±0.16	10.65±0.20	11.01±0.19	10.29±0.19	10.83±0.19
Bias on MATE	-0.23±0.39	-0.26±0.39	0.14±0.39	-0.17±0.39	-0.06±0.40	-0.26±0.40	0.15±0.40	0.11±0.41
τ 95% Coverage	**0.96±0.00**	**0.95±0.00**	0.97±0.00	0.96±0.00	0.98±0.00	0.98±0.00	0.98±0.00	0.98±0.00
MATE 95% Coverage	0.94±0.02	0.94±0.02	0.94±0.02	0.95±0.01	0.94±0.02	0.95±0.01	0.95±0.01	0.95±0.01
τ 95% CI Width	**39.91±0.20**	**38.62±0.20**	43.59±0.19	42.03±0.18	55.77±0.39	55.41±0.37	55.18±0.45	53.37±0.41
MATE 95% CI Width	22.85±0.12	22.64±0.11	22.92±0.10	22.63±0.10	23.68±0.11	23.57±0.10	24.23±0.11	23.98±0.12
DGP2								
PEHE on τ	36.04±0.31	**9.12±0.12**	36.25±0.32	9.43±0.12	37.45±0.32	10.29±0.15	37.51±0.32	10.21±0.15
Bias on MATE	35.09±0.32	1.02±0.29	35.19±0.32	1.13±0.29	35.97±0.33	1.51±0.30	36.12±0.33	1.57±0.30
τ 95% Coverage	0.07±0.01	0.93±0.00	0.10±0.01	**0.94±0.00**	0.27±0.01	0.98±0.00	0.26±0.01	0.97±0.00
MATE 95% Coverage	0.00±0.00	0.94±0.01	0.00±0.00	0.94±0.01	0.00±0.00	0.94±0.02	0.00±0.00	0.93±0.02
τ 95% CI Width	**39.09±0.18**	**33.14±0.15**	42.51±0.18	36.42±0.15	56.12±0.36	50.03±0.30	56.07±0.38	48.53±0.33
MATE 95% CI Width	21.38±0.09	18.92±0.07	21.24±0.07	18.79±0.07	21.72±0.08	19.26±0.07	22.01±0.08	19.49±0.07
DGP3								
PEHE on τ	9.59±0.15	9.78±0.15	9.91±0.15	9.90±0.15	11.08±0.19	11.10±0.18	11.00±0.19	10.89±0.17
Bias on MATE	-0.54±0.36	-0.15±0.35	-0.44±0.36	0.09±0.36	-0.48±0.37	-0.04±0.36	-0.22±0.36	0.21±0.37
τ 95% Coverage	**0.96±0.00**	**0.95±0.00**	0.97±0.00	0.96±0.00	0.98±0.00	0.98±0.00	0.98±0.00	0.98±0.00
MATE 95% Coverage	0.94±0.01	0.95±0.01	0.94±0.01	0.94±0.01	0.95±0.01	0.95±0.01	0.95±0.01	0.95±0.01
τ 95% CI Width	**42.41±0.21**	**38.97±0.18**	45.61±0.20	42.03±0.17	57.33±0.37	55.71±0.35	58.48±0.42	54.58±0.38
MATE 95% CI Width	23.03±0.11	22.80±0.10	22.97±0.09	22.78±0.09	23.70±0.10	23.63±0.10	24.64±0.11	24.21±0.10

Table 4.2: Simulation study results for y_1 and y_2 with a training data size of 500. Best results in bold where a clear winner exists. MVBCF is generally the top performer when predicting τ, achieving minimal bias with excellent coverage.

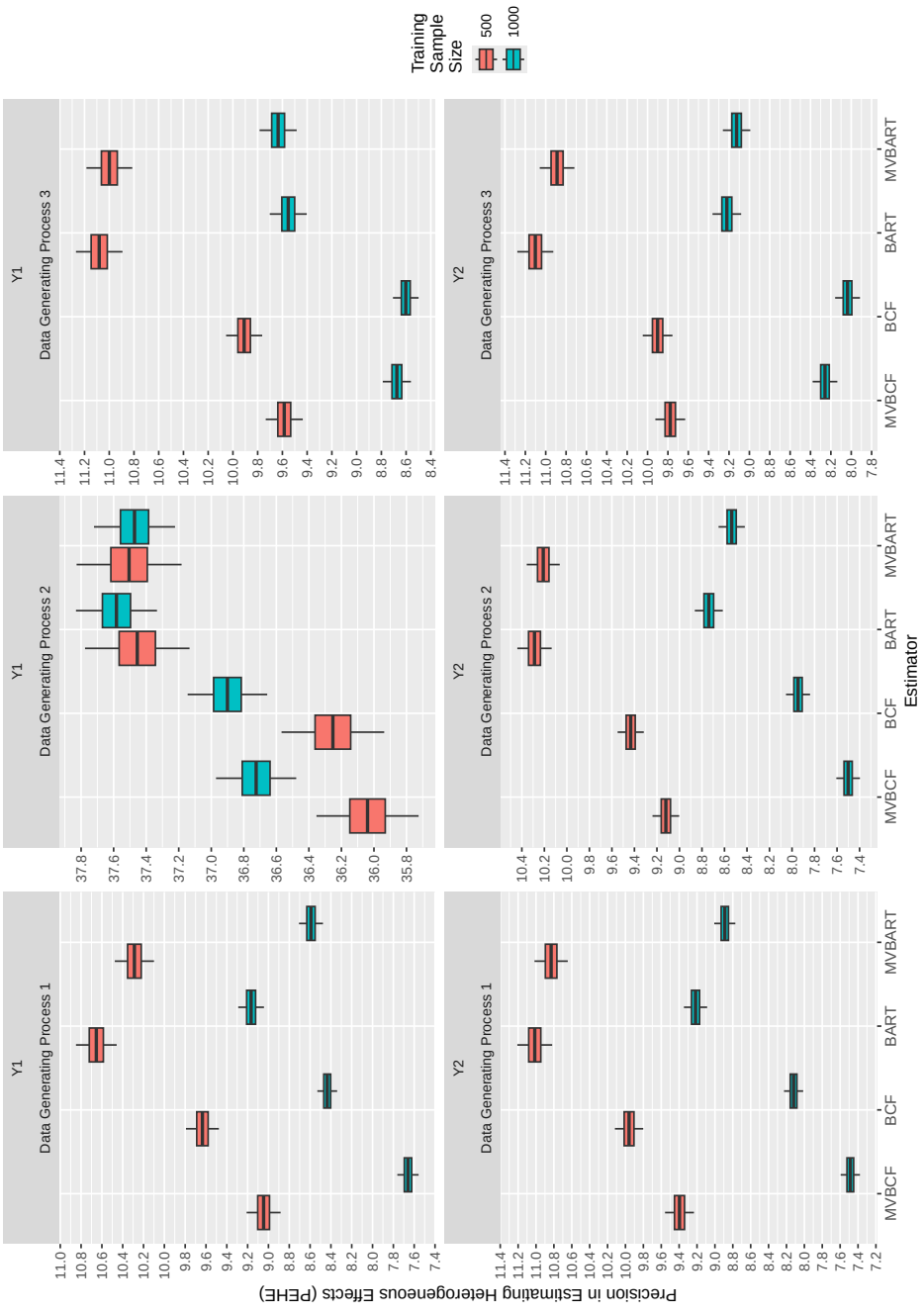

Figure 4.2: Simulation Study Results (Sample Size of 500 and 1000 Only). The central line of each modified candlestick plot shows the mean of the 1000 simulations, while the main body shows a 50% confidence interval, and the full width covers a 95% confidence interval. Results with a sample size of 100 are deferred to the supplementary material, as the larger scale of the y-axis makes visual comparison slightly harder.

approaches are quite consistent across the training sample sizes. We hypothesised that as the training sample size decreases, the benefits of the multivariate BCF approach may become clearer, however from this simulation there is no clear evidence of this, at least with these data set sizes.

Notably, in DGP1, where the multivariate model is ideally suited, the multivariate BART model also performs slightly better than its univariate counterpart. However, even the multivariate BART model falls short of matching the performance of the univariate BCF model, highlighting the advantages of the Bayesian Causal Forest approach.

The PEHE, bias, and coverage results from DGP1 in Table 4.2 tell a very similar story: In addition to the high accuracy in estimating heterogeneous treatment effects, multivariate BCF shows minimal bias in estimating the mixed average treatment effect, and the 95% coverage rate is close to ideal. Again, this is consistent across both outcome variables. The credible interval widths for the MATE are broadly similar across all methods, but the credible interval widths for the heterogeneous treatment effects are noticeably narrower for the multivariate BCF model, indicating the multivariate approach can achieve the same excellent coverage but with greater precision.

In DGP2 which included a confounded outcome Y_1 and an unconfounded outcome Y_2, it is clear that the presence of the unobserved confounding variable severely biased the treatment effect estimates of all models for Y_1. This is clearly reflected in the bias and PEHE rows of Table 4.2. Similarly, coverage rates are hugely reduced across all four methods tested. A potential concern was that given the shared dependency of outcomes Y_1 and Y_2 on the same tree structure, this reduction in performance may also be observed in Y_2. However, the results for Y_2 show that the multivariate BCF and multivariate BART approaches were unaffected by this and achieved strong performance on the unconfounded outcome Y_2. Multivariate BCF is again the strongest performer here in terms of PEHE, and achieves very good bias and coverage.

Our final data generating process, DGP3 was designed to be more challenging than DGP1 as it featured a different tree structure for both outcome variables. Reassuringly, while a very slight reduction in the performance of multivariate BCF is observed, the model maintains much of its impressive performance from

DGP1, while also maintaining minimal bias and excellent coverage. This is an important finding because it shows that even when the target variables are not ideally suited to the multivariate approach, MVBCF still performs at least as well as its competitors in this scenario. This may be attributable to the flexible tree prior $P(T_j)$ which allows the trees of the model to grow larger when necessary, and to adapt to the greater flexibility that is required in such instances.

However, we note that the Data Generating Processes used in this simulation were better suited to testing the model's capabilities in complex settings, rather than to directly mimicking the structure of real-world educational data. Therefore, a potential avenue for future research would be to incorporate a hierarchical structure into the simulated data, clustering data points within schools or classrooms. This would better reflect the data structures common in educational settings and provide a more comprehensive test of the model's hierarchical capabilities.

In future work, another interesting simulation study to perform would investigate how the multivariate normal assumption of the error term impacts model performance in cases where the error term for one outcome is normally distributed, but the error term for another outcome is not. This would provide valuable insights into the model's robustness when applied to only partially continuous outcome variables, and clarify the extent to which deviations from normality affect both predictive accuracy and inference quality.

To summarise, the results from this section have demonstrated the practical benefits of employing a multivariate BCF approach when estimating the causal effect of an intervention on two correlated outcomes. This is evident from the improved performance in multivariate BCF, which outperformed its univariate (and indeed multivariate) competitors when tasked with accurately predicting heterogeneous treatment effects. Good coverage and minimal bias was also achieved in each setting with the exception of the test involving a confounded outcome. Encouraged by the impressive performance of multivariate BCF, we now proceed to apply our model to a real dataset from the world of education in the next section.

4.5 Application to TIMSS 2019

In this section, we describe the data used from TIMSS 2019 before applying our multivariate BCF model to investigate the effect of three different treatments on student achievement in mathematics and science.

4.5.1 Data Description and Procedure

TIMSS 2019 is the seventh cycle of TIMSS to have taken place, with a total of 64 countries participating across the fourth and eighth grade components of the study, making TIMSS 2019 one of largest installments of the programme to date (Mullis et al., 2020). For the purposes of this study however, we will restrict our attention to the eighth grade subset of the data from Ireland. This subset of the data provides us with a representative sample of 4118 eighth grade secondary school students; 2118 male, 1948 female, and 52 who did not say with an average age of 14.42 years. In addition to this, the mathematics and science teachers of these students participated in the study, providing us with data on 565 mathematics teachers and 409 science teachers. The students' school principals participated too, giving us a total of 149 principal questionnaire observations.

After merging each student's data together with that of their mathematics teacher, science teacher, and school principal, the end product is a combined dataset of 4118 observations, each comprising 50 variables describing various student, teacher, and school characteristics. Important student level characteristics include gender, age, attitude towards and motivation for studying mathematics and science, as well as how many books are in their home, and the highest level educational qualification received by both parents. Important teacher level characteristics include number of years' teaching experience, area of study during their degree, perceptions of the school's level of emphasis on academic success, and teaching practices within the classroom. From the principal data we also have access to information such as the number of students in the school, a description of the average socioeconomic background of the students of the school, and a summary of how well resourced the school is in general. We will control for these variables as potential confounders as we investigate the three factors described in

Section 4.2. A complete list of all variables used can be found in the supplementary material.

TIMSS 2019 used a stratified two-stage cluster sample design to ensure that the data gathered can be used as a nationally representative sample of the population of eighth grade students within a country (Martin et al., 2020). As part of this complex survey design, students taking part in the study are assigned a sampling weight to indicate how many students in the total population they are representative of. These weights were accounted for in our study by appropriately weighting the treatment effect estimates of individual students when calculating the mixed average treatment effect for the student population.

One extra complicating factor that must be addressed when working with data from TIMSS is the use of plausible values for student achievement. It is difficult for TIMSS to accurately estimate student achievement with only a limited number of mathematics/science questions, and this is further complicated by the fact that not all students answer the same booklet of questions during the TIMSS study. Therefore, instead of providing a single achievement estimate of each student, the TIMSS study organisers have drawn five plausible values from the posterior distribution of each student's achievement level in order to better represent the underlying uncertainty that is present (Wu, 2005). In line with best practice, these five plausible values which are available in the public data were fully accounted for in our study. Five MCMC chains were run for every model, each corresponding to one of the five plausible values. These chains were pooled together after burn-in to capture the uncertainty in the provided estimates of student achievement.

Prior to this model fitting process, the overlap assumption, which is partially testable, was assessed by visually inspecting density plots of the propensity scores for each of the "Has Study Desk", "Often Absent", and "Often Hungry" factors, for treated and control units to confirm there were no regions lacking common support. Additionally, similar visual checks were conducted for all continuous covariates, while for categorical variables, contingency tables were created to ensure that no levels contained only treated or only control units.

Finally, there is a clear hierarchical structure to the TIMSS data as students are nested within classrooms. Failure to account for such designs can lead to inaccuracies, and underestimation of model uncertainty. Therefore the multivariate

BCF model we have applied to the data has been equipped with an extra random intercept parameter to account for the between classroom variation:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}(x_i) + \boldsymbol{\tau}(x_i)Z_i + \boldsymbol{\alpha}_{class,i} + \boldsymbol{\epsilon}_i$$

where $\boldsymbol{\alpha}_{class}$ is an added multivariate random intercept term for each classroom. We assume that each $\boldsymbol{\alpha}_{class}$ comes from a multivariate normal distribution with a population mean $\boldsymbol{\mu}_\alpha$ and population covariance matrix $\boldsymbol{\Sigma}_\alpha$: $\boldsymbol{\alpha}_{class} \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$, where the prior on $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Sigma}_\alpha$ is: $\boldsymbol{\mu}_\alpha \sim N(\boldsymbol{m} = \boldsymbol{0}, \boldsymbol{s} = 0.01\boldsymbol{I})$, and $\boldsymbol{\Sigma}_\alpha \sim \mathcal{W}^{-1}(a = 1, \boldsymbol{\Omega_0} = 0.1\boldsymbol{I})$.

To obtain the results in the following section, three separate multivariate BCF models were applied to the TIMSS data, one for each of the treatments under investigation. For each model, the five plausible value chains were run for 3000 burn-in and 2000 post burn-in iterations, with every second iteration kept to reduce memory costs. A total of 100 BART trees were used for the prognostic component $\mu()$, while a smaller number of 50 BART trees were used for estimating the treatment effects $\tau()$. Satisfactory convergence was assessed via visual inspection of samples from the variance covariance matrix Σ, predicted values $\hat{\mu}$ and $\hat{\tau}$ for a random sample of individuals in the dataset, and the MATE estimates themselves. For comparison, we have also obtained results using a univariate BART, and univariate BCF approach. The `bartCause` package used for fitting the univariate BART model included functionality for working with hierarchical data, however, this feature was not present in the univariate `bcf` package, so to obtain the univariate BCF results we instead applied our multivariate BCF model to each outcome separately in order to make use of the random intercept feature available with the model. The multivariate BART model (Um et al., 2023) was not employed here as it does not include functionality for working with hierarchical data.

4.5.2 Results

In the left column of Figure 4.3, we show density plots of the posterior distributions of the mixed average treatment effects for each of the treatments we have investigated. Detailed credible intervals for each of these treatments can be found in Table 4.3 which also provides the control and treatment group sizes for each

intervention under investigation. For comparison, results from the other three methods used in the simulation study have also been included. The treatment group size for "Has Study Desk" is 3672, indicating that 89% of the students in the sample did report having a study desk at home, while the remaining 11% did not. The control and treatment group sizes for often being hungry when arriving at school or often being absent have similar interpretations. A broader discussion of the wider context of these results can be found in the following section, but for now we will focus only on a summary.

Table 4.3 summarises the mixed average treatment effect results obtained from each of the methods applied to the TIMSS data. Reassuringly, the MATE estimates from each of the three methods are very similar, displaying a high level of agreement for every treatment under investigation. The credible interval widths are also of a very similar size and exhibit a high degree of overlap in each case. This agrees with what we saw in the simulation study results, which showed that the clearest improvements in performance related mainly to the estimation of the heterogeneous treatment effects. It is worth noting that despite the high level of agreement among all methods compared here, the true treatment effects remain an unobserved quantity, so the results from the real TIMSS data can not be used to compare the performance of the models - this is a task best suited to the simulation study results. As a further comparison, we have included Table 4.4, which presents the results of a 10-fold cross-validation for each of the three methods, assessing their predictive performance on student achievement in mathematics and science. Here, the multivariate BCF model displays a modest improvement in performance relative to the univariate BCF model. The univariate BART model is the best performer here, and this is expected given the primary focus of BART on predicting y rather than directly targeting treatment effects. Again, it is important to clarify that these results do not necessarily imply strong predictive performance in estimating the treatment effects (τ), as knowing the ground truth of the treatment effects is not possible.

To assist in the interpretation of the results that follow, consider that student achievement at the eighth grade in mathematics in Ireland is approximately normally distributed with mean 524, standard deviation 73, and students at the 10th and 90th percentiles scoring approximately 432 and 614 respectively. Therefore,

TIMSS Average Treatment Effect Results			
	Has Study Desk	Often Hungry	Often Absent
Control/Treatment Group Sizes:			
Treatment Group Size (Yes)	3672 (89%)	954 (23%)	503 (12%)
Control Group Size (No)	446 (11%)	3164 (77%)	3615 (88%)
Multivariate BCF: MATE 95% CI			
Mathematics	4.61 (-0.50, 10.14) [10.64]	-5.27 (-08.86, -1.56) [07.30]	-7.20 (-11.88, -2.27) [09.61]
Science	-3.39 (-11.13, 05.09) [16.22]	-5.73 (-10.35, -0.94) [09.41]	-3.45 (-09.17, 2.95) [12.12]
Standard BCF: MATE 95% CI			
Mathematics	4.03 (-1.63, 09.32) [10.95]	-5.22 (-08.99, -1.50) [07.49]	-7.02 (-11.65, -2.37) [09.28]
Science	-3.92 (-12.07, 05.02) [17.09]	-5.99 (-10.41, -0.64) [09.77]	-3.21 (-08.92, 3.09) [12.01]
Standard BART: MATE 95% CI			
Mathematics	3.71 (-01.42, 09.01) [10.43]	-5.22 (-09.14, -1.44) [07.70]	-7.07 (-11.88, -2.06) [09.82]
Science	-3.56 (-11.12, 04.52) [15.64]	-6.05 (-10.75, -0.89) [09.86]	-3.36 (-09.60, 2.51) [12.11]

Table 4.3: Application to TIMSS Results. 95% Credible intervals shown in round brackets. Credible interval widths shown in square brackets.

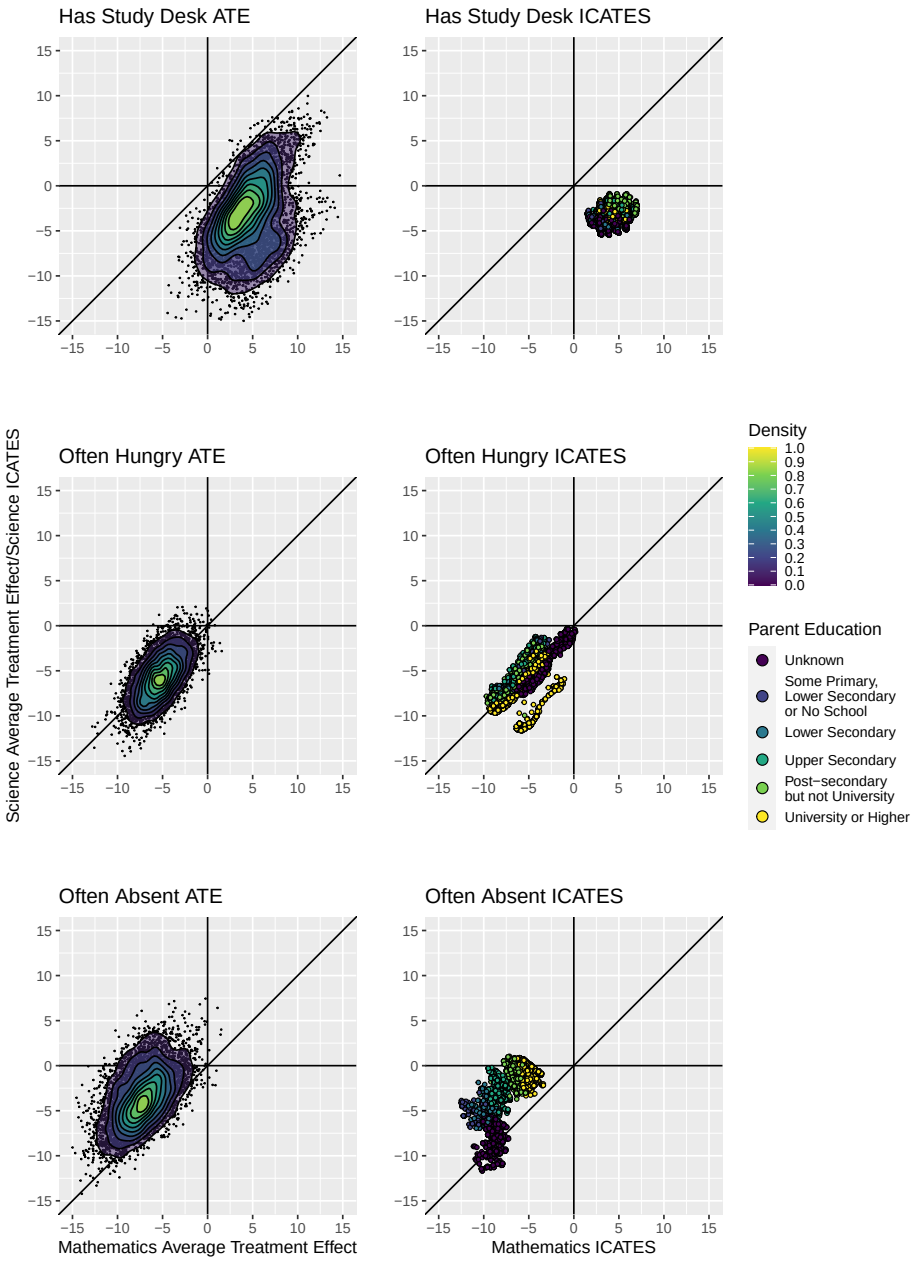

Figure 4.3: Plot of mixed average treatment effects and individual conditional average treatment effects (ICATEs). Each section in the left column displays a density plot of the sampled posterior distribution of the mixed average treatment effects for mathematics (x-axis) and science (y-axis). In the right column, the ICATEs are shown coloured by the parental education variable to help highlight some of the heterogeneity.

	TIMSS 10-Fold Cross Validation Results (RMSE)					
	MVBCF		BCF		BART	
	Mathematics	Science	Mathematics	Science	Mathematics	Science
Has Study Desk	43.46	52.82	45.10	53.90	43.00	52.27
Often Hungry	43.70	53.15	45.12	53.94	43.03	52.08
Often Absent	44.27	53.89	45.20	54.00	43.41	52.74

Table 4.4: 10-Fold cross validation results for each method applied to TIMSS. MVBCF performs strongly with marginally better results than the standard univariate BCF model.

a treatment effect magnitude of 7.3 would correspond to a 0.1 standard deviation increase/decrease in student achievement in mathematics. Effect sizes of this magnitude are common in educational studies and can be thought of as being "medium" in size (Kraft, 2020). Science achievement follows a very similar distribution with mean 523 and standard deviation 83.

Multivariate BCF clearly identifies access to a study desk at home as having a positive impact on student achievement in mathematics. The effect of having a desk on student achievement in science is less clear however, and the mixed average treatment effect is centred very close to zero. Our results for the second treatment under investigation, often being hungry at school, show that this factor is associated with a very negative impact on both mathematics and science achievement. The magnitude of the effect identified is almost identical for both mathematics and science achievement. Finally, often being absent from school is also identified as having a negative impact on achievement in both mathematics and science. The impact on science achievement however, as was the case with having access to a desk, is slightly less clear than for mathematics achievement. The effects in mathematics and science achievement are positively correlated for all three treatments under investigation. This agrees with our intuition that the effect any of these three factors may have is likely to be similar on achievement in both subjects. We did observe some differences in magnitude however, most notably in relation to the "Has Study Desk" treatment.

Given these findings, it is important to consider the implications of multiple

comparisons. First, there is a multiple comparisons concern arising from testing the effects of three different factors. This is acknowledged as a limitation. Then there is a second concern arising from analysing their effects on two outcomes, which we believe should be mitigated by the multivariate nature of the model. By jointly estimating the effects of each factor on both mathematics and science achievement, the model explicitly accounts for the relationship between these outcomes, and incorporates this dependence into the posterior of the effects on both outcomes.

Often, the effects of a treatment felt by an individual may be adjusted by one or more moderating variables. Thus in addition to the density plots for the MATE, we have included a scatterplot of the individual conditional average treatment effects (ICATEs) in the right column of Figure 4.3, coloured by the parental education variable to help highlight some of the underlying heterogeneity. This variable was chosen because it appeared quite frequently in the splitting rules of the τ part of the model, indicating the model had identified this variable as a source of heterogeneity in the treatment effects. This can be seen visually in Figure 4.3, as children with parents of a similar education level tend to be clustered quite close together, albeit less so in the "Has Study Desk" treatment where there is considerably less heterogeneity and all of the ICATE estimates are concentrated quite closely together.

While parental education is used here as a categorical variable to aid in the visualisation of treatment effect heterogeneity, it is important to acknowledge its close and well-documented association with broader indicators of socioeconomic status (SES). Within the TIMSS dataset, higher levels of parental education are consistently correlated with increased access to home educational resources, more advantaged school environments, and other key markers of socioeconomic advantage. As such, the patterns of heterogeneity observed in the treatment effects likely reflect underlying SES-related variation, the exact nature of which is beyond the scope of this study.

Home resources and school resources also appeared quite frequently in the set of splitting rules. To investigate the moderating effect of these variables we have created individual conditional expectation (ICE; Goldstein et al., 2015) plots which visualise the dependency of the treatment effects on these covariates. Figure 4.4 shows the results for the treatment "Often Hungry" which exhibits an interesting

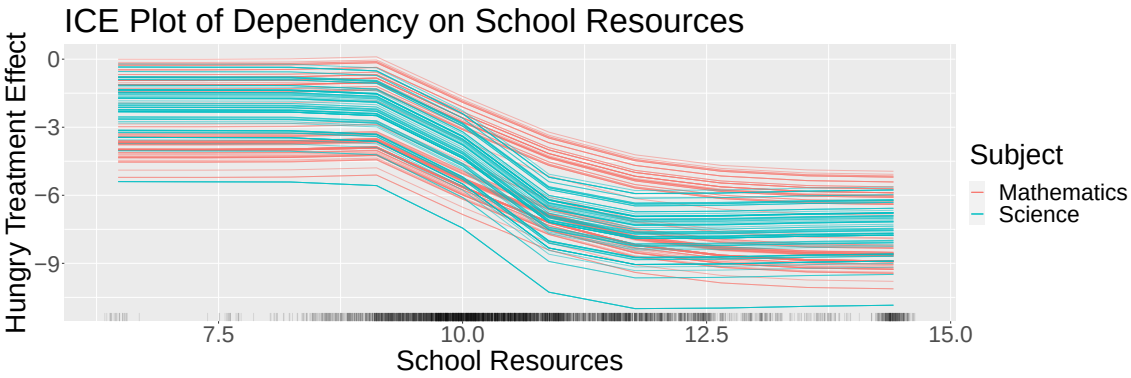

Figure 4.4: ICE Plot of the moderating role of school resources on the "Often Hungry" treatment effect. (Random sample of 100 students to avoid overprinting). A jittered rug has been added to the x-axis to display the distribution of the average school resources variable. Students in schools with fewer resources appear to be less negatively affected by arriving to school feeling hungry.

trend, as it would appear students in schools with fewer resources tend to experience a less negative treatment effect. Schools in disadvantaged areas with fewer resources are more likely to receive access to free school meal programmes in Ireland (Department of Social Protection, 2023), so this is possibly an indication that free school meal programmes are successfully mitigating the negative consequences of students often arriving at school feeling hungry. Without knowing which schools do in fact participate in free school meal programmes, however, we can only speculate on the true moderating role of school resources here. A different pattern is observed in Figures 4.5 and 4.6 which show that students with more educated parents and more home resources are less negatively affected by frequent absences from school. These students may be in a better position to "catch up" on missed school work due to the physical and parental resources available to them. However, due to the black box like nature of BART and BCF, we must acknowledge that our understanding of the genuine moderating role of home educational resources and parental education in the context of frequent school absences remains speculative.

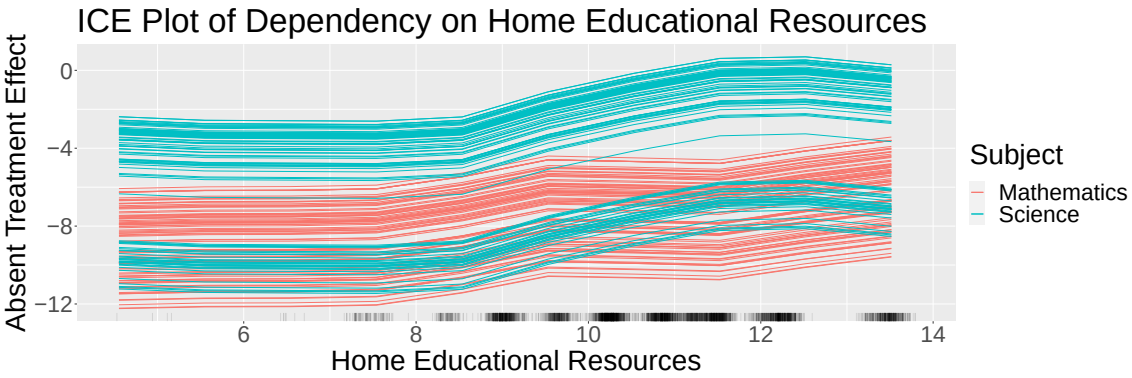

Figure 4.5: ICE Plot of the moderating role of home educational resources on the "Often Absent" treatment effect. (Random sample of 100 students to avoid overprinting). A jittered rug has been added to the x-axis to display the distribution of home resources variable. Students with greater educational resources at home appear to be less negatively affected by regular absences. Notice the two clusters of blue lines which correspond to students who know and don't know their parent's education level (See Figure 4.6).

4.6 Discussion

Motivated by data from the Trends in International Mathematics and Science Study which includes information on student achievement in both mathematics and science, we have developed a multivariate extension of Bayesian Causal Forests which can be used to estimate the causal effect of an intervention on two or more outcome variables simultaneously. The key advantage of our approach is the use of the same tree structure for both outcome variables. This enables us to leverage the shared variance across the outcomes, resulting in improved predictive performance, as demonstrated in our simulation study.

The results from our simulation study indicate that the multivariate BCF model is capable of estimating causal effects with a greater level of accuracy than the univariate BCF model and both BART based approaches. Our model was robust to violations of the model assumptions affecting only one of the outcome variables of interest. In the case where the target variables were not well suited to a multivariate approach it was observed that MVBCF performed equally as well as the univariate BCF model, showing that there was no reduction in performance

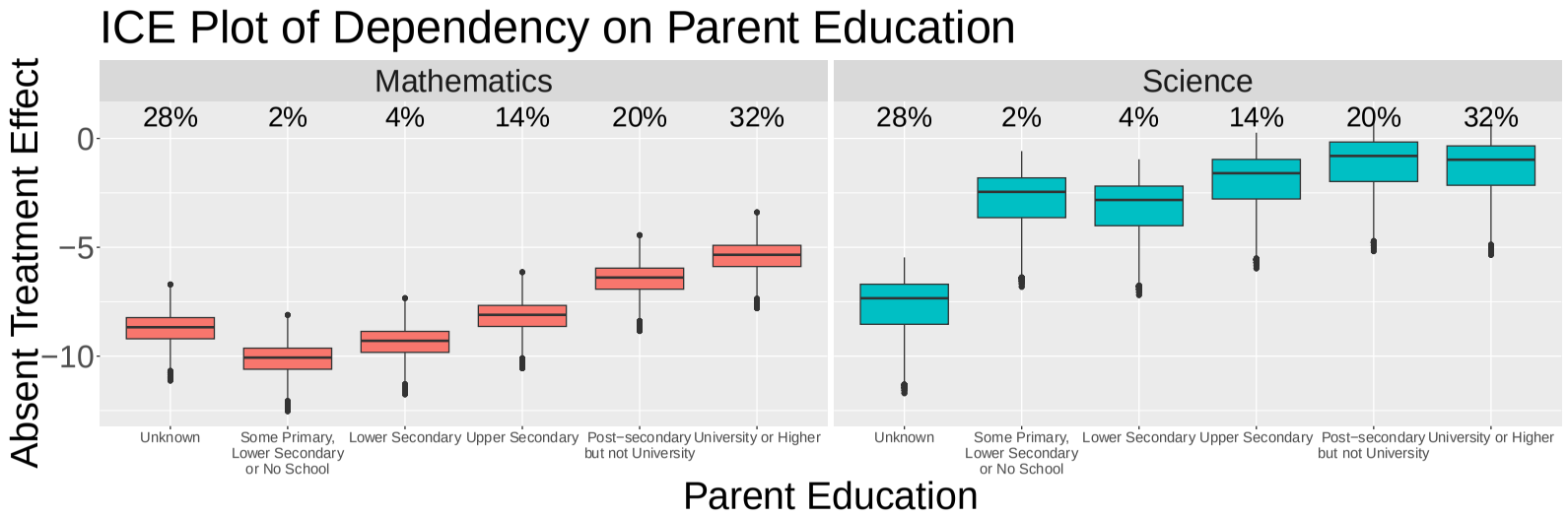

Figure 4.6: ICE plot of the moderating role of parent education on the "Often Absent" treatment effect. Students with highly educated parents appear to be less negatively affected by regular absences. Percentages indicate proportion of students belonging to each category.

below that of a univariate model. As well as observing an increase in predictive performance, we noted that our multivariate BCF was able to attain satisfactory coverage and minimal bias.

In our application of the multivariate BCF model to the motivating TIMSS dataset we found that access to a study desk at home is associated with a clear increase in mathematics achievement, but with no discernible change in science achievement. Unsurprisingly, often being hungry at school was identified as having a very negative impact on achievement in both mathematics and science. A similar negative effect was found to be associated with often being absent from school. These results agree with findings from previous studies within the field of education which have identified the importance of home-related factors for predicting student achievement in mathematics and science (e.g. Tsai and Yang, 2015; Vik et al., 2022; Vesić et al., 2021). Our study therefore makes an important contribution by verifying these results within a causal inference framework.

Our results provide further evidence of the potential for targeted interventions such as free school meal programmes to tackle the negative consequences of students frequently lacking a healthy breakfast in the morning, or a lunch while they are at school. Notably, 23% of students in our dataset reported frequently experiencing hunger when arriving at school, and this hunger had a significant impact, reducing student achievement by approximately five units in both mathematics and science. The positive effect of having a study desk (Mathematics MATE $= +4.61$) may also indicate an opportunity to inform parents about the importance of students having dedicated study spaces at home, as 10% of students in the data did not report having access to a study desk. Finally, the clear negative impact that was observed from students often being absent (Mathematics MATE $= -7.20$, Science MATE $= -3.45$) may highlight the potential for schools to investigate these absences, and to prepare extra supports for the 12% of students affected by this.

An important ethical consideration arising from this work concerns the treatment of socioeconomic status (SES) as a confounding variable. While this approach allows for the estimation and isolation of treatment effects separately from broader SES influences, it also carries the risk of obscuring the deeper, structural inequalities that shape educational opportunity and achievement. In particular, isolating

the effects of variables such as hunger, absence, or lack of study space may understate the role of broader systemic disadvantage, which often manifests through and is reinforced by socioeconomic conditions. It is therefore important to interpret these findings within the wider context of educational inequality, recognising that the variables analysed here may act as surface-level indicators of more entrenched social and economic disparities.

One limitation of multivariate tree based models is that they can struggle when the outcome variables of interest are weakly or not at all correlated. In some settings it may be the case that a very different tree structure is appropriate for the outcome variables, and in these situations the requirement that both outcome variables share the same tree structure can be quite restrictive, leading to a reduction in overall performance. This is evident from our simulation study, in which a slight reduction in performance was observed in MVBCF when applied to simulated data of this type. We note, however, that limitations of this kind are unlikely to apply to our investigation of the TIMSS data because there is a strong positive correlation between student achievement in mathematics and science, $r \approx 0.85$, and it is likely that variables related to a student's family's socioeconomic status or school will have a similar effect on achievement in both subjects, thus making a shared tree structure very appropriate.

A second limitation arises from the cross-sectional nature of the TIMSS data, which makes it difficult to establish the temporal ordering of variables with certainty. For instance, in the analysis of school absences, we included students' sense of school belonging as a potential confounder. While this variable could plausibly influence student absences, it may also reflect a student's feelings about school after a pattern of regular absences has already developed — thereby acting more as a mediator than a confounder. If this is the case, our estimates capture the direct effect of absence on achievement, rather than the total causal effect. Future work using longitudinal or experimental data would help to further explore these temporal relationships and strengthen the results.

While not emphasised, the hierarchical extension of the model adopted in this study is likely to hold significant value across many domains. Multilevel datasets are common in many settings, and nowhere more so than in the context of education, where accounting for hierarchical structures is crucial. Despite this, the only

existing multilevel BART and BCF applications we are aware of prior to this study are those of Dorie et al. (2022), and Yeager et al. (2019), who also applied BCF to a multilevel dataset with different schools. Therefore, the development of more advanced BCF models with the capability to account for even more complicated hierarchical data structures, perhaps with multiple levels of data would be very valuable. In particular, extending the work of Wundervald et al. (2022) which proposes a hierarchical extension of BART would be of interest.

A second advantage of the multivariate model, which was not emphasised in this chapter, is its usefulness in tackling the multiple comparisons problem. By jointly modeling all outcome variables and explicitly capturing their relationships, the model may reduce the risk of false discoveries that can arise from conducting separate analyses. This approach allows for more precise estimation of treatment effects, while also accounting for the correlations between outcomes, leading to more robust and reliable inferences. Therefore, the multivariate approach adopted in this Chapter can provide a principled way to address concerns related to multiple hypothesis testing, strengthening the validity of the results.

The measures of mathematics and science achievement used in this study are both continuous outcome variables. Often in education research, a question of interest is whether or not students have attained a specific level of mastery or ability in a given topic. These levels of mastery may be a simple binary outcome (achieved or has not achieved mastery), or may be ordinal (for example low, medium, or high level of mastery). Therefore, while some work has been conducted in the area of BART/BCF and binary outcomes (Hu et al., 2020; Starling et al., 2021), an extension of the work by Murray (2021) to multinomial or ordinal outcomes with BCF also presents a very exciting area for future research. Combining this with the multivariate BCF extension in this study would also potentially allow for a mixture of continuous, binary, or other types of outcomes.

To enhance the flexibility of the multivariate BCF approach, one potential avenue for further development is to enhance its flexibility by partitioning the ensemble of BART trees, denoted as $\mu()$, into subsets. Specifically, one subset, denoted as $\mathcal{M}1$, could be dedicated to independently predicting the first outcome, while another subset, $\mathcal{M}2$, could focus on predicting the second outcome. Additionally, a third subset $\mathcal{M}3$ could be designed to address all outcome variables collectively.

This increased adaptability would empower the multivariate BCF method to better accommodate distinct tree structures that might be well-suited for handling highly uncorrelated outcome variables, by allowing it to benefit from any correlation in the outcome variables through $\mathcal{M}3$, while also allowing the model to construct separate tree structures $\mathcal{M}1$ and $\mathcal{M}2$ for both outcomes if necessary. A similar modification could also be applied to the ensemble of trees responsible for predicting τ if there is reason to believe that the treatment effect is likely to have a very different impact on both outcomes, and the effect is likely to be moderated by a distinct set of variables for each outcome. In fact, since authoring this Chapter, other researchers have begun work on implementing a similar approach for use with BART (Esser et al., 2024). One potential drawback of this approach, however, is that by devoting a separate ensemble of decision trees to each outcome, the computational cost of this approach may limit its applicability to relatively small numbers of outcome variables.

Finally, although our motivating dataset came from the world of education in this study, it is also likely that our multivariate approach would be useful in other fields such as economics or medicine. A researcher may be interested for example in the effect of a drug D on both the systolic and diastolic blood pressure of patients who have been prescribed it by their doctor. Areas for future research therefore also include the application of multivariate BCF to other disciplines with multivariate outcomes of interest.

5

Little and Often: Causal Inference Machine Learning Demonstrates the Benefits of Homework for Improving Achievement in Mathematics and Science

5.1 Introduction

Homework plays a significant role in the daily lives of many students. Defined as the tasks assigned to be completed outside of school hours (Cooper, 1989), its role in shaping academic outcomes is far from understood. The conflicting perspectives surrounding the benefits and drawbacks of homework (Marzano and Pickering, 2007) continue to permeate educational literature, sparking ongoing discourse that shows no signs of abating. Some argue that homework is necessary for reinforcing learning and developing important skills such as time management and responsibility (Palardy, 1988), while others claim that it can be detrimental to students' mental health, and reduce the amount of time available for other important activities (Galloway et al., 2013).

Much of the research on homework indicates a modest yet positive impact on academic outcomes. However, this assertion demands critical examination. Notably, there is a growing awareness of the need to scrutinise how homework is distributed throughout the week (Trautwein et al., 2002), how its effects may dif-

fer by subject (Fernández-Alonso et al., 2017), and how students from different socioeconomic backgrounds can be impacted by homework (Eren and Henderson, 2011; Rønning, 2011). Compounding these issues is the ubiquitous challenge of confounding which is inherent in observational data, rendering causal interpretations elusive. However, causal inference methods remain a neglected avenue in the study of homework effects.

Motivated by these gaps in the research, this study employs a new approach to unravel the complex relationship between homework frequency, duration, and student achievement. Unlike previous research which has typically focused on a single subject at a time, our investigation extends to both mathematics and science. Critically, we leverage a recent extension (McJames et al., 2024, Chapter 4) of a causal inference machine learning model called Bayesian Causal Forests (Hahn et al., 2020), marking a departure from conventional methodologies and allowing a more precise understanding of causal relationships.

The key advantage of this model is its flexibility. While simpler models often rely on rigid assumptions, such as assuming a linear relationship between explanatory variables and the outcome, the model employed in this study utilises a set of decision trees. These decision trees can adapt automatically to complex features, such as non-linear relationships or interaction terms, without requiring the researcher to specify these relationships in advance. This is an important feature, because complex factors are known to influence the amount of homework students are likely to receive, and the benefit they may derive from it (Corno, 1996).

By accurately accounting for these often complicated relationships, the model separates the estimation of student achievement into two distinct parts: a level of achievement which would be observed without any homework, and a second part which estimates the change in achievement directly attributable to the homework. This isolation of the causal impact of homework from other associations allows for a causal interpretation of the results. Specifically, the causal estimates provided by the model are known as conditional average treatment effects: they estimate for each individual the amount by which their achievement would change by if they were moved from a control group (no homework) to a hypothetical treatment group (homework with a defined level of frequency and duration).

We apply this model to Irish eighth grade data from the Trends in International

Mathematics and Science Study (TIMSS 2019, Mullis et al., 2020): A nationally representative sample of Irish eighth grade students, with an average age of 14.4 years, 51.2% of whom identified as male. The data provides us with an estimate of academic achievement in both mathematics and science, as measured by a standardised assessment. Important contextual information related to the backgrounds of the students, their parents, their attitudes towards studying these subjects, and insights into their school learning environment are also provided. This data, combined with details of how often the students receive homework, and how long this homework usually lasts, makes TIMSS very well suited to the current investigation.

A more detailed description of our model, and the Irish data to which it is applied can be found in the methodology section, but now we review some key findings from the literature, and the unanswered questions motivating our study.

5.1.1 Factors Influencing Homework Efficacy

Frequency, Duration, and Time Spent on Homework

Studies exploring the impact of homework typically report a modest yet positive association with student achievement (Cooper et al., 2006). However, the literature also reveals a less optimistic side, indicating that an excess of homework can yield diminishing returns or, in some cases, even detrimentally affect student performance (Cooper, 1989). Despite these findings, a concerning feature of numerous investigations examining the homework and student achievement relationship is the reliance on models which assume a linear relationship (Zhou et al., 2023). This reliance on a linear framework oversimplifies the dynamics at play, potentially obscuring crucial insights.

Moreover, only a limited number of studies have ventured into determining the optimal amount of time students should dedicate to homework. The precise identification of a threshold, beyond which additional homework yields no discernible benefit, holds substantial value. Such insights could empower educators to streamline homework assignments, optimising efficiency and benefits, while mitigating potential adverse effects such as heightened stress levels, reduced time for extracurricular activities, and strains on families (Galloway et al., 2013; Pressman et al., 2015).

An often overlooked facet in homework studies is the strategic allocation of homework time throughout the week. Trautwein et al. (2002) underscored the importance of considering not only the total time spent on homework but also the frequency at which it is given. Their study revealed that homework frequency is a strong predictor of academic achievement, while overall time spent on homework is not. This highlights the need to dissect homework into its frequency and duration components for a more nuanced understanding. The primary importance of homework frequency rather than duration is further supported by Fernández-Alonso et al. (2019) and Trautwein (2007) who arrive at similar conclusions. However, Zhu and Leung's (2012) contrary findings complicate this narrative, emphasising the necessity for further research into the nuanced effects of homework frequency and duration to reconcile these conflicting perspectives.

Subject Specific Homework Effects

In addition to the frequency and duration considerations discussed above, a number of studies have indicated that the effects of homework on student achievement may vary depending on subject. Results in this area are quite limited however, and studies which have compared effect sizes in mathematics to other subjects such as history or English have returned conflicting results. Some studies report a notably stronger relationship between homework and mathematics achievement (Eren and Henderson, 2011), while other studies have characterised the relationship within mathematics as the weakest when compared to other subjects (Fernández-Alonso et al., 2017). Subject specific differences are therefore likely to exist, but the precise nature of these differences is difficult to determine without further research.

Further emphasising the need to differentiate between different academic subjects, research has shown that subject specific motivational factors can significantly contribute to the effort students put into completing their homework (Trautwein and Lüdtke, 2009), thus influencing the benefit they can derive from it (Flunger et al., 2015). Notably, an extra difference commonly reported in the literature is that homework time commitment expectations tend to be quite high in mathematics relative to other subjects (Fan et al., 2017). These variations in time expectations indicate that not only can the effects of homework differ by subject,

but the nature and extent of the academic demands imposed on students can vary significantly between subjects as well.

Socioeconomic Background and Homework Benefits

A common finding from homework studies is that students from socioeconomically advantaged backgrounds can benefit more from homework than their peers (Patall et al., 2008; Tan et al., 2020). This apparent divide could stem from many reasons such as greater access to study aids at home, a more suitable study environment with a desk, or the extra guidance that highly educated parents might be able to provide to their children. Findings such as this could also potentially stem from unobserved confounding if students from advantaged backgrounds are more likely to be provided with higher quality or better targeted homework assignments from school teachers (Dettmers et al., 2010).

Among these factors, one of the most explored is the role of parental involvement in homework (Patall et al., 2008; Gonida and Cortina, 2014; Viljaranta et al., 2018), and how its impact may be moderated by socioeconomic status. Studies such as those by Daw (2012) and Eren and Henderson (2011) have uncovered significant interactions between parental income or education and the efficacy of homework. This has led some authors to conclude that homework may be partially responsible for perpetuating social inequalities by widening the achievement gap between those from advantaged and disadvantaged socioeconomic backgrounds (Rønning, 2011). On a more positive note, some research suggests that these adverse effects may be partially mitigated by carefully designing homework assignments that confer no advantage to privileged students (Edwards, 2018), and educating teachers about the struggles faced by disadvantaged students (McCrory Calarco et al., 2022). Despite the importance of this issue however, we did not find any causal studies that examined the moderating role of socioeconomic status in homework effects, so there remains room for further exploration.

In addition to leading to a widening achievement gap, some authors have also explored the added stress that homework can cause parents when they feel unable to help their children effectively (Lutz and Jayaram, 2015; Solomon et al., 2002). Therefore, while this study focuses on exploring the relationship between

homework frequency, duration, and academic outcomes, it is important to recognise the broader implications and potential trade-offs associated with homework assignments.

5.1.2 Causal Inference and Advanced Modelling Techniques in Homework Studies

A common thread that runs through almost all homework studies is the challenge of working with observational data. Randomised controlled trials are the gold standard in all areas of research, but factors such as cost and complexity usually make experiments of this type very difficult, necessitating the use of observational data (West et al., 2008). While rare, however, a number of homework studies have employed experimental designs: randomly assigning students to homework versus no homework groups, or to groups using traditional versus online homework systems (Grodner and Rupp, 2013; Foyle et al., 1990; Roschelle et al., 2016). Reassuringly, results from experiments such as these have often been in agreement with the majority of the literature, affirming a small but positive effect of homework. Due to the difficulty of organising these studies, however, they are often limited to relatively small sample sizes and rarely involve representative samples, somewhat limiting the generalisability of these findings (Daw, 2012).

Observational data on the other hand is often much easier to collect, allowing rich data sets of representative samples to be collected as part of large scale national or international studies. Nevertheless, the drawback lies in the non-random selection of treatment groups, influenced by individual characteristics, commonly known as confounding variables. These variables may create an illusion of causality between two factors, introducing potential biases into statistical analyses (Greenland et al., 1999). Causal inference is the field of study focused on addressing these concerns, and various methods have been developed to mitigate these issues. Techniques range from controlling for confounding variables and matching individuals based on similar characteristics, to understanding how these traits impact group membership propensity (Hill, 2011; Stuart, 2010; Pan and Bai, 2018).

Despite often relying on observational data, however, few studies have actively integrated these methods into their investigations. In fact, to our knowledge, the

work by Gustafsson (2013) stands alone as an example. Moreover, despite the intricate nature of the relationship between homework and student achievement (Corno, 1996), and the myriad factors interacting with homework to influence its effectiveness, modern machine learning-based methods that can account for these complex patterns within the data are seldom utilised. A notable exception is the study by Eren and Henderson (2008), which linked homework to academic achievement using both parametric and non-parametric statistical models. The non-parametric approach used in their study demonstrates the potential of flexible and accurate regression models in homework related studies, which are seen too rarely in education research.

5.1.3 The Current Study

Our study aims to reduce the remaining uncertainty surrounding the effects of homework on student achievement by combining three key features of studies conducted by Trautwein et al. (2002), Eren and Henderson (2008), and Gustafsson (2013). To accomplish this, we employ a causal inference machine learning model to answer the following research questions:

RQ1: What is the effect of homework frequency on Irish students' mathematics and science achievement in TIMSS 2019? (Where homework frequency is defined to be the number of days each week a student is normally given a homework assignment of any duration)

RQ2: What is the effect of homework duration on Irish students' mathematics and science achievement in TIMSS 2019? (Where homework duration is defined to be the number of minutes typically required by a student to complete each homework assignment)

Given our dual focus on student achievement in both mathematics and science, we also attempt to answer:

RQ3: Is the optimal distribution of homework frequency and duration different in mathematics and science?

Finally, informed by prior research which indicates that students from advantaged socioeconomic backgrounds may benefit more from homework than other students, we ask:

> **RQ4:** Do advantaged students benefit more from homework than their peers?

The remainder of this chapter is structured as follows: First, we introduce the TIMSS 2019 dataset, and the Irish sample of students who are the subject of our study. We also describe our analytical strategy, based on multivariate Bayesian Causal Forests. We then present the results of our study, highlighting how they address our research questions. This is followed by a broader discussion, explaining how our results relate to previous research, outlining the main contribution of our study, and some limitations. Finally, we conclude this chapter by considering the implications of our research for homework policy, and provide suggestions for future research.

5.2 Method

5.2.1 Data and Sample

Our study uses data from the seventh cycle of the Trends in International Mathematics and Science Study, TIMSS 2019, organised by the International Association for the Evaluation of Educational Achievement (IEA, Mullis et al., 2020). TIMSS 2019 took place in 64 countries, which represented the largest cycle of this study to date. TIMSS surveys students in the fourth and eighth grades but we focus our attention on the Irish eighth grade subset of the data, comprising a nationally representative sample of 4118 students. For context, these students are in their second year of secondary school in Ireland, and are preparing for a national set of examinations called the Junior Cycle Examinations. This is not considered a high stakes exam, but it serves as a way to assess students' progress, and will help to determine the level of the courses the students will undertake for their final examinations, known as the Leaving Certificate. Additionally, the Junior Cycle Examinations provide valuable examination experience without the high-stakes

implications. To ensure a nationally representative sample, TIMSS 2019 employed a stratified two stage cluster sampling design which sampled complete classes from a list of all suitable schools within a country. The sampling weights resulting from this design were fully accounted for in our analysis by appropriately weighting our results.

As part of TIMSS, the fourth and eighth grade students from participating countries complete a short assessment in mathematics and science to measure their achievement levels in these subjects. Additionally, the eighth grade students answer a survey consisting of various questions designed to gauge their motivation, confidence, and how much they like studying mathematics or science. Questions related to their socioeconomic status such as how many books they have at home, and the education level of their parents are also included. To offer a more comprehensive view of the students' educational experiences, TIMSS extends its data collection efforts to encompass input from their mathematics and science teachers, as well as school principals. Mathematics and science teachers participate by completing a questionnaire that covers various aspects of classroom teaching practices, their perception of discipline, homework assignments, and other pertinent factors. Meanwhile, school principals provide valuable insights into school-related matters such as available resources, the emphasis on academic achievement, and various school characteristics.

Summary statistics on important characteristics related to the Irish eighth grade sample of students can be found in Table 5.1. Further statistics on all variables collected during TIMSS, including other countries which were not examined as part of this study can be found in the code books and almanacs published by the IEA (Fishbein et al., 2021). Note that in order to interpret the variables associated with the teacher or the school in Table 5.1, it is necessary to view them as characteristics of the students. For example, the percentages in the teacher gender section mean that 34.4% of students have a male mathematics teacher, not that 34.4% of mathematics teachers in Ireland are male.

Aside from the authors' familiarity with the Irish school system, our decision to focus on Ireland was based primarily on two reasons. Firstly, science is taught as part of an integrated curriculum in Ireland, which includes biology, chemistry, physics, and earth and space. Junior Cycle Science teachers in Ireland are trained

Variable	Proportion	Mathematics Achievement	Science Achievement
Student Gender			
Male	51.2%	523.7	521.0
Female	48.8%	524.9	526.1
Born in Ireland			
Yes	86.1%	525.1	523.2
No	13.9%	522.1	522.7
Parent's Education			
University	32.2%	552.4	555.3
Post-Secondary	19.6%	539.2	542.2
Upper-Secondary	13.7%	506.8	503.5
Lower-Secondary	3.9%	477.2	469.6
Primary/No School	2.3%	464.8	456.8
Student Unsure	28.2%	503.8	499.5
Mathematics Teacher Gender			
Male	34.4%	530.2	530.3
Female	65.6%	521.4	520.3
Science Teacher Gender			
Male	32.4%	519.8	521.2
Female	67.6%	529.8	530.5
School Average SES			
More Affluent	29.7%	544.9	546.9
Neither	44.1%	530.8	528.9
More Disadvantaged	26.2%	499.7	497.5

Table 5.1: Summary statistics for selected variables from the TIMSS 2019 data. The proportion column provides the proportion of students belonging to each category, while the mathematics achievement and science achievement columns provide the mean achievement level within each group.

in, and responsible for teaching all aspects of the science curriculum to their students. This makes Ireland well-suited to our study of homework, as there is no need to consider variations in homework requirements across the different science strands, and students are typically taught by the same teacher for the whole course which is taught as a single subject. Secondly, the sample size of 4118 students is well-suited to our machine learning approach, as it is large enough to enable the model to detect relationships and interactions between variables in the data.

5.2.2 Measures and Variables

Homework

The homework indicators we will use in this study to measure the frequency and duration of homework assignments come from the student questionnaire (TIMSS & PIRLS International Study Center, 2020). Students were asked "How often does your teacher give you homework in the following subjects?", and the possible responses for both mathematics and science were 1) "Every day", 2) "3 or 4 times a week", 3) "1 or 2 times a week", 4) "Less than once a week", or 5) "Never". The inclusion of "Every day" as a possible response introduces a minor ambiguity here, as some students might interpret it as "every day we have class", which may not equate to five days per week. Nevertheless, due to the clear and sequential structure of the response options, we believe it is reasonable to expect that most students would have correctly understood "Every day" to mean five days per week. For homework duration, the question asked was "When your teacher gives you homework in the following subjects, about how many minutes do you usually spend on your homework?", with possible responses being 1) "My teacher never gives me homework in (mathematics / science)", 2) "1-15 minutes", 3) "16-30 minutes", 4) "31-60 minutes", 5) "61-90 minutes", and 6) "More than 90 minutes". This measure of duration refers to the average time spent on each assigned homework assignment. Therefore, for a student who reports receiving mathematics homework every day, with a duration of 16-30 minutes on average, their total weekly time spent on mathematics homework is likely to be between 80, and 150 minutes.

Student interpretations of these questions could also impact results in another way. If students delay homework and complete multiple assignments in one sitting,

their reported duration may reflect accumulated work rather than the time spent on a single assigned task. This could confound the results in a complicated way that would depend upon both 1) how students interpret and respond to the TIMSS question, and 2) their own decision-making regarding homework scheduling. Additionally, there is the possibility that students who achieve lower academically might be more likely to misunderstand the question, resulting in systematic differences in reporting accuracy. This could introduce additional biases in the observed relationships between homework patterns and achievement, making it even more challenging to draw accurate conclusions. While a full investigation of these issues was beyond the scope of this study, it is important to acknowledge these potential limitations when interpreting the results.

The student responses to the TIMSS 2019 student questionnaire indicating how regularly they receive homework in mathematics and science, and how much time they normally spend on this homework are summarised in Table 5.2. Mean achievement scores observed in each category are also provided. Note that the percentages in the homework duration categories do not add to 100% as some students reported never receiving homework. An interesting observation from Table 5.2 is that homework frequency tends to be much lower in science than in mathematics, with 67% of students reporting receiving mathematics homework every day, while only 16% of students report receiving science homework with the same frequency. In contrast, homework duration appears to be remarkably similar across both subjects. A tile plot displaying the popularity of each frequency and duration combination is shown in Figure 5.1. A very popular combination in mathematics appears to be homework every day, with a duration of 16-30 minutes.

A plausible reason for the higher frequency of homework in mathematics relative to science is that it is common for more instructional time to be devoted to mathematics than to science in Ireland. Schools can exercise autonomy in this regard in Ireland (Prendergast and O'Meara, 2017), so variations will exist from school to school, but the TIMSS 2019 data showed that the most common instructional time per week in mathematics was 200 minutes, while for science the most common instructional time was lower at 160 minutes. Clearly, instructional time is likely to be strongly related to the amount of homework assigned in each subject, so this is controlled for as a confounding variable.

Homework Frequency and Duration

Frequency:	Never	Less Than Once a Week	1 or 2 Times a Week	3 or 4 Times a Week	Every Day
Mathematics	1.0% (434.4)	1.0% (484.3)	5.3% (502.6)	25.7% (513.6)	66.9% (534.9)
Science	4.2% (479.2)	16.0% (539.8)	38.7% (535.0)	25.3% (533.1)	15.8% (514.1)

Duration:	1-15 Minutes	16-30 Minutes	31-60 Minutes	61-90 Minutes	More Than 90 Minutes
Mathematics	34.8% (517.3)	43.6% (533.8)	16.7% (533.1)	2.5% (524.7)	1.4% (511.4)
Science	40.0% (531.6)	41.0% (537.1)	12.3% (526.9)	1.5% (510.9)	1.0% (458.7)

Table 5.2: Reported frequency and duration of homework - observed data from the TIMSS 2019 student questionnaire. The mean achievement of students belonging to each category is shown in brackets after the percentage. Note that as these achievement scores come directly from the observed data, care should be taken when interpreting them, as other factors are likely to be influencing the achievement levels. Homework frequency is much higher in mathematics. Homework duration is quite consistent across both subjects.

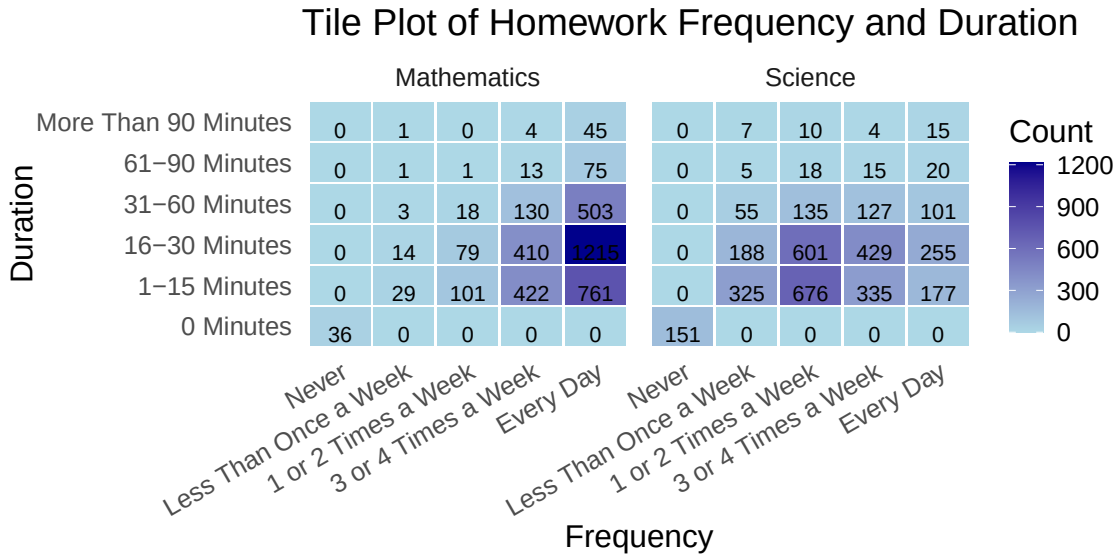

Figure 5.1: Tile plot of homework frequency and duration. The plot shows, for mathematics and science, the number of students who reported each combination of frequency and duration.

Later in our analysis, given the very small proportion of students reporting a homework frequency of less than once or twice per week, and the small proportion of students reporting a homework duration of greater than 30 minutes, we will focus on the aggregated categories of "up to once or twice per week", "three or four times per week" and "every day" for frequency, while for duration we will focus on the categories of "1-15 minutes", "16-30 minutes", and "greater than 30 minutes". This will help to ensure a more parsimonious model, allow for a simpler interpretation, and reduce the computational burden on our model.

The mathematics and science teachers were also asked how often they give homework to their students, and how long they believe the homework should normally take the "average" student in their class. Therefore, we could have used the teacher reports as our data instead, but decided the student reported version would be most appropriate for two reasons. Firstly, the teacher reported duration of homework only indicates how long the homework should take for the "average" student in the class, and fails to capture individual level differences which are bound to exist, and which the students themselves are best placed to report. Second, we wanted to be able to account for the possibility of homework differentiation, whereby a teacher assigns different amounts of homework to individual students based on their ability (see for example Keane and Heinz, 2019). This can also only be accomplished with the student version of the reports.

Student Achievement

Student achievement in TIMSS is measured using a short assessment in both mathematics and science. This standardised assessment is carefully designed and validated by the TIMSS study organisers to ensure its validity and reliability for measuring student achievement (Martin et al., 2020). The assessment is carefully structured to cover a diverse range of topics, including algebra, geometry, and various scientific strands. However, to assign every single item from the pool of available questions to every student would make the assessment unduly long. Therefore a rotated booklet design is employed whereby each student answers a booklet of questions in mathematics and science, each containing a small subset of the entire pool of questions. By ensuring an overlap exists between assessment

booklets, i.e., booklet one shares common questions with booklet two, and booklet two shares common questions with booklet three, it is possible to estimate the relative ability levels of students whilst ensuring that a wide variety of topics are covered without burdening students with a very long test. For the first time in the history of TIMSS, some countries issued the assessment electronically via tablets or other means in 2019, but Ireland opted for a traditional paper assessment.

Given the difficulty of accurately estimating any student's achievement level based on a limited subset of questions, assigning a single point estimate of achievement to any one student would be misleading. Therefore, to better capture the uncertainty that is present in the achievement estimates of the students, TIMSS supplies five plausible values for each student's achievement in mathematics and science. These five plausible values were fully accounted for in our analysis by appropriately pooling the estimates from five separate models, each of which were applied to one of the five plausible value estimates. Figure 5.2 shows a scatter plot of student achievement in both subjects using the first of these plausible values. A very strong positive correlation is evident ($\rho \approx 0.85$).

Control Variables

In addition to homework and student achievement in mathematics and science, TIMSS collects data on many other important factors such as socioeconomic status, teacher experience, and school emphasis on academic success. These factors can act as confounding variables, simultaneously influencing both the likelihood of receiving treatment and the outcome of interest, potentially biasing results (Greenland et al., 1999). For instance, teachers in academically focused schools with already high performing students may assign more homework, believing it enhances achievement. Alternatively, teachers under pressure to improve student performance due to low instructional quality may assign more homework as a means of compensation, obscuring the true relationship between homework and achievement (Rønning, 2011). Given the ability of confounding variables to lead to unexpected findings, it is important to control for their effects wherever possible. Therefore, we have taken steps to control for the influence of many potential confounding variables from the TIMSS study. A full list of the TIMSS variables

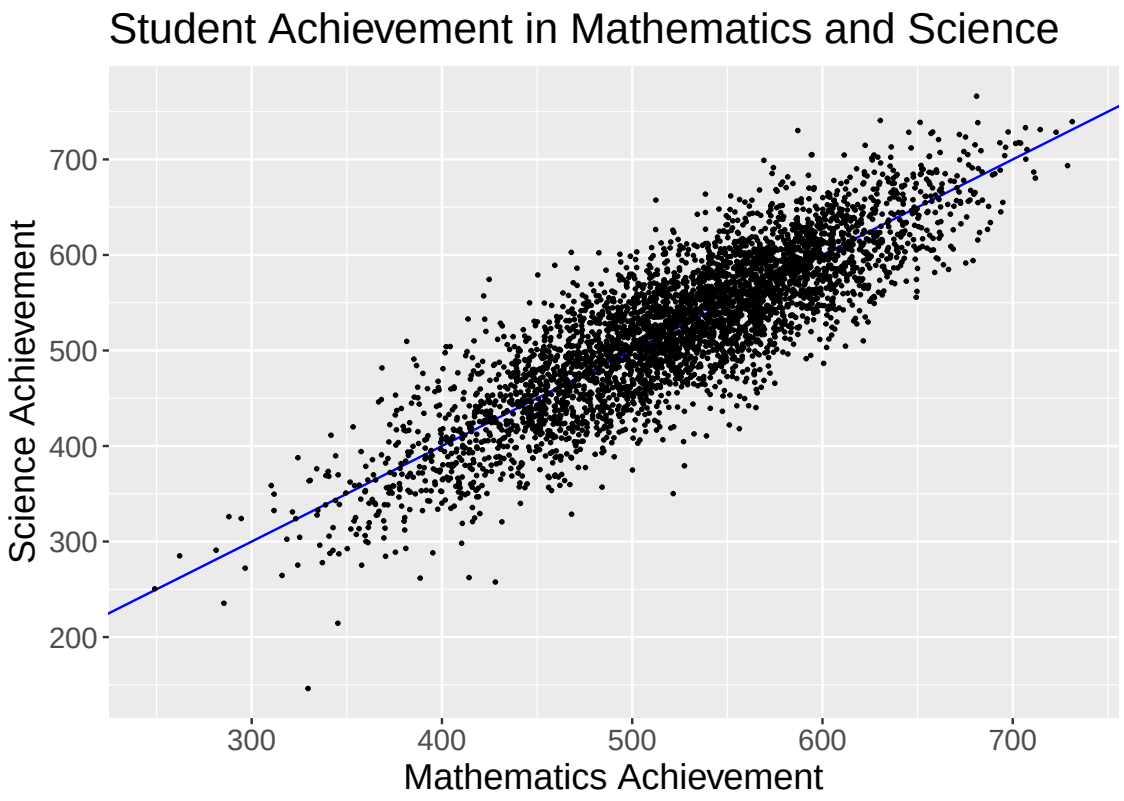

Figure 5.2: Scatter plot of student achievement in mathematics and science, using the first plausible value. There is a very strong positive correlation between student achievement in mathematics and science ($\rho \approx 0.85$). The blue line added as a visual aid is the line $y = x$.

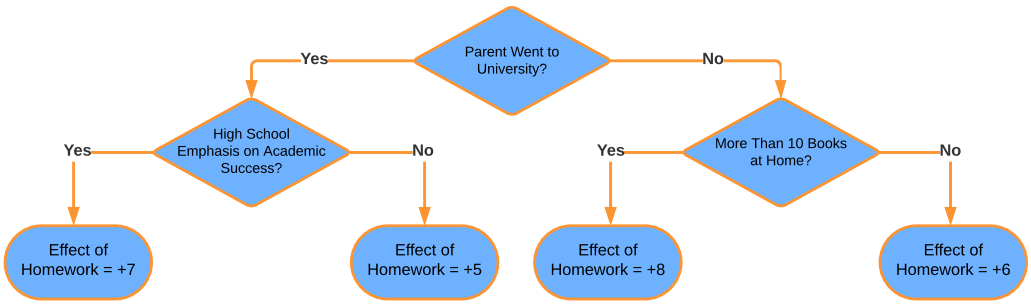

Figure 5.3: Example of a BCF model with a single decision tree used to make treatment effect predictions. The decision rules direct observations from the root of a tree to its terminal nodes where each observation is assigned a prediction. In this purely illustrative example, the rules say that for a student with a parent who went to university, and a high school emphasis on academic success, homework increases achievement by 7 units.

used in this study can be found in the supplementary materials.

5.2.3 Modelling Approach

Our study makes use of Bayesian Causal Forests (BCF, Hahn et al., 2020), a causal inference machine learning algorithm based on the highly flexible modelling tool called Bayesian Additive Regression Trees (BART, Chipman et al., 2010). BART and BCF are both tree based models, meaning they make predictions by creating a set of decision rules. When followed, these decision rules form a pathway directing observations from the root, down to the leaves of a tree, where all observations belonging to a leaf are assigned a prediction. See Figure 5.3 for an example. The example used in Figure 5.3 has deliberately been made very simple for illustration, but in practice, the predictions from BART or BCF are usually made by combining the outputs from multiple decision trees. BART and BCF can also adapt to the complexity of the data by increasing or decreasing the tree sizes, and modifying their decision rules as necessary. In this way, BART and BCF are capable of capturing very complicated patterns in the data such as interactions between variables and non-linear relationships.

A particularly powerful feature of BART and BCF is their ability to perform automatic variable selection. These models automatically assess the importance of different variables while building the trees, allowing them to prioritise and select the most relevant variables for making accurate predictions. This feature allows researchers to include a wide range of predictors that might not be feasible in more standard approaches, such as linear models, which often require manual selection to avoid issues like multicollinearity. For these reasons, BART and BCF have become increasingly popular and have been used in a diverse set of fields such as medicine, economics, and education (e.g., Spanbauer and Sparapani, 2021; Pierdzioch et al., 2016; Suk et al., 2021; Prado et al., 2021a).

To separate causation from mere correlation, the BCF model relies on a statistical framework known as the Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Sekhon, 2008). This framework, also known as the potential outcomes framework, is based on the idea that for every observation, or student in the dataset, there are two potential outcomes, or values of student achievement that may be observed: One achievement score that would be observed if the student was impacted by some treatment or intervention (such as homework), and one that would be observed if the student was not.

Estimation of causal effects using BCF requires several key assumptions:

(1) **Stable Unit Treatment Value Assumption (SUTVA).** Each student's potential outcomes are unaffected by the treatment assignment of other individuals (no interference), and the treatment is well-defined without multiple versions.

(2) **Ignorability (Unconfoundedness).** Conditional on observed covariates x_i, the treatment assignment Z_i is independent of the potential outcomes: $y_i(Z_i = 1), y_i(Z_i = 0) \perp\!\!\!\perp Z_i \,|\, x_i$. This implies that, after adjusting for covariates, there is no unmeasured confounding.

(3) **Overlap (Positivity).** For all individuals, the probability of receiving treatment given covariates must be strictly between 0 and 1: $0 < P(Z_i = 1 \,|\, x_i) < 1$. This ensures that comparisons between treated and untreated groups are possible across the range of covariates.

Provided these assumptions hold, the BCF model uses two sets of decision trees to predict these outcomes. The first set, often labeled as μ, predicts the achievement of each student i without the treatment based on their characteristics x_i. The second set, labeled as τ, predicts the change in achievement directly attributable to the treatment. By combining these predictions, the BCF model estimates both potential outcomes and the causal effect of homework from other associations in the data:

Predicted achievement of student i without homework (Potential Outcome 1) $= \mu(x_i)$

Predicted achievement of student i with homework (Potential Outcome 2) $= \mu(x_i) + \tau(x_i)$

Using $\tau(x_i)$ to represent our estimate of what is known as a conditional average treatment effect (CATE), the estimate of the effect of receiving treatment for all individuals with covariates x_i, we may also calculate the mixed average treatment effect (MATE), which is the average of these estimates over the population of interest.

Specifically, this study adopts a multivariate extension of BCF developed by McJames et al. (2024) (Chapter 4), which allows BCF to be applied to multiple outcome variables such as mathematics and science achievement simultaneously. The structure of the model remains the same, but instead of providing a single prediction at the leaves of the tree as shown in Figure 5.3, a separate prediction is provided for each outcome, or in our case, subject of interest.

In addition to allowing the model to provide predictions for multiple outcomes, given that the TIMSS homework data contains multiple levels of frequency, ranging from never to every day, and multiple categories of duration, ranging from less than fifteen minutes to greater than ninety minutes, we also allow the model to provide different estimates of the effect of homework based on the given frequency and duration. In summary, while the standard BCF model would be limited to analysing one subject at a time, and to estimating the impact of receiving any amount of homework relative to receiving no homework, the extended multivariate BCF model used in this study allows us to estimate the varying effects of homework frequency and duration in both mathematics and science.

As with all chapters in this thesis, we assessed the overlap assumption before fitting the model to the data. Given the multiple levels of homework frequency and duration in this study, this was performed by treating each additional homework

category as a distinct treatment and assessing overlap for each one separately. As before, this was evaluated by visually inspecting density plots of the propensity scores for belonging to each each homework treatment category relative to the reference group, and ensuring there were no regions lacking common support. Similar visual checks were also performed for all continuous covariates, while contingency tables were created for categorical variables to confirm that no levels contained only units belonging to the reference group of homework frequency and duration, or the level of frequency or duration for which overlap was being assessed.

The TIMSS data employed in this study possesses a clear hierarchical structure, whereby students are nested within schools. To account for this, our model is equipped with a random intercept term, which serves to shift the model estimates for each individual up or down, depending on the class they belong to. The inclusion of this random intercept term, which acknowledges the hierarchical nature of the data, introduces an extra level of robustness to the model. Full technical details associated with all aspects of the model can be found in the supplementary material.

Across all the variables used in our study, 6% of the data was missing. In order to avoid losing data by completely deleting rows with missing values, we imputed our dataset using the R package `missRanger` (Mayer, 2019). This procedure involves replacing missing entries with a value based on the data that is available for each observation. With this approach we are able to maximise the data available for use, but we acknowledge there is a degree of uncertainty associated with these imputed values which is not accounted for in our main study. Therefore, the true 95% credible intervals for our results reported later are likely to be slightly broader than displayed. As an alternative approach for handling missing data, the supplementary material also includes results from a version of the analysis that did not rely on imputation of missing data, and instead relied on complete cases only. This approach led to broadly similar results, but we will focus on the imputation based approach as it maintains a nationally representative sample of students.

5.3 Results

A summary of our results concerning the effect of homework on student achievement can be found in Table 5.3. Here we report the estimated mixed average treatment effects (MATEs) of providing homework with a specified frequency or duration relative to a reference level of up to one or two times per week for frequency, and up to 15 minutes each time for duration. The uncertainty in these values is reflected by the 95% credible intervals provided alongside each effect estimate. The 95% credible intervals represent the range of values within which the MATE lies with 95% probability.

For context, it helps to know that student achievement in mathematics at the eighth grade in Ireland follows an approximately normal distribution with a mean of 524 and a standard deviation of 73. Student achievement in science also follows an approximately normal distribution, but with a mean of 523 and a standard deviation of 83. Therefore, for ease of interpretation, in addition to the original effect estimates which are reported on the TIMSS achievement scale, we also provide normalised versions. These normalised values indicate how many standard deviations the effect estimates correspond to on the mathematics and science achievement scales.

Returning first to research question one, which concerns the effect of homework frequency on student achievement in mathematics and science, we see that in mathematics the MATE and 95% credible interval for providing homework every day is 7.51 and (1.63, 16.57). This MATE represents the average change in mathematics achievement that we would expect to see after increasing homework frequency from up to one or two times per week to every day. The 95% credible interval can be interpreted as meaning that we are 95% sure the true effect is in the range 1.63 to 16.57. In science, increasing homework frequency to every day does not have a clear benefit. Instead, the optimal homework frequency for science is identified as three or four times per week. Providing homework at this frequency is expected to improve student achievement by 5.31 points on the TIMSS scale, and by between 2.35 and 8.25 points with 95% probability.

Our second research question focuses on the effect of homework duration on student achievement in mathematics and science. Looking at the results for mathe-

Average Treatment Effect Results

Frequency:	3 or 4 Times a Week		Every Day	
	Achievement Points	Scaled Version	Achievement Points	Scaled Version
Mathematics	2.88 (-3.61, 10.88)	0.04 (-0.05, 0.15)	7.51 (1.63, 16.57)	0.10 (0.02, 0.23)
Science	5.34 (2.35, 8.25)	0.07 (0.03, 0.10)	-1.94 (-5.96, 2.28)	-0.02 (-0.07, 0.03)
Duration:	16-30 Minutes		More Than 30 Minutes	
	Achievement Points	Scaled Version	Achievement Points	Scaled Version
Mathematics	-0.98 (-3.85, 1.94)	-0.01 (-0.05, 0.03)	0.12 (-3.37, 3.72)	0.00 (-0.05, 0.05)
Science	1.80 (-1.37, 5.32)	0.02 (-0.02, 0.06)	3.27 (-2.20, 9.30)	0.04 (-0.03, 0.11)

Table 5.3: Average treatment effect results from the multivariate BCF model. The results show the estimated effect of providing homework with the specified frequency or duration on student achievement in mathematics and science. The reference level for homework frequency is up to one or two times per week, and the reference level for homework duration is up to 15 minutes. 95% credible intervals are shown in brackets. Scaled versions of the results indicating how many standard deviations the effect sizes correspond to are also provided. Mathematics achievement benefits the most from a frequency of every day. Science achievement, on the other hand, benefits most from a frequency of 3/4 times per week. Increasing homework duration beyond 15 minutes each time does not provide clear benefits in either mathematics or science.

matics, we see that relative to only providing homework that lasts up to 15 minutes each time, there is no clear benefit to increasing homework duration to 15-30, or even more than 30 minutes. Similarly, in science, as the 95% credible intervals for all effect estimates overlap with 0, there is insufficient evidence to say that longer duration homework assignments will lead to increased achievement. Therefore, in the absence of stronger evidence, homework assignments lasting up to 15 minutes each time may be considered equally as beneficial as longer ones.

Based on the above analysis, we can also answer our third research question, which asks if the optimal distribution of homework throughout the week may be different in mathematics and science. With respect to homework duration, given the lack of a clear benefit from increasing homework duration beyond 15 minutes, there is no indication that any duration should be preferred over any other. However, due to the negative impact that long homework hours may have on non-academic outcomes, it may be reasonable to suggest that adopting a policy of up to 15 minutes in both subjects may be the best option, as it is equally as beneficial as the longer homework durations. For frequency, the picture becomes more nuanced, because in mathematics the greatest gains were made by increasing homework frequency to every day, while in science the best results stemmed from a frequency of three or four times per week. This indicates that the optimal distribution of homework throughout the week is in fact subject dependent.

The effect sizes reported in Table 5.3 would be considered by some traditional measures to be very small (LeCroy and Krysik, 2007). However, it is important to note that in contrast to conventional effect sizes commonly reported in terms of Cohen's d, which are based purely on the raw difference in outcomes between two groups, and susceptible to the effects of confounding, this effect size has a causal interpretation. Furthermore, while larger effect sizes are commonly reported in the literature (e.g. Hattie, 2008), they are often based on factors that are not easily modified, and consequently provide no route to encouraging or enabling higher student achievement. The effect sizes reported here however, result from homework which can be easily modified by teachers and incorporated into updated education policy. Therefore, given the causal interpretation of the results, and the straightforward route they provide to enhancing student achievement, they should not be considered unimportant. We also emphasise that while some of the effect

estimates may be considered quite small, the uncertainty in the estimates allows for the possibility that they may be slightly larger and the values reported at the upper limits of the credible intervals are much more substantial. After taking these factors into account, an effect size of the magnitude of 0.1 can be thought of as a small to medium effect size (Kraft, 2020).

Finally, motivated by our fourth research question, we explored the potential moderating role of parental education level and home educational resources on the effect of receiving homework. These variables were chosen because they are useful and commonly adopted proxies for socioeconomic background when analysing TIMSS data (e.g. Broer et al., 2019; Heppt et al., 2022). A visual representation of our results can be found in Figure 5.4. The results show no clear trend, indicating that regardless of the number of books a student has at home, or the level of education of their parents, all students are predicted to benefit by approximately the same amount from the homework they are assigned. The lack of a clear trend here is evident in both subjects and at both frequencies, suggesting students from advantaged socioeconomic backgrounds do not benefit significantly more from homework than their peers, at least in the eighth grade.

5.4 Discussion

Our study advances the existing literature on the effect of homework by applying a causal inference machine learning approach to investigate its relationship with academic achievement. To our knowledge, ours is the only study to have employed such an approach. We have examined the impact of homework frequency and duration without making any assumptions about their relationship with achievement. In this way we are able to account for complicated interactions between variables and non-linear relationships which might otherwise go undetected by standard analytical techniques. Moreover, our model accounts for a wide range of confounding variables, eliminating the potential bias they may introduce to our results. Furthermore, the multivariate nature of our model enables us to jointly investigate both mathematics and science, which is important as subject specific differences are known to exist in the effect of homework on academic achievement (Eren and Henderson, 2011; Fernández-Alonso et al., 2017).

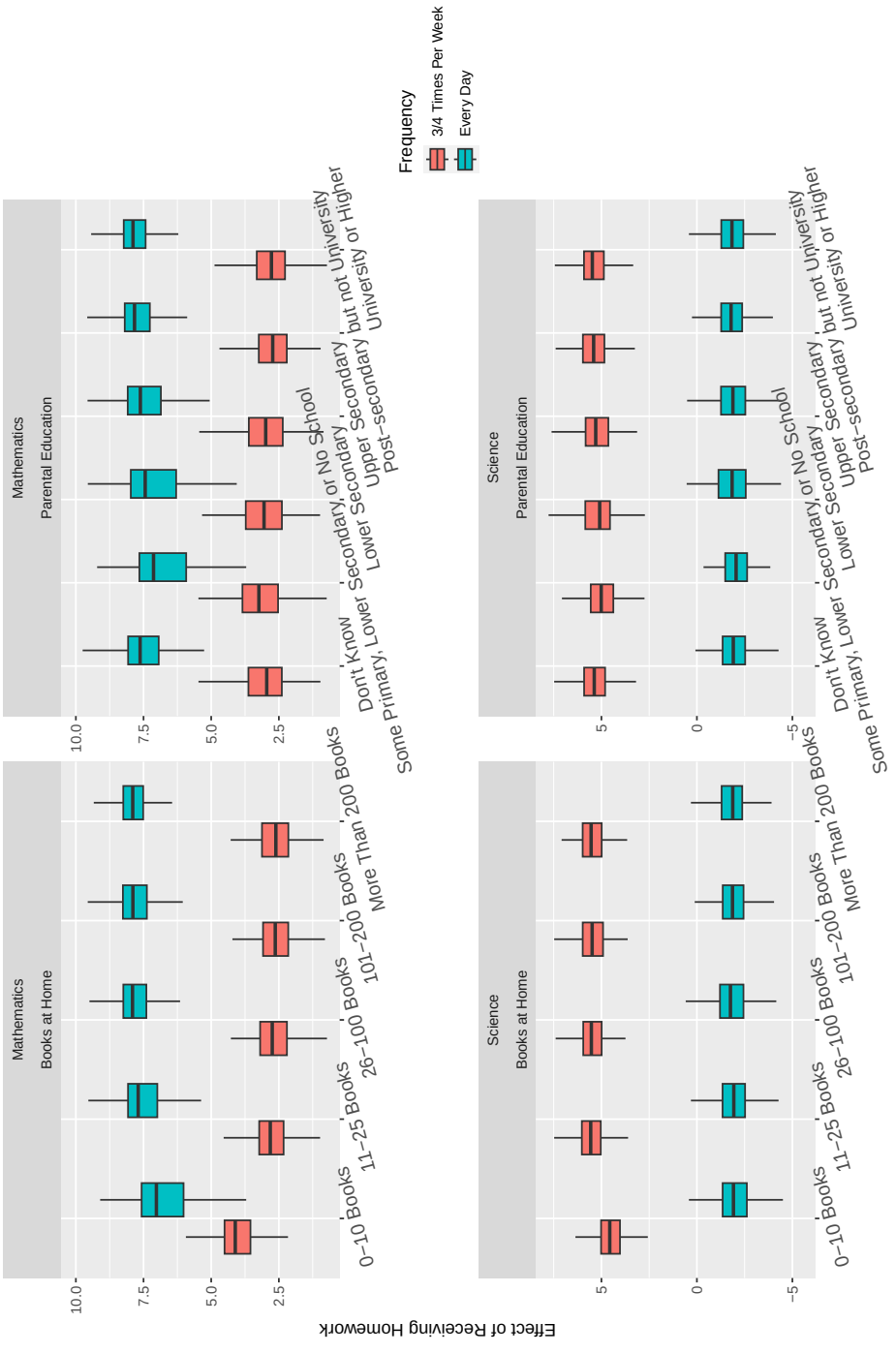

Figure 5.4: Plot of the variation in the magnitude of individual treatment effect estimates. The lack of a clear trend in either subject or at either frequency indicates that students with more books at home, and highly educated parents, do not benefit significantly more from homework, at least in the eighth grade level.

Our main finding from this study is that homework frequency is more important than homework duration for improving student achievement in mathematics and science. This finding is supported by previous research which shows that regular homework assignments are predictive of increased achievement, whereas weekly time spent on homework is not (e.g., Trautwein et al., 2002; Trautwein, 2007; Fernández-Alonso et al., 2019). It is also in line with the work of Kang (2016) who argues that policy changes promoting spaced repetition learning strategies can help to improve academic outcomes. As a result, educators may consider breaking down large homework assignments into smaller, daily tasks that offer regular and spaced opportunities to engage with the material, rather than assigning a single, long homework problem set.

Our finding that spending up to fifteen minutes on homework each time is equally as effective as spending up to an hour or more on homework is also in agreement with previous studies which have found that spending too much time on homework can lead to diminishing returns (Cooper, 1989). Researchers have long been wary that too much time spent on homework by students can have negative effects (e.g., Kohn, 2006; Buell, 2008). Our results can therefore help to bridge the gap between those who advocate for homework as a tool for improving academic outcomes (e.g., Cooper et al., 2006), and those who are concerned by the negative effects that unduly long homework assignments can have on aspects such as social development (e.g., Bennett and Kalish, 2007). By providing a middle ground based on frequent homework assignments of short duration, our study's recommendation allows students to benefit from homework while also enabling them to spend time on other pursuits in the evenings.

In line with previous research, our study revealed differences in the effect of homework on student achievement between mathematics and science (Eren and Henderson, 2011; Fernández-Alonso et al., 2017). However, in contrast to findings by Eren and Henderson (2011), who found homework only to be beneficial in science, our study found that homework can have a positive effect in both subjects. Specifically, assigning homework every day in mathematics was found to be associated with the largest increase in student achievement, while a frequency of three to four times per week was identified as being the most beneficial in science. These findings have important implications for educators and policymakers who

can use this information to tailor their homework assignments to each subject and optimise the potential benefits for students.

As discussed earlier, mathematics is typically assigned more instructional time per week than science in schools in Ireland. This difference in time allocation may have an impact on homework frequency and duration by affecting the number of classes taught per week in each subject. Due to limitations of the data, a detailed exploration of these subject specific differences was not possible. We note that while this is a limitation in terms of contextual understanding, it is unlikely to impact the results from our model as we have controlled for the total instructional time per week, and also accounted for the hierarchical nature of the data with our model. However, further research should continue to explore these subject-specific differences related to homework efficacy and identify the underlying mechanisms that contribute to them.

Our secondary exploration of the moderating role of home resources and parental education found no evidence of students from advantaged socioeconomic backgrounds benefiting more from homework than other students. This is in contrast to previous findings from Daw (2012), and Rønning (2011) who both found that advantaged students experienced a greater increase in achievement as a result of completing homework. However, we believe that the different ages of the students involved in these studies could offer a plausible explanation for this disparity. In the study conducted by Rønning (2011), the average age of the students was ten, which is nearly five years younger than the eighth-grade sample used in our study. This age difference could partially account for Rønning's finding that students with highly educated parents derive greater benefits from homework, as younger students tend to gain more from parental involvement in homework due to their ongoing development of study habits, as found by Patall et al. (2008). On the other hand, Daw's study involved older students from grade eight to twelve. At higher grade levels, as students delve into more advanced topics, their ability to utilise their home educational resources for research purposes or seek assistance from their parents may become more crucial. Therefore, the strength of the moderating role of home resources and parental education may differ depending on the ages of the students involved. Such differences warrant further investigation but were beyond the scope of our study.

While our study focuses on academic outcomes, it is important to acknowledge the non-academic considerations associated with homework. From a beneficial perspective, homework has been reported to promote time management skills and responsibility among students (Cooper and Valentine, 2001). However, it can also contribute to increased stress, reduced motivation, and limited free time for extracurricular activities (Galloway et al., 2013). Parents may also feel pressure to assist their children with homework (Pressman et al., 2015). Therefore, the potential positive effects of homework on student achievement must be weighed against these potential negative consequences.

A limitation of our study is that our investigation of homework was limited to the effect of homework frequency and duration. Other characteristics of the homework such as its quality and suitability for a given student's ability level are known to be important, however. Dettmers et al. (2010), for example, found that students who perceived their homework to be well selected for them had higher levels of motivation to complete their homework. In the same study it was also found that students reporting their homework to be challenging put less effort into their homework than students who found the difficulty more manageable, thus highlighting the importance of homework being well aligned with a student's ability level to be most beneficial. One possibility for addressing this additional factor may be to combine each student's version of the homework report, with that of their teacher, in order to identify students who spend much more or less time on homework than expected. Investigating how these discrepancies in expected and actual time spent on homework may impact achievement is an intriguing area for future work.

An important ethical consideration arising from this study concerns the broader implications of promoting homework as a means to improve academic achievement. While socioeconomic status (SES) was treated as a confounding variable, given its potential to influence the likelihood of students receiving homework, this methodological decision may inadvertently obscure deeper structural inequalities that shape students' educational experiences. Moreover, although our findings support the effectiveness of frequent, short homework assignments, it is essential to recognise that students' capacity to benefit from homework is not equally distributed. Factors such as access to a quiet study environment, availability of

parental support, and freedom from household responsibilities can significantly influence a student's ability to engage with homework meaningfully. Although our analysis did not identify differential effects of homework based on SES, this absence of evidence should not be taken as evidence of equality in homework conditions. On the contrary, the strong positive effects identified in this study underscore the need for future research to explore how homework practices might be better adapted to support all students, particularly those facing structural disadvantages, in order to ensure that the benefits of homework are equitably distributed.

A second important limitation is that the BCF model employed in this study makes a number of important assumptions. Notably, the model assumes that we are able to account for all possible sources of confounding. Whilst we have endeavoured to control for as many potential sources of confounding as possible in our model, it is certainly possible that there may be some unaccounted for confounding variables that were not controlled for or not collected as part of the TIMSS study.

Another limitation of our study is that the model specification does not explicitly acknowledge the potential for interaction effects between the different levels of homework frequency and duration that we have investigated. For example, it may be possible that the effect of homework duration may be dependent upon the frequency at which homework is given, but such effects are not estimated by the model we have used. Including additional parts in the model which would capture these interaction effects is possible, but would come with a significant overhead of increased computational demand, both in terms of running time and memory usage, so we decided to maintain a simpler model for this study. Therefore, investigating interaction effects between homework frequency and duration remains an interesting area for future research.

Finally, our results are limited to an eighth grade sample of mathematics and science students from Ireland. As demonstrated by our results, however, the effect of homework is different in mathematics and science, and therefore effect size differences in other subjects such as English or foreign languages can be expected. Grade level differences are also known to be very important, with homework effects typically found to be more positive for older students than younger students (Cooper, 1989; Cooper et al., 2006). Furthermore, education policy in different

countries can have a significant effect on homework and student achievement. Fan et al. (2017), for example, found that the positive effects of homework were greater in the US than in East Asian countries, despite the students in Asia often having higher motivation and a better attitude towards homework. Future work could therefore include an application of a similar method to ours to data from different subjects, grade levels, or other countries to improve the generalisability of our findings. Notably, as TIMSS provides data on many countries, many of which also teach science as part of an integrated curriculum as is the case in Ireland, an application of our methodology to countries such as Singapore, the US, England, or others would be of interest.

5.5 Conclusion

This study investigated the impact of homework frequency and duration in both mathematics and science, employing a new modelling approach specifically designed to address the research questions of our study. Our results demonstrate a clear positive effect of increasing homework frequency, but not homework duration. We therefore recommend that frequent homework assignments of short duration may be most effective for improving student outcomes. This strategy can help to promote academic achievement whilst avoiding the potential drawbacks associated with many hours spent on homework (Zhao et al., 2024; Galloway et al., 2013). Our second research finding was that students from disadvantaged socioeconomic backgrounds did not benefit less from homework than their peers. This has important implications for homework policy, as it challenges the suggestion from some researchers that homework may contribute to widening achievement gaps across socioeconomic groups. As a result, we also recommend that homework should continue to play an important role in the learning process, as we find no evidence of disparate impacts across socioeconomic groups. Finally, given our results pertain only to eighth grade students studying mathematics and science in Ireland, we strongly advocate for additional studies focused on data from other countries, subjects, or grade levels.

6

Bayesian Causal Forests for Longitudinal Data: Assessing the Impact of Part-Time Work on Growth in High School Mathematics Achievement

6.1 Introduction

For many high school students, part-time jobs have become an integral part of their daily routine, just as important as homework, studying, and completing assignments (Singh and Ozturk, 2000). The reasons for seeking part-time work can vary widely among students. Some work to support their families financially, others to develop their character, gain maturity, or simply to earn spending money (Kablaoui and Pautler, 1991). Regardless of the reasons for students choosing to work part-time, however, this work can have a significant impact on their educational journey (Bachman and Schulenberg, 2014). Our study introduces a new approach for modelling individual level growth in student achievement, and explores the causal effect of intensive part-time work on this growth, where intensive part-time work is defined as upwards of 20 hours of work per week during the school year (Lee and Staff, 2007).

Estimating causal effects from longitudinal data is a challenging but essential task. Established methods include inverse probability weighting (Hogan and

Lancaster, 2004), two-way fixed effects (Imai and Kim, 2021), and difference-in-differences (DiD, Donald and Lang, 2007). A key limitation of many of these approaches is that they often rely on strong assumptions that may not be appropriate for the target data. The parallel trend assumption of the difference-in-differences method, for example, assumes that the treatment group would have followed a similar trajectory to the control group had they not received treatment (Roth et al., 2023). This can easily be violated in practice, as confounding variables may influence both the probability of receiving treatment and the trajectories in the outcome of interest. Students who self select into part-time work, for example, may experience less growth than their peers even without part-time work (Monahan et al., 2011). Some work has been conducted to tackle this limitation by relaxing the assumption of parallel trends conditional on covariates (Abadie, 2005; Callaway and Sant'Anna, 2021), but important limitations remain.

Other methods rooted in structural equation modelling such as G-estimation (Robins, 1997), and longitudinal extensions of targeted minimum loss based estimation (Lendle et al., 2017) excel in estimating causal effects from longitudinal data when faced with challenges such as drop-out, time varying covariates, and dynamic treatment regimes. A weakness of these methods, however, is that they are often restricted to estimating average causal effects, without the ability to explore individual level variations or heterogeneity in responses to treatment. This is an important limitation, especially in the context of part-time work, as there is research to show that the effects of part-time employment can vary significantly depending on factors such as gender, motivations for working part-time, and socioeconomic backgrounds (Entwisle et al., 2000).

When understanding heterogeneity in causal effects is important, Bayesian nonparametric methods based on Bayesian Additive Regression Trees (Hill, 2011; Chipman et al., 2010) and Bayesian Causal Forests (Hahn et al., 2020) have emerged as leading approaches. The default implementations of these methods are only applicable to single time point observational data, however, precluding the study of trends in educational outcomes over time. Our study extends BART and BCF to the setting of longitudinal data. By combining the flexibility of these methods with the highly interpretable structure of the difference-in-differences model, we simultaneously relax the parallel trend assumption of the DiD methods,

while also allowing for the study of individual level variations in the growth curves of student achievement, and the heterogeneous impact of part-time work on this growth.

While other studies (Wang et al., 2024) have also introduced longitudinal extensions of BART and BCF, with a focus on situations where there is a staggered adoption of treatment, our proposed model assumes a very different structure, and includes three important features targeted specifically at our motivating data. First, our model places separate priors directly over the growth trajectories and the effects of treatment on this growth. This allows us to inform the model with prior information, and comes with the added advantage of allowing the incorporation of model explainability tools, and variable importance metrics directly associated with the parameters of interest (Inglis et al., 2022a,b). The model structure also accommodates time varying covariates, such as evolving levels of student motivation which are common in education studies. Finally, while not the main focus of the present study, an additional feature not included in BCF models before is the ability to handle missing data in the covariates or the treatment assignment. We tackle this issue with a feature borrowed from Kapelner and Bleich (2015), and a novel update step for the treatment status indicator.

The remainder of this chapter is structured as follows: In Section 6.2 we describe our motivating dataset, the High School Longitudinal Study of 2009 (HSLS), and outline some key features of the data. Section 6.3 introduces the proposed model, and shows how we extend BART and BCF to provide a foundation for estimating growth curves of student achievement and heterogeneous treatment effects of part-time work. To further support the credibility of our proposed methodology, Section 6.4 applies our model to simulated data designed to mirror the characteristics of the HSLS dataset. We benchmark our performance against other potential candidate models, showcasing the unique capabilities of our model in overcoming challenges that remain difficult for existing approaches. In Section 6.5, we deploy our model to the HSLS data and present the results of our study. Finally, we conclude this chapter in Section 6.6 with a discussion of our findings, implications for policy, and areas for future work.

6.2 Data

The High School Longitudinal Study of 2009 (Ingels et al., 2011) is an ongoing study of a nationally representative sample of high school students in the US. It is the most recent of a series of five longitudinal studies launched by the National Center for Education Statistics. The first wave of data collection for HSLS took place in the fall of 2009 at the start of the academic year when the students were in the ninth grade. More than 20,000 high school students took part in this part of the study. A follow up of these students then took place in the spring of 2012 when the students were in the eleventh grade. Further follow ups have also taken place in 2013, 2016, and 2017 to discover how the students are progressing in the years after high school, but did not involve mathematics achievement tests so we will focus solely on the first two waves of the data. Due to some students dropping out of high school, some schools closing, and others disagreeing to continue their participation in the study, the second wave of data collection involved just under 19,000 of the original Wave 1 sample.

Data collection during HSLS followed a similar procedure during both waves of the study. Mathematics achievement was assessed during both waves using a computer delivered assessment with questions designed to measure the algebraic reasoning abilities of the students. The resulting achievement estimates assigned to the students were calculated using Item Response Theory (Cai et al., 2016). The contextual data gathered as part of the study was based on a survey answered by the students, a parent, school administrators, and school teachers. Information collected from the student survey includes characteristics such as sex, age, ethnicity, self concept in mathematics, sense of school belonging, and other details such as participation in activities like part-time work. Data from the parent survey includes important socioeconomic variables such as family income, parental employment and education. School related data includes information such as the administrator's perception of the overall climate within the school, and the level of expectations of student academic success.

To ensure a representative sample of students, a stratified, two-stage random sampling design was employed by the study organisers. This involved first approaching eligible high schools, 944 of which agreed to participate in the study,

and then randomly sampling students from each of those schools, leading to a total sample of 21,444 participating students. Sampling weights resulting from this design are provided in the dataset to account for non-participation bias and were used to appropriately weight the results discussed later in the chapter. Table D.1 of the supplementary material provides weighted summary statistics for a subset of the categorical variables from the base year (Wave 1), and also provides mean achievement levels from both waves of the study.

Our study uses the public-use version of the HSLS data. Some of the data in this public use version of the dataset has been obfuscated or removed in order to maintain the anonymity of the students and the schools who took part in the study. Therefore, a school identifier indicating which students attend the same school is not available in this version of the dataset, precluding a hierarchical modelling strategy. The restricted use version of the dataset does include this information but is only available with strict controls in place. This is a limitation of our study, but ensures our results are more easily reproducible without requiring a restricted use version of the dataset. Furthermore, there is evidence to suggest that part-time work is more likely to be influenced by student and family related variables than school related variables (Howieson et al., 2012), partially mitigating the potential for unmeasured confounders to bias our results.

6.3 Methodology

6.3.1 The Model

Our motivating dataset consists of two waves, but for the sake of generality in this section we will describe how the model applies to datasets of up to T waves of student data. We are interested in modelling trajectories of student achievement where we have data on n_1 students participating in an initial base year assessment, and subsequent follow-ups on $n_2 \dots n_T$ of the same students during waves 2 to T. We allow for the possibility of drop-out, whereby $n_T \leq n_{T-1} \leq \ldots \leq n_1$. We will represent the contextual data associated with student i up to time t by $x_{i,t}$, where $t = 1$ indicates the data is from the base year (Wave 1), and subsequent values of t indicate the data encompasses extra information collected up to and including

Wave t. We will not distinguish between data from different surveys or questionnaires, so $x_{i,t}$ captures all of the student, parent, and school level data associated with student i up to time t. Given the accumulation of information on students over time as they complete more surveys from additional waves, the number of columns in $x_{i,t}$ will be less than the number of columns in $x_{i,t+1}$. To distinguish between students who do and do not work part-time, we will let $Z_{i,t+1}$ be a binary indicator of length n_{t+1} which indicates for each student if they reported having a part-time job which involved them working on average 20 hours or more per week during the period between Waves t and $t+1$. For the achievement data, let $y_{i,t}$ denote the observed mathematics achievement of student i recorded at time t.

Our research questions concern two quantities of interest. The first is related to the growth in mathematics achievement between Waves t and $t+1$, which we will denote by $G_{i,t+1} = y_{i,t+1} - y_{i,t}$. The second concerns the impact of part-time work on this growth. To understand this impact, we adopt the Neyman-Rubin causal model (Splawa-Neyman et al., 1990; Sekhon, 2008), and postulate that for each individual i, there are two potential growth values. One that would be observed if the student worked part-time, $G_{i,t+1}(Z_{i,t+1} = 1)$, and one that would be observed if the student did not, $G_{i,t+1}(Z_{i,t+1} = 0)$. With these quantities defined, the impact of part-time work on the growth in student achievement during this period is captured by $\tau_{i,t+1} = G_{i,t+1}(Z_{i,t+1} = 1) - G_{i,t+1}(Z_{i,t+1} = 0)$. Of course, we only ever observe one of these potential growth values, namely $G_{i,t+1} = G_{i,t+1}(Z_{i,t+1} = 1)Z_{i,t+1} + G_{i,t+1}(Z_{i,t+1} = 0)(1 - Z_{i,t+1})$, so we make the following assumptions:

Assumption 1: The Stable Unit Treatment Value Assumption. We assume that the potential growth values of every student i between periods t and $t+1$ are independent of whether or not any other student j worked part-time in any period. We also assume that the treatment, part-time work, is consistently defined for all individuals.

Assumption 2: The Sequential Ignorability Assumption. We assume that conditional on their observed characteristics and treatment history up to the period of interest, the potential growth values of student i are independent of whether or

not they worked part-time. Notationally, we assume that $G_{i,t+1}(Z_{i,t+1} = 0), G_{i,t+1}(Z_{i,t+1} = 1) \perp Z_{i,t+1}|x_{i,t+1}, Z_{i,t}$.

Assumption 3: The Overlap Assumption. We assume that for every observed covariate and treatment history, there is a non-zero probability of working, or not working part-time during any period of interest: $0 < P(Z_{i,t+1} = 1|x_{i,t+1}, Z_{i,t}) < 1$.

If these conditions hold (Angrist et al., 1996; Kurz, 2022; Myint, 2024; Hernán and Robins, 2020), then we may write that

$$E[G_{i,t+1}(Z_{i,t+1})|x_{i,t+1}] = E[G_{i,t+1}|Z_{i,t+1}, x_{i,t+1}].$$

Our model for student achievement across all waves of data then becomes:

$$y_{i,t} = \mu(x_{i,1}) + \sum_{w=1}^{T-1} G_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}, \hat{\pi}_{i,w+1}) I(t > w) + \epsilon_{i,t}$$

where

$$G_{w+1}() = \delta_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}, \hat{\pi}_{i,w+1}) + \tau_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}) Z_{i,w+1}$$

The different parts of the model work together in a cumulative fashion to predict different parts of a student's mathematics achievement. Predictions for achievement at Wave 1 are given by $\mu()$, while achievement at any subsequent Wave t is given by adding this to a cumulative sum of achievement growths, $G_{w+1}()$. Within each time period, $\delta_{w+1}()$ and $\tau_{w+1}()$ represent the growth that would have been realised without part-time work, and the expected impact of part-time work on this growth respectively.

Using standard terminology from the literature, we can say that $\tau_{w+1}(x_{w+1}, y_1 \dots y_w)$ represents the conditional average treatment effect (CATE) of the intervention between waves w and wave $w+1$ for observations with covariates x_{w+1} and outcome history $y_1 \dots y_w$ at wave $w+1$. While, $\delta_{w+1}(x_{w+1}, y_1 \dots y_w, \hat{\pi}_{w+1})$ represents what we will call a conditional average growth effect (CAGE) of the growth between waves w and wave $w+1$ for observations with covariates x_{w+1} and outcome history

$y_1 \dots y_w$ at wave $w+1$. Calculating the average of the CATEs over a population yields what Li et al. (2023) refer to as a mixed average treatment effect (MATE). Analogously, we will say that the average of the CAGEs is a mixed average growth effect (MAGE):

$$MATE_{w+1} = \frac{1}{n_{w+1}} \sum_{i=1}^{N} \tau_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w})$$

$$MAGE_{w+1} = \frac{1}{n_{w+1}} \sum_{i=1}^{N} \delta_{w+1}(x_{i,w+1}, y_{i,1} \dots y_{i,w}, \hat{\pi}_{i,w+1})$$

The additional covariate $\hat{\pi}_{i,w+1}$ included in the $\delta_{w+1}()$ part of the model is a propensity score, which estimates the probability of observation i receiving treatment during this period conditional on their covariates. This inclusion follows the advice of Hahn et al. (2020), who demonstrated that incorporating this "clever covariate" can help mitigate the issue of regularisation-induced confounding. Finally, $\epsilon_{i,t}$ represents the error term for student i at time t, which we assume to be normally distributed with mean 0 and variance σ^2, $\epsilon_{i,t} \sim N(0, \sigma^2)$.

In our model, the contributions made by $\mu()$ and each of $\delta_{w+1}()$ and $\tau_{w+1}()$ come from ensembles of n_μ, n_δ, and n_τ regression trees based on the BART model of Chipman et al. (2010). For ease of exposition as we discuss the Bayesian backfitting MCMC algorithm by which the regression trees fit to the data, let us consider the simplest scenario where $n_\mu = n_\delta = n_\tau = 1$ and $T = 2$, leaving the general case for the supplementary material. The MCMC sampler begins with each of $\mu()$, $\delta_2()$, and $\tau_2()$ initialised as stumps (decision trees where the root is also the sole terminal node, and the terminal node parameter of each tree is set to zero). Next, each iteration starts by selecting at random one of four possible operations (grow, prune, change, or swap) to apply to the $\mu()$ tree in order to propose a new tree structure. This proposal is then accepted or rejected with a Metropolis-Hastings step before the terminal node (or now possibly nodes) of the $\mu()$ tree are updated via a Gibbs sampling step which attempts to explain any leftover variation in the partial residual $y_{i,t}$ less the contribution from $\delta_2()$ and $\tau_2()$. Analogous operations are then applied to the $\delta_2()$ and $\tau_2()$ trees before the residual variance parameter is also updated via Gibbs sampling. This cycle repeats for a specified number of

iterations, providing a desired number of posterior draws for the tree structure and terminal node parameters of $\mu()$, $\delta_2()$, and $\tau_2()$, as well as the residual variance parameter σ^2.

Overfitting is prevented through the use of the tree prior from Chipman et al. (2010) which specifies that the probability of any node at depth d being non-terminal is given by $\alpha(1+d)^{-\beta}$. Therefore, for a tree T with terminal nodes $h_1...h_K$, and non-terminal nodes $b_1...b_L$, we have that:

$$P(T) = \prod_{k=1}^{K} \alpha(1+d(h_k))^{-\beta} \prod_{l=1}^{L} [1-\alpha(1+d(b_l))^{-\beta}]$$

The strength of this prior can be adjusted through setting different values for α and β. For the $\mu()$ and $\delta()$ trees we adopt the default prior from Chipman et al. (2010), of $\alpha = 0.95$, $\beta = 2$, while for the $\tau()$ trees we impose stronger regularisation as we expect there to be less heterogeneity in the effects of part-time work than in y itself, choosing $\alpha = 0.25$, $\beta = 3$ as suggested by Hahn et al. (2020).

To ensure each tree contributes approximately equally to the overall prediction, the terminal node parameters of each tree are given a normal prior. In each type of tree, we have

$$\mu \sim N(0, \sigma_\mu^2),\ \delta \sim N(0, \sigma_\delta^2),\ \tau \sim N(0, \sigma_\tau^2)$$

After scaling y to follow a standard normal distribution prior to fitting the model, a sensible choice for σ_μ^2 is $1/n_\mu$, ensuring the terminal mode parameters in the $\mu()$ trees have adequate room to cover the range of the data. Similarly, we use a prior of $\sigma_\delta^2 = 1/n_\delta$, but given we expect the magnitude of the treatment effects to be relatively small in comparison to y we set $\sigma_\tau^2 = 0.5^2/n_\tau$. Finally, the conjugate prior for σ^2 is an inverse gamma distribution:

$$\sigma^2 \sim \text{Inverse-Gamma}(\nu/2, \nu\lambda/2),$$

for which we have found a reliable default choice is to set $\nu = 3$, and $\lambda = 0.1$.

6.3.2 Special Features

Two challenges related to missing data required us to build some extra functionality into our model. The first challenge was related to missing data in the covariates. Missing data in the covariates can arise for several reasons in the dataset. For example, some questions may have been purposely skipped by students, their parents, or teachers, and other times the answer to a particular question may not have been known. On average, 1.9% of the data was missing, and the most data missing for any particular variable was 19%. Common approaches for dealing with missing data in the covariates include single or multiple imputation (Lin and Tsai, 2020), but an extra possibility specific to tree based models is the approach developed by Kapelner and Bleich (2015), which involves treating missing data as an important feature of the data.

This approach differs from traditional imputation-based methods, which assume missingness is Missing at Random (MAR) and explicitly estimate missing values before modeling Y. Instead, the Kapelner and Bleich (2015) approach models $E[Y|X, M]$, where M is an indicator for missingness, allowing BART to learn patterns of missingness non-parametrically. A key consideration is whether this method imposes assumptions stricter than MAR. In standard multiple imputation frameworks, MAR is assumed to hold so that missing values can be estimated using observed covariates. The Kapelner and Bleich (2015) method does not impose stronger assumptions, rather, it allows missing values to be handled within the model structure without requiring explicit imputation. In the context of causal inference, an additional assumption regarding overlap is likely to be required, however. If missingness in a covariate X serves as a confounder affecting both the treatment Z and the outcome Y, sufficient overlap must exist in the missingness pattern to ensure comparability between treated and control groups.

The second challenge was related to missing data in the treatment $Z_{i,2}$ itself, as not all students answered the question on how many hours they worked part-time during school weeks. This type of missingness affected 3.4% of the observations in the data. Here we make the strong assumption of Missing at Random (MAR), meaning the presence of missingness in Z is unrelated to the actual Z values themselves, or any other unobserved covariates, and introduce an additional

Gibbs sampling step at the end of each iteration of the MCMC sampler, where the missing $Z_{i,2}$ values are themselves treated as parameters to be updated, with prior probability p_i, conditional on the rest of the data:

$$P(Z_{i,2} = 1|...) = \frac{1}{1 + \left(\frac{1-p_i}{p_i}\right)\left(e^{\frac{1}{2\sigma^2}[(y_{i,2}-\mu_i-\delta_i-\tau_i)^2-(y_{i,2}-\mu_i-\delta_i)^2]}\right)}.$$

These two added features allowed us to keep a full representative sample of students while accounting for the added uncertainty introduced into our results by the presence of missing data. While novel, we note that there is precedence for imputing missing treatment indicators in causal analyses, and highlight Mitra (2023) as an example.

One final challenge that is common when working with assessment data of student achievement is the use of plausible values (Wu, 2005; Khorramdel et al., 2020). In order to prevent the computer delivered assessment from taking unduly long, it was only possible for HSLS to present each student with a limited number of questions. This introduces some room for error in the achievement estimates of the students, and as a result, HSLS provides researchers with five plausible values of student achievement from the posterior of each student's achievement estimate. In line with best practice, we therefore ran five chains of our model, one applied to each plausible value of student achievement, and pooled them together after burn-in to appropriately handle this uncertainty.

6.3.3 Alternative Methodologies

Our work shares connections with several areas of Bayesian non-parametric modelling, and longitudinal methods for causal inference. First there is the clear connection to BART (Chipman et al., 2010), as this method provides a foundation for the different parts of our model. BART based methods have become popular in the area of causal inference (Hill, 2011) and have demonstrated impressive performance and reliable uncertainty quantification. A second strong connection is with the Bayesian Causal Forest model developed by Hahn et al. (2020), which also uses BART as a foundation for estimating causal effects. Both methods could potentially be applied to our research questions but fall short of offering the same

abilities in this context as our own model in several important ways which are worth discussing.

The most natural way for BCF to be applied to our problem setting would be to manually calculate the growth values $G_{i,t+1}$ for each student i, and each time period t to $t+1$. Applying the BCF model to a specific time period would then yield the following, allowing us to recover what our model captures with the $\delta_{t+1}()$ and $\tau_{t+1}()$ part of the model:

$$G_{i,t+1} = \delta_{t+1}(x_{i,t+1}, y_{i,t}, \hat{\pi}_{i,t+1}) + \tau_{t+1}(x_{i,t+1}, y_{i,t})Z_{i,t+1} + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2)$$

A key limitation of this approach is that it does not model the full data generating process, only the growth between Waves t and $t+1$. This means that students who participate in Wave t but not in Wave $t+1$ (and consequently have no calculable $G_{i,t+1}$) are excluded from the model and are unable to inform the predictions made by the model. Secondly, manually calculating the $G_{i,t+1}$ growth values (to be used as the response variable in this approach), is likely to lead to a smaller signal to noise ratio in the response as the error terms from $y_{i,t}$ and $y_{i,t+1}$ combine, making it more difficult for the model to detect the relationships it is trying to model. A BART only approach could also be applied to the data in a similar way using $G_{i,t+1}$ as the response, but this approach would share the same limitations. Of course, if treatment effects were the only quantity of interest then it would also be possible to apply BART or BCF directly to $y_{i,t+1}$, but this would preclude any inference on the growth values $G_{i,t+1}$, so would fail to address this aspect of our study.

Researchers more familiar with difference-in-differences (Roth et al., 2023) based approaches might like to think of our model as a Bayesian non-parametric DiD model where our $\delta_{t+1}()$ trees model the difference for the control group (non part-time workers), and our $\tau_{t+1}()$ trees model the difference in this difference experienced by the treatment group (the part-time workers). Crucially, our approach handles this situation much more flexibly than traditional DiD based methods, as the flexibility of the $\delta_{t+1}()$ trees means we can relax the assumption of parallel trends conditional on the covariates of the students, and the $\tau_{t+1}()$ trees also allow us to capture heterogeneity in the effects of part-time work which is often not pos-

sible with DiD based methods. See the supplementary material for an illustration of how our proposed model fits into this framework.

Finally, our method also shares similarities with causal methods applicable to longitudinal data such as G-estimation (Tompsett et al., 2022), or longitudinal extensions of targeted minimum loss based estimation (LTMLE, Lendle et al., 2017). Both methods have gained popularity owing to their ability to handle complex situations such as time varying confounding, situations where the primary interest is in the causal effect of a series of sustained or irregular treatments, and where the interest is in the lagged effects of a treatment. Our focus however, will be on heterogeneity in the direct effect of a single period of part-time work on the immediately following mathematics assessment, which is not achievable with the available implementations of these methods. Additionally, our model will also provide insights into the growth trajectories of student achievement, a feature that is not modelled by these other approaches.

6.4 Simulation Studies

In this section, we assess our proposed model's performance in a simulation study designed to match the features of the motivating HSLS data. We also compare our proposed model with alternative approaches in order to highlight the added performance offered by our method. Our simulation study consists of two data generating processes. DGP1 focuses on heterogeneity in treatment effects and growth curves, making it well-suited to flexible approaches based on BART and BCF. It features two waves of data to accommodate the alternative methods which can not handle multiple time periods. DGP2 is inspired by a synthetic dataset from the R package `gesttools`. This process includes more than two time points and features time-varying covariates. It focuses on estimating the mixed average treatment effect, enabling a fair comparison of our method with the `gesttools` and `ltmle` packages, which do not support the estimation of heterogeneous treatment effects, but are correctly specified for the features of this DGP.

6.4.1 Data Generating Process 1

Our first data generating process is based on a modified version of the first Friedman dataset (Friedman, 1991), a common benchmarking dataset featuring non linear effects and interaction terms. We will use this dataset to assess how well each of the flexible causal machine learning methods can capture heterogeneity in the growth curves of student achievement, and the treatment effects themselves. We simulate ten covariates measured at Wave 1: $x_1 \dots x_{10}$, and a second observation of each of these ten variables again at the final Wave 2: $x_{11} \dots x_{20}$, where the second observation of each variable is equal to the first plus a small amount of random noise, e.g., $x_{16} = x_6 + r$, with r a uniform random variable between 0 and 0.4. The structure of the simulated achievement level of each student is of the form described earlier:

$$y_{i,t} = \mu(x_{i,1}) + \underbrace{\delta_2(x_{i,2}, y_{i,1}, \hat{\pi}_{i,2})I(t>1) + \tau_2(x_{i,2}, y_{i,1})Z_{i,t+1}I(t>1)}_{G(x_{i,2}, y_{i,1}, \hat{\pi}_{i,2})} + \epsilon_{i,t}$$

where $\mu(x_{i,1}) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5$, $\delta_2() = \frac{1}{3}\mu(x_{i,1}) + 3x_{11}^2 + 2x_{15}^2$, $\tau_2() = -x_4 - x_{14}^2 - x_{15}^3$, $\sigma^2 = 9$, and , $\epsilon_{i,t} \sim N(0, \sigma^2)$. The true propensity scores are given by $p_i = P(Z_{i,2} = 1|x_{i,2}) = \text{Plogis}(\mu_i^* + \delta_i^*)$, where $\mu_i^* + \delta_i^*$ is a normally scaled version of each of the original $\mu_i + \delta_i$ values.

The compared methods are our longitudinal BCF model, BART using the approach outlined in Hill (2011), a standard BCF model from Hahn et al. (2020), and the causal Generalised Random Forest model (GRF) from Wager and Athey (2018). The recently proposed BCF extension by Wang et al. (2024) would also make an excellent method for comparison when a documented R package becomes available.

As outlined earlier, given that the longitudinal BCF model is the only one capable of directly modelling the growth curves, we will apply the other competing methods to the transformed outcome $y_{i,2} - y_{i,1}$, the difference in outcomes between Waves 1 and 2, to enable the prediction of growth using BART and BCF. For the longitudinal BCF model, we use 100 trees in the $\mu()$ part of the model responsible for predicting $y_{i,1}$ at Wave 1, 70 trees in the $\delta()$ part of the model responsible for

predicting the growth under control, and 30 trees in the $\tau()$ part of the model responsible for predicting the heterogeneous treatment effects. For the standard BCF approach, we use 170 trees in the prognostic part of the model which will provide estimates for the growth under control, and 30 trees for estimating the treatment effects. The BART and GRF approaches both use 200 trees in total. Each simulation consists of 500 training observations, and 1000 test observations. The Bayesian methods are run for 500 burn-in and 500 post burn-in iterations. Satisfactory convergence was assessed via visual inspection of the posterior samples for a small subset of the 1000 replications of the data generating process.

Table 6.1 summarises how the compared approaches perform when tasked with predicting δ_i and τ_i across 1000 replications of the simulation. For δ_i, this performance is evaluated using the average root mean squared error (RMSE) over the 1000 simulations: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\delta_i - \hat{\delta}_i)^2}$. The equivalent metric used for τ_i is the precision in estimating heterogeneous effects (PEHE): $PEHE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tau_i - \hat{\tau}_i)^2}$, also averaged over the 1000 simulations. Mean coverage rates of the 95% credible intervals, bias, and credible interval widths are also provided for both δ_i and τ_i. A visual representation of these results can be found in Figure 6.1.

The clearest differences in Figure 6.1 relate to model performance predicting the δ_i values, with our proposed LBCF model achieving much lower RMSE values. Comparison with the GRF model was not possible here, as the GRF model output only provides treatment effect estimates. In the right panel of Figure 6.1, the differences are more subtle, but the proposed model performs marginally better than the BART and BCF methods, which in turn both outperform the GRF based approach.

Finally, the LBCF estimates are the least biased of all the compared methods, and are accompanied by close to ideal coverage rates. The credible interval widths from the LBCF estimator are similar to the competing methods when estimating the treatment effects, but are considerably narrower than the competing methods when estimating the growth values, offering a high degree of precision.

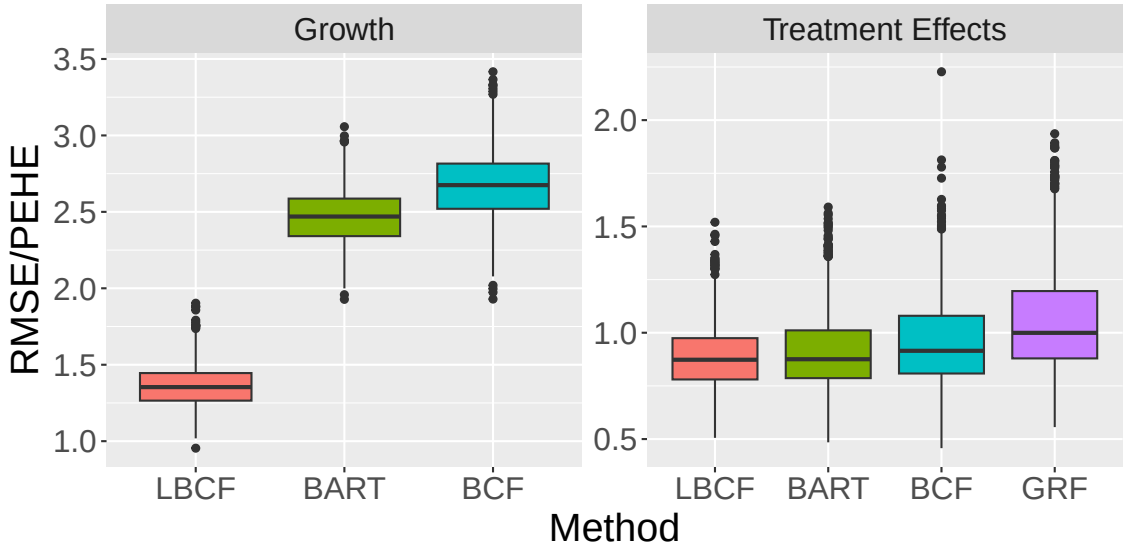

Figure 6.1: Visualisation of RMSE and PEHE metrics evaluated over 1000 replications of DGP1 for the BART, BCF, GRF, and LBCF models. In the left panel, which displays the RMSE of the δ_i predictions, the LBCF approach is clearly the strongest performer, with considerably lower RMSE values. In the right panel, which visualises the PEHE metrics, LBCF is again the strongest performer, but by a narrower margin.

6.4.2 Data Generating Process 2

Our second data generating process comes from the R package `gesttools` (Tompsett et al., 2022), which implements G-estimation for longitudinal data. Our focus here is on estimating the mixed average treatment effect. As described in Tompsett et al. (2022), the dataset includes:

- A baseline covariate $U \sim N(0, 1)$
- Covariates $L_t \sim N(1 + L_{t-1} + 0.5A_{t-1} + U), t = 1, 2, 3, A_0 = 0$
- Exposure $A_t \sim \text{Bin}(1, \text{expit}(1 + 0.1L_t + 0.1A_{t-1})), t = 1, 2, 3$
- Time varying outcome $Y_t \sim N(1 + A_t + \gamma_t A_{t-1} + \sum_{i=1}^{t} L_t + U, 1), t = 2, 3, 4$
- Constants $(\gamma_1, \gamma_2, \gamma_3) = (0, 0.5, 0.5)$

	δ_i predictions			τ_i predictions			
	LBCF	BART	BCF	LBCF	BART	BCF	GRF
RMSE/PEHE	**1.362**	2.470	2.671	**0.886**	0.907	0.958	1.057
Mean Absolute Bias	**0.265**	0.311	0.303	**0.324**	0.424	0.424	0.604
95% Coverage	**0.980**	0.996	0.891	**0.935**	0.921	0.911	0.769
95% CI Width	6.559	14.306	8.624	3.409	3.343	3.515	2.654

Table 6.1: Summary of important metrics measured for δ_i and τ_i predictions, averaged over 1000 simulations of DGP1. The proposed LBCF model performs competitively, achieving a lower mean RMSE and PEHE than the alternative models. Bias, coverage, and credible interval widths are also close to ideal. Best results are highlighted in **bold** where a clear winner exists.

In this simulation study, the baseline covariate U remains fixed, while the time varying covariates L_t change at each wave in response to the values of the preceding covariates, and whether or not treatment was received. The likelihood of receiving treatment also depends on previous covariates and treatments. Note that while the time varying outcome depends on the treatment status at the current and previous time points, we will only estimate the direct effect of treatment at time t on y_t.

The methods we will compare are G-estimation as implemented by `gesttools`, longitudinal targeted minimum loss based estimation from the `ltmle` package, and our proposed method. The `gesttools` and `ltmle` approaches will use the default settings of the R packages, which make them the correctly specified models, while our approach will use the same setup from the previous simulation study. As before, we will run 1000 replications of the simulation study, but will evaluate performance on the training sample of 500 observations (the `ltmle` and `gesttools` packages can not make predictions on unseen data).

Figure 6.2 visualises the MATE estimates from the proposed approach, the `gesttools` package, and the `ltmle` package. For the `gesttools` and `ltmle` results, only one boxplot is shown. In the case of the `gesttools` results, this is because the package assumes the treatment has the same effect at all time points. In this simulation, this assumption is valid, but in general, the ability of our model to provide separate estimates at each time point is likely to be valuable. With the

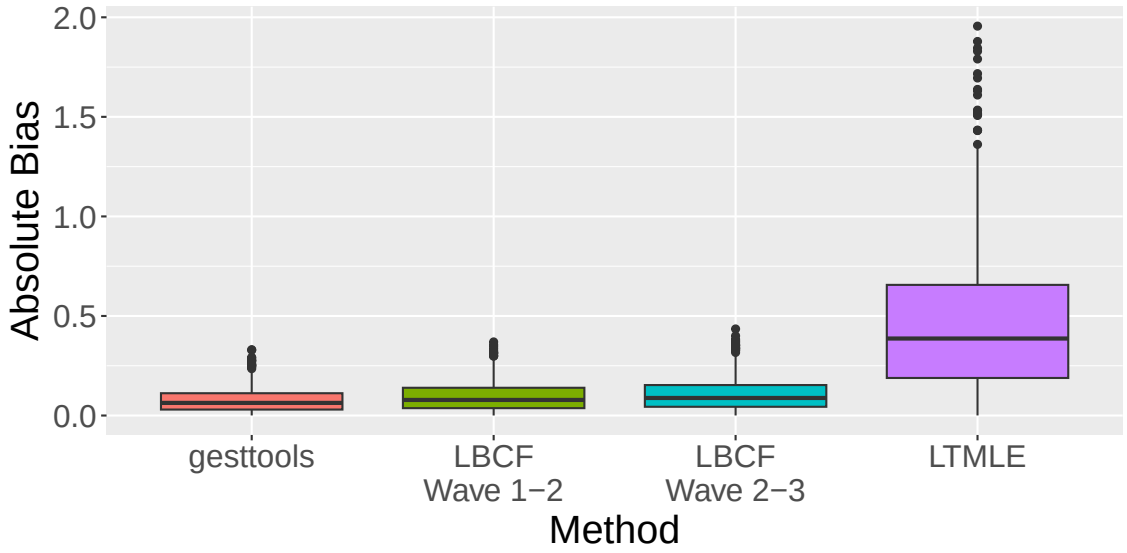

Figure 6.2: Visualisation of bias in MATE estimates over 1000 replications of DGP2 for the `gesttools`, LBCF, and `ltmle` models. The `gesttools` package, which assumes a constant treatment effect at all time points shows minimal bias. This strong performance is closely followed by the proposed LBCF model, which provides estimates for the treatment applied between Waves 1 and 2, and 2 and 3. The `ltmle` estimates appear to be much more biased.

`ltmle` package, it is necessary to define a contrast in order to estimate the effect of some sequence of treatments on the final observed outcome variable (in this case Y_3). For the simulation above, we tasked the `ltmle` package with estimating the effect of the treatment sequence $(A_1 = 0, A_2 = 0, A_3 = 1)$ relative to $(A_1 = 0, A_2 = 0, A_3 = 0)$. This will recover the direct effect of A_3 on Y_3, which is equal to the direct effect of A_t on y_t, consistent across time. In contrast, our proposed LBCF model is able to provide MATE estimates for both the effect of A_2 on Y_2, and the effect of A_3 on Y_3, offering a more detailed and flexible analysis.

Figure 6.2 visualises the absolute bias in estimating the MATE for each of the approaches over 1000 replications of DGP2. The `gesttools` package is the best performer here, closely followed by the LBCF estimates which are consistently accurate across both time periods. We note, however, that the `gesttools` package assumes the treatment effect is the same at all time points, and this is unlikely to always be valid. The bias of the `ltmle` package is consistently much higher,

Model/Metric	`gesttools`	LBCF Wave 1-2	LBCF Wave 2-3	`ltmle`
Mean Absolute Bias	**0.077**	0.097	0.107	0.476
95% Coverage	**0.950**	0.913	0.936	0.838
Mean 95% CI Width	**0.384**	0.438	0.502	1.688

Table 6.2: Absolute bias, coverage rates, and credible/confidence interval widths averaged over 1000 replications of DGP2. Coverage rates are very good for `gesttools` and LBCF, but less than ideal for `ltmle`. The `gesttools` package provides the most precise estimates, with slightly narrower confidence interval widths than LBCF. Best results are highlighted in **bold** where a clear winner exists.

indicating the model often struggled to identify the true MATE from the data.

A similar pattern is observed in Table 6.2, which provides additional information on the coverage rates, and mean credible/confidence interval widths. Here, the coverage achieved by the `gesttools` package and the two estimates provided by the proposed LBCF model are very close to ideal. The `ltmle` package appears to underestimate the uncertainty in its estimates, however, and only achieves 83.8% coverage. The LBCF model's credible interval widths at both time points are slightly wider than those of the `gesttools` package but remain significantly narrower than the `ltmle` package's confidence interval widths.

In summary, the results from both data-generating processes in our simulation study underscore the proposed model's ability to provide flexible and accurate predictions, even when confronted with highly non-linear growth patterns or heterogeneity in treatment effects. The model achieved near-ideal coverage rates, exhibited minimal bias, and produced narrower credible intervals compared to other non-parametric causal models. In the second data-generating process, where the proposed model was benchmarked against a correctly specified G-estimation model, the LBCF model matched its strong performance, without making the same assumption that the treatment effect was consistent over time. Encouraged by the robust performance of our proposed model, we proceed to the next section, where we apply the longitudinal BCF method to the motivating HSLS dataset to assess the impact of part-time work on student achievement.

6.5 Application to High School Longitudinal Study

Recall that HSLS includes two waves of data, with student achievement and other background characteristics measured at both time points. We are interested in understanding the amount by which the mathematics achievement of the students increases between these waves, how this growth depends on the characteristics of the students, the effect of part-time work on this growth, and how this effect is potentially moderated by other observed variables.

We apply our model to this dataset using the same model structure from the simulation study, with the same number of trees, but run a larger number of burn-in (3000) and post burn-in iterations (2000), to ensure satisfactory convergence. As described in the methodology, missing data is handled internally by the model, so there is no requirement for multiple imputation. The plausible values of student achievement are appropriately accounted for by pooling 5 separate chains, each of which were applied to one of the 5 sets of plausible values. Sampling weights are also accounted for by appropriately weighting the mixed average treatment effect results displayed below.

Our control variables in the $\delta()$ part of the model include potential confounders measured between Waves one and two to mitigate bias in the estimated treatment effects. However, since these variables were recorded at the same time as the treatment, some may lie on the causal pathway between the treatment and the outcome, potentially introducing over-control bias and attenuating the estimated effects. A more conservative approach would be to restrict controls to time-invariant covariates or those measured unequivocally before treatment to minimise this risk. Conversely, limiting controls to only time-invariant covariates could lead to omitted variable bias if key confounders that influence both the treatment and the outcome are excluded. Given this trade-off, in this case, we chose to control for potential confounders measured between Waves one and two in attempt to capture all potential sources of confounding.

Figure 6.3 shows the posterior distribution of the mixed average growth effect, and a histogram of the individual δ_i estimates for each student present in Wave 2 of the dataset. The average growth is close to 0.63, and the majority of the growth estimates are positive, indicating that most students are expected to increase their

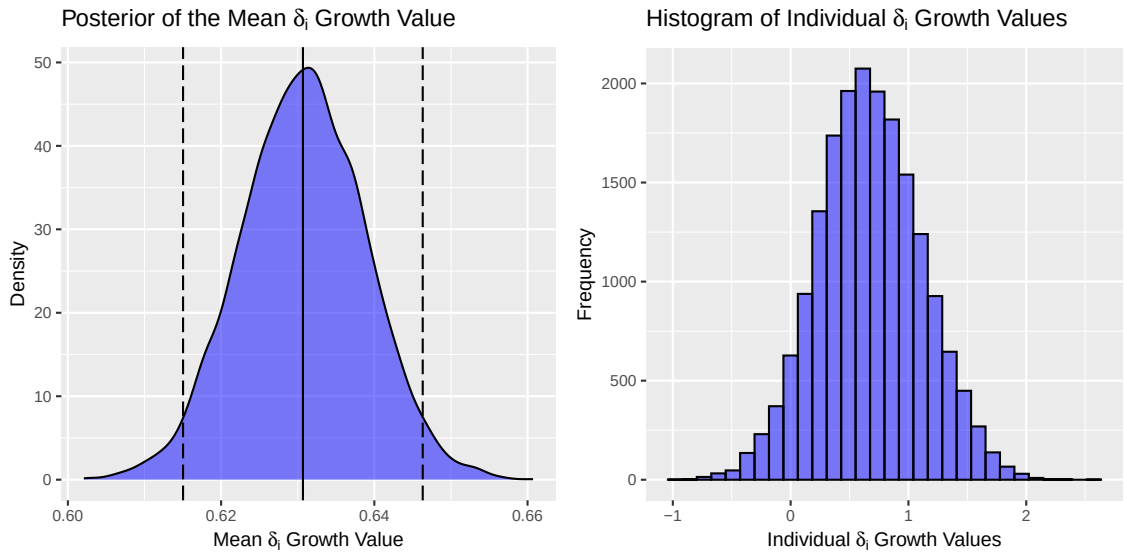

Figure 6.3: The left plot shows the posterior distribution of the mixed average growth effect, while the one on the right displays a histogram of the individual δ_i estimates. The solid line in the left plot shows the posterior mean, while the dashed lines indicate a 95% credible interval. Substantial variability is present in the δ_i values, indicating that some students are predicted to increase their achievement by much more than others who may even experience a decrease in achievement.

mathematics achievement between Waves 1 and 2. Within the sample there is large variation, however, with some students predicted to increase their mathematics achievement by up to 2 units on the achievement scale, while for a small number of students, mathematics achievement is actually predicted to decrease by a small amount. For context, achievement at Wave 1 was normally distributed with a mean of approximately 0, and a standard deviation of approximately 1. Therefore, an increase in achievement by two units, or two standard deviations, is quite significant.

To identify key moderating variables contributing to the variation in δ_i values, variable importance measures were calculated for the $\delta()$ trees by counting how often different variables were selected for the splitting rules used in this part of the model. This investigation identified the achievement of the students measured at Wave 1 as being highly influential. Prompted by this finding, we created Figure 6.4 which shows a scatter plot of the δ_i predictions versus the achievement of the

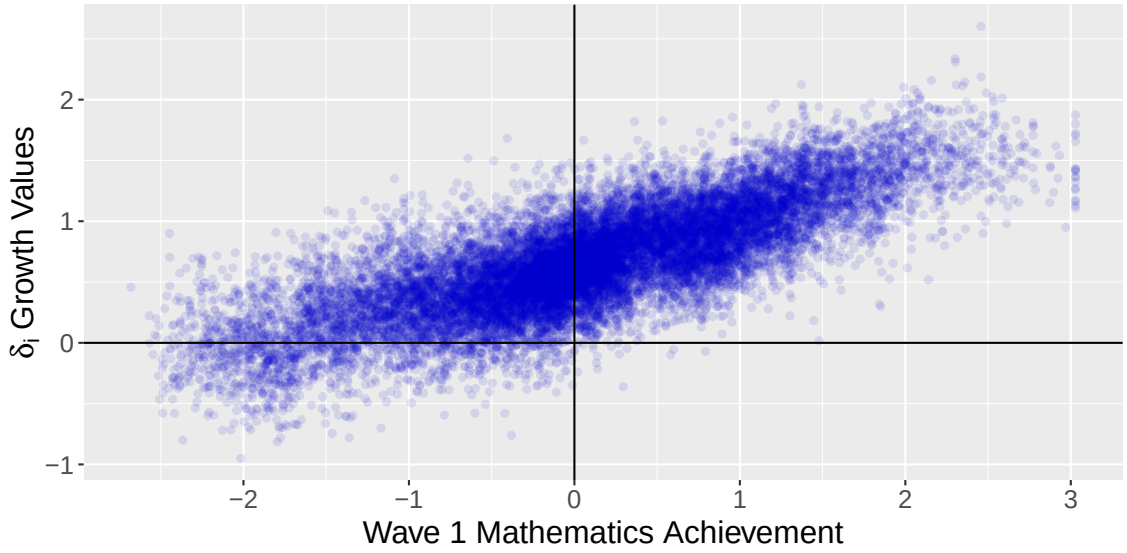

Figure 6.4: Scatterplot of the relationship between Wave 1 achievement and predicted δ_i values. Students with initially high levels of academic achievement are predicted to increase their achievement by higher amounts than their peers.

students measured at Wave 1. The very strong positive relationship between Wave 1 achievement and predicted growth indicates that students who initially perform well in mathematics are predicted to increase their achievement by substantially more than those with lower achievement levels. At the extremely high levels of Wave 1 achievement, students are predicted to increase achievement by 1.5 units on average, while for students at the opposite end of the spectrum, growth in achievement is minimal. This observation points to a widening achievement gap between students at the high and low ends of the achievement spectrum (McCall et al., 2006; Rowley et al., 2020).

The posterior distribution of the mixed average treatment effect for working part-time at an intensity of greater than 20 hours per week between Waves 1 and 2 is displayed in Figure 6.5. The posterior mean of the MATE is approximately -0.08, with a 95% credible interval ranging from -0.050 to -0.110. This indicates that on average, part-time work is expected to reduce the growth in student achievement between Waves 1 and 2 by between 0.050 and 0.110 units. To contextualise this effect size, note that the standard deviation of the δ_i growth values in achievement is approximately 0.44. Thus, the observed effect size corresponds to a decrease in

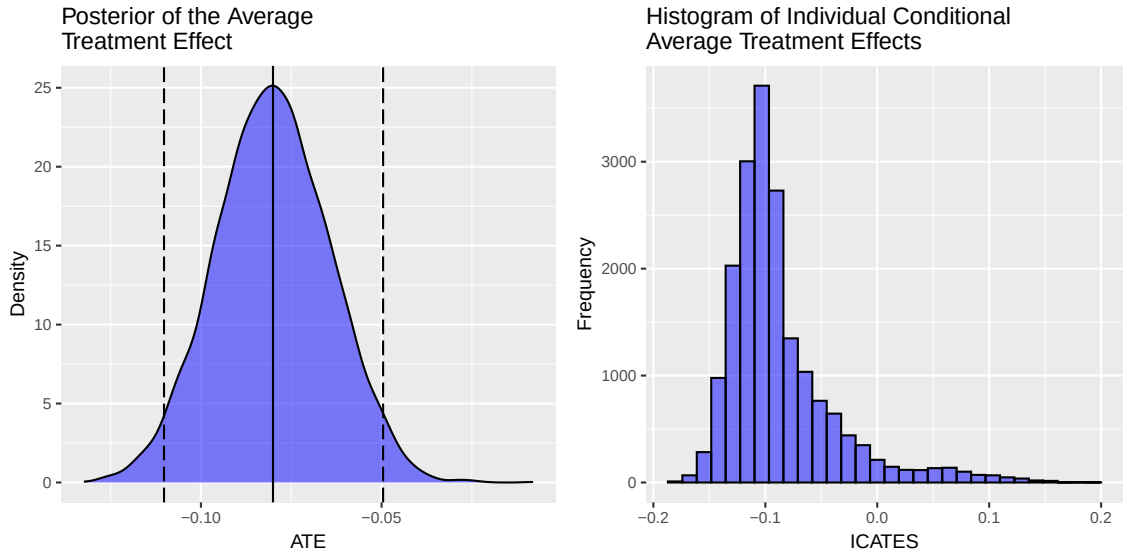

Figure 6.5: The posterior distribution for the Mixed Average Treatment Effect (MATE) is shown on the left, and a histogram of the individual conditional average treatment effects is provided on the right. The solid line shows the posterior mean, while the dashed lines indicate a 95% credible interval. An interesting subgroup of students on the right tail of the histogram are predicted to benefit from part-time work.

achievement growth by nearly 0.2 standard deviations, which can be considered a medium to large effect size (Kraft, 2020).

A histogram of the individual conditional average treatment effects (ICATEs) for each of the students in the sample can be found in Figure 6.5. The majority of the ICATEs are centered quite close to the MATE of -0.08, but there are also signs of heterogeneity. Notably, there is an interesting tail of the histogram stretching across into a positive area where the effect of part-time work is actually predicted to have a positive effect on achievement growth. To explore this finding further, we calculated variable importance metrics for the $\tau()$ trees in our model to identify any variables that might strongly moderate the treatment effect. Variables such as socioeconomic status (SES) and prior academic achievement were considered in this analysis. Although prior achievement at Wave 1 demonstrated a strong relationship with overall growth in mathematics achievement, it did not show a similarly strong association with variation in the effects of part-time work. In-

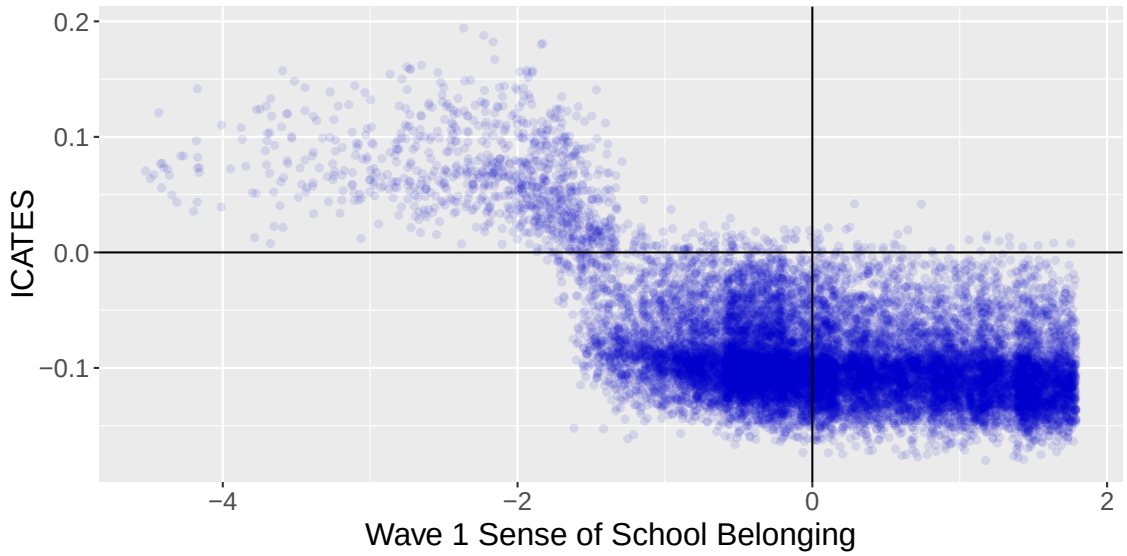

Figure 6.6: Scatterplot of the relationship between Wave 1 school belonging and the effect of working part-time. The effect of part-time work is negative for most students, but for a subgroup of students with low sense of school belonging the predicted effect is positive.

stead, the most influential effect moderator emerging from this investigation was students' sense of school belonging at Wave 1. Figure 6.6 visualises this variable's relationship with the ICATEs from the model. The results suggest that the students predicted to experience a positive effect from part-time work are those with an initially low sense of school belonging.

This interesting finding, which might initially appear quite strange, aligns well with some 'traditional' views that part-time work can benefit students. Early research has suggested, for example, that part-time employment can provide students with greater time management skills (Robotham, 2012), and other benefits such as a sense of purpose and responsibility. These benefits can be especially pronounced among students with low achievement or a diminished sense of belonging in school (King et al., 1989; Steinberg et al., 1982). This sense of purpose and responsibility acquired through part-time work could serve to re-focus students, leading to spillover effects benefiting their academic performance (Zimmerman and Kitsantas, 2005). Therefore, while part-time work may be associated with negative outcomes for the majority of students, there may be certain subgroups, such

as students experiencing a low sense of belonging in school, who may experience positive effects from employment.

In summary, this section presented two key findings from the analysis of the HSLS data. Firstly, substantial variation was observed in the extent to which students improved their achievement between Waves 1 and 2. Further analysis showed that this variation was driven primarily by the baseline achievement levels of the students, with initially high performing students showing much higher growth than their peers. These students, starting from a solid foundation of high achievement may find it easier to build upon their academic progress, as they are in a better place to acquire and digest new knowledge in class. The second key finding was that on average, part-time work had a modest, but negative effect on the growth of student achievement. This supports the "zero-sum" argument that part-time work detracts from study time, homework completion, and rest, hindering academic progress as a result. A notable exception was that students with initially low school belonging might actually benefit from part-time work, highlighting the ability of our model to capture complex relationships between student performance and employment.

6.6 Discussion

Drawing on longitudinal data from the High School Longitudinal Study of 2009, our study introduced an innovative method for modelling growth in student achievement. Our model also estimates the causal impact of interventions such as part-time work on this growth. By extending Bayesian Additive Regression Trees (Chipman et al., 2010) and Bayesian Causal Forests (Hahn et al., 2020), the primary strength of our model lies in its ability to flexibly capture both individual growth trajectories in student achievement and the potentially heterogeneous treatment effects of part-time work, which may be influenced by various covariates. This approach contrasts with many existing methods that either lack the flexibility to model individual variations or are confined to single time-point observational data, precluding an analysis of achievement growth over time.

Our model was also equipped with two special features that allowed it to handle missing data in the covariates and the treatment status indicator. Simulation

study results from Section 6.4 provide strong support for the impressive predictive performance of the model, which demonstrated clear advantages over three competing methods when tasked with predicting growth values at the individual student level, and heterogeneous treatment effects. Close to ideal coverage rates were also achieved. The proposed model also showed strong performance in a second simulation study, matching the performance of two correctly specified models designed specifically for use with longitudinal datasets.

The results from our model application to the motivating HSLS data produced some interesting findings. First, the model was able to reveal a large disparity in the predicted growth values among students with initially high and low levels of academic achievement. This finding of a widening achievement gap underscores the importance of early interventions in schools and academic institutions. By addressing achievement gaps at the elementary and middle school levels, policy decisions can prevent these disparities from becoming entrenched. This is especially important given previous research which indicates that it becomes much more challenging to effectively remedy these gaps by the ninth or eleventh grade (Morgan et al., 2016).

On average, part-time work was found to have a negative effect on student achievement, with the 95% credible interval for the MATE ranging from -0.050 to -0.110. This is important, as we calculated nearly 50% of students in our sample participated in some level of part-time work during high school, and more than 15% of students participated in intensive part-time work, requiring upwards of 20 hours of work a week. Large amounts of heterogeneity were apparent in the ICATEs, however, and an analysis of the variable importance metrics from the model identified sense of school belonging during Wave 1 as a significant contributor to this variation. The finding that students with a low sense of school belonging may actually be benefiting slightly from part-time work ties in with previous findings that show students can benefit from the routine, sense of purpose and responsibility that part-time work can provide (Robotham, 2012; King et al., 1989; Steinberg et al., 1982). From a policy perspective, however, we do not recommend that students beginning to disengage from the school system should take on intensive part-time work. Instead, we suggest that further research is needed to explore how disengaging students can be encouraged to find a sense of purpose or rou-

tine through other activities such as sports or youth programmes. Alternatively, part-time work with moderate hours may be a more balanced approach.

An important ethical consideration arising from this work concerns the treatment of ethnicity as a control variable. Ethnic background is often connected with systemic factors such as unequal access to academic support, employment opportunities, and safe working conditions. These are factors that may influence both participation in part-time work and academic outcomes. By adjusting for ethnicity, we aim to isolate the effect of part-time work from these broader systemic influences. However, it is still important to acknowledge that this does not erase the reality of these structural disadvantages. The experiences of students from different backgrounds may differ not only in their access to work opportunities but also in the quality of those experiences and their potential academic consequences. Findings from this study should therefore be interpreted within the wider context of educational and work inequalities, acknowledging that the variables analysed here such as part-time work may act as surface-level indicators of deeper social and structural disparities.

A limitation of the model proposed in our study is that owing to the fact each growth period and associated treatment effect is dedicated a separate BART model, the computational cost of running the model may become quite large in settings with many waves of data. Replacing the BART models with more efficient XBART models as in He and Hahn (2023) and Krantsevich et al. (2023) would therefore make a promising area for future work, widening the applicability of the proposed method.

A second limitation of the proposed model is that it is currently designed to estimate only the direct effect of a treatment on the outcome measured in the immediately following time period. The key challenge in extending the model to later outcomes is that treatment effects may propagate not only through direct causal pathways but also indirectly by influencing control variables measured after the intervention. Since our approach adjusts for these control variables to estimate causal effects, failing to account for these indirect pathways could introduce post-treatment bias. Adapting the method in order to accommodate the estimation of longer-term effects is therefore a promising avenue for future work.

Another challenge in applying the proposed model to datasets with multiple

growth periods would be in ensuring sufficient overlap in covariate distributions across treatment groups over time. Evaluating this assumption is more challenging with extra growth periods, as the overlap assumption must be assessed for each period, conditional on prior covariate histories. In cases where the overlap assumption is only met for a subset of the growth periods of interest, this may preclude a full exploration of the effects of treatment between certain periods, limiting analysis to time periods where the assumption does hold.

Given the flexibility and widely adopted nature of the underlying BART framework, a natural extension of the longitudinal causal model adopted in our study might be to survival data (Sparapani et al., 2016). Other natural extensions could include allowing multivariate (McJames et al., 2024, Chapter 4) or multinomial outcomes (Murray, 2021), or the incorporation of random effects (Wundervald et al., 2022; Yeager et al., 2022). Additionally, given the specificity of our results to a representative sample of ninth to eleventh grade high school students from the US, an application of a similar model to other countries or grade levels would be of interest. More generally, we expect that the model's flexibility will allow it to be applied to a wide variety of datasets across diverse fields and application areas.

7 Conclusion

This thesis has focused on the development of new extensions of Bayesian non-parametric causal inference machine learning methods, and their application to large scale education datasets to investigate important issues related to education policy in Ireland and internationally. In this concluding chapter, we summarise the contribution made by each of the research chapters of this thesis, consider the limitations of our research, and identify promising areas for future work.

7.1 Chapter Summaries

Chapter 3 of this thesis focuses on the application of Bayesian Additive Regression Trees to data from the Teaching and Learning International Survey (TALIS 2018) in order to investigate factors affecting teacher job satisfaction - an important research question given the high levels of teacher turnover and teacher shortages currently facing many countries. Ours is not the first study to investigate this problem of international interest, but the adoption of the BART based approach for causal inference outlined in Hill (2011) makes a significant contribution. This methodology allowed us to address the problem with a flexible causal inference approach, demonstrating the potential for advanced statistical methods in education research. Additionally, our focus in Chapter 3 on specific and implementable factors, such as mentoring schemes and continual professional development, helps to increase the relevance and value of our findings. These factors can be more readily incorporated into policy updates compared to less easily modified aspects,

such as the overall sense of safety and orderliness within a school.

Motivated by data from the Trends in International Mathematics and Science Study (TIMSS 2019) which includes data on student achievement in both mathematics and science, Chapter 4 of this thesis introduced a multivariate extension of Bayesian Causal Forests. This multivariate extension allows for the identification of the heterogeneous treatment effects of an intervention on multiple outcome variables simultaneously. The key advantage of this approach is that we allow the structures and decision rules of the trees to benefit from the correlation and shared information across all outcome variables. This allows us to achieve more accurate and more precise treatment effect estimates than similar univariate approaches, while achieving excellent coverage and minimal bias. As demonstrated by the simulation study contained in the chapter, the method was also robust to violations of the model assumptions affecting only one of the outcome variables of interest.

With this approach, we were able to identify the effect of different home related factors on student achievement in both mathematics and science. Our findings indicate that having access to a study desk at home can positively impact mathematics achievement, while often feeling hungry upon arriving at school and frequent absences can have negative consequences in both subjects. These results have important implications for government policies such as free school meal schemes, which can provide healthy meals to students at school, and also highlight the potential to inform parents of the importance of students having access to dedicated study spaces at home.

Chapter 5 is based on the use of the multivariate BCF model from Chapter 4 to investigate the effects of homework on student achievement, again using data from TIMSS 2019. An important contribution made by this chapter is separating the effects of homework frequency and duration in order to identify the optimal distribution of homework throughout the week. Additionally, the use of the multivariate model from Chapter 4 to investigate the effects of homework in both mathematics and science makes a valuable contribution, as subject specific differences are relatively under explored in this area. Indeed, subject specific differences were identified from our investigation, showing that daily homework benefited mathematics achievement the most, while three to four days per week was the most beneficial frequency in science. A finding that was common across both subjects,

however, was that short homework assignments lasting up to 15 minutes each time were equally as beneficial as longer ones.

While not emphasised in Chapter 5, this was only achievable with an important multi-treatment extension of BCF that allows for the consideration of multiple interventions, or variations of a single intervention simultaneously. This simple yet powerful extension, closely related to the "no multiple versions of treatment" aspect of the stable unit treatment value assumption, is likely to be equally as useful in many other settings. It is easy to imagine, for example, a medical study in which two different drugs, or versions of the same drug are being tested simultaneously on the same target population.

The final research chapter of this thesis, Chapter 6, extends BCF to longitudinal data in order to investigate the effect of part-time work on the growth in high school mathematics achievement between grades 9 and 11, by using data from the High School Longitudinal Study of 2009. Some might like to think of the model described in Chapter 6 as a flexible Bayesian non-parametric difference-in-differences model. Crucially, however, by using the flexible BART and BCF as a foundation for the model, we were able to relax the restrictive parallel trends assumption of the standard difference-in-differences approach. Additionally, the intuitive parameterisation which models the outcome variable as a cumulative sum of growths and treatment effects from consecutive time periods, meant we were able to place separate priors on the growth in student achievement and the effects of part-time work on this growth.

With this model we were able to identify a small but negative effect of part-time work on student achievement. Interestingly, owing to the flexible BART and BCF foundation, we were also able to identify heterogeneity in the predicted growth in student achievement, and the estimated effects of part-time work. These results pointed to a widening achievement gap, as well as potential benefits of part-time work for students with an initially low sense of school belonging. These results have policy implications, by highlighting the need for early interventions in school, in order to help prevent large achievement gaps from developing before they are given a chance to widen. They also highlight the need for policy makers to remain vigilant of the negative effects that part-time work can have on student achievement, and potentially tighten regulations surrounding young labour, especially intensive

work requiring greater than 20 hours of work per week. Lastly, the finding that students with an initially low sense of school belonging may be benefiting from part-time work highlights an opportunity to explore other ways in which students disengaging from school may be encouraged to engage in school again.

7.2 Limitations and Future Work

The research carried out in Chapter 3 is primarily motivated by high rates of teacher turnover and attrition, as well as teacher shortages. However a limitation of Chapter 3 is that the measure of job satisfaction used as an outcome variable does not measure those outcomes directly. Other research shows that job satisfaction is an important predictor of teacher intentions to remain teaching, but in order to provide a direct link to teacher shortages and attrition, our study would have had to use a direct measure of these outcomes, which were not available in the TALIS data. As such, external factors could weaken the strength of the relationship between job satisfaction and actual turnover behavior, meaning that job satisfaction may not consistently serve as a reliable proxy for predicting turnover and attrition. Therefore, future work could include the application of a similar method to ours to a dataset with direct measurements of teacher turnover rates, if such data sources become publicly available. Alternatively, future work which examines the mediating role of teacher job satisfaction on turnover or other outcomes of interest could help to strengthen the link between our findings and rates of attrition by demonstrating a causal link between job satisfaction and rates of teacher turnover etc.

A limitation of the multivariate shared tree structure adopted in Chapter 4 is that in situations where the outcome variables of interest may experience the influence of different predictor variables in substantially different ways, such as through distinct interaction terms, or possibly even distinct sets of covariates, the tree structure adopted in Chapter 4 which uses the same tree structure and decision rules for all outcome variables may be inappropriate. In fact, the simulation study included in Chapter 4, which incorporates three different data generating processes with differing levels of suitability for a shared tree structure, provides evidence of this potential drawback. An additional limitation of the shared tree structure

used by all outcome variables, is that model interpretability and explainability tools will not easily be able to untangle the separate variable importance metrics and interaction terms associated with each outcome individually. Instead, variable importance metrics based on the multivariate BCF model can only be interpreted as the overall effect of the predictor variables on all outcome variables.

To address the above issues related to shared tree structures not being ideal in certain circumstances, an interesting avenue for future work would be to allow separate ensembles of decision trees to independently focus on unique outcome variables, while maintaining a shared tree structure ensemble focused on all outcome variables. This would allow the model to benefit from any shared information across all outcome variables, while also providing added flexibility in situations where separate tree structures would be preferred. This may also allow a deeper understanding of how the covariates jointly and separately influence the outcome variables of interest. Since publishing Chapter 4, other researchers have already begun work independently on a similar extension applied to BART (Esser et al., 2024), but a similar extension applied to BCF would also likely be very useful.

An important limitation of Chapter 5 is that by only considering an academic outcome variable - student achievement, it is not possible to directly weigh the pros and cons of assigning homework to students, as the potentially negative effects of homework on non-academic outcomes such as stress levels, and the necessarily reduced time available for important extracurricular activities such as sport and recreation were not considered. Furthermore, simply assigning homework with the identified optimal frequency and duration is unlikely to be beneficial without also ensuring the homework is well suited to the ability of the students, and of a sufficiently high quality. Therefore, a promising area for future work would be to apply the multivariate BCF model to any available non-academic measures of student welfare and well-being, while also examining the moderating role of homework quality and appropriateness for the ability levels of the students, in order to more accurately evaluate the benefits and drawbacks of homework in different levels of quality.

A limitation of Chapter 6 is that as separate BART ensembles are devoted to each of the growth and treatment effect estimates from every time period, the computational burden of the model is likely to become very large when applied

to datasets with many waves of data, unless a small number of trees is used. To tackle this issue, a very useful area for future work would be to incorporate the accelerated Bayesian Additive Regression Tree (XBART) model of He et al. (2019) into the longitudinal BCF framework, thus improving computational efficiency, and possibly allowing the LBCF model to larger datasets without the need to sacrifice on number of trees or posterior samples.

A general limitation that applies to all chapters of this thesis, is that each of the models employed rely on a number of important assumptions. Notably, the ignorability assumption requires that we have controlled for all sources of confounding. While we have endeavoured to do so in our analyses by controlling for all variables available to us that we believe may act as confounders, it would be irresponsible to disregard the possibility that some important confounding variables may have been missing from the data sources employed during the writing of this thesis. As a result, extending the models developed in this thesis to include ideas from Bargagli-Stoffi et al. (2019), who develop an instrumental variable version of BCF would make a significant contribution, by helping to relax the strong ignorability assumption.

A further limitation that applies to all chapters of this thesis is that the results are always specific to data from one country or grade level. A replication of the studies conducted in this thesis, performed with data from other countries and grade levels would therefore be necessary to further generalise the results reported in this thesis.

Given the many BART and BCF extensions developed by other researchers, an area of future work that would be applicable to all chapters of this thesis is incorporating these extensions into the models we have developed to further improve their applicability to a wider selection of datasets. For example, the shrinkage BCF extension by Caron et al. (2022b) could be incorporated into our work in order to improve performance in high dimensional settings, the work by Starling et al. (2021) could be incorporated to allow for the investigation of treatment effects that are expected to exhibit smooth variations across a covariate, or the work by Prado et al. (2021a) could be incorporated to take advantage of local linearities present among the relationships in the data.

Finally, while code is available online at https://github.com/Nathan-McJames

for reproducing the results, and fitting the models developed in this thesis, a dedicated R package that combines these BCF extensions in one user friendly package would be an excellent avenue for future work, enabling other researchers from other disciplines to more easily benefit from the valuable contributions developed in this thesis. An early version of such a package, implementing the multivariate BCF model, is now available at https://nathan-mcjames.github.io/mvbcf/.

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Allen, M. B. (2005). Eight questions on teacher recruitment and retention: What does the research say? *Education Commission of the States (NJ3)*.

Allen, R. and Sims, S. (2017). Improving science teacher retention: Do national STEM learning network professional development courses keep science teachers in the classroom. *Education Research in STEM Subjects: Main Collection*.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Bachman, J. G. and Schulenberg, J. (2014). How part-time work intensity relates to drug use, problem behavior, time use, and satisfaction among high school seniors: Are these consequences or merely correlates? In *Risks and Problem Behaviors in Adolescence*, pages 198–213. Routledge.

Bargagli-Stoffi, F. J., De Witte, K., and Gnecco, G. (2019). Heterogeneous causal effects with imperfect compliance: A novel Bayesian machine learning approach. *arXiv preprint arXiv:1905.12707*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300.

Bennett, S. and Kalish, N. (2007). *The case against homework: How homework is hurting children and what parents can do about it*. Harmony.

Bennour, K. (2021). Multilevel modeling of the effect of bullying on absenteeism and performance in Saudi schools. *Munich Personal RePEc Archive*, pages 1–15.

Berliner, D. C. (2011). The context for interpreting PISA results in the USA: Negativism, chauvinism, misunderstanding, and the potential to distort the educational systems of nations. In *PISA under examination*, pages 75–96. Brill.

Blette, B. S., Granholm, A., Li, F., Shankar-Hari, M., Lange, T., Munch, M. W., Møller, M. H., Perner, A., and Harhay, M. O. (2023). Causal Bayesian machine learning to assess treatment effect heterogeneity by dexamethasone dose for patients with COVID-19 and severe hypoxemia. *Scientific Reports*, 13(1):6570.

Boe, E. E., Bobbitt, S. A., Cook, L. H., Whitener, S. D., and Weber, A. L. (1997). Why didst thou go? Predictors of retention, transfer, and attrition of special and general education teachers from a national perspective. *The Journal of Special Education*, 30(4):390–411.

Bowman, N. A., Preschel, S., and Martinez, D. (2023). Does supplemental instruction improve grades and retention? A propensity score analysis approach. *The Journal of Experimental Education*, 91(2):205–229.

Broer, M., Bai, Y., and Fonseca, F. (2019). *Socioeconomic Inequality and Educational Outcomes: Evidence from Twenty Years of TIMSS*. Springer.

Buell, J. (2008). *Closing the book on homework: Enhancing public education and freeing family time*. Temple University Press.

Burić, I. and Kim, L. E. (2021). Job satisfaction predicts teacher self-efficacy and the association is invariant: Examinations using TALIS 2018 data and longitudinal Croatian data. *Teaching and Teacher Education*, 105.

Cai, L., Choi, K., Hansen, M., and Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, 3(1):297–321.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Capone, V. and Petrillo, G. (2020). Mental health in teachers: Relationships with job satisfaction, efficacy beliefs, burnout and depression. *Current Psychology*, 39(5):1757–1766.

Carnegie, N., Dorie, V., and Hill, J. L. (2019). Examining treatment effect heterogeneity using BART. *Observational Studies*, 5(2):52–70.

Caron, A., Baio, G., and Manolopoulou, I. (2022a). Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1115–1149.

Caron, A., Baio, G., and Manolopoulou, I. (2022b). Shrinkage Bayesian causal forests for heterogeneous treatment effects estimation. *Journal of Computational and Graphical Statistics*, 31(4):1202–1214.

Chen, X. (2022). The effects of individual-and class-level achievement on attitudes towards mathematics: An analysis of Hong Kong students using TIMSS 2019. *Studies in Educational Evaluation*, 72:101113.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Coldwell, M. (2017). Exploring the influence of professional development on teacher careers: A path model approach. *Teaching and Teacher Education*, 61:189–198.

Cooper, H. (1989). *Homework.* Longman.

Cooper, H., Robinson, J. C., and Patall, E. A. (2006). Does homework improve academic achievement? A synthesis of research, 1987–2003. *Review of Educational Research*, 76(1):1–62.

Cooper, H. and Valentine, J. C. (2001). Using research to answer practical questions about homework. *Educational Psychologist*, 36(3):143–153.

Corno, L. (1996). Homework is a complicated thing. *Educational Researcher*, 25(8):27–30.

Corral, D. and Yang, M. (2024). An introduction to the difference-in-differences design in education policy research. *Asia Pacific Education Review*, pages 1–10.

Dahler-Larsen, P. and Foged, S. K. (2018). Job satisfaction in public and private schools: Competition is key. *Social Policy & Administration*, 52(5):1084–1105.

Daly, C., Gandolfi, H., Pillinger, C., Glegg, P., Hardman, M., Stiasny, B., and Taylor, B. (2021). The early career framework–a guide for mentors and early career teachers. www.ucl.ac.uk/ioe-cttr.

Daw, J. (2012). Parental income and the fruits of labor: Variability in homework efficacy in secondary school. *Research in Social Stratification and Mobility*, 30(3):246–264.

Dee, T. S. and Cohodes, S. R. (2008). Out-of-field teachers and student achievement: Evidence from matched-pairs comparisons. *Public Finance Review*, 36(1):7–32.

Delahunty, T. (2024). The convergence of late neoliberalism and post-pandemic scientific optimism in the configuration of scientistic learnification. *Educational Review*, pages 1–23.

Department for Education (2019). Early career framework. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/978358/Early-Career_Framework_April_2021.pdf.

Department of Social Protection (2023). School meals scheme. Available at: https://www.gov.ie/en/service/29a3ff-school-meals-scheme/.

Dettmers, S., Trautwein, U., Lüdtke, O., Kunter, M., and Baumert, J. (2010). Homework works if homework quality is high: Using multilevel modeling to predict the development of achievement in mathematics. *Journal of Educational Psychology*, 102(2):467–482.

Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, 89(2):221–233.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., and Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, 59(10):1087–1091.

Dorie, V. and Hill, J. (2020). Package 'bartCause'. https://cran.r-project.org/web/packages/bartCause/bartCause.pdf.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Dorie, V., Perrett, G., Hill, J. L., and Goodrich, B. (2022). Stan and BART for causal inference: Estimating heterogeneous treatment effects using the power of Stan and the flexibility of machine learning. *Entropy*, 24(12):1782.

d'Agnese, V. (2015). PISA's colonialism: Success, money, and the eclipse of education. *Power and Education*, 7(1):56–72.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Edwards, H. (2018). *Homework Without Barriers* [Master's thesis, California State University]. Digital Commons. https://digitalcommons.csumb.edu/caps_thes_all/366/.

Eivers, E. (2010). PISA: Issues in implementation and interpretation. *The Irish Journal of Education/Iris Eireannach an Oideachais*, pages 94–118.

Engel, L. C., Rutkowski, D., and Thompson, G. (2019). Toward an international measure of global competence? A critical look at the PISA 2018 framework. *Globalisation, Societies and Education*, 17(2):117–131.

Entwisle, D. R., Alexander, K. L., and Olson, L. S. (2000). Early work histories of urban youth. *American Sociological Review*, 65(2):279–297.

Eren, O. and Henderson, D. J. (2008). The impact of homework on student achievement. *The Econometrics Journal*, 11(2):326–348.

Eren, O. and Henderson, D. J. (2011). Are we wasting our children's time by giving them more homework? *Economics of Education Review*, 30(5):950–961.

Esser, J., Maia, M., Parnell, A. C., Bosmans, J., van Dongen, H., Klausch, T., and Murphy, K. (2024). Seemingly unrelated Bayesian additive regression trees for cost-effectiveness analyses in healthcare. *arXiv preprint arXiv:2404.02228*.

Fan, H., Xu, J., Cai, Z., He, J., and Fan, X. (2017). Homework and students' achievement in math and science: A 30-year meta-analysis, 1986–2015. *Educational Research Review*, 20:35–54.

Ferguson, K., Frost, L., and Hall, D. (2012). Predicting teacher anxiety, depression, and job satisfaction. *Journal of Teaching and Learning*, 8(1).

Fernández-Alonso, R., Álvarez-Díaz, M., Suárez-Álvarez, J., and Muñiz, J. (2017). Students' achievement and homework assignment strategies. *Frontiers in Psychology*, 8:1–11.

Fernández-Alonso, R., Woitschach, P., Álvarez-Díaz, M., González-López, A. M., Cuesta, M., and Muñiz, J. (2019). Homework and academic achievement in Latin America: A multilevel approach. *Frontiers in Psychology*, 10:1–10.

Fishbein, B., Foy, P., and Yin, L. (2021). TIMSS 2019 user guide for the international database (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-database/.

Flunger, B., Trautwein, U., Nagengast, B., Lüdtke, O., Niggli, A., and Schnyder, I. (2015). The janus-faced nature of time spent on homework: Using latent profile analyses to predict academic achievement over a school year. *Learning and Instruction*, 39:97–106.

Foyle, H., Lyman, L., Tompkins, L., Perne, S., and Foyle, D. (1990). Homework and cooperative learning: A classroom field experiment. ERIC ED350285.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Galloway, M., Conner, J., and Pope, D. (2013). Nonacademic effects of homework in privileged, high-performing high schools. *The Journal of Experimental Education*, 81(4):490–510.

Gil-Flores, J. (2017). The role of personal characteristics and school characteristics in explaining teacher job satisfaction. *Revista de Psicodidáctica (English ed.)*, 22(1):16–22.

Gill, P. K., Steele, K. M., Donelan, J. M., and Schwartz, M. H. (2023). Causal modelling demonstrates metabolic power is largely affected by gait kinematics and motor control in children with cerebral palsy. *PLOS One*, 18(5):e0285667.

Gold, M. S. and Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, 7(3):319–355.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65.

Gonida, E. N. and Cortina, K. S. (2014). Parental involvement in homework: Relations with parent and student achievement-related motivational beliefs and achievement. *British Journal of Educational Psychology*, 84(3):376–396.

Gray, L. and Taie, S. (2015). Public school teacher attrition and mobility in the first five years: Results from the first through fifth waves of the 2007-08 beginning teacher longitudinal study. First look. *National Center for Education Statistics*.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, 14(1):29–46.

Grodner, A. and Rupp, N. G. (2013). The role of homework in student learning outcomes: Evidence from a field experiment. *The Journal of Economic Education*, 44(2):93–109.

Guarino, C. M., Santibanez, L., and Daley, G. A. (2006). Teacher recruitment and retention: A review of the recent empirical literature. *Review of Educational Research*, 76(2):173–208.

Guo, B., Holscher, H. D., Auvil, L. S., Welge, M. E., Bushell, C. B., Novotny, J. A., Baer, D. J., Burd, N. A., Khan, N. A., and Zhu, R. (2021). Estimating heterogeneous treatment effect on multivariate responses using random forests. *Statistics in Biosciences*, pages 1–17.

Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: A comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, 24(3):275–295.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Han, D. and Hur, H. (2021). Managing turnover of STEM teacher workforce. *Education and Urban Society*.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

He, J., Buchholz, J., and Fischer, J. (2022). Cross-cultural comparability of latent constructs in ilsas. In *International handbook of comparative large-scale studies in education: Perspectives, methods and findings*, pages 845–870. Springer.

He, J. and Hahn, P. R. (2023). Stochastic tree ensembles for regularized nonlinear regression. *Journal of the American Statistical Association*, 118(541):551–570.

He, J., Yalov, S., and Hahn, P. R. (2019). XBART: Accelerated Bayesian additive regression trees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1130–1138. PMLR.

Henderson, M. (2019). *Does taking dual enrollment on a college campus improve student outcomes? A quasi-experimental approach using inverse probability of treatment weighting*. North Carolina State University.

Heppt, B., Olczyk, M., and Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education*, 25(4):903–928.

Hernández-Torrano, D. and Courtney, M. G. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-Scale Assessments in Education*, 9(1):17.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Application*, 7:251–278.

Hill, J. G. and Dalton, B. (2013). Student math achievement and out-of-field teaching. *Educational Researcher*, 42(7):403–405.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hilton, G. L. (2017). Disappearing teachers: An exploration of a variety of views as to the causes of the problems affecting teacher recruitment and retention in England. *Bulgarian Comparative Education Society*.

Hobbs, L. and Törner, G. (2019). Teaching out-of-field as a phenomenon and research problem. In *Examining the Phenomenon of "Teaching Out-of-field"*, pages 3–20. Springer.

Hogan, J. W. and Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research*, 13(1):17–48.

Howieson, C., McKechnie, J., Hobbs, S., and Semple, S. (2012). New perspectives on school students' part-time work. *Sociology*, 46(2):322–338.

Hu, L. and Gu, C. (2021). Estimation of causal effects of multiple treatments in healthcare database studies with rare outcomes. *Health Services and Outcomes Research Methodology*, pages 1–22.

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. (2020). Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical Methods in Medical Research*, 29(11):3218–3234.

Hulme, M. and Wood, J. (2022). The importance of starting well: The influence of early career support on job satisfaction and career intentions in teaching. *Journal of Further and Higher Education*, 46(4):504–521.

Imai, K. and Kim, I. S. (2021). On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis*, 29(3):405–415.

Ingels, S. J., Pratt, D. J., Herget, D. R., Burns, L. J., Dever, J. A., Ottem, R., Rogers, J. E., Jin, Y., and Leinwand, S. (2011). High school longitudinal study of 2009 (HSLS: 09): Base-year data file documentation. NCES 2011-328. *National Center for Education Statistics*.

Ingels, S. J., Pratt, D. J., Herget, D. R., Dever, J. A., Fritch, L. B., Ottem, R., Rogers, J. E., Kitmitto, S., and Leinwand, S. (2013). High school longitudinal study of 2009 (HSLS: 09): Base-year to first follow-up data file documentation. *National Center for Education Statistics*.

Ingersoll, R. and Kralik, J. M. (2004). The impact of mentoring on teacher retention: What the research says. *Education Commission of the States*.

Ingersoll, R. M. and Strong, M. (2011). The impact of induction and mentoring programs for beginning teachers: A critical review of the research. *Review of Educational Research*, 81(2):201–233.

Inglis, A., Parnell, A., and Hurley, C. (2022a). Visualizations for Bayesian additive regression trees. *arXiv preprint arXiv:2208.08966*.

Inglis, A., Parnell, A., and Hurley, C. B. (2022b). Visualizing variable importance and variable interaction effects in machine learning models. *Journal of Computational and Graphical Statistics*, 31(3):766–778.

Kablaoui, B. N. and Pautler, A. J. (1991). The effects of part-time work experience on high school students. *Journal of Career Development*, 17(3):195–211.

Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1):12–19.

Kapelner, A. and Bleich, J. (2015). Prediction with missing data via Bayesian additive regression trees. *Canadian Journal of Statistics*, 43(2):224–239.

Keane, G. and Heinz, M. (2019). Differentiated homework: Impact on student engagement. *Journal of Practitioner Research*, 4(2):1.

Kennedy, P. (2013). *Key themes in social policy*. Routledge.

Khorramdel, L., von Davier, M., Gonzalez, E., and Yamamoto, K. (2020). *Plausible values: Principles of item response theory and multiple imputations*. Springer International Publishing.

King, A. J. et al. (1989). *Improving Student Retention in Ontario Secondary Schools. Student Retention and Transition Series.* ERIC.

Klassen, R. M. and Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: Teacher gender, years of experience, and job stress. *Journal of Educational Psychology*, 102(3):741.

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., and Baumert, J. (2008). Teachers' occupational well-being and quality of instruction: The important role of self-regulatory patterns. *Journal of Educational Psychology*, 100(3):702.

Kohn, A. (2006). Abusing research: The study of homework and other examples. *Phi Delta Kappan*, 88(1):9–22.

König, C. and van de Schoot, R. (2018). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, 70(4):486–509.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4):241–253.

Krantsevich, N., He, J., and Hahn, P. R. (2023). Stochastic tree ensembles for estimating heterogeneous effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6120–6131. PMLR.

Kretschmann, J., Vock, M., and Lüdtke, O. (2014). Acceleration in elementary school: Using propensity score matching to estimate the effects on academic achievement. *Journal of Educational Psychology*, 106(4):1080.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. and Liddell, T. M. (2018). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25:178–206.

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167.

Ladd, G. W., Ettekal, I., and Kochenderfer-Ladd, B. (2017). Peer victimization trajectories from kindergarten through high school: Differential pathways for children's school engagement and achievement? *Journal of Educational Psychology*, 109(6):826.

LeCroy, C. W. and Krysik, J. (2007). Understanding and interpreting effect size measures. *Social Work Research*, 31(4):243–248.

Lee, J. C. and Staff, J. (2007). When work matters: The varying impact of work intensity on high school dropout. *Sociology of Education*, 80(2):158–178.

Lendle, S. D., Schwab, J., Petersen, M. L., and van der Laan, M. J. (2017). LTMLE: An R package implementing targeted minimum loss-based estimation for longitudinal data. *Journal of Statistical Software*, 81:1–21.

Levy, A. J., Joy, L., Ellis, P., Jablonski, E., and Karelitz, T. M. (2012). Estimating teacher turnover costs: A case study. *Journal of Education Finance*, 38(2):102–129.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001.

Li, F., Ding, P., and Mealli, F. (2023). Bayesian causal inference: A critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153.

Li, W. and Konstantopoulos, S. (2017). Does class-size reduction close the achievement gap? Evidence from TIMSS 2011. *School Effectiveness and School Improvement*, 28(2):292–313.

Lin, W.-C. and Tsai, C.-F. (2020). Missing value imputation: A review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509.

Linero, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636.

Linero, A. R., Sinha, D., and Lipsitz, S. R. (2020). Semiparametric mixed-scale models using shared Bayesian forests. *Biometrics*, 76(1):131–144.

Linero, A. R. and Zhang, Q. (2022). Mediation analysis using Bayesian tree ensembles. *Psychological Methods*.

Locke, E. A. (1969). What is job satisfaction? *Organizational Behavior and Human Performance*, 4(4):309–336.

Long, R. and Danechi, S. (2021). Teacher recruitment and retention in England. *House of Commons Briefing Paper*, Number 07222.

Lopes, J. and Oliveira, C. (2020). Teacher and school determinants of teacher job satisfaction: A multilevel analysis. *School Effectiveness and School Improvement*, 31(4):641–659.

Lunsford, L., Baker, V., and Pifer, M. (2018). Faculty mentoring faculty: Career stages, relationship quality, and job satisfaction. *International Journal of Mentoring and Coaching in Education*, 7(2):139–154.

Lutz, A. and Jayaram, L. (2015). Getting the homework done: Social class and parents' relationship to homework. *International Journal of Education and Social Science*, 2(6):73–84.

Madigan, D. J. and Kim, L. E. (2021). Towards an understanding of teacher attrition: A meta-analysis of burnout, job satisfaction, and teachers' intentions to quit. *Teaching and Teacher Education*, 105.

Martin, M. O., von Davier, M., and Mullis, I. V. (2020). Methods and procedures: TIMSS 2019 technical report. *International Association for the Evaluation of Educational Achievement*.

Marzano, R. J. and Pickering, D. J. (2007). Special topic: The case for and against homework. *Educational Leadership*, 64(6):74–79.

Mayer, M. (2019). Package 'missRanger'. https://cran.r-project.org/web/packages/missRanger/missRanger.pdf.

McCall, M. S., Hauser, C., Cronin, J., Kingsbury, G. G., and Houser, R. (2006). Achievement gaps: An examination of differences in student achievement and growth. The full report. *Northwest Evaluation Association*.

McCormick, M. P., O'Connor, E. E., Cappella, E., and McClowry, S. G. (2013). Teacher–child relationships and academic achievement: A multilevel propensity score model approach. *Journal of School Psychology*, 51(5):611–624.

McCrory Calarco, J., Horn, I. S., and Chen, G. A. (2022). "You need to be more responsible": The myth of meritocracy and teachers' accounts of homework inequalities. *Educational Researcher*, 51(8):515–523.

McInerney, D. M., Korpershoek, H., Wang, H., and Morin, A. J. (2018). Teachers' occupational attributes and their psychological wellbeing, job satisfaction, occupational self-concept and quitting intentions. *Teaching and Teacher Education*, 71:145–158.

McJames, N., O'Shea, A., Goh, Y. C., and Parnell, A. (2024). Bayesian causal forests for multivariate outcomes: Application to Irish data from an international large scale education assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, pages 1–23.

McJames, N., Parnell, A., and O'Shea, A. (2023a). Investigating the effect of creative mathematical reasoning tasks on student achievement: A causal inference machine learning approach. In Twohill, A. and Quirke, S., editors, *Proceedings of the Ninth Conference on Research in Mathematics Education in Ireland (MEI 9)*.

McJames, N., Parnell, A., and O'Shea, A. (2023b). Factors affecting teacher job satisfaction: A causal inference machine learning approach using data from TALIS 2018. *Educational Review*, pages 1–25.

McKelvie-Sebileau, P., Swinburn, B., Glassey, R., Tipene-Leach, D., and Gerritsen, S. (2023). Health, wellbeing and nutritional impacts after 2 years of free school meals in New Zealand. *Health Promotion International*, 38(4):daad093.

McNealis, V., Moodie, E. E., and Dean, N. (2024). Revisiting the effects of maternal education on adolescents' academic performance: Doubly robust estimation in a network-based observational study. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(3):715–734.

Mitra, R. (2023). A latent class model to multiply impute missing treatment indicators in observational studies when inferences of the treatment effect are made using propensity score matching. *Biometrical Journal*, 65(3):2100284.

Monahan, K. C., Lee, J. M., and Steinberg, L. (2011). Revisiting the impact of part-time work on adolescent adjustment: Distinguishing between selection and socialization using propensity score matching. *Child Development*, 82(1):96–112.

Morgan, P. L., Farkas, G., Hillemeier, M. M., and Maczuga, S. (2016). Science achievement gaps begin very early, persist, and are largely explained by modifiable factors. *Educational Researcher*, 45(1):18–35.

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., and Fishbein, B. (2020). TIMSS 2019 international results in mathematics and science. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results.

Mullis, I. V. S. and Martin, M. O. (2017). TIMSS 2019 assessment frameworks. Retrieved from Boston College, TIMSS & PIRLS International Study Center website.

Murphy, D. (2014). Issues with PISA's use of its data in the context of international education policy convergence. *Policy Futures in Education*, 12(7):893–916.

Murray, J. S. (2021). Log-linear Bayesian additive regression trees for multinomial logistic and count regression models. *Journal of the American Statistical Association*, 116(534):756–769.

Myint, L. (2024). Controlling time-varying confounding in difference-in-differences studies using the time-varying treatments framework. *Health Services and Outcomes Research Methodology*, 24(1):95–111.

Newhouse, J. P. and McClellan, M. (1998). Econometrics in outcomes research: The use of instrumental variables. *Annual Review of Public Health*, 19(1):17–34.

Nguyen, T. D. (2021). Linking school organizational characteristics and teacher retention: Evidence from repeated cross-sectional national data. *Teaching and Teacher Education*, 97.

Niemann, D., Martens, K., and Teltemann, J. (2017). PISA and its consequences: Shaping education policies through international comparisons. *European Journal of Education*, 52(2):175–183.

OECD (2018). TALIS 2018 teacher questionnaire. https://www.oecd.org/education/school/TALIS-2018-MS-Teacher-Questionnaire-ENG.pdf.

OECD (2019a). TALIS 2018 and TALIS starting strong 2018 user guide. https://www.oecd.org/education/talis/TALIS_2018-TALIS_Starting_Strong_2018_User_Guide.pdf.

OECD (2019b). TALIS 2018 technical report. Chapter 11. Validation of scales and construction of scale scores. https://www.oecd.org/education/talis/TALIS_2018_Technical_Report.pdf.

Olmos, F. (2010). *Square peg in a round hole: Out-of-field teaching and its impact on teacher attrition*. University of California, Irvine and California State University, Los Angeles.

O'Neill, S., Grieve, R., Singh, K., Dutt, V., and Powell-Jackson, T. (2024). Persistence and heterogeneity of the effects of educating mothers to improve child immunisation uptake: Experimental evidence from Uttar Pradesh in India. *Journal of Health Economics*, 96:102899.

O'Doherty, T. and Harford, J. (2018). Teacher recruitment: Reflections from Ireland on the current crisis in teacher supply. *European Journal of Teacher Education*, 41(5):654–669.

Palardy, J. M. (1988). The effect of homework policies on student achievement. *NASSP Bulletin*, 72(507):14–17.

Pan, W. and Bai, H. (2018). Propensity score methods for causal inference: An overview. *Behaviormetrika*, 45(2):317–334.

Parker, M. A., Ndoye, A., and Imig, S. R. (2009). Keeping our teachers! Investigating mentoring practices to support and retain novice educators. *Mentoring & Tutoring: Partnership in Learning*, 17(4):329–341.

Patall, E. A., Cooper, H., and Robinson, J. C. (2008). Parent involvement in homework: A research synthesis. *Review of Educational Research*, 78(4):1039–1101.

Pearl, J. (1995). From Bayesian networks to causal networks. In *Mathematical models for handling partial knowledge in artificial intelligence*, pages 157–182. Springer.

Perrachione, B. A., Rosser, V. J., and Petersen, G. J. (2008). Why do they stay? Elementary teachers' perceptions of job satisfaction and retention. *Professional Educator*, 32(2).

Pierdzioch, C., Risse, M., and Rohloff, S. (2016). Are precious metals a hedge against exchange-rate movements? An empirical exploration using Bayesian ad-

ditive regression trees. *The North American Journal of Economics and Finance*, 38:27–38.

Ponzo, M. (2013). Does bullying reduce educational achievement? An evaluation using matching estimators. *Journal of Policy Modeling*, 35(6):1057–1078.

Prado, E. B., Moral, R. A., and Parnell, A. C. (2021a). Bayesian additive regression trees with model trees. *Statistics and Computing*, 31(3):1–13.

Prado, E. B., Parnell, A. C., McJames, N., O'Shea, A., and Moral, R. A. (2021b). Semi-parametric Bayesian additive regression trees. *arXiv preprint arXiv:2108.07636*.

Prendergast, M. and O'Meara, N. (2017). A profile of mathematics instruction time in Irish second level schools. *Irish Educational Studies*, 36(2):133–150.

Pressman, R. M., Sugarman, D. B., Nemon, M. L., Desjarlais, J., Owens, J. A., and Schettini-Evans, A. (2015). Homework and family stress: With consideration of parents' self confidence, educational level, and cultural background. *The American Journal of Family Therapy*, 43(4):297–313.

Provasnik, S. and Dorfman, S. (2005). Mobility in the teacher workforce findings from the condition of education 2005. *National Center for Education Statistics*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reeves, P. M., Pun, W. H., and Chung, K. S. (2017). Influence of teacher collaboration on job satisfaction and student achievement. *Teaching and Teacher Education*, 67:227–236.

Reeves, T. D., Hamilton, V., and Onder, Y. (2022). Which teacher induction practices work? Linking forms of induction to teacher practices, self-efficacy, and job satisfaction. *Teaching and Teacher Education*, 109:103546.

Renbarger, R. and Davis, B. (2019). Mentors, self-efficacy, or professional development: Which mediate job satisfaction for new teachers? A regression examination. *Journal of Teacher Education and Educators*, 8(1):21–34.

Richter, D., Kunter, M., Lüdtke, O., Klusmann, U., Anders, Y., and Baumert, J. (2013). How different mentoring approaches affect beginning teachers' development in the first years of practice. *Teaching and Teacher Education*, 36:166–177.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Robotham, D. (2012). Student part-time employment: Characteristics and consequences. *Education+ Training*, 54(1):65–75.

Ronfeldt, M. and McQueen, K. (2017). Does new teacher induction really improve retention? *Journal of Teacher Education*, 68(4):394–410.

Rønning, M. (2011). Who benefits from homework assignments? *Economics of Education Review*, 30(1):55–64.

Roschelle, J., Feng, M., Murphy, R. F., and Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, 2(4):1–12.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Rowley, K. J., Edmunds, C. C., Dufur, M. J., Jarvis, J. A., and Silveira, F. (2020). Contextualising the achievement gap: Assessing educational achievement, inequality, and disadvantage in high-income countries. *Comparative Education*, 56(4):459–483.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.

Rutkowski, D. J. (2007). Converging us softly: how intergovernmental organizations promote neoliberal educational policy. *Critical Studies in Education*, 48(2):229–247.

Rutkowski, L., Gonzalez, E., Joncas, M., and Von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2):142–151.

Saloviita, T. and Pakarinen, E. (2021). Teacher burnout explained: Teacher-, student-, and organisation-level variables. *Teaching and Teacher Education*, 97.

Samartsidis, P., Seaman, S. R., Montagna, S., Charlett, A., Hickman, M., and Angelis, D. D. (2020). A Bayesian multivariate factor analysis model for evaluating an intervention by using observational time series data on multiple outcomes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1437–1459.

Sarti, D. A., Prado, E. B., Inglis, A. N., Dos Santos, A. A., Hurley, C. B., Moral, R. A., and Parnell, A. C. (2023). Bayesian additive regression trees for genotype by environment interaction models. *The Annals of Applied Statistics*, 17(3):1936–1957.

Schwartz, M. H., Kainz, H., and Georgiadis, A. G. (2021). Estimating causal treatment effects of femoral and tibial derotational osteotomies on foot progression in children with cerebral palsy. *medRxiv*, pages 2021–03.

Sekhon, J. S. (2008). The Neyman-Rubin model of causal inference and estimation via matching methods. *The Oxford Handbook of Political Methodology*, 2:1–32.

Sellar, S. and Lingard, B. (2014). The OECD and the expansion of PISA: New global modes of governance in education. *British Educational Research Journal*, 40(6):917–936.

Shen, J. (1997). Teacher retention and attrition in public schools: Evidence from SASS91. *The Journal of Educational Research*, 91(2):81–88.

Sims, S. (2017). TALIS 2013: Working conditions, teacher job satisfaction and retention. Statistical working paper, UK Department for Education, Castle View House, East Lane, Runcorn, Cheshire, WA7 2GJ, UK.

Singh, K. and Ozturk, M. (2000). Effect of part-time work on high school mathematics and science course taking. *The Journal of Educational Research*, 94(2):67–74.

Skaalvik, E. M. and Skaalvik, S. (2015). Job satisfaction, stress and coping strategies in the teaching profession-what do teachers say? *International Education Studies*, 8(3):181–192.

Small, C. (2020). *A Comparison of Public and Private School Teachers' Job Satisfaction When Controlling for Policy Perspectives, Individual, and Workplace Characteristics*. Doctor of education in educational leadership dissertation, Kennesaw State University, Kennesaw, Georgia, USA.

Smethem, L. (2007). Retention and intention in teaching careers: Will the new generation stay? *Teachers and Teaching: Theory and Practice*, 13(5):465–480.

Solomon, Y., Warin, J., and Lewis, C. (2002). Helping with homework? Homework as a site of tension for parents and teenagers. *British Educational Research Journal*, 28(4):603–622.

Sönmezer, M. G. and Eryaman, M. Y. (2008). A comparative analysis of job satisfaction levels of public and private school teachers. *Journal of Theory & Practice in Education (JTPE)*, 4(2).

Sorensen, L. C. and Ladd, H. F. (2020). The hidden costs of teacher turnover. *AERA Open*, 6(1).

Spanbauer, C. and Sparapani, R. (2021). Nonparametric machine learning for precision medicine with longitudinal clinical trials and Bayesian additive regression trees with mixed models. *Statistics in Medicine*, 40(11):2665–2691.

Sparapani, R. A., Logan, B. R., McCulloch, R. E., and Laud, P. W. (2016). Nonparametric survival analysis using Bayesian additive regression trees (BART). *Statistics in Medicine*, 35(16):2741–2753.

Spilt, J. L., Koomen, H. M., and Thijs, J. T. (2011). Teacher wellbeing: The importance of teacher–student relationships. *Educational Psychology Review*, 23(4):457–477.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472.

Starling, J. E., Murray, J. S., Carvalho, C. M., Bukowski, R. K., and Scott, J. G. (2020). BART with targeted smoothing: An analysis of patient-specific stillbirth risk. *The Annals of Applied Statistics*, 14(1):28–50.

Starling, J. E., Murray, J. S., Lohr, P. A., Aiken, A. R., Carvalho, C. M., and Scott, J. G. (2021). Targeted smooth Bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous vs. interval medical abortion regimens over gestation. *The Annals of Applied Statistics*, 15(3):1194–1219.

Steinberg, L. D., Greenberger, E., Garduque, L., and McAuliffe, S. (1982). High school students in the labor force: Some costs and benefits to schooling and learning. *Educational Evaluation and Policy Analysis*, 4(3):363–372.

Stekhoven, D. J. and Bühlmann, P. (2012). missForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 25(1):1.

Suk, Y., Kim, J.-S., and Kang, H. (2021). Hybridizing machine learning methods and finite mixture models for estimating heterogeneous treatment effects in latent classes. *Journal of Educational and Behavioral Statistics*, 46(3):323–347.

Tan, C. Y., Lyu, M., and Peng, B. (2020). Academic benefits from parental involvement are stratified by parental socioeconomic status: A meta-analysis. *Parenting*, 20(4):241–287.

Tang, A., Li, W., and Liu, D. (2022). The impact of teachers' professional development in science pedagogy on students' achievement: Evidence from TIMSS 2019. *Journal of Baltic Science Education*, 21(2):258–274.

Taras, H. (2005). Nutrition and student performance at school. *Journal of School Health*, 75(6):199–213.

TIMSS & PIRLS International Study Center (2020). TIMSS 2019 context questionnaires. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/questionnaires/.

Tompsett, D., Vansteelandt, S., Dukes, O., and De Stavola, B. (2022). gesttools: General purpose G-estimation in R. *Observational Studies*, 8(1):1–28.

Toropova, A., Myrberg, E., and Johansson, S. (2021). Teacher job satisfaction: The importance of school working conditions and teacher characteristics. *Educational Review*, 73(1):71–97.

Trautwein, U. (2007). The homework–achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction*, 17(3):372–388.

Trautwein, U., Köller, O., Schmitz, B., and Baumert, J. (2002). Do homework assignments enhance achievement? A multilevel analysis in 7th-grade mathematics. *Contemporary Educational Psychology*, 27(1):26–50.

Trautwein, U. and Lüdtke, O. (2009). Predicting homework motivation and homework effort in six school subjects: The role of person and family characteristics, classroom factors, and school track. *Learning and Instruction*, 19(3):243–258.

Tsai, L.-T. and Yang, C.-C. (2015). Hierarchical effects of school-, classroom-, and student-level factors on the science performance of eighth-grade Taiwanese students. *International Journal of Science Education*, 37(8):1166–1181.

Um, S., Linero, A. R., Sinha, D., and Bandyopadhyay, D. (2023). Bayesian additive regression trees for multivariate skewed responses. *Statistics in Medicine*, 42(3):246–263.

UNESCO (2015). The challenge of teacher shortage and quality: Have we succeeded in getting enough quality teachers into classrooms?

Upsing, B. and Hayatli, M. (2021). The challenges of test translation. *International Perspectives on School Settings, Education Policy and Digital Strategies: A Transatlantic Discourse in Education Research*, page 373.

Vesić, D., Džinović, V., and Mirkov, S. (2021). The role of absenteeism in the prediction of math achievement on the basis of self-concept and motivation: TIMSS 2015 in Serbia. *Psihologija*, 54(1):15–31.

Vik, F. N., Nilsen, T., and Øverby, N. C. (2022). Aspects of nutritional deficits and cognitive outcomes–triangulation across time and subject domains among students and teachers in TIMSS. *International Journal of Educational Development*, 89:102553.

Viljaranta, J., Silinskas, G., Lerkkanen, M.-K., Hirvonen, R., Pakarinen, E., Poikkeus, A.-M., and Nurmi, J.-E. (2018). Maternal homework assistance and children's task-persistent behavior in elementary school. *Learning and Instruction*, 56:54–63.

Volante, L., Fazio, X., and Ritzen, J. (2017). The OECD and educational policy reform: International surveys, governance, and policy evidence. *Canadian Journal of Educational Administration and Policy*, (184).

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, H., Hall, N. C., and Rahimi, S. (2015). Self-efficacy and causal attributions in teachers: Effects on burnout, job satisfaction, illness, and quitting intentions. *Teaching and teacher education*, 47:120–130.

Wang, K., Li, Y., Luo, W., and Zhang, S. (2020). Selected factors contributing to teacher job satisfaction: A quantitative investigation using 2013 TALIS data. *Leadership and Policy in Schools*, 19(3):512–532.

Wang, M., Martinez, I., and Hahn, P. R. (2024). Longbet: Heterogeneous treatment effect estimation in panel data. *arXiv preprint arXiv:2406.02530*.

West, S. G., Duan, N., Pequegnat, W., Gaist, P., Des Jarlais, D. C., Holtgrave, D., Szapocznik, J., Fishbein, M., Rapkin, B., Clatts, M., et al. (2008). Alternatives to the randomized controlled trial. *American Journal of Public Health*, 98(8):1359–1366.

Wiggan, G., Smith, D., and Watson-Vandiver, M. J. (2021). The national teacher shortage, urban education and the cognitive sociology of labor. *The Urban Review*, 53(1):43–75.

Woessmann, L. and West, M. (2006). Class-size effects in school systems around the world: Evidence from between-grade variation in TIMSS. *European Economic Review*, 50(3):695–736.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, 31(2-3):114–128.

Wundervald, B., Parnell, A., and Domijan, K. (2022). Hierarchical embedded Bayesian additive regression trees. *arXiv preprint arXiv:2204.07207*.

Yeager, D. S., Bryan, C. J., Gross, J. J., Murray, J. S., Krettek Cobb, D., HF Santos, P., Gravelding, H., Johnson, M., and Jamieson, J. P. (2022). A synergistic mindsets intervention protects adolescents from stress. *Nature*, 607(7919):512–520.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., et al. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774):364–369.

Yoon, I. and Kim, M. (2022). Dynamic patterns of teachers' professional development participation and their relations with socio-demographic characteristics, teacher self-efficacy, and job satisfaction. *Teaching and Teacher Education*, 109.

Zhang, S., Shi, Q., and Lin, E. (2020). Professional development needs, support, and barriers: TALIS US new and veteran teachers' perspectives. *Professional Development in Education*, 46(3):440–453.

Zhao, L., Yuan, H., and Wang, X. (2024). Impact of homework time on adolescent mental health: Evidence from China. *International Journal of Educational Development*, 107:103051.

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, 21(2):245–266.

Zhou, H., Liu, J., Shao, Y., and Yanyan, T. (2023). How much is too much time spent on homework: An exploratory study based on a Bayesian multilevel piecewise model with a random change point. *Educational Studies*, pages 1–13.

Zhu, Y. and Leung, F. K. S. (2012). Homework and mathematics achievement in Hong Kong: Evidence from the TIMSS 2003. *International Journal of Science and Mathematics Education*, 10:907–925.

Zimmerman, B. J. and Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, 30(4):397–417.

A

Appendix for Chapter 3

A.1 Definitions of Key Terms Given in TALIS Questionnaire

Key Term	Definition
CPD	In this section, 'professional development' is defined as activities that aim to develop an individual's skills, knowledge, expertise and other characteristics as a teacher.
Induction	'Induction activities' are designed to support new teachers' introduction into the teaching profession and to support experienced teachers who are new to a school, and they are either organised in formal, structured programmes or informally arranged as separate activities.
Mentoring	'Mentoring' is defined as a support structure in schools where more experienced teachers support less experienced teachers. This structure might involve all teachers in the school or only new teachers. It does not include mentoring of student teachers doing teaching practice at this school.
Public School	This is a school managed by a public education authority, government agency, municipality, or governing board appointed by government or elected by public franchise.
Private School	This is a school managed by a non-government organisation; e.g. a church, trade union, business or other private institution.

A.2 Questions Used to Define Treatment Groups

Treatment	Question	Condition
CPD	During the last 12 months, did you participate in any of the following professional development activities?	Teachers who responded "yes" to any 4 of the 10 available options.
Induction	Did you take part in any induction activities?	Teachers who responded "yes" to either taking part in a formal or informal induction programme at their current school.
Observing	On average, how often do you do the following in this school?	Teachers who did not respond "never" to the option "Observe other teachers' classes and provide feedback."
Team Teaching	On average, how often do you do the following in this school?	Teachers who did not respond "never" to the option "Teach jointly as a team in the same class."
Has Mentor	Are you currently involved in any mentoring activities as part of a formal arrangement at this school?	Teachers who responded "yes" to having a mentor.
Is Mentor	Are you currently involved in any mentoring activities as part of a formal arrangement at this school?	Teachers who responded "yes" to being a mentor.
Public	Is this school publicly or privately-managed?	Teachers with a principal who indicated their school is publicly-managed.
30+ Students	How many students are currently enrolled in this class?	Teachers who answered 30 or more students.
Out-of-field	Were the following subject categories included in your formal education or training, and do you teach them during the current school year to any students in this school?	Teachers who indicated that at least one option given was not included in their education, but that they do currently teach it.
Part-Time	What is your current employment status as a teacher, in terms of working hours?	Teachers who indicated they do not have a full time contract at their current school.

A.3 List of Potential Confounders Used

TALIS Variable Code	Description	Removed from X for treatment
IDCNTPOP	Primary/Secondary School.	
TT3G01	Gender.	
TT3G03	Highest level of formal education completed.	
TT3G04	How did you receive your first teaching qualification?	
TT3G05	Year of Qualification.	
TT3G08	Was teaching your first choice as a career?	
TT3G09	Permanent/Fixed-Term Contract.	
TT3G10A	Working hours at this school.	Part-Time Contract.
TT3G10B	Working hours altogether.	Part-Time Contract.
TC3G12	School publicly/privately managed	Public School.
TT3G11A	Year(s) working as a teacher at this school.	
TT3G11B	Year(s) working as a teacher in total.	
TT3G11C	Year(s) working in other education roles.	
TT3G11D	Year(s) working in non education roles.	
TT3G12	Do you currently work as a teacher at another school?	
TT3G14	Number of students in class with special needs.	
TT3G37	Subject taught.	
TT3G38	Number of students in class.	30+ Students.
TT3G39A	% of time spent on administrative tasks.	
TT3G39B	% of time spent keeping order in classroom.	
TT3G39C	% of time actually spent teaching.	
T3STBEH	Student behaviour stress.	
T3CLAIN	Clarity of instruction.	
T3CLASM	Classroom management.	
T3COGAC	Cognitive activation.	
T3COLES	Professional collaboration in lessons among teachers.	
T3EFFPD	Effective professional development.	
T3EXCH	Exchange and co-ordination among teachers.	
T3PDBAR	Professional development barriers.	
T3DISC	Teachers' perceived disciplinary climate.	
T3PERUT	Personal utility motivation to teach.	
T3PDIV	Needs for professional development for teaching for diversity.	
T3PDPED	Needs for professional development in subject matter and pedagogy.	
T3VALP	Perceptions of value and policy influence.	
T3SATAT	Satisfaction with target class autonomy.	
T3SECLS	Self-efficacy in classroom management.	
T3SEINS	Self-efficacy in instruction.	
T3SEENG	Self-efficacy in student engagement.	
T3SEFE	Self-related efficacy in multicultural classrooms.	
T3SOCUT	Social utility value.	
T3STAKE	Participation among stakeholders, teachers.	
T3TEAM	Team innovativeness.	
T3STUD	Teacher-student relations.	
T3WELS	Workplace well-being and stress.	
T3WLOAD	Workload stress.	
T3TPRA	Teaching practices, overall.	
T3COOP	Teacher cooperation.	
T3SELF	Teacher self-efficacy.	
T3DIVP	Diversity practices.	
T3JOBSA	Overall job satisfaction.	All.

B

Appendix for Chapter 4

B.1 Multivariate BCF Updates

B.1.1 Log-Likelihood of a μ Tree

Let T_j denote the j^{th} μ tree in the ensemble with partial residuals R_j. Also suppose that tree T_j has K terminal nodes $h_{j,1}...h_{j,K}$, and L non-terminal nodes $b_{j,1}...b_{j,L}$. Furthermore, let $R_{j,k,1}...R_{j,k,n_k}$ denote the partial residuals which fall into the k^{th} terminal node of tree T_j. Then given the residual covariance matrix Σ, the tree priors α and β, and the prior covariance matrix Σ_μ for terminal node parameters, we have that:

$$\ell(T_j|R_j,\Sigma) = \ell(R_j|T_j,\Sigma) + \ell(T_j) + C$$

$$\ell(T_j) = \sum_{k=1}^{K} \log\left(1-\alpha(1+d(h_{j,k}))^{-\beta}\right) + \sum_{l=1}^{L} \log(\alpha) - \beta\log(1+d(b_{j,l}))$$

$$\ell(R_j|T_j,\Sigma) = \sum_{k=1}^{K} \ell(R_{j,k,1},\ldots,R_{j,k,n_k}|T_j,\Sigma)$$

$$= \sum_{k=1}^{K}\left\{-\frac{n_k}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\log\left(|\Sigma_\mu|\right) + \frac{1}{2}\log\left(|\Sigma_{k,0}|\right) - \frac{1}{2}\sum_{i=1}^{n_k}\left(R_{j,k,i}^T\Sigma^{-1}R_{j,k,i} - \mu_{k,0}^T\Sigma_{k,0}^{-1}\mu_{k,0}\right)\right\} + C$$

where

$$\Sigma_{k,0}^{-1} = n_k\Sigma^{-1} + \Sigma_\mu^{-1}$$

and

$$\mu_{k,0} = \Sigma_{k,0}\Sigma^{-1}\left(\sum_{i=1}^{n_k} R_{j,k,i}\right)$$

B.1.2 Posterior Distribution of Terminal Node Parameters in a μ Tree

For the k^{th} terminal node of any tree T_j, the posterior distribution of the terminal node parameter $\mu_{j,k}$ with prior mean μ_0 is given by:

$$\mu_{j,k}|\ldots \sim N\left(\mu_n, \Sigma_n\right)$$

where

$$\mu_n = \left(\Sigma_\mu^{-1} + n_k\Sigma^{-1}\right)^{-1}\left(\Sigma_\mu^{-1}\mu_0 + n_k\Sigma^{-1}\overline{R}\right)$$

and

$$\Sigma_n = \left(\Sigma_\mu^{-1} + n_k\Sigma^{-1}\right)^{-1}$$

B.1.3 Log-Likelihood of a τ Tree

Given the prior covariance matrix for terminal node parameters Σ_τ, and vector Z_k which holds the treatment status of each observation in terminal node k, we obtain:

$$\ell(T_j|R_j,\Sigma) = \ell(R_j|T_j,\Sigma) + \ell(T_j) + C$$

$$\ell(T_j) = \sum_{k=1}^{K}\log\left(1-\alpha(1+d(h_{j,k}))^{-\beta}\right) + \sum_{l=1}^{L}\log(\alpha) - \beta\log(1+d(b_{j,l}))$$

$$\ell(R_j|T_j,\Sigma) = \sum_{k=1}^{K}\ell(R_{j,k,1},\ldots,R_{j,k,n_k}|T_j,\Sigma)$$

$$= \sum_{k=1}^{K}\left\{-\frac{n_k}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\log\left(|\Sigma_\tau|\right) + \frac{1}{2}\log\left(|\Sigma_{k,0}|\right) - \frac{1}{2}\sum_{i=1}^{n_k}\left(R_{j,k,i}^T\Sigma^{-1}R_{j,k,i} - \tau_{k,0}^T\Sigma_{k,0}^{-1}\tau_{k,0}\right)\right\} + C$$

where

$$\Sigma_{k,0}^{-1} = \sum_{i=1}^{n_k}Z_{j,k,i}\Sigma^{-1} + \Sigma_\tau^{-1}$$

$$\tau_{k,0} = \Sigma_{k,0}\Sigma^{-1}\left(\sum_{i=1}^{n_k}R_{j,k,i}Z_{j,k,i}\right)$$

B.1.4 Posterior Distribution of Terminal Node Parameters in a τ Tree

Analogously to the terminal node parameters in a μ tree, the posterior distribution of the terminal node parameter $\tau_{j,k}$ in the k^{th} leaf of the j^{th} τ tree T_j, with prior mean τ_0 is:

$$\tau_{j,k}|\ldots \sim N\left(\tau_n, \Sigma_n\right)$$

where

$$\tau_n = \left(\Sigma_\tau^{-1} + \sum_{i=1}^{n_k} Z_{j,k,i} \Sigma^{-1} \right)^{-1} \left(\Sigma_\tau^{-1} \tau_0 + \Sigma^{-1} \sum_{i=1}^{n_k} R_{j,k,i} Z_{j,k,i} \right)$$

and

$$\Sigma_n = \left(\Sigma_\tau^{-1} + \sum_{i=1}^{n_k} Z_{j,k,i} \Sigma^{-1} \right)^{-1}$$

B.1.5 Posterior Distribution of Residual Covariance Parameter Σ

Given observed and predicted values y and $\hat{y}$, the posterior distribution of the covariance matrix Σ for the n residuals, with prior scale matrix Σ_0 and ν_0 degrees of freedom is given by:

$$\Sigma | \ldots \sim \mathcal{W}^{-1} \left(\nu_0 + n, [S_0 + S_\theta]^{-1} \right)$$

where

$$S_\theta = \sum_{i=1}^{n} (y_i - \hat{y}_i)^T \Sigma^{-1} (y_i - \hat{y}_i).$$

B.2 TIMSS Variables Used in Study

TIMSS Variable Codes		
Variable Code	Obtained From	Description
BSDAGE	Student Questionnaire	Student Age
BSBG01	Student Questionnaire	Student Gender
BSBG03	Student Questionnaire	How often student speaks English at home
BSBG04	Student Questionnaire	Number of books at home
BSBG07	Student Questionnaire	How far in education student expects to go
BSBG08A	Student Questionnaire	Was parent/guardian A born in Ireland
BSBG08B	Student Questionnaire	Was parent/guardian B born in Ireland
BSBG09A	Student Questionnaire	Was student born in Ireland
BSBG10	Student Questionnaire	How often student is absent
BSBG11A	Student Questionnaire	How often student feels hungry when arriving at school
BSBG11B	Student Questionnaire	How often student feels tired when arriving at school
BSDGEDUP	Student Questionnaire	Parent's highest education level
BSBGHER	Student Questionnaire	Number of home educational resources
BSBGSSB	Student Questionnaire	Sense of school belonging
BSBGSB	Student Questionnaire	School bullying
BSBGSCM/BSBGSCS	Student Questionnaire	Confidence in mathematics/science
BSBGSVM/BSBGSVS	Student Questionnaire	Student values mathematics/science
BSBGICM/BSBGICS	Student Questionnaire	Instructional clarity in mathematics/science
BSBG05A	Student Questionnaire	Has computer/tablet at home
BSBG05B	Student Questionnaire	Has study desk at home
BSBG05C	Student Questionnaire	Has own bedroom
BSBG05D	Student Questionnaire	Has home internet connection
BSBG05E	Student Questionnaire	Has own mobile phone
BSBG05F	Student Questionnaire	Has gaming system
BSBG05G	Student Questionnaire	Home TV has "premium" TV channels
BTBG01	Teacher Questionnaire	Number of years teaching
BTBG02	Teacher Questionnaire	Teacher gender
BTBG03	Teacher Questionnaire	Teacher age
BTBG10	Teacher Questionnaire	Number of students in class
BTBGTJS	Teacher Questionnaire	Teacher job satisfaction
BTBGSOS	Teacher Questionnaire	Safe and orderly school
BTBGLSN	Teacher Questionnaire	Teaching is limited by students not ready for instruction
BTBGEAS	Teacher Questionnaire	Emphasis on academic success
BTDMME	Teacher Questionnaire	Type of degree
BCBGDAS	Principal Questionnaire	School discipline
BCBGEAS	Principal Questionnaire	Emphasis on academic success
BCBGMRS/BCBGSRS	Principal Questionnaire	Resource shortages in mathematics/science
BCDGSBC	Principal Questionnaire	School average socioeconomic background

Table B.1: Variable codes for the treatment variables and control variables used from the TIMSS 2019 data. The same control variables and treatment effect moderators are used in all three models. Of the control variables used, only home resources, parental education, school average socioeconomic status, and school resources were examined as potential effect moderators.

B.3 Sensitivity to σ_μ and σ_τ

The plots below visualise the sensitivity of the model performance as measured by the Precision in Estimating Heterogeneous Effects (PEHE) to the specific choice of σ_μ and σ_τ. This was accomplished by completing many simulations from Data Generating Process 1 (DGP1), with a sample size of 500, and varying the values of σ_μ and σ_τ in order to build a picture of how these choices affect model performance. In the upper plot which visualises the sensitivity of the PEHE to σ_μ, the value for σ_τ was fixed at $\frac{3}{8\sqrt{J_\tau}}$, where J_τ is the number of τ trees used in the ensemble, while the value for σ_μ was chosen uniformly at random from between $\frac{1}{100\sqrt{J_\mu}}$ and $\frac{3}{\sqrt{J_\mu}}$ in each unique simulation. In the lower plot which visualises the sensitivity of the PEHE to σ_τ, the value for σ_μ was fixed at $\frac{1}{\sqrt{J_\mu}}$, while the value for σ_τ was chosen uniformly at random from between $\frac{1}{100\sqrt{J_\tau}}$ and $\frac{3}{\sqrt{J_\tau}}$ in each unique simulation. In both cases, the performance is quite insensitive to the specific values of σ_μ and σ_τ, except for very low values which impose excessively strong regularisation on the terminal node parameters.

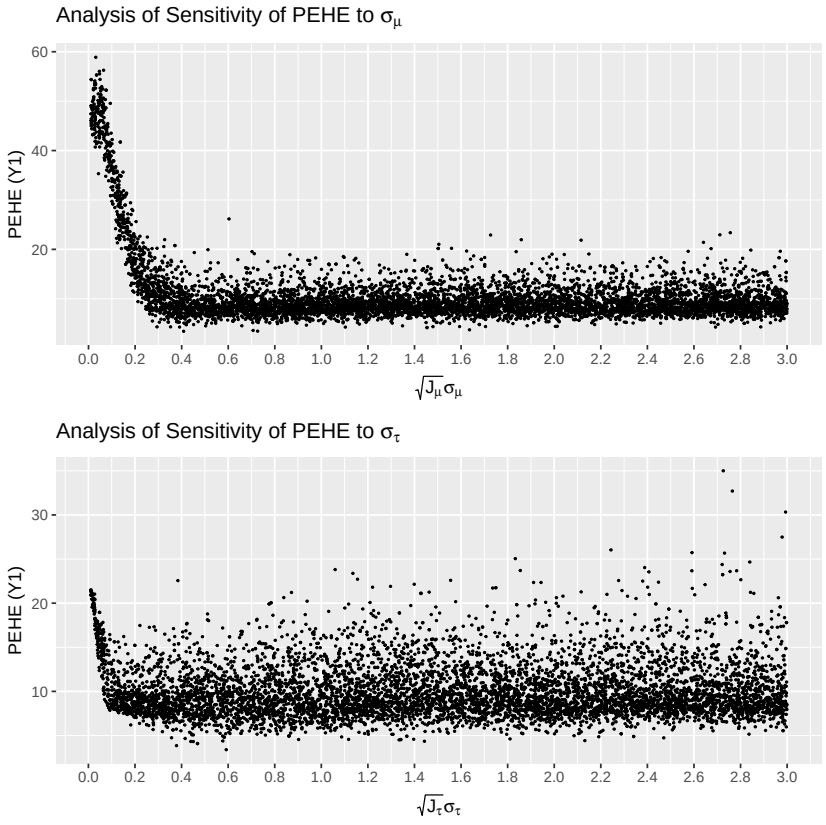

Figure B.1: Investigation of model sensitivity to σ_μ and σ_τ.

B.4 Simulation Study Results For All Sample Sizes

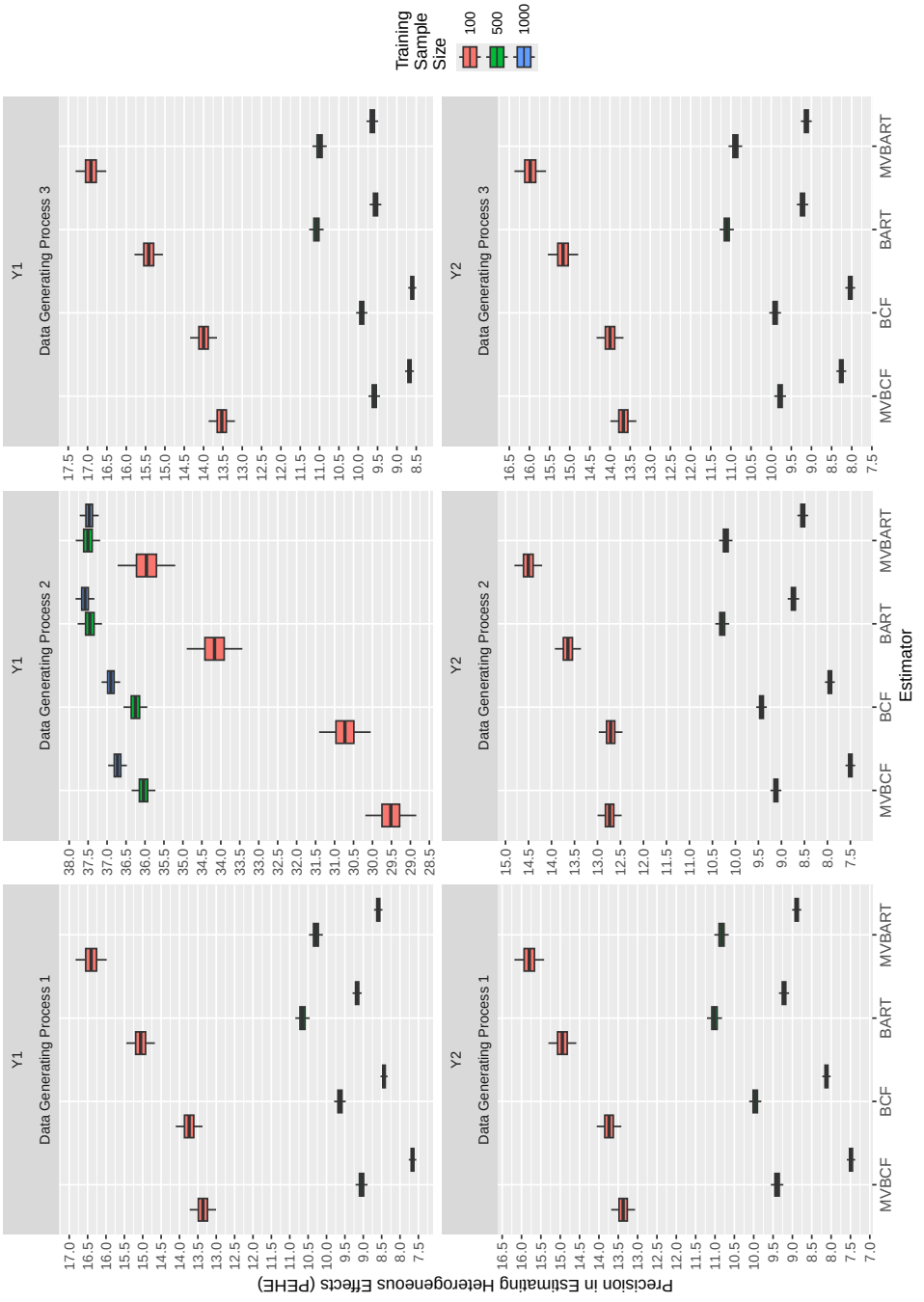

Figure B.2: Simulation study results for all sample sizes.

	Simulation Study Results, Training Sample Size = 100							
	MVBCF		BCF		BART		MVBART	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
DGP1								
PEHE on τ	13.36±0.36	13.37±0.31	13.74±0.36	13.74±0.32	15.06±0.39	14.94±0.36	16.40±0.42	15.80±0.38
Bias on MATE	-0.98±0.76	0.01±0.70	-0.26±0.78	0.56±0.72	-0.17±0.83	0.53±0.79	-0.26±0.89	0.39±0.83
τ 95% Coverage	0.98±0.00	0.97±0.00	0.98±0.00	0.97±0.00	0.99±0.00	0.99±0.00	0.99±0.00	0.99±0.00
MATE 95% Coverage	0.97±0.01	0.98±0.01	0.96±0.01	0.98±0.01	0.96±0.01	0.97±0.01	0.96±0.01	0.97±0.01
τ 95% CI Width	65.63±0.33	63.05±0.32	67.60±0.33	65.01±0.32	86.17±0.51	83.61±0.49	100.8±0.61	96.12±0.56
MATE 95% CI Width	53.14±0.35	52.15±0.34	51.68±0.35	51.10±0.34	53.63±0.38	53.37±0.37	58.49±0.45	58.12±0.42
DGP2								
PEHE on τ	29.52±0.67	12.74±0.26	30.73±0.68	12.71±0.26	34.17±0.73	13.64±0.28	35.96±0.76	14.50±0.30
Bias on MATE	27.97±0.73	-2.62±0.60	29.20±0.73	-1.59±0.60	32.22±0.79	-0.28±0.65	33.56±0.83	0.26±0.68
τ 95% Coverage	0.61±0.02	0.95±0.01	0.60±0.02	0.96±0.01	0.76±0.02	0.99±0.00	0.82±0.01	0.99±0.00
MATE 95% Coverage	0.41±0.03	0.97±0.01	0.35±0.03	0.96±0.01	0.30±0.03	0.96±0.01	0.32±0.03	0.97±0.01
τ 95% CI Width	63.31±0.28	52.80±0.23	65.46±0.28	54.51±0.22	88.15±0.48	72.80±0.41	102.5±0.55	84.43±0.44
MATE 95% CI Width	50.33±0.28	43.44±0.24	48.56±0.27	42.16±0.22	50.75±0.28	44.13±0.24	54.18±0.33	47.69±0.26
DGP3								
PEHE on τ	13.53±0.34	13.67±0.32	14.00±0.35	14.00±0.33	15.42±0.37	15.17±0.38	16.91±0.40	15.98±0.39
Bias on MATE	0.36±0.76	-0.31±0.73	0.80±0.78	0.54±0.75	0.83±0.82	0.59±0.81	0.30±0.87	0.21±0.85
τ 95% Coverage	0.98±0.00	0.97±0.01	0.98±0.00	0.97±0.01	0.99±0.00	0.99±0.00	1.00±0.00	0.99±0.00
MATE 95% Coverage	0.98±0.01	0.97±0.01	0.97±0.01	0.96±0.01	0.96±0.01	0.96±0.01	0.97±0.01	0.96±0.01
τ 95% CI Width	68.96±0.36	62.86±0.32	70.93±0.35	64.53±0.31	89.26±0.52	83.56±0.49	106.2±0.64	95.82±0.56
MATE 95% CI Width	55.05±0.39	51.99±0.34	53.10±0.38	50.82±0.32	54.81±0.39	53.28±0.35	60.04±0.46	57.80±0.40

Table B.2: Simulation study results for y_1 and y_2 with a training data size of 100. No method can be said to exhibit much better performance than all other methods here, so no results are highlighted in bold.

	Simulation Study Results, Training Sample Size = 1000							
	MVBCF		BCF		BART		MVBART	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
DGP1								
PEHE on τ	**7.66±0.10**	**7.49±0.11**	8.44±0.10	8.12±0.11	9.17±0.12	9.22±0.13	8.59±0.12	8.89±0.12
Bias on MATE	-0.11±0.24	-0.18±0.23	-0.04±0.24	-0.29±0.23	-0.07±0.25	-0.30±0.24	0.11±0.24	0.03±0.24
τ 95% Coverage	**0.95±0.00**	**0.95±0.00**	0.96±0.00	0.97±0.00	0.96±0.00	0.97±0.00	0.96±0.00	0.95±0.00
MATE 95% Coverage	0.94±0.01	0.95±0.01	0.94±0.01	0.95±0.01	0.94±0.01	0.95±0.01	0.95±0.01	0.95±0.01
τ 95% CI Width	**32.00±0.17**	**31.68±0.16**	35.61±0.17	35.56±0.16	41.76±0.30	43.01±0.29	39.90±0.31	39.83±0.30
MATE 95% CI Width	15.89±0.06	15.79±0.06	16.06±0.06	15.94±0.05	16.49±0.06	16.49±0.06	16.81±0.06	16.69±0.06
DGP2								
PEHE on τ	36.72±0.25	**7.50±0.10**	36.90±0.24	7.95±0.11	37.58±0.25	8.74±0.12	37.47±0.25	8.54±0.12
Bias on MATE	35.91±0.25	1.20±0.21	35.95±0.25	1.21±0.21	36.37±0.26	1.44±0.22	36.40±0.26	1.45±0.21
τ 95% Coverage	0.02±0.00	0.92±0.00	0.03±0.00	0.94±0.00	0.10±0.01	0.96±0.00	0.08±0.00	0.94±0.00
MATE 95% Coverage	0.00±0.00	0.93±0.02	0.00±0.00	0.94±0.01	0.00±0.00	0.93±0.02	0.00±0.00	0.92±0.02
τ 95% CI Width	**31.27±0.18**	**27.31±0.14**	34.68±0.18	30.95±0.15	41.71±0.30	38.40±0.26	39.81±0.32	35.53±0.28
MATE 95% CI Width	14.77±0.06	12.98±0.05	14.73±0.05	13.01±0.04	14.87±0.05	13.19±0.04	14.95±0.05	13.24±0.05
DGP3								
PEHE on τ	8.67±0.11	8.26±0.12	8.60±0.10	**8.04±0.12**	9.55±0.15	9.22±0.14	9.63±0.15	9.13±0.13
Bias on MATE	-0.20±0.26	-0.37±0.25	-0.25±0.26	-0.29±0.25	-0.20±0.27	-0.46±0.26	0.10±0.27	-0.13±0.26
τ 95% Coverage	**0.95±0.00**	**0.94±0.00**	0.97±0.00	0.97±0.00	0.97±0.00	0.97±0.00	0.96±0.00	0.95±0.00
MATE 95% Coverage	0.94±0.02	0.94±0.02	0.95±0.01	0.94±0.01	0.95±0.01	0.94±0.01	0.96±0.01	0.95±0.01
τ 95% CI Width	**34.56±0.21**	**31.97±0.18**	37.38±0.20	35.49±0.17	43.64±0.34	42.85±0.30	42.90±0.37	40.68±0.34
MATE 95% CI Width	15.97±0.07	15.97±0.07	16.00±0.06	15.88±0.06	16.63±0.07	16.47±0.06	16.99±0.08	16.82±0.07

Table B.3: Simulation study results for y_1 and y_2 with a training data size of 1000. Best results in bold where a clear winner exists. MVBCF performs very strongly again. All methods perform better with the larger sample size.

C

Appendix for Chapter 5

C.1 Technical Details

C.1.1 Mathematical Description of BART, BCF, and MVBCF

BART

Given an outcome variable y of length n, and a covariate matrix X consisting of n observations of d variables, the BART model (Chipman et al., 2010) can be written as follows:

$$y_i = \sum_{j=1}^{J} g(T_j, M_j, x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

In the equation above, the function $g()$ calculates the individual contribution of each tree, T_j, out of a total of J trees. The parameters M_j represent the terminal nodes associated with the j-th tree, T_j. The residuals, ϵ_i, are assumed to follow a normal distribution with a mean of 0 and a variance of σ^2. Since the BART model is Bayesian, appropriate priors need to be specified for T_j, M_j, and σ^2.

BCF

BCF (Hahn et al., 2020) expresses the outcome y as:

$$y_i = \mu(x_i, \hat{\pi}_i) + \tau(x_i) Z_i + \epsilon_i$$

In the equation above, $\mu()$ and $\tau()$ are both BART ensembles that work together to estimate two distinct components of y: a prognostic effect, μ, which represents the expected outcome under control when the treatment variable Z is coded as 1 for treatment and 0 for control, and a

treatment effect, τ, which indicates the impact on y resulting from receiving the treatment. The additional covariate, $\hat{\pi}_i$, included in the $\mu()$ part of the model, is called the propensity score. This propensity score, denoted as $\hat{\pi}_i = P(Z_i = 1)$, represents the estimated probability of individual i receiving treatment. The propensity score can be estimated using logistic regression, BART, or any other appropriate classification model capable of providing estimated probabilities.

Multivariate BCF

In the case of multiple outcomes, the MVBCF model (McJames et al., 2024, Chapter 4) specification becomes:

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\tau}_i \circ \boldsymbol{Z}_i + \boldsymbol{\epsilon}_i$$

In the equation above, $\boldsymbol{Y}_i$ represents a vector of length p, denoting the i^{th} observation of the p-dimensional outcome variable $\boldsymbol{Y}$. $\boldsymbol{\mu}_i$ and $\boldsymbol{\tau}_i$ correspond to the i^{th} predictions from the functions $\boldsymbol{\mu}(x_i)$ and $\boldsymbol{\tau}(x_i)$, respectively. $\boldsymbol{\epsilon}_i$ represents the i^{th} residual, and $\circ$ denotes the Hadamard product operator. This formulation can be extended to incorporate multiple treatment groups, such as different frequencies or durations of homework, by including additional $\boldsymbol{\tau}$ components in the model.

C.1.2 Log-Likelihood and Posterior Distribution of MVBCF Model Parameters

Log-Likelihood of a μ Tree

Let T_j denote the j^{th} μ tree in the ensemble with partial residuals R_j. Also suppose that tree T_j has K terminal nodes $h_{j,1}...h_{j,K}$, and L non-terminal nodes $b_{j,1}...b_{j,L}$. Furthermore, let $R_{j,k,1}...R_{j,k,n_k}$ denote the partial residuals which fall into the k^{th} terminal node of tree T_j. Then given the residual covariance matrix Σ, the tree priors α and β, and the prior covariance matrix Σ_μ for terminal node parameters, we have that:

$$\ell(T_j|R_j,\Sigma) = \ell(R_j|T_j,\Sigma) + \ell(T_j) + C$$

$$\ell(T_j) = \sum_{k=1}^{K} \log\left(1-\alpha(1+d(h_{j,k}))^{-\beta}\right) + \sum_{l=1}^{L}\left[\log(\alpha) - \beta\log(1+d(b_{j,l}))\right]$$

$$\ell(R_j|T_j,\Sigma) = \sum_{k=1}^{K} \ell(R_{j,k,1},\ldots,R_{j,k,n_k}|T_j,\Sigma)$$

$$= \sum_{k=1}^{K}\left\{-\frac{n_k}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\log\left(|\Sigma_\mu|\right) + \frac{1}{2}\log\left(|\Sigma_{k,0}|\right) - \frac{1}{2}\sum_{i=1}^{n_k}\left(R_{j,k,i}^T\Sigma^{-1}R_{j,k,i} - \mu_{k,0}^T\Sigma_{k,0}^{-1}\mu_{k,0}\right)\right\} + C$$

where

$$\Sigma_{k,0}^{-1} = n_k\Sigma^{-1} + \Sigma_\mu^{-1}$$

and

$$\mu_{k,0} = \Sigma_{k,0}\Sigma^{-1}\left(\sum_{i=1}^{n_k} R_{j,k,i}\right)$$

Posterior Distribution of Terminal Node Parameters in a μ Tree

For the k^{th} terminal node of any tree T_j, the posterior distribution of the terminal node parameter $\mu_{j,k}$ with prior mean μ_0 is given by:

$$\mu_{j,k}|\ldots \sim N\left(\mu_n,\Sigma_n\right)$$

where

$$\mu_n = \left(\Sigma_\mu^{-1} + n_k\Sigma^{-1}\right)^{-1}\left(\Sigma_\mu^{-1}\mu_0 + n_k\Sigma^{-1}\overline{R}\right)$$

and

$$\Sigma_n = \left(\Sigma_\mu^{-1} + n_k\Sigma^{-1}\right)^{-1}$$

Log-Likelihood of a τ Tree

Given the prior covariance matrix for terminal node parameters Σ_τ, and $n_k \times p$ matrix Z_k which holds the treatment status of each observation in terminal node k, who's transposed i^{th} row we denote by $Z_{k,i}$, we obtain:

$$\ell(T_j|R_j,\Sigma) = \ell(R_j|T_j,\Sigma) + \ell(T_j) + C$$

$$\ell(T_j) = \sum_{k=1}^{K} \log\left(1-\alpha(1+d(h_{j,k}))^{-\beta}\right) + \sum_{l=1}^{L} \log(\alpha) - \beta\log(1+d(b_{j,l}))$$

$$\ell(R_j|T_j,\Sigma) = \sum_{k=1}^{K} \ell(R_{j,k,1},\ldots,R_{j,k,n_k}|T_j,\Sigma)$$

$$= \sum_{k=1}^{K}\left\{-\frac{n_k}{2}\log\left(|\Sigma|\right) - \frac{1}{2}\log\left(|\Sigma_\tau|\right) + \frac{1}{2}\log\left(|\Sigma_{k,0}|\right) - \frac{1}{2}\sum_{i=1}^{n_k}\left(R_{j,k,i}^T\Sigma^{-1}R_{j,k,i} - \tau_{k,0}^T\Sigma_{k,0}^{-1}\tau_{k,0}\right)\right\} + C$$

where

$$\Sigma_{k,0}^{-1} = Z_{j,k}^T Z_{j,k}\Sigma^{-1} + \Sigma_\tau^{-1}$$

and

$$\tau_{k,0} = \Sigma_{k,0}\sum_{i=1}^{n_k} Z_{j,k,i} \circ \Sigma^{-1}R_{j,k,i}$$

Posterior Distribution of Terminal Node Parameters in a τ Tree

Analogously to the terminal node parameters in a μ tree, the posterior distribution of the terminal node parameter $\tau_{j,k}$ in the k^{th} leaf of the j^{th} τ tree T_j, with prior mean τ_0 is:

$$\tau_{j,k}|\ldots \sim N\left(\tau_n, \Sigma_n\right)$$

where

$$\tau_n = \left(\Sigma_\tau^{-1} + Z_{j,k}^T Z_{j,k} \circ \Sigma^{-1}\right)^{-1}\left(\Sigma_\tau^{-1}\tau_0 + \sum_{i=1}^{n_k} Z_{j,k,i} \circ \Sigma^{-1}R_{j,k,i}\right)$$

and

$$\Sigma_n = \left(\Sigma_\tau^{-1} + Z_{j,k}^T Z_{j,k} \circ \Sigma^{-1}\right)^{-1}$$

Posterior Distribution of Residual Covariance Parameter Σ

Given observed and predicted values y and $\hat{y}$, the posterior distribution of the covariance matrix Σ for the n residuals, with prior scale matrix Σ_0 and ν_0 degrees of freedom is given by:

$$\Sigma|\ldots \sim \mathcal{W}^{-1}\left(\nu_0 + n, [S_0 + S_\theta]^{-1}\right)$$

where

$$S_\theta = \sum_{i=1}^{n}(y_i - \hat{y_i})^T \Sigma^{-1}(y_i - \hat{y_i}).$$

C.1.3 Computation of Results

The model in the paper uses multivariate BART and BCF as a foundation to estimate the effect of different levels of homework frequency and duration as follows:

$$y_{i,j} = \mu_j(x_i) + \tau_{j,1}(x_i)Z_{i,j,1} + \tau_{j,2}(x_i)Z_{i,j,2} + \tau_{j,3}(x_i)Z_{i,j,3} + \tau_{j,4}(x_i)Z_{i,j,4} + \alpha_{class,i,j} + \epsilon_{i,j}$$

where $y_{i,j}$ is the achievement of student i in subject j ($j = 1$ for mathematics or $j = 2$ for science). For a student who receives homework up to one or two times per week with a duration of up to fifteen minutes, their achievement in subject j is given by $\mu_j(x_i)$, where x_i denotes the characteristics associated with student i. Students who receive homework with a greater frequency or duration belong to the treatment groups $Z_{i,j,1} \ldots Z_{i,j,4}$:

- Frequency of three or four times per week $\rightarrow Z_{i,j,1} = 1$,
- Frequency of every day $\rightarrow Z_{i,j,2} = 1$,
- Duration of fifteen to thirty minutes $\rightarrow Z_{i,j,3} = 1$,
- Duration of greater than thirty minutes $\rightarrow Z_{i,j,4} = 1$.

The causal effect of belonging to these groups is given by $\tau_{j,k}(x_i)$, where $k = 1 \ldots 4$ as above. To account for the hierarchical nature of the data, $\alpha_{class,i,j}$ is a random intercept which captures the between classroom variation in the data for student i and subject j. The remaining variation is represented by the error term for student i in subject j, denoted as $\epsilon_{i,j}$.

For the random intercept term, we assume that each $\boldsymbol{\alpha}_{class}$ comes from a multivariate normal distribution with a population mean $\boldsymbol{\mu}_\alpha$ and population covariance matrix $\boldsymbol{\Sigma}_\alpha$: $\boldsymbol{\alpha}_{class} \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha)$, where the prior on $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\Sigma}_\alpha$ is: $\boldsymbol{\mu}_\alpha \sim N(\boldsymbol{m} = \mathbf{0}, \boldsymbol{s} = 0.01\boldsymbol{I})$, and $\boldsymbol{\Sigma}_\alpha \sim \mathcal{W}^{-1}(a = 1, \boldsymbol{\Omega_0} = 0.1\boldsymbol{I})$.

In the μ part of the model, 50 trees were used to estimate the expected achievement of students in mathematics and science if they only receive homework up to one or two times per week, at a duration of up to fifteen minutes each time. In each of the τ parts of the model

which are responsible for estimating the treatment effects of receiving homework with a greater frequency and/or duration, 20 trees were used. We ran a total of 5000 iterations of a statistical method called Markov Chain Monte Carlo (MCMC), which allowed our model to gradually update its estimates in a step by step manner. In line with best practice, we discarded the results from the first 3000 of these iterations, as these were considered "burn-in" samples.

(1) Five MVBCF models were fitted to the TIMSS data, each corresponding to one of the five available plausible values for student achievement: Model one used mathematics plausible values from BSMMAT01, and science plausible values from BSSSCI01, while Model 2 used BSMMAT02 and BSSSCI02 etc.

(2) Each model used 50 trees for the μ component of the model, and 20 trees in each of the τ treatment effect ensembles.

(3) All models used the same treatment indicator Z, to indicate if student i received treatment k in subject j.

(4) Each model ran for 3000 burn in iterations and 2000 post burn in iterations. Thinning was used to save memory, so every second post burn in sample was saved, leading to 1000 posterior samples for each parameter of interest.

(5) Upon collection of the posterior samples from each of the five models, the posterior samples from all five models were pooled together to reflect the uncertainty in the plausible values of student achievement.

(6) Satisfactory convergence was assessed via visual inspection of the resulting MCMC chains, and metrics including RMSE were calculated to ensure model accuracy.

C.1.4 Variance Explained By Model

The table below summarises the total variance explained by the model, how much of this explained variance is due to the class specific random intercepts, and the Intraclass Correlation Coefficients (ICCs) for both mathematics and science. The residual variance for each subject is the posterior mean from the relevant diagonal entry of the residual covariance matrix Σ. The explained variance refers to the variance of the predictive mean values of mathematics and science achievement, and the random intercept variance refers to the posterior mean of the relevant diagonal entry from Σ_α, the covariance matrix of the class effects. The R^2 for mathematics is 0.636, indicating that 63.6% of the variation in mathematics achievement is explained by the model. Similarly, the R^2 for science indicates the model has explained 61.7% of the variation in science achievement. The ICC measures what proportion of the variation explained by the model that is attributable to the class specific random intercept terms. In mathematics, the ICC is 0.230, indicating that 23.0% of the variation explained by the model is due to these random intercept terms, while in science the figure is 21.8%.

Variance Explained By Model

Variance/Subject	Mathematics	Science
Residual Variance (σ^2)	1831.2	2515.6
Residual Standard Deviation (σ)	42.8	50.2
Explained Variance ($\sigma^2_{\hat{y}}$)	3202.7	4055.3
Standard Deviation of Predictive Means ($\sigma_{\hat{y}}$)	56.6	63.7
Random Intercept Variance (σ^2_α)	547.8	702.4
Random Intercept Standard Deviation (σ_α)	23.4	26.5
$R^2 = \frac{\text{Explained Variance}}{\text{Explained Variance + Residual Variance}}$	0.636	0.617
$\text{ICC} = \frac{\text{Random Intercept Variance}}{\text{Random Intercept Variance + Residual Variance}}$	0.230	0.218

Table C.1: A summary of the variance explained by model. The model explains 63.6% of the variation in mathematics achievement, and 61.7% of the variation in science achievement. Of this variation, 23.0% is attributable to the classroom specific random effects in mathematics, while in science the classroom specific random effects account for 21.8% of this variation.

C.2 Complete Case Version of Results

As mentioned in the main paper, an alternative to imputing missing data is to use the complete cases, removing rows with missing data. Our preference in the main paper was to use the imputation based approach which maintained a nationally representative dataset. The results below show the estimated effects of homework frequency and duration using complete cases only. Once again, the optimal frequency in mathematics appears to be every day (albeit with some greater uncertainty), and the optimal frequency in science is identified as three or four times per week. Similarly to the imputation based approach, increasing homework duration beyond 15 minutes each time is not identified as yielding significant improvements.

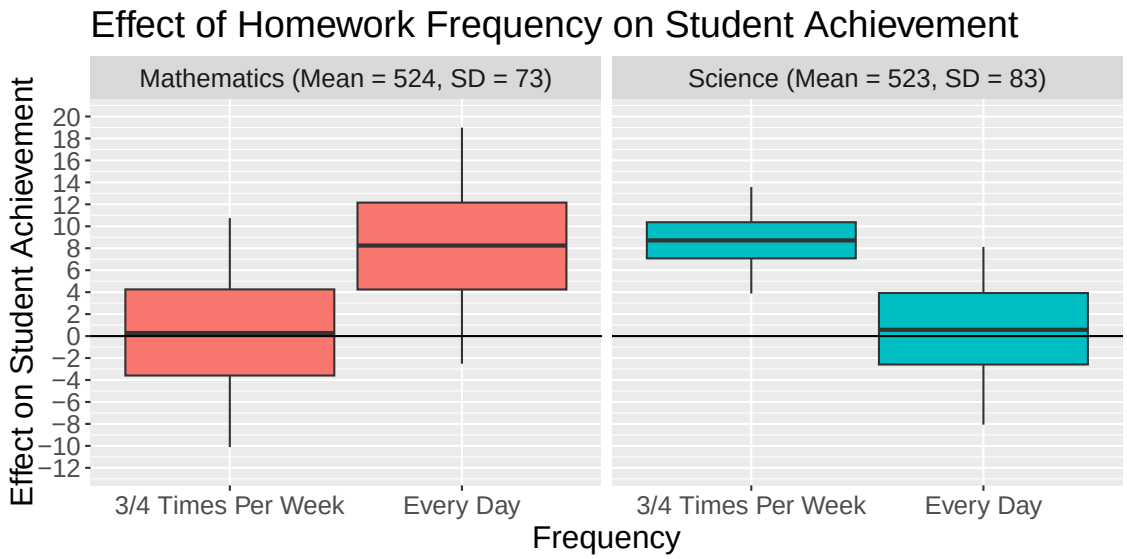

Figure C.1: Effect of homework frequency on student achievement using the complete cases of the dataset only. Every day is the best frequency in mathematics, but with more uncertainty than the results from the imputation based approach. Once again, three or four times per week is the best frequency in science.

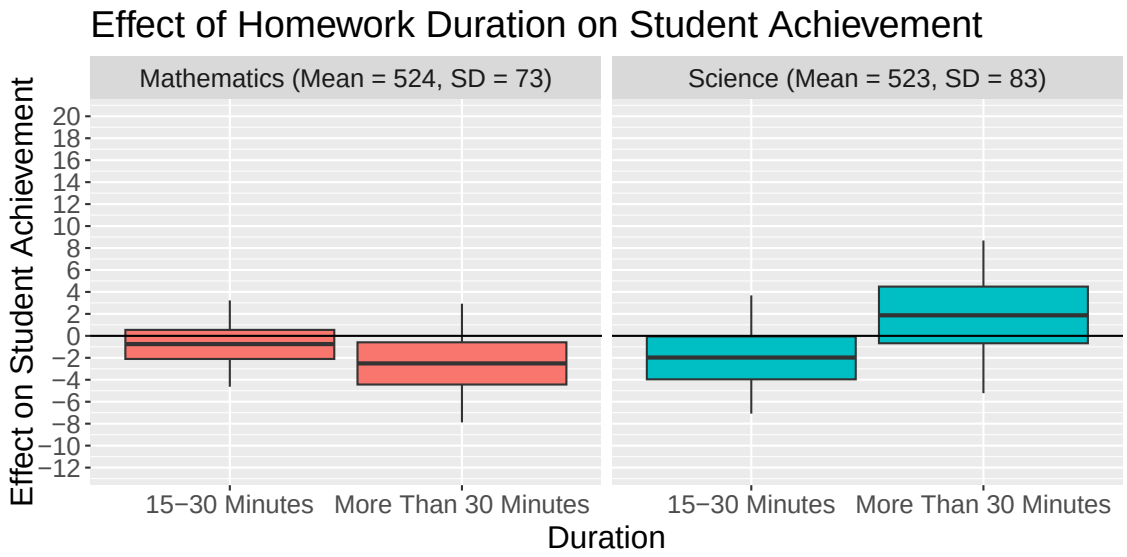

Figure C.2: Effect of homework duration on student achievement using the complete cases of the dataset only. The results show that increasing the duration of homework assignments beyond 15 minutes does not yield significant improvements in achievement. This agrees with the results from the imputation based approach.

C.3 TIMSS Variables Used

TIMSS Variables Used		
Variable Code	Obtained From	Description
BSDAGE	Student Questionnaire	Student Age
BSBG01	Student Questionnaire	Student Gender
BSBG03	Student Questionnaire	How often student speaks English at home
BSBG04	Student Questionnaire	Number of books at home
BSBG07	Student Questionnaire	How far in education student expects to go
BSBG08A	Student Questionnaire	Was parent/guardian A born in Ireland
BSBG08B	Student Questionnaire	Was parent/guardian B born in Ireland
BSBG09A	Student Questionnaire	Was student born in Ireland
BSBG10	Student Questionnaire	How often student is absent
BSBG11A	Student Questionnaire	How often student feels hungry when arriving at school
BSBG11B	Student Questionnaire	How often student feels tired when arriving at school
BSDGEDUP	Student Questionnaire	Parent's highest education level
BSBGHER	Student Questionnaire	Number of home educational resources
BSBGSSB	Student Questionnaire	Sense of school belonging
BSBGSB	Student Questionnaire	School bullying
BSBGSCM/BSBGSCS	Student Questionnaire	Confidence in mathematics/science
BSBGSVM/BSBGSVS	Student Questionnaire	Student values mathematics/science
BSBGICM/BSBGICS	Student Questionnaire	Instructional clarity in mathematics/science
BSBG05A	Student Questionnaire	Has computer/tablet at home
BSBG05B	Student Questionnaire	Has study desk at home
BSBG05C	Student Questionnaire	Has own bedroom
BSBG05D	Student Questionnaire	Has home internet connection
BSBG05E	Student Questionnaire	Has own mobile phone
BSBG05F	Student Questionnaire	Has gaming system
BSBG05G	Student Questionnaire	Home TV has "premium" TV channels
BTBG01	Teacher Questionnaire	Number of years teaching
BTBG02	Teacher Questionnaire	Teacher gender
BTBG03	Teacher Questionnaire	Teacher age
BTBG10	Teacher Questionnaire	Number of students in class
BTBM14/BTBS14	Teacher Questionnaire	Instructional time with class per week
BTBGTJS	Teacher Questionnaire	Teacher job satisfaction
BTBGSOS	Teacher Questionnaire	Safe and orderly school
BTBGLSN	Teacher Questionnaire	Teaching is limited by students not ready for instruction
BTBGEAS	Teacher Questionnaire	Emphasis on academic success
BTDMME	Teacher Questionnaire	Type of degree
BCBGDAS	Principal Questionnaire	School discipline
BCBGEAS	Principal Questionnaire	Emphasis on academic success
BCBGMRS/BCBGSRS	Principal Questionnaire	Resource shortages in mathematics/science
BCDGSBC	Principal Questionnaire	School average socioeconomic background

Table C.2: Variable codes of potential confounders controlled for as part of the study. All variables listed were used in both the μ and τ parts of the multivariate BCF model.

D

Appendix for Chapter 6

D.1 Table of Summary Statistics

Variable	Proportion Wave 1	Achievement Wave 1	Achievement Wave 2
Student Gender			
Male	50.3%	-0.08	0.63
Female	49.7%	-0.06	0.60
First Language			
English Only	82.3%	-0.05	0.63
English and Other	6.1%	-0.15	0.58
Other	11.5%	-0.16	0.58
Family Setup			
Live With Both Biological Parents	43.2%	0.21	0.93
Other Arrangement	32.9%	0.02	0.71
No Response	23.9%	-0.38	0.25
Parent Education Level			
Less Than Bachelor's Degree	47.7%	-0.23	0.39
Bachelor's Degree or Higher	28.4%	0.45	1.23
No Response	23.9%	-0.38	0.25
Student Future Expectations			
Less Than Bachelor's Degree	21.5%	-0.56	0.05
Bachelor's Degree or Higher	56.8%	0.18	0.89
Student Not Sure	21.7%	-0.25	0.41
School Type			
Public	92.8%	-0.11	0.57
Catholic or Private	7.2%	0.35	1.21

Table D.1: Summary statistics for categorical variables. The proportion column provides the proportion of students belonging to each category in Wave 1, while the achievement columns provide the mean achievement level within each group.

D.2 LBCF Diagram

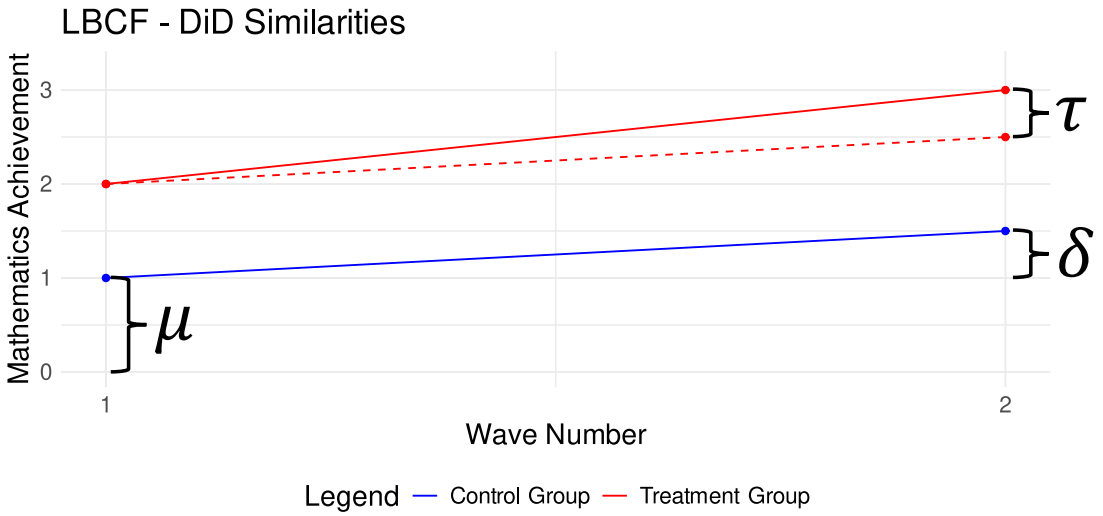

Figure D.1: Diagram of how the proposed LBCF model fits into the framework of the difference-in-differences approach. Two observations are shown for the purposes of illustration - one from an imaginary control group, and one from a corresponding treatment group. The solid lines indicate the realised achievement trajectories, while the dashed line in red indicates a counterfactual trajectory for the treated unit had it actually not received treatment. Initial achievement estimates at Wave 1 are provided by μ. The expected growth (difference) in achievement without treatment is provided by δ, while the effect of treatment on this growth (the difference-in-differences) is captured by τ. Note that while only one μ value is indicated in the diagram to avoid overprinting, the model does in fact provide individual μ estimates for every observation. Similarly, individual estimates are provided for each of the δ and τ estimates as well.

D.3 LBCF Algorithm

Algorithm 2 LBCF MCMC Algorithm

Require: Outcome variable $y_{i,t}$ (response for individual i at time t of T time periods), Time varying covariates $x_{i,t}$ (covariates collected on individual i up to time t), Treatment variable $Z_{i,t+1}$ (to indicate if individual i received treatment between periods t and t+1: 1 for treatment, 0 for control)

Ensure: Posterior list of trees, values of σ^2, fitted values $\hat{\mu}_i$, $\hat{\delta}_{i,t}$, and $\hat{\tau}_{i,t}$

Initialise hyper-parameter values of α_μ, β_μ, α_δ, β_δ, α_τ, β_τ, σ^2_μ, σ^2_δ, σ^2_τ, ν, λ, Number of μ trees n_μ, Number of δ trees n_δ, Number of τ trees n_τ, Number of iterations N, Initial value $\sigma^2 = 1$, Set μ trees T_j; $j = 1, \ldots, n_\mu$ to stumps, Set δ trees T_j; $j = 1, \ldots, n_\delta$ to stumps, Set τ trees T_j; $j = 1, \ldots, n_\tau$ to stumps, Set terminal node parameters of all μ, δ, and τ trees to 0

for iterations i from 1 to N **do**

for μ trees j from 1 to n_μ **do** Compute partial residuals from y minus predictions of all trees except μ tree j, Grow a new tree T_j^{new} based on grow/prune/change/swap, Accept/Reject tree structure with Metropolis-Hastings step using $P(T_{\mu,j}|R_{\mu,j}, \sigma^2) \propto P(T_{\mu,j})P(R_{\mu,j}|T_{\mu,j}, \sigma^2)$, Sample μ values from normal distribution using $P(M_{\mu,j}|T_{\mu,j}, R_{\mu,j}, \sigma^2)$

end for

for Time periods t from 2 to T **do**

for δ_t trees j from 1 to n_δ **do** Compute partial residuals from y minus predictions of all trees except δ_t tree j, Grow a new tree T_j^{new} based on grow/prune/change/swap, Accept/Reject tree structure with Metropolis-Hastings step using $P(T_{\delta_t,j}|R_{\delta_t,j}, \sigma^2) \propto P(T_{\delta_t,j})P(R_{\delta_t,j}|T_{\delta_t,j}, \sigma^2)$, Sample δ_t values from normal distribution using $P(M_{\delta_t,j}|T_{\delta_t,j}, R_{\delta_t,j}, \sigma^2)$

end for

for τ_t trees j from 1 to n_τ **do** Compute partial residuals from y minus predictions of all trees except τ_t tree j Grow a new tree T_j^{new} based on grow/prune/change/swap Accept/Reject tree structure with Metropolis-Hastings step using $P(T_{\tau_t,j}|R_{\tau_t,j}, \sigma^2) \propto P(T_{\tau_t,j})P(R_{\tau_t,j}|T_{\tau_t,j}, \sigma^2)$ Sample τ_t values from normal distribution using $P(M_{\tau_t,j}|T_{\tau_t,j}, R_{\tau_t,j}, \sigma^2)$

end for

end for

Get predictions $\hat{y}$ from all trees, Update σ^2 with Inverse-Gamma distribution using $P(\sigma^2|\hat{y})$

end for