Essays on Family, Education, and Public Health Policies

Grant Nippler

Presented in fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Economics
University of Strathclyde
Glasgow, UK

December 2024

Declaration

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree. The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis.

Signed:

Date: December 23, 2024

Statement of Co-Authorship

Chapter 3 and chapter 4 are works which involved the contribution of co-authors. Chapter 3, “The Effects of School Discipline Reform: Insights from Oklahoma City,” was coauthored with Noah Spencer at the University of Toronto. He contributed to the development of the paper by conceptualizing the central idea, sourcing secondary data, and conducting preliminary coding and analyses. Chapter 4, “Supervised Consumption Sites in Toronto: Did Harm Reduction Work?” was coauthored with my supervisor and colleague Dr. Agnese Romiti.

Acknowledgements

I would like to begin by thanking my primary supervisor, Dr. Agnese Romiti for her guidance, support, and mentorship throughout my PhD studies. I am deeply indebted to you for the time, care, and encouragement you devoted to nurturing my growth as a researcher, without which this thesis would not have been possible. I look forward to exploring new and interesting ideas together in the future. Special thanks to the Institute for Inspiring Children's Futures for their financial support of this thesis and to Professor Jennifer Davidson for fostering connections among the IICF scholars. The insights from our interdisciplinary space have been invaluable, constantly inspiring me with new ideas, and I feel very fortunate to be connected to the IICF.

I would also like to thank the wider Economics Department at Strathclyde, with its continuously open doors, for fostering a welcoming environment where students feel like part of the department. The staff's willingness to listen, attend seminars, and provide thoughtful feedback has been invaluable. I would like to acknowledge PGR Directors Julia Darby, Alex Dickson, and Agnese Romiti for fostering and maintaining a culture of collaboration and engagement among the PhD students. Through this supportive environment, I have been fortunate to build not only fruitful and enriching professional connections but also lifelong friendships. Thank you PhD colleagues, Annie, Arnold, Baland, Ben, Beth, Césarine, Geoff, Iswat, Jon, Joe, Kat, Lateef, Rory, Sabin, Sam, and Zhan for being an integral part of my journey.

I would also like to thank the Mairi Spowage and the Fraser of Allander institute for providing me with the opportunity to gain valuable practical research experience during my master's and PhD by working on exciting and engaging projects. Additional thanks go to my FAI colleagues for their valuable feedback on my work and for being so kind and welcoming, creating an environment where I felt supported

and encouraged throughout my journey. Additional thanks to Professor Graham Roy, for meeting with me at crucial points throughout my PhD and for your professional mentorship.

I would also like to acknowledge my undergraduate institution, St. John's University, for the foundational education and support that shaped my early academic journey and set me on the path toward pursuing a PhD. Special thanks to Dr. Aleksandr Gevorkyan, Dr. Gary Mongiovi, and Dr. Kirstin Munro for opening my eyes to new perspectives and inspiring my passion for research and learning.

Finally, a heartfelt thank you to my mother, Mimi Tanner; my father, John Nippler; my stepfathers, Jon Harmon and Tom Haydukovich; and the wider Wolford and Nippler families for their love and support throughout this process. To my brother, Logan, for keeping me sane with his comedic relief, and to my partner, Zoë McClellan, whose strength uplifted me countless times and whose unwavering support made this journey possible, thank you.

Abstract

This thesis consists of three self-contained essays in applied microeconomics.

Chapter two investigates the effects of UK austerity measures on individuals and families using data from the UK Household Longitudinal Study. By applying difference-in-differences and event study models, I find that austerity significantly increased parental employment and reduced unemployment. However, we present suggestive evidence showing that mothers may have experienced a 20 percentage point drop in income, amounting to £182 per month, while fathers' incomes were unaffected. These gendered income changes shaped relationship dynamics, with mothers showing a decline in divorce rates and an increase in new cohabitations, patterns not observed among fathers. Austerity also reduced parent-child interaction, as time was reallocated to meet rising labor demands. By examining the effects of benefit reduction policies, this paper adds to the literature by demonstrating their far-reaching implications for labor markets, family relationships, and inner-household relationships.

Chapter three examines the 2016 school reform policies implemented by the Oklahoma City School District, which sought to replace exclusionary discipline practices with more inclusive approaches. Using a Synthetic Difference-in-Differences model, I analyse school administrative records alongside FBI crime data. The reforms resulted in a 50 percent drop in school suspensions, improvements in math and science scores, mixed effects on reading performance, and a 22.8 percent reduction in youth arrests. These findings highlight the trade-offs between academic outcomes and broader youth impacts, showing how inclusive policies can shape both classroom and community dynamics. This study contributes to the literature by providing evidence on the short-term effects of discipline reforms on student achievement and juvenile justice outcomes.

Chapter four estimates the impact of supervised consumption sites (SCS) on opioid-related emergency callouts, crime, and mental health incidents, using neighborhood-level data from the Toronto Police. By a difference-in-differences approach, I analyse the staggered implementation of 10 Health Canada-approved SCS in Toronto between 2017 and 2021. The results show no significant changes in overdose callouts, assault rates, or mental health apprehensions, though a rise in break-ins near SCS locations is observed. These findings suggest that SCS do not substantially reduce emergency service demands or mental health issues but may bring localised property crime concerns. This study contributes to understanding the trade-offs of SCS and emphasizes

the importance of embedding them in broader public health strategies that address community-level impacts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Public policies aimed at addressing pressing social and economic challenges often yield multifaceted and far-reaching impacts. From austerity measures reshaping household dynamics to school discipline reforms influencing youth behavior, and harm reduction initiatives tackling public health crises, the ripple effects of such interventions extend well beyond their intended scope. This thesis examines three critical policy areas, economic austerity, educational reform, and harm reduction, through a series of empirical studies that highlight the interplay between policy interventions and individual, familial, and community outcomes.

Chapter 2 investigates the consequences of austerity policies implemented in the United Kingdom during the 2010s. I conduct a comprehensive analysis of the effects of UK austerity measures on individuals and families, drawing on data from the UK Household Longitudinal Study (UKHLS). Using an event study model, I show that austerity policies implemented in the 2010s led to significant increases in parental employment, with both mothers and fathers intensifying their participation in the labor market. This increase in labor supply was accompanied by a corresponding reduction in unemployment. However, while more parents entered or remained in the workforce, the financial outcomes differed significantly by gender. We find some suggestive evidence showing that Mothers experienced reductions in total income, corresponding to a 20% percentage point decrease or approximately £182 per month due to cuts in welfare benefits, particularly the Council Tax Benefit and Housing Benefit. Though we note that these particular results were sensitive to alternative inferencing procedures namely, Rambachan and Roth's (2023) test for parallel trends. In contrast, we found no evidence of men facing income reductions, maintaining their pre-austerity total income levels.

This suggested gender-specific impact on income could have had an impact on relationship dynamics. For mothers, the reduction in total income was associated with a 5 percentage point decrease in divorce rates and a 10 percentage point increase in the frequency of entering new relationships or cohabiting. These trends were not observed for men, whose relationship statuses remained largely unaffected by the austerity measures. The divergent effects suggest that economic pressures reshaped household decision-making in ways that disproportionately affected women.

Furthermore, austerity led to a decline in parental interaction with children. As parents reallocated their time to meet increased labor demands, household activities such as shared meals and general family time were deprioritized. This reduction in parental engagement, particularly in the context of family meals, may have longer-term implications for child development, especially for boys, who tend to be more sensitive to changes in household dynamics.

Overall, this chapter highlights the far-reaching and gendered impacts of austerity policies, demonstrating how fiscal tightening can reshape not only economic conditions but also family structures and interpersonal relationships. By offering a detailed analysis of labor market responses, gendered income effects, and their cascading influence on household dynamics, this chapter contributes to the broader literature on austerity by uncovering the nuanced ways in which fiscal policies interact with social and relational behaviors. Furthermore, it provides policymakers with critical insights into the unintended consequences of economic reforms, particularly their disparate impacts across gender lines, which can inform the design of more equitable welfare interventions.

Chapter 3 examines the effects of the 2016 Oklahoma City School District (OKCSD) reforms on disciplinary practices, academic performance, and adolescent arrests. The reforms, which included the introduction of Positive Behavioral Interventions and Supports (PBIS) and comprehensive staff retraining, were aimed at reducing exclusionary discipline measures, such as suspensions, and creating a more inclusive school environment. To assess the impact, we apply a Synthetic Difference-in-Differences (SDID) methodology, leveraging K-means clustering to select a well-matched control group for comparison.

The analysis reveals a substantial 38% reduction in school suspensions, suggesting that OKCSD took considerable steps towards limiting the pervasiveness of exclusionary discipline policies. In tandem, we find evidence that OKC increased its counseling staff by 17%, which may have contributed to the decrease in suspensions by providing additional student support. However, the academic effects were mixed. The OKCSD policy reforms produced mixed academic outcomes, with notable declines in 6th-grade reading performance (-6.51 points, or -0.276 standard deviations) but significant improvements in structured subjects like math and science. Specifically, 7th-grade math scores increased by 7.31 points (0.261 standard deviations), a 52% improvement, while 8th-grade science scores rose by 7.01 points (0.320 standard deviations), representing a 61% gain. These results suggest that while the reforms introduced challenges in less structured subjects such as reading, they enhanced student outcomes in structured disciplines, likely due to improved classroom management and inclusivity. Potential mechanisms include gendered classroom dynamics and changes in teacher and peer effects, which merit further investigation to better understand these heterogeneous impacts.

These findings underscore the importance of inclusive educational policies in reducing disciplinary actions and promoting better social outcomes, though careful attention is needed to address the varied effects on student performance across age groups. This chapter makes two key contributions. It provides the first empirical insights into how school discipline reforms influence juvenile outcomes outside of school, such as arrests, and introduces an innovative methodology that leverages K-means clustering to enhance Synthetic Difference-in-Differences (SDID) estimation for large datasets. These advancements offer valuable tools and perspectives for policymakers and researchers addressing the multifaceted impacts of educational reforms.

Chapter 4 analyses the staggered rollout of opioid harm reduction facilities in Canada. The opioid crisis continues to pose significant public health challenges across the globe and particularly in large urban centers like Toronto, where opioid-related deaths and overdose incidents have surged. In response, policymakers have implemented harm reduction interventions, including Supervised Consumption Sites (SCS), to mitigate the adverse health effects of opioid use. However, the overall impact of these sites on surrounding communities remains a point of contention. This study aims to provide empirical evidence on the effectiveness of SCS in reducing overdose callouts, crime, and mental health crises in Toronto neighborhoods.

We employ a spatial Difference-in-Differences (DID) approach, relying on the framework developed by Callaway and Sant'Anna (2021), to evaluate the impact of the staggered rollout of 11 SCS across Toronto from 2017 to 2021. We merge our unique primary SCS data which includes geographic information, site openings and site closures with three datasets from Toronto Police's open data portal to assess the effect of SCS on emergency callouts related to drug overdoses, reports of assault and breaking and entering, and mental health apprehensions Our findings suggest that SCS do not significantly influence overdose-related callouts or crime rates in the areas surrounding these facilities. Specifically, we find no statistically significant changes in the frequency of overdose callouts or incidents of assault. We find some evidence indicative of an increase in the frequency of break-ins in neighborhoods close to SCS. Additionally, there is no evidence to suggest that the presence of SCS exacerbates mental health crises, as indicated by the absence of changes in mental health apprehensions following the opening of these sites.

These results contribute to the ongoing debate about the effectiveness of harm reduction policies. While SCS do not appear to worsen public safety or mental health conditions, their ability to alleviate the burdens on emergency services and crime prevention efforts seems limited. This chapter provides the first causal evidence on the community-level effects of supervised consumption sites (SCS) in a major urban setting, using a robust spatial Difference-in-Differences (DID) framework to assess their impact on public safety and health outcomes. These findings offer valuable insights for policymakers seeking to understand the role of SCS within broader harm reduction strategies.

Chapter 2

Cutting Benefits, Changing Lives: Austerity's Impact on Work, Income, and Families

2.1 Introduction

While there is a breadth of literature outlining the effect of changes in government benefit expenditures on employment and income (Eissa and Hoynes, 2006; Farber, Rothstein, and Valletta, 2015), there exists a significant gap in the literature regarding the effect of fiscal tightening programs, particularly concerning the wider social implications of these policies. Eissa and Hoynes (2006) explores the effect of the introduction of the Earned Income Tax Credits (EITC) on single mother labour participation, and found that the EITC was successful in encouraging employment on the extensive margin. Farber, Rothstein, and Valletta (2015) stands as one of the few recent examples in the literature which addresses the effect of benefit reduction programs on employment. This study seeks to contribute to this evolving body of literature by investigating the effects of austerity on parental employment and income and further examine the impact that these benefit reductions and changes in economic activity have had on various aspects of the family, including decisions about relationships and fertility, inter-household productivity, and child interaction.

The UK government's austerity policies, which came into force in the early 2010's, provides an excellent setting to research the role of sharp fiscal policy tightening on parental labour market outcomes, income, and the complex interplay between economic circumstances and decision making in the household. Further, the analysis is extended to explore parental decision making and child interaction. In discussing the impact of austerity policies, I primarily focus on the UK's expenditure-based approach which led to changes in key tax-based welfare benefits. In the UK, public cash benefit transfers make up the bulk of social welfare expenditure with adults of parenting age, in their 20s and 30s, claiming more public cash transfers proportionately, compared to other age groups (Gornick and Smeeding, 2018). With this in mind, UK austerity policies provide a unique setting for research on how fiscal tightening programs impact labour supply, trends in income levels, and unintended effects on relationship and fertility decisions (Bitler et al., 2004; Diedrick, 1991; Fisher and Zhu, 2019; He, 2016). Further, related to the supplementary analysis on child outcomes, existing literature has shown that children from low-income households are particularly sensitive to changes in household incomes when compared to their peers from middle to high-income households (Aizer et al., 2016; Milligan and Stabile, 2011). By addressing these topics, I contribute to the literature in three ways. First, I explore the effects of reduced welfare benefit spending on individual labour market outcomes, where previous literature primarily examines the effect of benefit expansions. Second, I analyse how benefit reduction may influence relationship decisions and family formation. Finally, I look into the intergenerational effect of austerity, by examining a number of child outcomes, including the frequency of parent and child interactions and the effect on adolescent mental health. In short, this paper seeks to outline the impact of Austerity policies, by focusing on the Welfare Reform Act, to gain a better understanding of how reductions in government spending influence parental employment, earnings, household relationships, and social decision making.

To capture the extent of family exposure to austerity policies, I primarily focus on the effects of the Welfare Reform Act of 2012 (WRA) and, more specifically, the abolishment of the Council Tax Benefit and the changes to the eligibility criteria for the Housing Tax Benefit.[1] After assigning the treatment based on exposure, I employ an event study model to estimate the effect of austerity on a number of outcomes, beginning with standard indicators for labour supply such as employment, unemployment and income. Here, my primary results show that individuals exposed to austerity measures increase their labour market activities, as parents seek for ways to compensate for their losses in benefit income. I also present evidence that the intended effect of austerity (Bank of England, 2014) was realised for parents, and led to significant increase in employment for fathers, by around ∼15% (from the baseline pre-treatment employment rate of 80%), a reduction in unemployment, which was reduced from 20% to 16% following the WRA and a subsequent increase in labour income for both mothers and fathers. However, I also observe some indirect effects of these policies.

One interesting finding suggests that post austerity incomes between mothers and fathers diverge,[2] with mother total incomes *decreasing* by up to 20 percentage points (ppts henceforth) due to the council tax benefit reductions, with father incomes *rising* by ∼15 ppts.[3] This effect corresponds to a decrease in total income for mothers of around £182 per month.[4] In tandem, I find that austerity caused gender specific changes in relationship decisions, where women exposed to austerity experience a decrease in the frequency of divorces (by 5 ppts against a baseline divorce rate of 20% in the pretreatment period) and an increase in the frequency of marriages and cohabitating relationships by around 8 ppts (where 48% of mothers were married in the pretreatment period).[5] Finally, I also present and briefly discuss potential mechanisms driving the effects of austerity on parent & child interaction and child outcomes.

[1] More information on these specific policies can be found in the institutional background section 2.2.

[2] In section 2.5.4 following Rambachan and Roth (2023), I provide a robustness check showing that this particular outcome may be sensitive to parallel trend violations, therefore these results should be interpreted with caution and considered as suggestive.

[3] See Figure 2.1 for the event study results outlining employment and income trends.

[4] In real terms relative to 2009 £s.

[5] See Figure 2.2 for the event study results concerning relationship decisions.

Where the results show that mothers and fathers might have faced differing trajectories on relationship decision making post-austerity, I turn to the theoretical framework developed by Browning and Chiappori (1998). This paper rejected the prevailing "unitary" model of household decision making where households are viewed as a single decision maker. My findings could be seen as providing support for a more nuanced "collective" framework for modeling household decision making where household members have individual preferences and bargaining power (Browning and Chiappori, 1998; Molina, Velilla, and Ibarra, 2023).[6] Many papers have looked into the role that economic conditions such as income and employment play in inter-household bargaining power (Agarwal, 1997; Basu and Maitra, 2020), where lower income and less labour market flexibility has been shown to reduce the inter-household bargaining power for women. Further, Chen, Conconi, and Perroni (2007) demonstrates how, even when female labour market participation increases, women, especially those who are specialized in household activities, often face a "double burden." In Chen, Conconi, and Perroni (2007), women face a dual dilemma where they must manage their increasing activity in the labour market while simultaneously addressing a sustained demand for household work in the home, thus further eroding their ability to re-negotiate their bargaining position due to time constraints. Recently Dong (2022), looked into the effects of changes to Chinese marriage laws and property division and showed that reductions in female inter-household property ownership reduced female bargaining power, these reductions led to significant reduction divorce rates. Here, we have limited ability to prove that women faced lower bargaining power, yet this is one theory that could explain why the trend in mothers getting divorced declined in the post-treatment period.[7]

Aligned with the previous concept, shifts in relative economic independence can help explain how austerity has led to an increase in the frequency of marriages. Here, the results showing diverging trends in total incomes by gender can be viewed as evidence of reductions in female economic independence (Périvier, 2018a,b). Previous literature has shown that female relationship decisions are closely tied to perceived economic independence, where women with lower incomes are less likely to divorce (No, Andrews, and Yigletu, 2007; Nunley and Zietz, 2012). A similar mechanism may be at play in my results, where I find that women who are experiencing systematic disadvantages may consider entering into new relationships for economic stability.

While exploring the impact of austerity on household dynamics, this paper also examines the potential effects on parent-child interactions and child outcomes. This research highlights parents as the main agents experiencing the intergenerational effects of austerity. It seeks to fill a gap in the literature on how fiscal tightening in-

[6] Here the collective framework could provide an explanation for why females are less likely to alter relationship status post-austerity, due to their now higher perceived costs of exiting relationships.

[7] This difference in outcomes highlights the role of bargaining power dynamics within couples, showing how economic and social factors can have distinct impacts on mothers and fathers. This is something the collective model is better equipped to account for than the unitary framework.

fluences family dynamics and child outcomes (Heckman and Mosso, 2014), building on earlier work that primarily examines the effects of welfare expansion, such as in Eissa and Hoynes (2006). Here, I build on existing work which outlines the transfer of human and social capital from parents to children (Carneiro et al., 2015), and the important role that parental interaction plays in the development of non-cognitive skills (Henney, 2016).

Cunha and Heckman (2008) positions parenting as a mechanism for intergenerational skill development, this suggests that alterations in parental investment due to austerity measures could have long-term repercussions. Here, the youth analysis is limited to adolescents (ages 10-16) due to the scope of the United Kingdom Household Labour Survey's (UKHLS) Youth Survey. It's worth noting the importance of this age range for the development of non-cognitive skills, which can significantly impact mental health (Heckman, 2005). Previous studies, like Agostinelli and Sorrenti (2021) have explored the impact of policy changes on child outcomes, focusing on maternal employment and income effects, while simultaneously recognizing the trade-off between parental time and financial resources. They describe this trade off as the "income vs substitution effect" and highlights the specific relevance of this trade-off for low-income families.

This paper contributes to existing literature by looking beyond the economic effects of fiscal tightening programs on labour market activities and incomes, but also by delving into the broader social context shaped by these economic measures. It outlines the complex interplay between reduced government spending, shifts in parental employment and income, and the consequent influences on household decision-making. By unraveling the diverse and far-reaching effects of austerity, this research provides context for understanding inter-household decision-making during periods of fiscal contraction. The paper also underscores the need for policy-making that adequately considers the wider social implications of benefit reduction programs. As such, the insights provided in this paper aim to inform future policy decisions, with the goal of mitigating the potential adverse consequences of such policies.

2.2 Institutional Background

In the wake of the 2008 financial crisis, slow economic recovery, stagnant GDP growth, and rising government debt, prompted many western economies to pursue various policies of austerity. Austerity policies generally aim to reduce government deficits by either generating revenues with increased taxation (taxed based approach), reducing government expenditure, or a combination of both approaches, among other cuts to local spending (Alesina, Favero, and Giavazzi, 2020). Driven by the desire to lower unemployment, and increase labour force participation, the Welfare Reform Act of 2012 had a significant impact on the provision of welfare support programs and included changes to the council tax, housing, and child tax credit benefits. This resulted in an estimated welfare expenditure reduction of nearly 23.4 percent per person in real terms between 2010 and 2015, with the largest measured reductions occurring in the poorest areas (Fetzer, 2019). I focus on two of the primary policy changes within the WRA, the abolition of the council tax benefit, and changes to the housing benefit. I focus on these specific laws because of their relative importance for low income families. Prior to 2012, both the council tax and housing benefit were means tested and only offered to low-income families.[8],[9].

The Council Tax is a localised tax on domestic property to pay for the allocation of some public goods (i.e., water and waste collection). Prior to the introduction of the Welfare Reform Act, low-income households could qualify for reductions or exemptions on levied council tax payments, through the council tax benefit. In April of 2013, with the introduction of the WRA, the council tax benefit was cancelled without replacement. This lead to an estimated 2.4 million households liable to pay the full council tax for the first time. While the total cost of the council tax varies from council to council, the cost for most households is typically at least £1,000 per household per year (Fetzer, 2019) [10].

Some recipients of the UK housing benefit were also impacted by the WRA. Housing benefits are offered to individuals and households that are either living on low-income, unemployed, or qualify for other forms of welfare benefits. This encompasses individuals residing in social or council housing who receive allocated funds to assist benefit recipients in subsidizing their rent. In accordance to the WRA, the housing tax benefit was changed to exclude the costs of an additional room, meaning that benefit recipients living in accommodation which could be defined as having excess bedrooms would see a reduction in their housing benefit. The new guidelines allowed housing benefit recipients to have one bedroom for each adult couple, and single inhabitant over the age of 16, one bedroom for every two children who are the same sex, and one bedroom for children under 10 regardless of their sex. Housing benefit recipients with excess bedrooms, experienced a benefit reduction of around 14 percentage points (ppt) in households with 1 extra bedroom and by 25 ppt in households with

[8] The WRA also made changes to the Child Tax Benefit by enforcing an income cap for eligibility, where families making over £50,000 per year (pre-tax) would no longer receive the benefit. Thus, the child tax benefit was not particularly relevant for low income families or the focus of this paper.

[9] The WRA included additional policies which are not the focus of this paper. For example the WRA marked the introduction of Personal Independence Payments (PIP) which replaced Disability Living Allowance (DLA) and only impacted those with long term health-issues.

[10] While London is an outlier, with city administering more local support than other councils, a report by the Child Poverty Action Group estimated an average liability per household of around £200 per annum, with 4 out of 10 affected Londoners receiving a court summons for non-payment (Ashton, 2014)

two or more excess bedrooms, leading to an income loss of around 56 - 100 pounds for effected households.[11]

2.3 Data

2.3.1 UK Household Longitudinal Study: Understanding Society

In this paper, I examine the impact of UK austerity policies on a number of outcomes by leveraging data from the UK Household Longitudinal Study (UKHLS). The UKHLS is an individual level longitudinal survey on the members of around 40,000 households which are observed annually. The portion that I use, the Understanding Society Dataset, is a continuation of the British Household Panel Survey, which began in 1991 and was subsequently continued and renamed to the UKHLS in 2009. The UKHLS uses a clustered and stratified sample that aims to represent the entire UK population.[12] Interviews are carried out in person when possible, with all household members selected in the first wave comprising of the core sample. This core is then followed throughout their lives as long as they remain in the UK; family members of core sample members then comprise the secondary sample, who are interviewed as long as they remain in the household of the core sample member. While, the household sample is the primary survey, household members between the ages of 10 – 16 are asked to complete a short self-completed youth questionnaire (with permission from their parents), once they reach the age of 16, they are then eligible for the full interview. Also included in the UKHLS is a child survey answered by parents on behalf of child household members who are younger than 9 years old. The Understanding Society dataset is unique in its inclusion of variables that provide information on the life experiences of both children and adults in the UK. With questions ranging from concrete indicators of education qualifications/employment, and income, and extending to personal questions about participant mental health, life ambitions, and behavioral habits.

2.3.2 Identifying Exposure to Changes in Benefit Incomes

In this section, I begin by describing treatment allocation and variable selection. Following this, I will then discuss the construction of the youth file (sample) used in the

[11] The WRA also carried out changes to the child benefit, most notably, increasingly stringent eligibility requirements and capped yearly benefit increases to 1% per year. Changes to the child benefit tax had the largest impact on households in which at least one household member earns at least £50,000 GBP per year, these households became ineligible for the child tax benefit after the WRA was implemented. Primary analysis on the changes to the child benefit tax saw little impact on parental outcomes and employment, which can be potentially explained by the relatively high income of the recipients who were impacted by this particular policy change, because of this, changes to total income resulting from changing benefit incomes caused by the child benefit are proportionately low when compared the financial impact of the other primary WRA changes. Keeping this in mind, the analyses primarily focus on the impact of changes to the council tax benefit, and the bedroom tax.

[12] The UKHLS includes a tandem of boost samples, the ethnic minority boost sample, and the immigrant boost sample seeks to ensure that minority populations are well represented.

analysis. Since this research is focused on the experiences of the family and household dynamics, I restrict the sample to only include households in which children (under the age of 16) are present. In the UKHLS, survey participants over the age of 16 are included in the main file. I use UKHLS waves 1-10[13],excluding 2020 to avoid distortions caused by COVID-19, which may affect the validity of pre-pandemic comparisons.

Our treatment variables aim to identify individuals and households that have been directly exposed to either of the two main austerity policies described in the background section, these being, council tax benefit (CTB) recipients and claimants of household benefit with spare bedrooms (i.e., "spare bedroom tax" or BTX henceforth). These measures were key provisions of the 2012 Welfare Reform Act and took effect in April 2013, aligning with the start of the UK financial year, which is a common implementation period for major fiscal policies. The timing was driven by broader austerity measures rather than immediate economic conditions, which has implications for the common trends assumption, discussed further below (White, 2016).

In order to identify individuals exposed to changes in tax credits I make use of the UKHLS data's "Unearned Income and State Benefits Module" in which survey respondents indicate if they are currently receiving any state benefits in that particular wave. Due to the research question, which focuses on household experiences and family dynamics, household married spouses and non-married couples were matched to generate more accurate benefit receipt indicators. For example if one matched spouse or partner in the same household meets the criteria for being included in the treatment group, the other spouse or partner in that household would be included in the treatment group. It is understood that in the scenario explained above, spouses or partners would be exposed to the WRA through changes in benefit incomes at the household level. In instances of a change in household makeup (i.e., marriage or divorce following April 1st 2013) future spouses or partners, who are normally new entrants into the survey, are also included in the treatment group because of their association with exposure to austerity through their spouse or partner. The exposure indicators described below are mirrored in the youth dataset, by extracting the treatment indicator and then matching these indicators to the youth dataset through the parent's unique identifier found in both datasets. Those who do not meet the set of treatment group criteria explained below are understood as being unexposed to austerity policies and act as pure controls in the model specification outlined in section 2.4.

While this classification ensures that treatment assignment accurately reflects exposure to the WRA through household benefit receipt, it is important to establish that the control group represents a reasonable counterfactual. Table A.1 presents a comparison between the treatment and control groups before the implementation of the policy. While some differences in demographic and control groups before the implementation of the policy exist, these differences are expected given the nature of bene-

[13] Wave 1 was collected in 2009 - 2010, wave 10 was collected between 2018 and 2020.

fit receipt and household composition. However, these differences do not necessarily threaten identification as the empirical strategy employed accounts for time-invariant and household characteristics.

Additionally, to further assess the suitability of the control group, a propensity scores matching (PSM) approach, shown in figure A.9, was implemented as a robustness check to re-weight the control sample to better resemble the treatment group in terms of observable characteristics. The main results remain consistent when using the PSM- reweighed sample, suggesting that pre-existing difference between groups do not drive the findings. Further, the DID approach inherently accounts for baseline disparities by differencing out time-invariant selection effects.

Council Tax Benefit Abolishment. Assigning treatment indicators for households and individuals who were exposed to the abolishment of the CTB is straight forward, with the UKHLS data providing an individual indicator for the CTB in the state benefits module. Knowing that the policy came into force in April of 2013, households and individuals who indicated that they had continually received the CTBs in at least 2 waves, or the 3 years leading up to April of 2013 are considered treated, this definition excludes up to 1 wave of non-response. Following this logic, households that indicated they had only received the benefit, in the final wave (i.e., 2012 observations) would not be included in the treatment group. Once treatment groups are established, indicators for exposure to the CTB would be matched to unique personal identifiers in all subsequent waves, regardless of their changing household compositions.

While this definition omits individuals who received benefits only in 2013, this choice is intentional. The goal is to estimate a conservative, lower-bound effect of the policy change by focusing on those with sustained exposure to the benefit. Including short-term or marginal recipients could introduce additional heterogeneity, potentially inflating estimated effects. A breakdown of pre-2013 benefit recipients shows that 13.8% of the sample is classified as treated under this definition. An alternative, less restrictive definition results in a treatment group comprising 14.5% of the sample. This means that approximately 0.7% of households in the sample are excluded due to having received benefits only in 2012. This exclusion ensures that the estimated effects reflect those most affected by the policy rather than capturing individuals who may not have continued receiving benefits in the absence of the reform. This treatment identification strategy follows the method used by Fetzer (2019) and is shown below.

$$T_{i,CTB} = \begin{cases} 1 & \text{if individual or spouse received the council tax benefit for two waves prior to April 20} \\ 0 & \text{else} \end{cases}$$

Bedroom Tax. Categorization into treatment groups indicating exposure to the BTX is less straight forward. Exposure to the BTX impacted families that had excess bedrooms with the following criteria, household members above the age of 15 are allocated one bedroom, children under 15 are permitted 1 room for 2 children, and couples are expected to share a bedroom. The WRA also excluded households where at least one household member makes over £50,000 per year.

I leverage household level data to identify households which were exposed to the "bedroom tax" first by using derived monthly gross income variables followed by an estimate in the number of "excess" bedrooms per household as defined by the WRA. Using information about the household makeup and the number of bedrooms in each household I compute the number of "allowed" bedrooms per person based on the number and age of children, the number of single household members over 15, and the number of couples within the household. This "allowed" number of bedrooms was then subtracted from the total number of bedrooms as given in the household data file. After estimating the number of excess bedrooms per household, I then define the treatment by identifying households with at least one "excess bedroom" and households that had identified that they were receiving the housing benefit in at least 2 waves leading up to the implementation of the WRA, on April 1^{st} 2013. Similar to the council tax benefit above, under this definition, 5.68% of the sample is classified as treated, compared to 5.8% under a less restrictive definition. This amounts to only 62 households, 2% of households are excluded in the more restrictive definition. Treatment indicators were then created following the strategy used to assign the council tax treatment, as seen below.

$$T_{i,BTX,t} = \begin{cases} 1 & \text{if individual or spouse benefitted from the council tax benefit prior to April 2013} \\ 0 & \text{else} \end{cases}$$

It is important to mention that the vast majority of those exposed to the bedroom tax changes were also exposed to the council tax benefit abolishment See table A.3 for the number of treated units in each treatment category in 2012 the last pre-treatment period.. A number of outcomes were tested using the specification outlined in section 2.4.1 and more detailed description of the dependent variables can be found 2.4.1. Selected outcomes aim to estimate the effects of austerity policy changes on parental employment trends, income levels, and family dynamics. These outcomes can be categorized into three major areas: changes in income and employment, alterations in family structures and relationships, and the impact on children.

A number of outcomes were tested using the specification outlined in section 2.4.1 and more detailed description of the dependent variables can be found 2.4.1. Selected outcomes aim to estimate the effects of austerity policy changes on parental employment trends, income levels, and family dynamics. These outcomes can be categorized into three major areas: changes in income and employment, alterations in family structures and relationships, and the impact on children.

2.4 Identification Strategy

2.4.1 Specification

To tease out the effect of the welfare reform act on labour market outcomes and family dynamics, I estimate an event study model to exploit the introduction of the Welfare Reform act which came into effect in April of 2013. The event study model can be understood as generalized extension on the standard difference-in-differences (DID) model, but also allows us to estimate the dynamic effects of the policy over time. Additionally, plotting event study coefficients along with their confidence intervals allows us to visually assess potential identification issues. The decision to focus on mothers and fathers separately in the main specification is grounded in prior research highlighting how welfare benefits are distributed differently between men and women (Ştefan and Avram, 2017). In the extended analysis on children, I also observe that parental interaction variables are sensitive to gender differences (Bastiaansen, Verspeek, and van Bakel, 2021; Chen, Huang, and Ye, 2020). Section 2.5.4 explores this further, where I present and discuss results that are not specific to gender. The event study specification for parents can be expressed as follows (equation 1) informed by Clarke and Tapia-Schythe (2021):

$$eq : 1 y^g_{i,t} = \lambda^g_t + \gamma^g_i + \sum_{q=1}^{m} \beta^g_{+q} \text{Austerity}_{i,t+q} + \sum_{\substack{q=1 \\ q \neq \text{ (reference)}}}^{p} \beta^g_{-q} \text{Austerity}_{i,t-q} + \delta^g X_{i,t} + \epsilon^g_{i,t} \tag{1}$$

Where $y^g_{i,t}$ denotes an outcome selected from the parent dataset, where $g \in \{M, F\}$ denotes gender groups (male/female), which pertains to labour market outcomes such as employment and wages, health, or parent child time investment. Individual fixed effect, γ_i capture time invariant differences within individuals, the inclusion of individual fixed effects are important to control for if individual traits that do not vary over time are correlated with selection into the treatment[14].

The coefficient of interest is β, corresponding to the interaction between exposure to Austerity measures, $\text{Austerity}_{i,t-q}$. Coefficients β_qs are to be interpreted relative to the commencement of the welfare reform act of 2013. The estimates are then plotted to visually display the dynamic trends in various outcomes. This implies that time dummy where $t = 0$ is the baseline period and is omitted, therefore estimated βs are interpreted relative to this time period, corresponding to the first year of the WRA, 2013. The coefficients $\beta_{+1it}, ..., \beta_{+m}$ measure the lags, or post-treatment effects, with $\beta_{-1}, ..., \beta_{-p}$ represent the leads, or pre-treatment effects. For instance, $\text{Austerity}_{i,t+2}$ would include a binary indicator equal to 1 in observations from 2011, or two years

[14] The year before treatment ($t = -1$) is the reference period and excluded from the summation.

before the introduction of the WRA for the treated group whereas the variable is always equal to zero for the control group.

Our specification for child outcomes uses the same equation shown in equation 7. In all iterations of equation 7, standard errors are clustered at the household level, to account for possible serial correlation within households over time.

Also included is both parent and child specifications are explanatory variables, $X_{i,t}$, which seek to control for time-invariant socioeconomic characteristics and increase precisions in the estimates. I include a second-order polynomial function for parental age to capture the non-linear relationship between age and the outcome variable of choice. Further, I include the covariates denoting the number of kids in the household, [15] and a discrete regional identifier (north England, England, London, Scotland, Northern Ireland, and Wales) which helps control for potential regional disparities or changes, such as families moving between regions.

Outcomes. I begin by estimating the direct effects from austerity policies with an analysis on the intended labour market impact of the WRA, namely that a reduction in welfare benefit incomes would increase labour market activities (Makowiecki, 2014). The idea is that these increases are driven by former benefit recipients who seek to replace losses in benefit incomes with labour market wages and earnings, where previous literature shows that changes in benefits can impact labour market activities (Eissa and Hoynes, 2006; Farber, Rothstein, and Valletta, 2015).

Our main results estimate the expected impact of austerity on the labour force participation by using the UKHLS' labour force status variables, I create a binary indicators for those in paid employment (including self-employed) and those who are unemployed. I then focus on the intensive margin of labour supply by using a log transformation of the hours worked. Importantly, I also estimate the effect of austerity policies on both labour and total personal income. I estimate annual income by scaling up the reported monthly gross income.

Following the analysis of direct labour market effects of austerity measures, I delve into the wider social implications of the policies. Recognizing that economic instability and changes in financial circumstances can significantly influence family structures and dynamics, I estimate the effects of austerity policies on relationship decisions and fertility intentions. The UKHLS provides detailed information on relationship status; using this, I construct relationship status categories. Here, I simplify the data by grouping married and cohabiting individuals into a single category, 'in a couple,' while leaving other statuses like 'single' and 'divorced' unchanged from the raw data. I then create dichotomous variables based on these categories to analyse trends in relationship statuses.

I also explore the potential effects of austerity measures on fertility. While the UKHLS does not include a direct measure of fertility, I proxy it by coding a binary

[15] See Figure A.10 in the Appendix for results which exclude this particular covariate.

variable that identifies individuals with a child less than one year old. To better align the timing of the birth with the period in which the decision to conceive likely occurred, I shift this indicator back by one wave. This approach, which follows Romiti (2018), accounts for the average nine-month gestation period and the fact that most births occur mid-year.

To address outcomes relating to mental health, I make use of the General Health Questionnaire (GHQ) measurements, which assess general wellbeing with "Likert" Scale responses rating from 0 (least distressed) to 4 (most distressed). There are 10 individual GHQ questions that seek to understand participant mental health, notable questions asked about the participants happiness, confidence, ability to concentrate, and ability to face day-to-day activities. Individual GHQ questions were then combined into a combined indicator for subjective wellbeing with scores range from 0-36.[16] I include this measure as a dependent variable after preforming a z-transformation. Following the discussion, I also include the plotted estimates of each individual GHQ component (found in appendix). To further assess the impact of austerity on parental mental health I use principal component analysis (PCA) to estimate a latent variable for mental health (Abdi and Williams, 2010; Brewer, Dang, and Tominey, 2024).

Parental investment into child development variables seek to understand the frequency and quantity of parental interactions with their children. Beginning with variables indicating discrete frequency (times per week) of family dinners, talking about things that matter, quarreling with children and asking about school, I create binary indicators denoting household with parental variable frequencies below the sample average. For instance, in the dichotomous variable capturing "parent talk,"" where parents talk with their children 2-3 times per week on average, parents that talk less 2 times per week with their children are assigned a value of 1.

Youth control variables are similar to the parental covariates introduced earlier and include indicators for age, sex, region, and ethnicity.

Covariates in the youth specifications are consistent with those explained previously for parents. Youth outcome variables can be roughly categorized into 2 main groups – parent time and health/mental health. Parent time outcome variables are nearly identical to the questions asked to parents. Health and mental health outcomes are computed by leveraging the strengths and difficulty questionnaire responses (SDQ). The SDQ is similar to the GHQ in that it addresses questions about the mental health of the youth participants, but also has questions about peer relationships, conduct, and hyperactivity. A composite score includes all these measures scaled from 0-36 (with 36 indicating the most distressed children). Following a similar analysis found in Brewer, Dang, and Tominey (2024) I also test an estimated latent SDQ variable by using PCA,

[16] 0 denoting the least distressed individuals.

without identifying significant changes. Results which use the PCA weights can be found in A.6 in Figure A.9.

2.5 Results

In the following section I present my results showing the estimated impact of austerity policies on the outcomes described in Section 2.3. These results are displayed graphically by plotting of the estimated coefficients from the model described in equation 7.[17]

This section will begin by presenting and briefly discussing the main results, which confirms the intended effect of the WRA as a motivation for increasing labour market activities. This includes the estimated impact of austerity policies leading to changing employment, income, and labour supply of treated individuals, compared to those who were not directly impacted by the WRA. Following the discussion on the intended effect of the WRA, I introduce outcomes related to the unintended consequences of austerity. In this section I will also briefly discuss my findings on the impact of the WRA on parental mental health outcomes.

2.5.1 Main Results

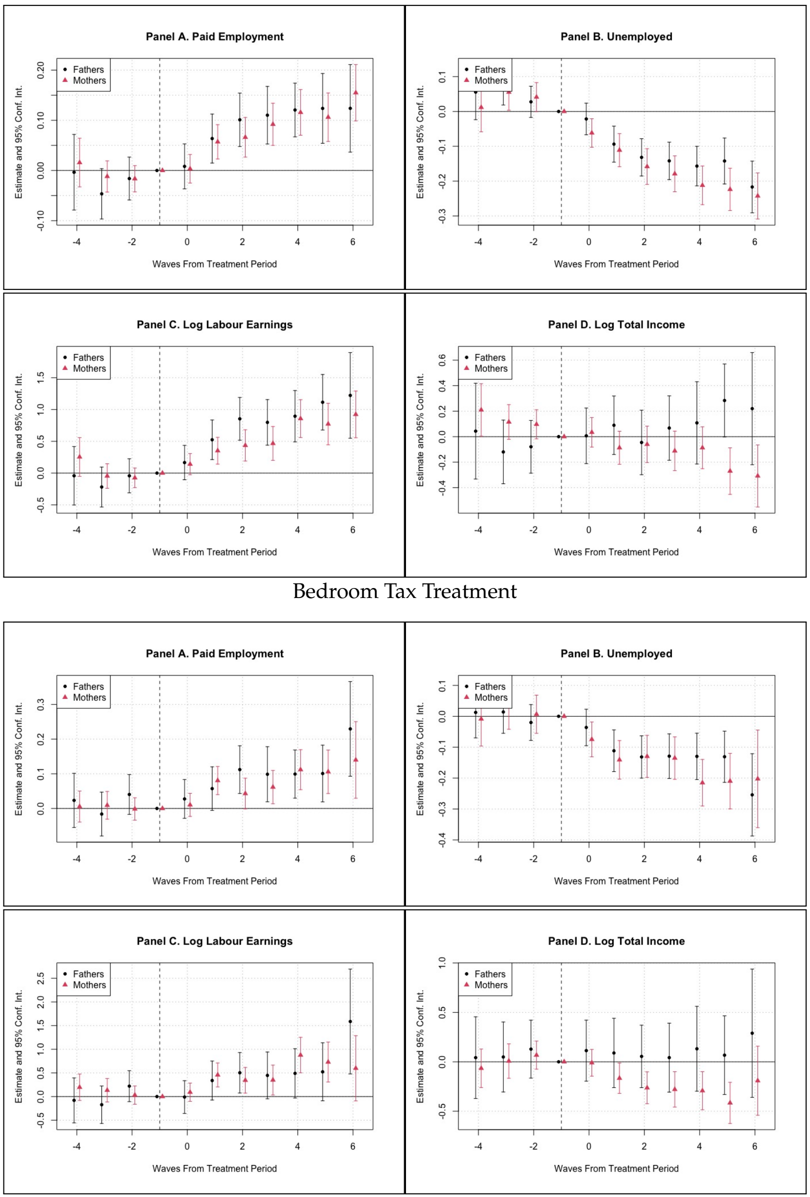
Below, Figure 2.1 presents the results of estimating the specification shown in equation 7 on indicators of labour market participation and earnings. The results are displayed in two subfigures, A and B, which isolates the effects of each individual policy. The coefficient estimates are displayed alongside the 95% confidence interval bands.

Parent Employment. Starting with Figure 2.1 section A, the results concerning employment and earnings begin with an inquiry into the primary intended impact of austerity policies for the benefit claimants who were subjected to changes in the WRA. A notable justification for the WRA was the notion that the government could both reduce fiscal expenditure while encouraging a reduction in the headline unemployment rate (Alesina, Favero, and Giavazzi, 2020). As seen in Section A, Panel A. and interpreting the plot in line with the guidelines outlined in Freyaldenhoven et al. (2021). Section A. panel A. shows a "smooth event time trend" suggesting that both mothers and fathers, impacted by the abolishment of the council tax, were indeed driven into the labour force. These effects culminate in a 19 ppt increase in employment for mothers (with 67% of mothers in employment) and a 15 ppt increase in paternal employment (in this sample, 80% of men in the sample are employed).

In the pre-treatment year of 2009 is suggestive of some marginal degree of diverging employment trends in the treated group compared to the untreated. It is entirely

[17] I also pay close attention to the pre-period estimates and confidence intervals as a way to visually assess the validity of the parallel trend assumption. Here, I look for common stable pre-trend differences between the treated and control group (Freyaldenhoven et al., 2021). The parallel trend assumption will be revisited in section 2.5.4.

Figure 2.1: Mother and Father Labour Participation



Council Tax Treatment

Bedroom Tax Treatment

Note: Figure 2.1 sections A and B show the effect of austerity policies on being in paid employment (panel a), being unemployed (panel b), earnings from labour (panel c), and total income (Panel d). Figure 2.1 is broken into two sections with section A displaying the impact of the council tax benefit abolishment and section B displaying the effect of the "spare bedroom tax." Mother and Father results are plotted together, with the coefficients associated with mothers shown in red and the results associated with fathers shown in black.

plausible that this effect may have been caused by a delayed reaction to the financial crisis of 2008-2009. Noting that high unemployment in the aftermath of the great recession developed slowly (Bank of England, 2020), and that lower paying jobs were initially less affected by the crisis compared to other job categories (Verdugo and Allègre, 2020). This situation would have been particularly relevant for the treated group in the sample, as the majority of benefit recipients who were employed during the pre-treatment period were classified within the lowest socioeconomic category. I also find that mothers and fathers share similar employment trajectories with significant increases in employment. Still, these results are indicative of the effect of the council tax abolishment on parents entering into employment.

Panel B's trend plot of unemployment mirrors the employment findings with a similarly smooth event time trend, though I see a slightly higher reduction in unemployment, peaking at around 20 ppts in 2017 for fathers where 8% of men in the sample were unemployed,[18] results for mothers are similar, peaking at a 20 ppt reduction in unemployment with 7% of women in the sample being unemployed.[19] Similarly to the employment results in Panel A., these findings show a clear impact of the council tax benefit abolishment on unemployment trends for both mothers and fathers.

Not surprisingly, Panel C shows that as trends in employment have risen in the post-treatment period, labour earnings also significantly rise, and in a large way. These large coefficients are primarily driven by the large share of treated individuals who were either not in the labour market or were unemployed in the pre-treatment period, with just 28% of the treated group stating that they were in employment. As I have shown in figure A.5 in the appendix, when the model is restricted to exclude individuals who were unemployed or outside of the labour force in the pre-treated period, the upward trend in labour earnings is all but cancelled out. Where any additional increase in labour income constitutes an increase of 100%, inflating the estimated effect of austerity on labour earnings. Still, the effect of austerity on labour market incomes is substantial and goes hand in hand with increased employment. I can then look to Panel D, for evidence of the impact of austerity policies on total income.

Labour Earnings and Total Incomes. For fathers, post-treatment income differences look to be increasingly significant in the long run, seeing around a 20 ppt increase in earnings compared to the counterfactual trend. It can be speculated that the Panel D results for fathers may be suggestive of an increase in total incomes, but with few statistically significant coefficients. Knowing that total incomes for fathers have either seen no change in trajectory, if not a slight increase, I find that most fathers in the treatment group were able to replace their lost benefit incomes by labour earnings, and that are no pre-trend concerns as coefficients remain close to null in all pre-treatment periods. For mothers in Panel D, there seems to be little significant impact of the Council

[18] Compared with a 10% unemployment rate in the pre-treatment period.

[19] Compared with a 9% unemployment rate in the pre-treatment period.

Tax Benefit abolishment on total income, though there is some evidence of a downward trend in the very long run, consistent with reducing incomes associated with exposure to austerity. While most pre-treatment coefficients are statistically insignificant. Some, such as the estimate four years prior, are non-trivial in magnitude (0.2 for log income), warranting a cautious interpretation of parallel trends. Taken with appropriate caution given the potential for non-negligible pre-trend differences, the post-treatment estimates suggest economically meaningful effects. To contextualize the size of these estimated effects, the policy resulted in an average monthly income loss of £182 for mothers; equivalent to a 17% drop relative to pre-treatment income levels (£1,070/month). The associated 19 ppt drop in employment thus represents a substantial behavioral response to a moderate but not overwhelming income shock. For fathers, whose earnings largely replaced lost benefit income, the 15 ppt employment effect aligns with a more successful adjustment to the policy change.

Figure 2.1 Section B results covering the impact of the so called "bedroom tax" on labour force participation outcomes and earnings largely mirror the results in section A discussed above, though the impact of the policy seems to be more subdued. Pre trends for all four panels are not indicative of a violation in the parallel trend assumption. Panel A shows an upward trend in employment for the treated group, peaking at an increase of employment of about 20 ppts for men and 12 ppts for women. Though there are no statistically significant differences between mother and father outcomes, coefficients on the mother subset suggest that the bedroom tax seemed to have had less of an effect on mother employment than the council tax benefit. As with the council tax results, I again see evidence of significant divergent trends in the unemployment for both mothers and fathers. Once again, I see a disconnect between female employment and unemployment explained by the increases in mothers staying at home in those waves. Panel C results show that while an increase in labour earnings can be linked to the bedroom tax, the effect is less pronounced when compared to section A panel B. Differences in the log of total income for fathers shows a null effect, with decreasing precision in the later waves which can be associated with sample attrition. I see similar income trends for women when compared to the council tax results discussed previously.

In summary, figure 2.1 results suggest that the WRA had an impact on employment trajectories, including increases in working hours and overtime hours, shown in figure A.6, results that can be found in the supplement results section in the appendix. Compared to the hypothetical counterfactual trend without austerity policies in place, the council tax and the bedroom tax have had a considerable impact on the treatment group, along with growing employment and shrinking unemployment, I see increases in labour earnings, and no large significant effect on total earnings for fathers. The employment results suggest that there may be gendered differences on the impact of the WRA, particularly concerning total income, in which our results suggest that mothers may have been worse off overall post treatment. The Roth pre-trend test, which assesses the validity of the parallel trends assumption in difference-in-differences analysis, indicates potential violations in this case. As such, these findings should be considered as suggestive. While I do not formally test for statistically significant differences across gender, all models are estimated with household-level clustering to account for within-household correlation. Visual inspection of the confidence intervals suggests potentially meaningful differences, particularly in the ability of fathers versus mothers to offset lost benefit income through increased earnings, but these comparisons are interpreted as suggestive rather than definitive.

The results for total benefit incomes are more ambiguous, particularly for single women, as child support payments have increased (distinct from the child tax credits altered under the WRA), while other social benefits have decreased. Despite the jump in child support, austerity's effect on mother total income was still negative, as shown in figure 2.1.

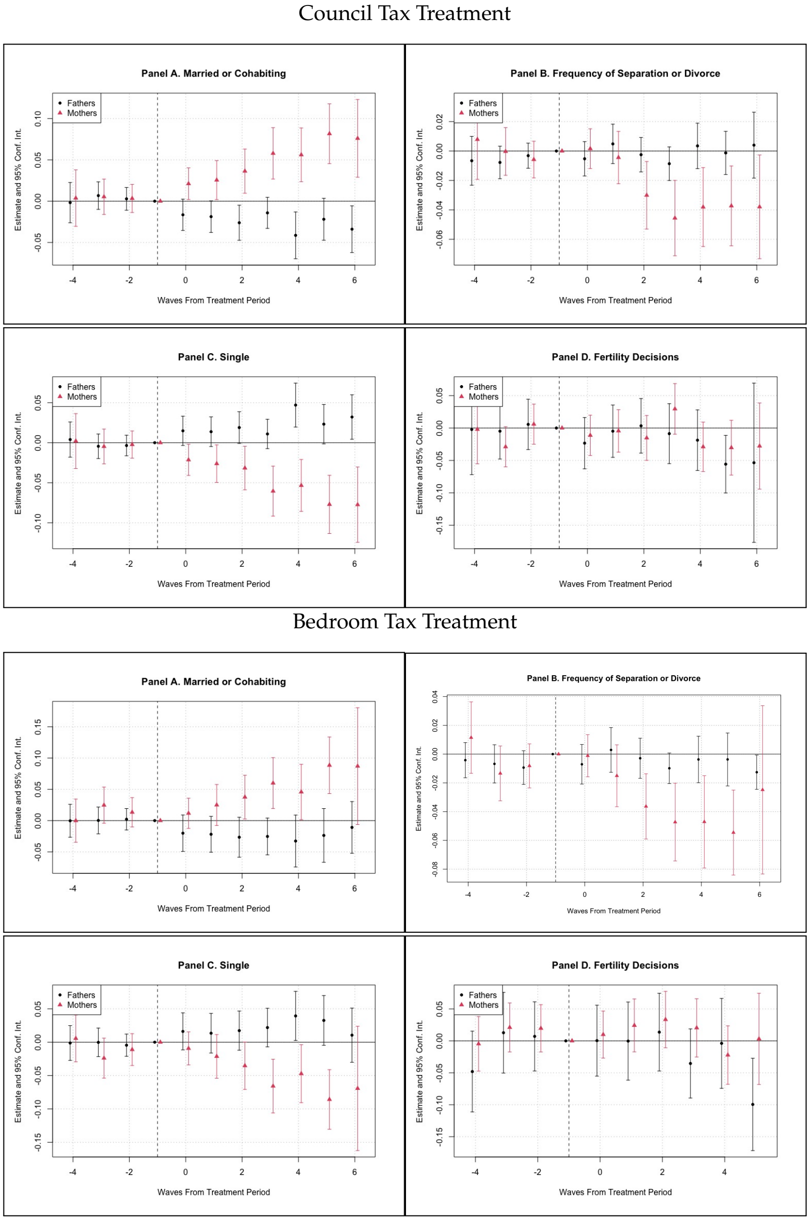
2.5.2 Household Dynamics

So far, the results have primarily focused on the direct consequences of the WRA on employment and income. However, the impacts of economic policies also often extend beyond the immediate realm of economic measures. In this section, I turn our attention to a less immediately observable but equally consequential aspect where I provide evidence to show that austerity has changed dynamics within households. Above, in figure 2.1 I presented a particularly interesting finding where I show that austerity has led to women experiencing negative, divergent trends in total income compared to fathers. These unintended effects prompted us to look into some of the potential externalities of these findings, where the differential impact on income between genders has the potential to significantly reshape family dynamics in a number of ways.

In the following section, I will explore these effects, as well as other implications of my findings. The aim here is to paint a more holistic picture of the impact of the WRA, not just on the economic lives of individuals, but also on the social fabric within which these individuals are situated. I believe that understanding these indirect effects is crucial for comprehending the full scope of austerity's influence and its long-lasting socio-economic implications.

Relationships and Fertility Figure 2.2, Section A, Panel A, presents compelling results that point to an increased rate at which mothers are entering into new marriages or relationships in which they cohabit with partners. This effect is not mirrored for fathers, where their results show that austerity did not impact their trends in relationships. I can also see that mother and father trends are divergent, shown where the confidence intervals separate. The pre-treatment point estimates suggest no evidence of differential pre-trends between treatment and control groups. While this is encouraging,

Figure 2.2: Mother and Father Relationships and Fertility


Council Tax Treatment

Bedroom Tax Treatment

Note: Figure 2.2 above plots the estimated effect of austerity on variables related to household dynamics. These results are broken into two sections, a & b, where section a presents the effect of changes to the Council Tax Benefit and section b refers to the“bedroom tax.” Here results are sub-setted by gender, with mothers shown in red and fathers shown in black.

it should be interpreted as suggestive rather than definitive, since the parallel trends assumption fundamentally relates to unobserved counterfactual outcomes in the post-treatment period.. The plot displayed in section b Panel A shows results for the same outcome but displays the effect of the "bedroom tax." I can see that these results follow a similar trend but the effect is not as strong. For the "bedroom tax" results I still find significant gendered differences between mother and father relationship formation, where pre-treatment estimates are stable. Note that the 6th post-treatment period is omitted from the fertility analysis because there are no treated observations available in that period. This is due to attrition, which is more pronounced for outcomes like fertility that are only asked of a subsample, and particularly so for respondents affected by the Bedroom Tax. As shown in the appendix balance table, the number of Bedroom Tax treated observations is substantially lower overall, making this combination especially vulnerable to sample loss over time. While this limits the ability to observe longer-run effects for this subgroup, there is no evidence that attrition is systematically related to treatment timing or outcomes. Nevertheless, I interpret the fertility estimates, especially in later periods, with caution, and acknowledge that sample composition may partially influence observed trends.

It's important to note that in cases where the panel is gender-balanced, we might assume that relationships status' by gender to trend together. The sample has an unequal gender distribution, with fewer men than women (58% women). On average, men in the sample are more likely to be married, with 66% of males being married in the pre-treatment period compared to 52% of women.[20] Further, women make up roughly 65% of the treatment group. These data features make it possible to capture the gendered effect of the WRA on relationship status that this paper highlights.

Following this, I turn to the implications of the WRA on the incidence of divorce. Figure 2.2 Panel B plots the estimated effect of the WRA on a binary variable which notes if the individual is divorced. These results mirror the effect of the WRA on marriage and cohabitating relationships and shows that changes to the council tax benefit likely led to a decline in divorces for mothers. In Section B, Panel B, noting the effect of the "bedroom tax" on divorce, the effect is again less strong, with wider confidence intervals, but also provides supportive evidence of significant divergent trends between women and men.

Next, Panels C in both sections of figure 2.2 re-affirm the previous results on changes in family formation behavior. Here I find that women are increasingly less likely to remain single in the post-austerity period. Again, I find that this trend diverges from men, whose likelihood of remaining single in the aftermath of austerity measures remained largely unchanged, or even slightly increasing in later periods. Once again, this suggests that the effect of the WRA on relationship behavior demonstrates significant changes and significant differences between genders. From these results, I

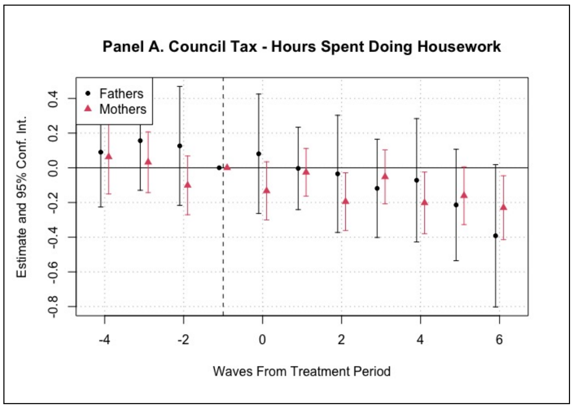
[20] This rose to around 56% in the post-treatment period.

can infer that the socio-economic changes brought about by austerity have not only impacted the financial and employment aspects of people's lives but have also significantly influenced their personal relationships and family structures. This paper is the first to show the impact of fiscal tightening policies on marital decision making. Finally, Panel D in both sections presents my results outlining the impact of UK austerity on fertility decisions. Despite the marked impact of austerity measures on relationship dynamics as previously outlined, I find no statistically significant effects on fertility decisions for either mothers or fathers. However, the confidence intervals around these estimates are relatively wide, particularly in later periods, meaning that I cannot rule out the possibility of sizable positive effects in the short run and negative effects in the long-run. As such, these null findings should be interpreted with caution. The absence of statistically significant variation in fertility amid austerity may suggest that, while economic conditions influence relationship status and household dynamics, their direct bearing on parenthood decisions remains less clear in this sample.

The findings above indicate that austerity may have influenced certain aspects of personal and familial relationships. For mothers, there is evidence of an increase in new partnerships or cohabitations, a gradual decline in divorce rates, and a reduction in the share remaining single following the onset of austerity. In contrast, men's relationship trajectories appear relatively stable. These patterns point to potential gender-specific responses to economic pressure, suggesting that the effects of austerity may extend beyond employment and income, subtly shaping relationship dynamics as well. While not definitive, the results highlight how broader socio-economic shifts can interact with intimate aspects of individuals' lives.

Mental Health. Building on Section 2.5.2, which examines shifts in family formation decisions, I now turn to explore the impact of austerity measures on mental health. Figure A.11 in the Appendix presents results that outline the estimated effect of austerity policies on the 12 mental health components that make up the GHQ. Here I do not find any indication of austerity driven trends, or structural patterns for parental GHQ responses. Some of this is likely due to the relatively large portion, roughly 1/3 of treated individuals, who were non-responsive to the GHQ portion of the UKHLS. Father GHQ questions often have divergent pre-trends, raising specification validity concerns related to the common trend assumption. For instance, coefficients for GHQ questions associated with female decision making seem to be increasing, but this is only speculative and statistically insignificant. Similarly, mother coefficients in the post-treatment period for GHQ questions about feeling useful also appear to be consistently increased in the post treatment period without being significant. I also tested a comprehensive mental health score constructed using principal component analysis. Similar to the standard comprehensive measure of mental health included in the understanding society dataset, the constructed comprehensive mental health scores do not seem to be influenced by austerity policies.

Figure 2.3: Parent Log Hours of Housework Per Week



Note: Figure 2.3 plots the estimated coefficients of the dynamic changes in the number of hours mothers and fathers spend doing housework for the council tax treatment group.

Housework. Figure 2.3 provides a continuation of the analysis on household dynamics in presenting the estimated impact of austerity on the frequency (number of hours) that household members spend doing housework. The housework hours outcomes stem from a variable that asks participants how many hours per week they spend doing housework, I then log transformed the original variable.[21] This effect of austerity on the frequency of housework can be understood as closely linked to the results presented in figure A.6 where I show that both mothers and fathers significantly increase their work hours post-austerity and subsequently have less time for household chores. While imprecise and only suggestive, Figure 2.3 suggests that both mothers and fathers may be spending less time doing housework in the post-treatment period. Here, parents are reducing the amount of time they spend on housework by around 20 to 30 ppts, meaning that hours spent on housework decreased by around 3 to 4 hours.[22]

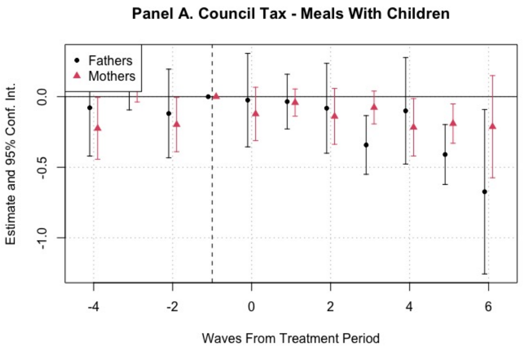
2.5.3 Austerity and Children

Parent Time Use. The previous section outlines that in the aftermath of the introduction of the WRA, benefit claimants increased their employment, a reduction of unemployment, and an increase in the number of hours that mothers and fathers work.

[21] Excluding null or zero observations

[22] Parents exposed to austerity spent on average 15 hours per week on housework in the pre-treatment period, a 20% reduction in 15 leaves us with 12 hours per week spent on housework, this corresponds to a 3 hour reduction. Here the results are slightly more precise for women

I then shift the focus onto the effect of these changes to the quantity of child interactions, as parents are forced to adjust their time use in response to their increases in labour market activities. As previously discussed, there is a large body of literature discussing the tradeoffs that parents make in their time use decisions (Cunha and Heckman, 2008; Doepke, Sorrenti, and Zilibotti, 2019; Doepke and Zilibotti, 2017). The UKHLS's Understanding Society Dataset is limited in its inclusion of variables that can be used to proxy child and parent interactions, with low response rates for some of the most relevant questions. Figure 2.4 below outlines the impact of austerity on the frequency of parental interaction proxied by the frequency of parent and child meals. Here "meals with children" is the average number of times per week parents share a meal with their children. Family meals, especially between parents and children, have been found to play an important role in child development, and have been linked extensively to important child outcomes (Horning et al., 2017; Jones, 2018; Price, Rodgers, and Wikle, 2021). These results also closely relate to figure 2.3 discussed above. While housework in itself cannot be seen as directly related to parenting, I can view this variable as a proxy which is indicative of home engagement and responsibility (Norman and Elliot, 2015).



Parent Frequency of Having Dinner with Children

Note: This figure plots parental frequency of sharing a meal with children in the household. The figure focuses on the impact of changes to the council tax benefit. The dinner frequency outcome is a binary outcome indicating households below and above the pre-treatment average number of shared meals per week.

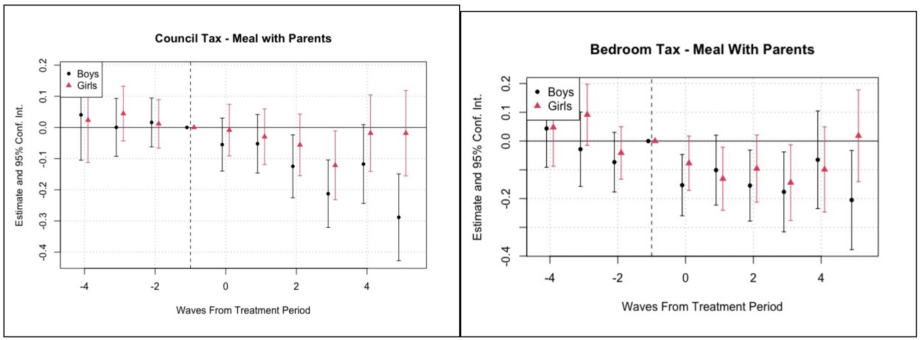
Beginning with the impact of the council tax treatment on the binary variable for frequency of sharing a meal with their children, shown in figure 2.4 , I can see that overall, the pre-treated periods are not suggestive of differential trends, which offers some support for the parallel trends assumption. However, this evidence is inherently suggestive rather than conclusive. The effect seems to be stronger and more significant for fathers, specifically in the long run. While being careful to point out that these results do not show significant difference in the post-treatment period, compared to the pre-treatment period they do still provide some evidence to show a reduction in the frequency of father child interactions.

The divergent trends in housework hours, shown in figure 2.3 can be interpreted in a similar manner to the dinner frequency results displayed in figure 2.4. Here, both father and mother pre-treatment estimates are centered around zero, providing suggestive evidence in support of the common trends assumption during the pre-treatment period. Despite no periods of significant changes from the baseline period, the estimates are consistently lower in the post-treatment period. When observing these results as a whole, the trajectory for trends in housework are suggestive of a negative impact of austerity on housework hours.

When considering results for proxy measures of parent investment and household engagement as a whole, despite not identifying significant and large negative divergent trends, the results are suggestive of some negative impact of austerity policies on parental and child interaction. As discussed further in Section 2.6, I should consider survey limitations when exploring results on parental interaction and relatively lower frequency of parental interaction and household engagement present in the treated group during the pre-trend period. Due to the lack of granularity in the discrete frequency variables used to proxy parental interaction, I can only capture certain specific movements in the level of parental interaction. For instance, a common response to the dinner frequency variable notes that parents may eat with children 3-5 times per week, therefore I miss reductions of parental meals between these frequencies, this effect is further compounded by the fact that many treated individuals respond in the second lowest category of parental interaction in the pre-treatment period.

Evidence from the Youth File. The previous section shows that the WRA led to significant increases in parental employment, a reduction in unemployment, and an increase in the number of hours that parents work. The previous chapter also provides some evidence of a reduction in household engagement, as mothers and fathers seem to be decreasing the amount of time they spend doing housework, results also suggest that fathers in the treated group may be decreasing the number of weekly meals they have with children, especially in later post-treatment periods. While section 2.5.3 gives us some insight into the impact of austerity policies on parental time investment, testing observations from the child's perspective allows us to more directly capture how children's lives changed as a consequence of austerity. This section addresses the effect on parent child interactions from the child's point of view. These results are subsetted by gender. There is a wide body of literature outlining gendered differences in educational development (Beaman, Wheldall, and Kemp, 2007; Entwisle, Alexander, and Olson, 2007) and importantly Bertrand and Pan (2013) found evidence to support that boys are more responsive to being in "broken families" when compared to girls, specifically related to the development of non-cognitive skillsets. This is particularly relevant when considering my findings related to boy peer relationships, where boys seem to be struggling to make and keep friends post-austerity.

Figure 2.5: Boys and Girls Dinner with Parents



Note: Figure 2.5 plots the estimated coefficients of boys and girls having an average, or above average number of meals with their parents per week. The plots cover the effect of the council tax benefit abolishment and the bedroom tax benefit changes.

Primarily for boys, I can see that children who are members of households that were exposed to austerity policies are significantly moving towards eating less than 5 meals per week with their parents. I can see that the pre-treatment period for boys and girls displays no diverging trends for boys exposed to the council tax abolishment, with post-treatment period coefficients consistently declining, suggesting that there is a causal relationship between the introduction of the WRA and decreasing frequency of sharing a meal with parents. The effect is not as strong for girls with the estimates being less precise.

In Figure 2.5 the estimates for boys are noticeably stronger and more precise than the estimates for girls. Recent literature provides some insights into this, with Choe and Yu (2022) reporting that adolescent boys are more likely to suffer negative effects, such as deteriorating relationships and increased isolation, in response to neglect, when compared to girls. One stark limitation of using the UKHLS for exploring child outcomes is that only children over the age of 10 are observed in their own survey, during ages in which parental interaction inputs play less of a role than in younger age groups. The possible impact of reductions in welfare income on child non-cognitive behaviors, such as social interaction and trust, warrants further in-depth investigation to fully understand and address the potential long-term consequences. It is important to emphasize that these effects were identified in spite of data limitations. The youth file included in the UKHLS lacks granularity and precision, with many missing observations for the SDQ outcomes discussed above. This is combined with pre-existing skew in the data towards high SDQ scores in the pre-treatment period for the treatment group. Here, boys in the treatment group have on average higher SDQ scores in the pre-treatment period, this makes identifying incremental increases in poor outcomes difficult.

2.5.4 Sensitivity and Heterogeneity

This section seeks to outline a number of sensitivity and robustness checks which have been used to test the model assumptions. Here I focus on the main findings, discussed in 2.5.1. The placebo section outlines and discusses a series of placebo tests used to test the sensitivity of the results. The heterogeneity section explores the topic of treatment effect heterogeneity and provides insight into sources of treatment effect variation.

Placebo Test. In this section, I conduct a series of placebo tests to assess the validity of the common trends assumption and to check whether the estimated effects could be driven by underlying trends unrelated to austerity. I change the treatment definitions by adjusting the treatment timing to periods in which austerity policies had not yet been passed. I preform the sensitivity analyses based on a set number of quarters removed from the treatment period in increments of 3, 4, 8, & 12 quarters removed from the treatment period. For instance, since the WRA was passed in April of 2013 (quarter 2 of 2013) the first round of sensitivity analysis results test the specified outcome with an adjusted treatment exposure time 3 quarters prior, in quarter 3 of 2012. Note that the reference quarter displayed in the results will also be adjusted in accordance with the quarter change increments explained above. Further sensitivity analyses follow this logic with 4 quarters prior to the treatment exposure identified as quarter 2 of 2012, etcetera. These tests are informed and follow the guidelines for sensitivity analyses outlined in (Rambachan and Roth, 2020). Results for the sensitivity analyses can be found in the appendix A.2.

I preform the sensitivity tests outlined above on important labour force participation outcomes outlined in section 2.5. The analysis was performed for each treatment separately, i.e., the council tax and the benefit tax and split by male and female. The results stemming from the sensitivity analyses show that the alternative treatment date has very little to no effect on diverging trends post treatment. I can see that in the alternative treatment exposure periods of 3, 4, 6, 8 and 12 quarters before April 2013, estimated coefficients and confidence intervals show very little change or sensitivity to the "fake treatment date." One notable result that should be considered is displayed in figure A.1 showing the placebo test for father employment. While the coefficients aren't statistically significant from zero, the last few quarters show a slight upward trend in the coefficients, but not enough to cause great concern as these effects do not carry into later waves.

This series of sensitivity analysis provides plausible evidence to suggest that parallel trend violations or potential pre-trend confounding variables were likely mitigated in the specification. As I have shown in the appendix section A.2 alternative treatment definitions show no significant changes to the estimated outcomes.

Treatment Effect Heterogeneity and Sources of Treatment Variability. I also performed a number of checks to probe treatment effect heterogeneity through various controls. Along with the parallel trend assumption, the possibility of heterogeneous treatment effects across groups is important to consider, as it can complicate the interpretation of average treatment effects in standard DID models (De Chaisemartin and D'Haultfoeuille, 2022). However, certain features of the WRA's implementation make standard DID estimation more robust to treatment effect heterogeneity. De Chaisemartin and D'Haultfoeuille (2022) outline a set of conditions under which variation in treatment effects across groups does not bias the estimated average treatment effect on the treated (ATT). My specification which outlines the impact of the WRA on parent and child outcomes meets these three pronged sets of criteria being (i) treated individuals can only be assigned to the treatment group and this treatment status cannot change, (ii) the treatment is binary and (iii) there is no variation in treatment timing. This means that despite within-group differences, estimates of the average treatment effect on the treated group are robust to heterogeneous effects. Still, I have completed a number of additional analyses on specific sub-groups within the population to clarify the effects of austerity.

Heterogeneity Analysis by Subgroups. I probe into the heterogeneity of treatment effects on different demographic characteristics, such age sex, age, and ethnicity. First, I tested the combined effect of the WRA on mothers and fathers together, this notes a different approach from the gender specific estimates displayed in the results section 2.5. Figure A.4 in the appendix shows these main labour force and earnings results on parents, which are not sub-grouped by gender, and include a gender specific covariate. I can see that these results are very similar to the ones displayed in section 2.5 with the exception of a few outcomes. Labour force participation variables, including employment, unemployment, and working hours show the same smooth time trends shown in the main specification findings. Results displaying austerity's impact on income mirrors the effect on women shown in the main specification but differs from the male only subset results.

I also consider the role that ethnicity plays in the effects of austerity and in particular the differences between white and non-white survey participants. This angle was informed by a recent paper, Pilkauskas et al. (2022), that highlights some important differences on the effect of income changes for white vs. non-white subsets of the population, where they find the effect of income losses due to universal credit were significantly stronger for black families, when compared to white or Hispanic families. When estimating the effect of austerity on white vs non-white observations, the results are similar to those found in the main specification but are consistently lower, also notable, those results are not significantly different from the white subgroup, see Figure A.7.

Next, I examine the effects of austerity based on household indicators for low socio-economic status by testing sample subsets of household members below and above the sample median income in the pre-treatment period.[23] I again see that using these subsets have little effect on the overall impact of austerity policies. When comparing the two groups, the above-median group's employment trends do not look to be increasing at the same rate as the below median subset. This is to be expected, as those with above median income were more likely to be employed in the pre-treatment periods. Similarly, the above-median income group has a much smaller sample size; $\sim$80% of treated individuals within the sample that fall in the "below median" SEC category, see Figure A.8.

Propensity Score Weighting. As noted earlier, in section 2.3, I assess the comparability of treatment and control groups using both pre-treatment balance checks and a propensity score weighting approach. While the main DID specification already accounts for time-invariant group differences, this re-weighting serves as a functional form robustness check. Propensity scores estimate the likelihood of being assigned to the treatment group based on observed characteristics (Olmos and Govindasamy, 2015), and allow for re-weighting the control sample to more closely resemble the treated group. This helps address potential imbalances in observable covariates that might otherwise lead to confounding (Austin, 2011).

The selection model includes all covariates from the main specification, along with additional controls such as education, income, and employment. Estimated scores are converted to weights and applied to the main specification. Results from this re-weighted model remain closely aligned with those from the unweighted model, suggesting that the findings are not sensitive to the method of covariate adjustment. The full set of coefficient plots from this robustness check is shown in Figure A.9.

Rumbachan & Roth's (2023) Check. As an additional robustness check, I applied the sensitivity analysis outlined in Rambachan and Roth (2023) to the main difference-in-differences results. This method relaxes the standard parallel trends assumption by allowing for limited deviations in trends across groups. Specifically, it assumes that the size of any deviation in the post-treatment period does not exceed a multiple (denoted as $\bar{M}$) of the largest observed deviation between adjacent pre-treatment periods. In this context, $\bar{M}$ serves as a bound on the extent of potential violations of the parallel trends assumption.

Rather than producing a single point estimate under the strict assumption of parallel trends, this approach yields a set of plausible treatment effects (a partially identified region) based on different levels of $\bar{M}$. By examining how these intervals change as $\bar{M}$ increases, one can assess how sensitive the conclusions are to potential violations. If the confidence intervals for a given $\bar{M}$ remain bounded away from zero, it indicates that the treatment effect remains statistically distinguishable from zero even under

[23] Here, binary variables for below and above pre-treatment medians were created

some degree of trend misspecification. I report these robustness intervals to evaluate whether my main results hold under increasingly relaxed assumptions.

The results of this test can be found in Section A.5. For most outcomes, the test demonstrates that the effects remain robust under modest assumption violations, particularly for smaller *M-bar* values (i.e., 0.5 or 1.0). Confidence intervals in these cases are consistent and significant, meaning that the estimated effects remain stable and unlikely to be driven by minor assumption violations. When examining outcomes like mothers' paid employment and father unemployment, the results show strong robustness under smaller violations. Confidence intervals remain stable and significant, suggesting that the treatment had a meaningful impact even when allowing for slight deviations from strict parallel trends. Similarly, for outcomes such as mothers' marital status and labor earnings, robustness is observed under lower *M-bar* values. However, sensitivity increases as deviations grow larger, with confidence intervals eventually including zero.

In contrast, outcomes related to total income (for both fathers and mothers) and father labor earnings exhibit greater sensitivity to larger *M-bar* values. While the effects are detectable under smaller deviations, allowing for larger violations leads to significant widening of confidence intervals, often including zero. These patterns underscore the challenges of maintaining robust causal interpretations for income-related outcomes under more relaxed assumptions.

Overall, the sensitivity analyses shows that most treatment effects hold up well under stricter assumptions about parallel trends. However, the results for mothers' marital status and labor earnings stand out as less stable. When allowing for larger deviations from the parallel trends assumption, the confidence intervals for these outcomes widen enough to include zero, which suggests the effects might not be as reliable under more relaxed assumptions.

While smaller deviations align with what we'd expect based on economic intuition, the sensitivity in these specific outcomes is a reminder to interpret them with some caution. Even though this doesn't take away from the main conclusions about the treatment's effects overall, it's worth noting that these particular results for mothers might need a closer look in future analyses.

2.6 Discussion

This section seeks to provide a detailed discussion, to complement the results previously displayed in section 2.5. In approaching questions about the impact of austerity policies on employment outcomes and the wider social impact of these policies, the approach was two-fold. First, I began with an inquiry into the intended effects of the policy. Following this, I then used these findings to inform the analyses that outline some of the unintended effects associated with austerity, including the impact that the policy had on wider social dynamics.

The analysis on the intended effects of austerity policies further support the growing body of evidence showing that austerity policies have been successful in driving employment and increasing work hours (Alesina, Favero, and Giavazzi, 2019; Fetzer, 2019). Here, I also find that austerity led to significant measurable increases in employment and a reduction in unemployment. While this general story aligns with that of the previous literature, these results provide insight into the effect of austerity on parents and presents new evidence for this demographic.

On the topic of the divergent trends in total incomes by gender, shown in Panel D of Figure 2.1, the effect can can be viewed as a direct consequence of diminishing benefit incomes. Leading up to austerity, mothers significantly benefited from various social programs designed to support children which were subsequently cut. For example, In the UK context, austerity led to cuts in child benefits and a reduction in pregnancy-related grants intended for mothers (Adam and Browne, 2013; Karamessini and Rubery, 2013; Perugini, Žarković Rakić, and Vladisavljević, 2019). With this in mind, it makes sense that when those benefits are reduced, mothers, have more difficulty replacing their lost benefit incomes. While this paper assigns treatment based on benefit receipt at the household level, this potential explanation is particularly relevant for single mothers.

In addition to diverging income trends, austerity measures appear to have also influenced mothers' decisions on marriage and divorce. The unintended effects of austerity measures could have led to significant disruptions in household dynamics, primarily through the mechanisms of bargaining power and economic independence. Where evidence shows that total income trends for men and women diverged under these measures, it is possible that women's bargaining position within households was affected. This effect has can have implications for women's decision-making capabilities, particularly impacting their freedom to end marital relationships. Simultaneously, austerity measures undermined women's economic independence.[24] This loss of financial autonomy pushed women into a precarious position. As their economic independence declined, women may have been more likely to enter relationships for economic stability rather than personal preference.

Finally, this research also unveils another interesting result that indicates that austerity resulted in a decline in the frequency of housework undertaken by parents. As introduced in section 2.6, figure 2.3 suggests that while this effect is true for both mothers and fathers. I can view this result as a confirmation of the results I previously discussed. When parents are increasing their employment along with the hours they spend at work, this can act as a time constraint, leaving less time for other activities (Lenhart, 2019), including household activities such as household chores and parenting (Carlson, Petts, and Pepin, 2021; Crandall-Hollick, 2018). When viewing figure 2.3 it is important to consider the gender discrepancy of time spent doing housework

[24] See Figure 2.1

in the pre-treatment period. In the pre-treatment period mothers, on average, spent approximately 16 hours per week on housework, significantly more than fathers, who only contribute around 7 hours weekly. In 2.3, we see similar incremental changes in the frequency of doing housework (of around 20-30 ppts), both mother and fathers, meaning that although both genders are spending less time doing household chores than before, women are still shouldering more of the burden post-treatment, despite increasing their labour market activities.

In reference to section 2.5.3 I also examine how austerity can influence children's lives. My inquiry into the effect of austerity on children stems from the evidence showing how austerity policies have pushed parents into the workforce, effectively acting as a time constraint. Previous literature on parental time use notes that there are roughly four ways that parents can choose allocate their time: employment activities, leisure, housework, and parenting (Carlson, Petts, and Pepin, 2021). With this in mind, I would expect that the increase in working hours displayed in Figure 2.1 lead to a reassessment of the ways that parents allocate their time. It's clear that understanding the impact of austerity measures on parental interaction is notable because of the significant role these interactions play in child development and wellbeing. A reduction in these interactions due to increased parental work hours can potentially lead to long-term negative effects on a child's educational outcomes, emotional health, and overall life prospects (Cunha and Heckman, 2007; Doepke, Sorrenti, and Zilibotti, 2019).

In Figure 2.5 we show that austerity has caused a decrease in the quantity of time that they spend with children, as shown by the reduction in parents sharing meals with children. It can be argued that these estimates should be seen as a conservative estimate. For instance, despite the high levels of unemployment, those who were identified as benefit recipients in the pre-treatment period consistently responded to having lower household interaction in the pre-treatment period. I used a simple two-way t-test to compare the pre-treatment responses between the treatment and untreated groups and found that untreated individuals shared an average of six meals per week with their children, as compared to three to five meals per week in the treatment group ($p < 0.05$). This has implications for estimating the effect on this outcome i.e., if parents in the pre-treated period were having dinner five times a week with their children, and subsequently reduce this frequency to three times a week, the observed frequency of dinners in the survey would be the same due to the low granularity of the data. Meaning that, these results only capture families that reduce their dinners from 3-5 times per week to 2 times per week, barring those parents that do not share meals at all. This is a theme that also had a considerable impact on the GHQ results, where results on mental health outcomes (including GHQ scores) were imprecise and unclear. These results have important implications for children, as several papers in the child development literature solely focus on the importance of family meals, finding that regular family meals can lead to improved child wellbeing, including better peer relationships (Horning et al., 2017; Mishra et al., 2021; Price, Rodgers, and Wikle, 2021). Along with this, my evidence on reductions in time spent doing housework can be viewed as a proxy for parental interactions (Norman and Elliot, 2015). With more detailed data on parent-child interaction, I could further explore not only the quantity, but also the quality of parent-child interaction.

Outcomes observations from the youth file provide even more compelling and substantial evidence for this decrease in parental interaction. Figure 2.4 outlines the results from the child's perspective where the post-austerity estimates consistently point to declining frequency in sharing meals with their parents. Once again, with children in the youth sample having on average a lower frequency in the number of meals and the frequency of talking with their parents in the pre-treatment period, it can be argued that these parental interaction estimates could be considered lower bound estimates. This issue effectively caps the available range in which I can capture downwards trends in measured parental interaction.

I explored a wide range of child outcomes to assess the impact of austerity on children. Notably, I explored the effect of austerity on risk behavior, mental health, and peer relationships.[25] While some of these results seemingly point to a decline in peer relationships for boys, ultimately the results were too imprecise with wide confidence intervals.[26] This theme raises an important question: given the observed changes in household dynamics and parental interaction, why is there not more conclusive evidence of the WRA's impact on child outcomes?

One explanation lies in the literature on child development, which highlights the tradeoff between time investments and financial resources. Increased labor supply may raise earnings but also reduces time available for parenting. The income effect suggests that higher earnings allow for greater investment in child development, while the substitution effect reflects the cost of reduced parental time (Agostinelli and Sorrenti, 2021; Pieters and Rawlings, 2020). In this context, however, I do not observe clear increases in household income despite higher employment rates, limiting the potential for greater monetary investment. Thus, the substitution effect may dominate, but without corresponding income gains, making it difficult to observe short-term developmental improvements.

That said, the evidence is not definitive. Some child outcomes such as peer relationship scores for boys display wide confidence intervals. While these are not statistically significant, the imprecision means I cannot rule out potentially meaningful effects. This underlines the limitations of the current analysis: although the point es-

[25] See Figures A.12 and A.13 in the Appendix for the effect of the WRA on overall SDQ scores for girls and boys.

[26] In total, I tested 21 youth related outcomes, these included the 5 component SDQ breakdown, outcome variables on relationships with peers and parents, and questions which relate to risky behavior. As explained, these were all insignificant with wide confidence intervals, meaning that the policy had no strong or significant effects. For brevity, these results are excluded from the paper.

timates suggest limited movement in most child outcomes, the uncertainty around some estimates prevents strong conclusions. My findings therefore reflect both the theoretical ambiguity in expected effects and the empirical constraints of the data.

In conclusion, this study highlights the potential for austerity measures to generate unintended consequences within households. While these policies were implemented with economic objectives in mind, the findings suggest they may have contributed to changes in familial structures and personal relationships, particularly for mothers. These patterns may reflect shifts in economic independence or bargaining dynamics within households, though I do not directly test these mechanisms. Although I do not find clear evidence of impacts on most child outcomes, this does not rule out the possibility of longer-term consequences—especially given observed changes in parental behavior and the potential for reduced parental interaction. This is particularly relevant for younger children, who may be more sensitive to changes in their caregiving environment. Given the uncertainty in some estimates and the potential for delayed effects, further research is needed to assess the longer-run implications of these policy changes.

Our research underscores the importance of carefully considering the full range of potential consequences, particularly the effects on household dynamics and child development when implementing significant policy changes such as austerity measures. I highlight the complex interplay between macroeconomic policy, personal economics, and social outcomes, reinforcing the notion that economic decisions are rarely isolated from the broader fabric of societal dynamics. By bringing these nuanced effects to light, I hope to contribute to a more holistic approach to policy evaluation and development, one that consider the intersectionality of economic, social, and familial factors.

2.7 Conclusion

In this paper I delve into the multifaceted implications of fiscal tightening policies on household dynamics, familial relationships, and family planning, by analyzing the effect of austerity policies in the UK. While a substantial library of literature exists on the impact of austerity policies on labour force participation and earnings, literature focusing on the wider social implications of fiscal tightening programs is sparse. The UK government's Austerity policies, and notably, the Welfare Reform Act of 2012 (WRA), provides an interesting application for this research, with advantageous characteristics for causal inference techniques, including binary treatment assignment and timing. I implement an event study model specification to capture the dynamic trend effects on parents and children who were exposed to austerity through the implementation of the WRA by making use of the Understanding Society Dataset, a comprehensive panel survey spanning from 2010 - 2019. I extend the event study model to study three main themes, beginning with the intended economic effects of austerity policies, followed by the estimated impact on family dynamics and finally, I research the impact on children by making use of the understanding society's youth survey. Here, Exposure to austerity is measured by being exposed to either the council tax benefit abolishment or the so called "bedroom tax." These two policies made up the bulk of reductions in social welfare benefits leading to a combined income reduction of around 1,100 GBP per year on average.

I find significant and considerable increases in the margin and intensity of labour market activities for both mothers and fathers, as parents seek to replace benefit incomes lost by austerity. With employment trends around 12 ppt higher in the post-treatment period, unemployment trends down at around 15 ppts and labour income rises considerably, over 50 ppts for fathers and mothers. We find suggestive evidence showing that effect on total income is different for mothers and fathers, with slight increases in total income, but some reductions in incomes for mothers, peaking at a 20 ppt total reduction. In addition to the economic effects, I find that austerity policies have had implications for household dynamics and familial relationships, where women in the treatment group significantly reduce the rate of divorce (by around 5 ppts) and increasingly enter into new cohabiting relationships and marriages ($\sim$10 ppt increase in post-treatment marriages). I recognize that these effects can largely be attributed to a single primary mechanism in each case. Results concerning a decrease in the number of divorces can be attributed to changes in household bargaining power as a result of the diverging trends in total incomes between women and men, where women feel trapped into relationships. Further, the results which outline increases in getting into relationships are likely driven by decreases in economic independence also associated with these divergent income trends. As their economic independence decreased, women were found to consider entering into relationships more for economic stability than personal preference. This shift in relationship dynamics underscores the far-reaching social implications of austerity policies, extending beyond their primary financial intent.

Furthermore, austerity resulted in a decline in the frequency of housework undertaken by parents. When parents are increasing their employment along with the hours they spend at work, this can act as a time constraint, leaving less time for other activities, including household activities such as household chores and parenting. This has led to significant decreases in household activities, such as sharing meals with children, and spending time at home doing housework, directly caused by the WRA.

In terms of the impact on children, I have shown evidence of well identified decreases in parental interaction, such as sharing meals with their children ($\sim$20 ppt decrease in family meals for boys). This decrease in interaction is a direct consequence of austerity policies that have pushed parents, into the workforce, effectively acting as a time constraint. Observations from the youth file confirm the results from the main file as these results are also indicative of a decrease in the frequency of family meals.

Parental interaction plays a crucial role in child development and is closely linked to the quality of peer relationships (Carneiro et al., 2015; Henney, 2016). However, these results do not indicate significant changes in child outcomes, suggesting that the decreased frequency of parental interaction may not have been substantial enough to produce measurable effects

In conclusion, this study reveals the some of the unintended consequences that Welfare Reform Act had on the intricate dynamics within households. We've shown that these policies, while designed with economic objectives in mind, led to significant disruption in familial structures and personal relationships. I also extend the analysis to examine impacts on children and parental interactions. While increased parental employment is observed, the findings suggest no significant increase in total household income. This may explain the limited short-term effects on children, as any potential benefits from employment are not accompanied by additional financial resources. The long-term consequences, particularly for younger children who are more sensitive to changes in their environments, remain an important area for further exploration. This research underscores the importance of carefully considering the full range of potential consequences, particularly the effects on household dynamics and child development, when implementing significant policy changes such as austerity measures.

Chapter 3

School Discipline Reforms and Their Impact in Oklahoma City

3.1 Introduction

Over the past two decades exclusionary discipline practices, such as out-of-school suspensions and alternative placements, have drawn significant criticism. Despite the growing evidence that deterrence-based discipline approaches in school settings can be linked to multiple negative outcomes, including declines in academic achievement (Anderson, Ritter, and Zamarro, 2019), increased dropout rates (Bennett, Contreras, and Cerda, 2022), students' feelings of alienation, and the widening achievement gap among minority students (Gopalan and Nelson, 2019; Gregory, Skiba, and Noguera, 2010), nearly 5% of U.S. public school students face out-of-school suspensions each year (De Brey et al., 2019). Within this context, compared to white students, African American students are nearly three times more likely to be suspended or expelled (Nowicki, 2018). Here, the consequences can be far reaching, with growing concern over the so-called "school-to-prison pipeline" (Skiba, Arredondo, and Williams, 2014). In response to this, policy makers have started to implement less punitive discipline approaches, which aim to emphasize support and inclusion over punishments like school suspensions.

Recent literature has linked less punitive discipline approaches to more inclusive school environments, reduced instances of school violence, and improvements in academic achievement (Craig and Martin, 2023; Gallego, Oreopoulos, and Spencer, 2023; Perry and Morris, 2014). On the other hand, this approach has been associated with negative externalities. A separate branch of the literature suggests that exclusionary discipline policies can benefit better performing students (Carrell and Hoekstra, 2010; Pope and Zuo, 2023) by protecting them from negative spillovers associated with students from less stable home environments. For example, Carrell and Hoekstra (2014), concluded that children from troubled families and those who have experienced domestic violence significantly decrease classroom reading and math test scores and increase incidents of school misconduct. Further, Pope and Zuo (2023) provides a framework for understanding how policies that lead to reductions in school suspensions can drive declining test scores. The paper suggests that without exclusionary discipline policies, the additional time spent on children who are more prone to misbehavior or need additional support can act as a time constraint for teachers[27]. It should be mentioned that these effects could be purely mechanical or compositional, where the stu-

[27] These findings suggest that teachers may spend less quality time with better performing students.

dents that would have been excluded had the policy not taken place might have lower test scores on average, leading to the estimated negative impacts that they find.[28]

We focus on a comprehensive discipline reform that occurred in Oklahoma City School District (OKCSD). Here, the policy intended to implement a shift away from zero-tolerance punitive measures in favor of promoting inclusion and student support. OKCSD implemented a package of policy interventions that were designed to foster positive school climates and address disproportionate exclusionary discipline practices for sets of marginalized students. Using administrative data from Oklahoma, this paper employs a Synthetic Difference in Differences (DID) approach to understand the policy effect on a range of student outcomes at the school level including rates of student suspensions, absenteeism, math, reading, and sciences test scores. Additionally, by leveraging the FBI's Uniform Crime Reporting (UCR) Arrests by Sex and Race (ASR) datasets, we explore the impact of these policies on crime.

Evaluating large, national level FBI crime data using SDID served as a challenge because control units outnumbered our single treated unit by 4000 observations.[29] Thus, informed by recent computer science literature (Agarwal, Shah, and Shen, 2020; Bayani, 2021; Kinn, 2018), we improved the comparability of our control group by implementing the K-means clustering algorithm described in Hartigan (1979). To do so, we used sub-county and metropolitan area census data to identify cities with characteristics comparable to Oklahoma City, to ensure the control group includes suitable comparisons. After applying the K-means clustering algorithm, we reduced the control group to 201 unique police agencies with complete data, improving the balance and relevance of control units and enhancing the stability of SDID unit and time weight estimation. This approach provides a straightforward solution to a novel problem, as the SC and SDID methods gain prominence.[30]

In the short term, OKCSD's reforms led to a nearly 50% reduction in total school suspensions, which serves as evidence that the district's shift away from punitive discipline policies was implemented as intended. However, the policy had mixed effects in other areas. I interpret this change primarily as a first-stage check confirming the policy was enacted and had immediate effects on disciplinary practices, rather than as a standalone behavioral outcome. Absenteeism rose by 4% and the effects on academic performance varied by subject and grade. After the reforms, 7th grade math scores rose by about 7%, as did 8th grade science scores. In contrast, 6th grade reading scores declined by around 11%. Our most notable findings provide evidence of significant spillover effects on adolescent lives outside of school, where we show total youth crime rates within the police agency covering OKCSD fell by nearly 23% as an effect

[28] This is a phenomenon outlined in Angrist (2014) in a discussion about the pitfalls of the researching peer effects.

[29] In setups with such large disparities between the number of treated and control units, synthetic control settings struggle with overfitting, see section 3.5.2 for an overview of this problem.

[30] This method only applies to our outcome on youth and adolescent arrests, as this is the only outcome that makes use of the FBI's ASR dataset.

of the policy. To support our findings, we provide a set of robustness checks that include SDID event study output and additional placebo tests. Amid ongoing debate in the literature regarding the effects of more inclusive discipline approaches, this paper demonstrates that inclusive school discipline approaches, such as those implemented in OKCSD, result in an immediate reduction in adolescent arrests.

3.2 Literature

We contribute to the ongoing discussion on the impact of inclusive discipline policies on school-level outcomes, as the limitations of exclusionary discipline techniques are increasingly recognized. These limitations include the trade-offs posed by potential negative spillovers in more inclusive school environments and the challenge of balancing effective discipline with fostering a supportive learning atmosphere. While some studies emphasize the benefits of limiting exclusionary discipline (Craig and Martin, 2023; Gallego, Oreopoulos, and Spencer, 2023), such as Lincove, Mata, and Cortes (2024), which examined Maryland's statewide ban on suspensions for young children and linked these changes with a reduction in overall infractions[31]. Others highlight unintended consequences associated with permissive discipline, including declining test scores and more disruptive classroom dynamics (Pope and Zuo, 2023). The findings in the literature remain fragmented, reflecting both promise and caution. Our paper addresses a critical gap by offering a comprehensive analysis that not only explores the complex interactions between reducing school exclusions, disciplinary practices, and student outcomes but also examines the broader implications for students' lives beyond the school environment, including arrests.

Here, where the direct routes of policy implementation are mostly untestable, we lean on the previous literature and economic theory to guide our discussion. While our effects on suspensions can be seen as a direct effect of the policy, which implemented several reforms which have previously been linked with lowering the suspension rates, including implementing Positive Behavioral Intervention Strategies (PBIS) and the hiring of school guidance counselors (Carrell and Hoekstra, 2014; Gage et al., 2020; Gallego, Oreopoulos, and Spencer, 2023), the effects on absenteeism, test scores, and crime require more nuanced analysis.

The significant increase in absenteeism could be due to several factors, including the idea that students who might have been suspended before now feel less discouraged from skipping class because the consequences are less severe.[32] Alternatively, these might be purely substitution effects, where students who would have previously faced suspensions might be substituting suspensions with absenteeism (*i.e.*, students disengaging from school but avoiding overt misbehavior).

[31] This study focuses on children in kindergarten through 2nd grade.

[32] This idea links back to Becker (1968) and his model on crime and incentives

We observe differential effects on school academic performance, and find evidence of declining reading scores but improvements in math scores. Here, some of these effects might be driven by shifting social and teaching dynamics within schools.

Given that OKCSD primarily targeted a reduction in suspensions, which disproportionately affect boys (Kothari et al., 2018a), a potential unintended consequence may be an increased focus on more structured subjects, such as math and science. This emphasis could come at the expense of less structured subjects, like reading, possibly due to classroom disruptions or challenges in maintaining student interest. The literature on the gender gap in math scores is helpful in explaining how social structures within the classroom can give comparative advantages to distinct groups of students (Breda, Jouini, and Napp, 2018; García-Echalar, Poblete, and Rau, 2024; Guiso et al., 2008). In OKCSD's case, where teachers were subjected to re-training programs, shifts in how teachers address classes might explain the observed gains in math performance. Where boys could have disproportionately benefited from the reform's focus on classroom inclusion, at the expense of classes with less structured teaching. This mechanism can be understood as complementary to the scenario in which reading scores drop due to class disruption. Still, we expect some of this composition effect to be cancelled out by the increasing rate of absences.

There are currently no papers that we are aware of that discuss the short-term effects of school suspension reduction and adolescent arrests. Still, there is some preexisting evidence on long term effects of exclusionary discipline policies. Wald and Losen (2003) finds that being subjected to school suspensions increases the probability of being arrested as an adult by up to as much as 20% Bacher-Hicks, Billings, and Deming (2019). Further, descriptive evidence shows that school suspensions are associated with crime outside of school (Wald and Losen, 2003). This paper is the first papers to use quasi-experimental and causal inference techniques to describe the impact of school discipline reform on youth crime. We suggest the effects on crime can be understood as either spillover-effects led by the benefits of creating a more inclusive and supportive educational framework or a purely mechanical effect, where adolescents kept in school do not commit crimes. This former idea is relates to Alan et al. (2021) which studies the effect of teacher re-training programs in Turkey and find that the policies had largely positive effects and led to a reduction in school violence, bullying, and social ethnic segregation among students. They suggest that these findings were primarily driven by more school cohesion

These results have important implications for both education and criminal justice policy. Here, there is little evidence on the mechanisms through which schools can have an influence on youth behavior, suggesting that policies aimed at fostering a more inclusive, supportive, and nurturing learning environment have the capacity to yield benefits that extend beyond academic achievements, shaping youth behavior and reducing adolescent involvement with the criminal justice system.

3.3 Institutional Background

The early 2010s marked a significant turning point in the approach to school discipline in the United States. Skyrocketing suspension rates sparked growing concern over the prospect of a 'school to prison pipeline' (STPP) (Skiba, Arredondo, and Williams, 2014). In short, the STPP refers to a framework describing policies and practices, particularly in school discipline and juvenile justice systems, that reduce students' academic success and increase their risk of negative life outcomes, including involvement in the juvenile justice system. Given these concerns, the trends in rising suspensions raised concerns about the fairness of disciplinary practices and their broader consequences, such as higher dropout rates and increased contact with the law (Mowen and Brent, 2016). Juvenile arrests have been linked with negative long-term outcomes, including lower economic attainment later in life, lower high school completion rates, and an increased probability of being arrested as an adult (Aizer and Doyle Jr, 2015; Siennick and Widdowson, 2020). In 2014, these concerns manifested in the form of the Department of Education and the Department of Justice's 'Dear Colleague' letter which aimed to address widespread racial disparities in student discipline practices. The letter positioned the growing problem as a potential Title VI violation under the Civil Rights Act of 1964, which prohibits race, color, or national origin discrimination in federally funded programs. The letter included a series of key recommendations to address the STPP by promoting safe, inclusive, and positive learning environments. Following the letter, the Office for Civil Rights (OCR) within the U.S. Department of Education began a formal inquiry to assess the scope of the problem.

3.3.1 Civil Rights Project Report

In this context, OKCSD found itself at a crossroads. In 2015, the UCLA's Civil Rights Project released a study Losen et al. (2015), that offered compelling descriptive evidence of widening disparities in student discipline and academic achievement across the 2000s to the early 2010s. The analysis showed a steady increase in the use of suspensions over this period. The report also provided an analysis on the growing discrepancies in the use of suspensions by race, where the disciplinary gap between white and black student suspensions was shown to be around 7% in Oklahoma.[33]

OKCSD was mentioned in the report numerous times and was found to have the 10th highest student suspension rate in the country, out of all 13,300 school districts, at 45.2%.[34] In the appendix, school districts problematic suspensions were outlined. Here they show that OKCSD suspended 75% of all black male students at least once in the 2011-2012 school year. During this same time period, the ethnic discipline gap

[33] For reference, the highest disciplinary gap between white and black students was found in Missouri, at 12.5%.

[34] This statistic denotes the proportion of the total student body in OKCSD who were subject to out of school suspension (OSS) in 2011.

widened substantially for black children, by 12.6% for elementary school students and by 21.7% for middle and high school students between 2009 to 2012. In effect, the report identified OKCSD as a district with profound systemic issues, as evidenced by Black students being approximately 63% more likely to be suspended than their peers from other ethnic backgrounds.

3.3.2 OKCSD's Federal Investigation

In the lead-up to the UCLA Civil Rights Project's study, OKCSD was already being monitored by the U.S. Department of Education's Office for Civil Rights (OCR). The scrutiny began on February 19, 2014, when OCR received a complaint alleging discrimination on the basis of race and disability within OKCSD. The complaint initiated a formal investigation (OCR Docket 07141086), detailed in a letter to district officials. The allegations, though redacted for privacy, led to a thorough review. The review was conducted under Title VI of the Civil Rights Act of 1964, Section 504 of the Rehabilitation Act of 1973, and Title II of the Americans with Disabilities Act of 1990, which collectively prohibit discrimination based on race, color, national origin, and disability by recipients of federal financial assistance.

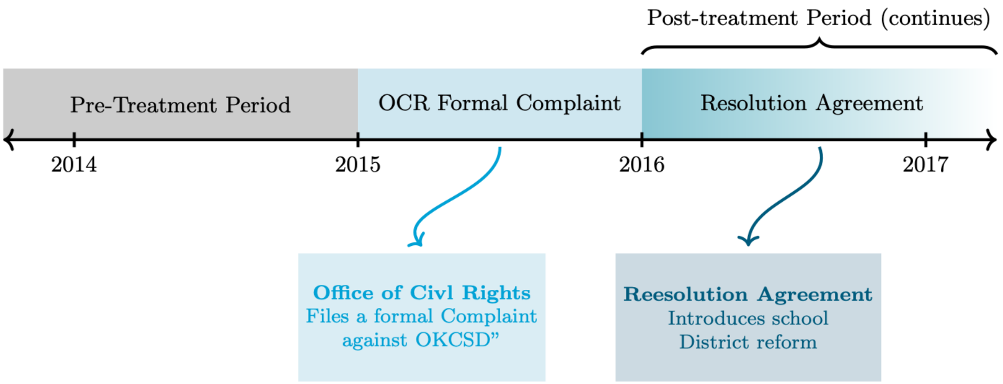
By 2015, with the release of the UCLA study highlighting severe disparities, the context and urgency for the OCR's investigation became even more pronounced. The investigation and the subsequent policy changes it catalyzed were pivotal, marking a significant effort by OKCSD to overhaul its discipline policies and practices to ensure a more equitable and inclusive educational environment. Rob Neu, OKCSD's superintendent at the time, stated

> "What we've found through this audit is that we've kind of relied on out-of-school suspension as our imbedded way of handling student discipline and that needs to stop... the national data suggests that if we don't change our trajectory for how we are handling these students, how we are engaging these students, then they're not only going to be dropouts, but they're likely going to be prematurely incarcerated, or dead" (Wendler, 2015).

This set the stage for the comprehensive policy changes analysed in the study, demonstrating OKCSD's attempt to rectify the systemic issues highlighted by both the OCR investigation and the UCLA report.

3.3.3 Timeline of Policy Changes

Figure 3.6: Timeline Of Events



Note: Figure 3.6 above provides a timeline of events leading up to OKCSD's resolution agreement with the Office of Civil Rights.

As a consequence of the 2014 investigation, on April 7, 2016, the Office for Civil Rights (OCR) and Oklahoma City Public Schools District (OKCSD) formally agreed to undertake significant reforms to ensure compliance in the form of a resolution agreement[35]. Key among these reforms was the adoption of Positive Behavioral Interventions and Supports (PBIS) aimed at fostering a more inclusive and equitable school environment. Additionally, the agreement mandated comprehensive retraining of staff to ensure disciplinary actions are fair and do not disproportionately impact students based on race. The agreement required OKCSD to revise their student code of conduct to more clearly define offenses and disciplinary procedures, thereby reducing subjective interpretation and ensuring consistent disciplinary action was taken across the district. Finally, the agreement stipulated that the district carry out staff level assessment of counselors and student support staff.

The primary changes outlined in the resolution agreement that comprise of the policy change which first occurred in the 2015-2016 school year include:

1. **The Adoption of Positive Behavioral Interventions and Supports (PBIS) strategies in OKCSD:** A cornerstone of the reform is the implementation of PBIS, an evidence-based framework designed to enhance school climate and culture.

[35] This agreement was legally binding and OKCSD was required to undertake the measures outlined below.

PBIS aims to reduce disciplinary incidents through proactive strategies for defining, teaching, and supporting appropriate student behaviors. By adopting PBIS, the Oklahoma City Public Schools District (OKCSD) commits to creating a more inclusive and supportive environment that promotes positive behavior and minimizes the need for disciplinary actions that remove students from the educational setting. In essence, the adoption of PBIS strategies created a structured approach to cultivating supportive classroom settings, where before implementation, OKCSD had no specific protocols or structures. As an example, the PBIS approach has a 4 step guide which begins with step 1 "creating positive teaching and learning environments," with 5 sub actions to implement the aforementioned step. One sub-step in this category, titled "Establish Positive Connections" gives a number of steps which must be taken to achieve this goal. These steps include "aligning and integrating family engagement in PBIS," suggesting a number of approaches including "home-visits" and a though exercise for how to better work with parents of in need children[36].

2. **Staff Re-training Carried Out by OKCSD and an External Advisor, “The Learner First”:** Recognizing the importance of staff competence in the successful implementation of PBIS and the new code of conduct, the agreement stipulates comprehensive retraining for teachers, administrators, and school aides. This retraining, supported by both OKCSD and the external educational services company“The Learner First” focuses on equipping staff with the skills necessary to administer discipline fairly and equitably, manage classrooms effectively, and apply de-escalation techniques. The training also emphasizes the importance of maintaining a safe and orderly educational environment that is conducive to learning for all students. Some notable examples of approaches which were outlined in the Learners first program include slogans such as "Equal treatment leads to unequal outcomes," and "Parents are by far the best experts of their own children, and we need their expertise to get this right."

3. **A Redefined Code of Conduct to Reduce Discretionary Suspensions:** To address the issue of discriminatory discipline practices and reduce the reliance on subjective judgment, the resolution agreement includes a comprehensive review and revision of the district’s Student Code of Conduct. The redefined code aims to eliminate vague and subjective categories of offenses, thereby reducing discretionary suspensions and ensuring a more objective and consistent approach to discipline. By clarifying definitions of misconduct, establishing clear criteria for disciplinary actions, and promoting alternatives to suspension or expulsion, OKCSD seeks to ensure that disciplinary practices are fair, equitable, and do not disproportionately affect students of any race or background. One notable ex-

[36] See more information on PBIS steps in (PBIS.org, 2020)

ample of the changes which took place revolve around attendance related discipline, prior to the policy changes, students which leave ”the classroom, assigned area, or campus without prior consent from appropriate school personnel.” were subject to school suspension, an updated 2016 code of conduct from the district states in bold that ”students may not be suspended for walking out of class, assigned area, or campus” [37].

4. **An Assessment of Staffing Levels of School Mental Health Workers:** Importantly, OKCSD committed to assess the staffing levels of school counselor's and administrators in each school. Following this, the district agreed to carry out staff hires that will be specifically tailored to the with the unique challenges facing specific schools. This is to ensure that all children in the district have access to mental health and social care resources within the school in order to decrease behavioral difficulties and increase the students' abilities to benefit from a learning environment.

As evidence of the magnitude of these changes, we reference quotes from OKCSD staff. Cherie Owen, a school district counselor, announced at a school board meeting in 2015 that she was changing her approach in:

> “learning to be positive in what we're doing... to be proactive... they're calling it pre-correction, where we actually teach them the skills before the negative behaviors happen. That's a complete shift from what we've been doing before,” (Willert, 2015).

Tori Bell, a long-time OKCSD school administrator teared up as she explained:

> “This is truly the most exciting thing I have seen the district do... We are taking time to see kids for who they are and what they need,” (Willert, 2015).

As we move to the data section, we examine the impacts of these reforms on the school district's disciplinary outcomes and school climate to better understand the effects of these policy changes.

3.4 Data

To investigate the effects of OKCSD's interventions on in-school outcomes we use administrative school data from Oklahoma. To study the impact of these measures on adolescent crime, we link Oklahoma's school data to the FBI's Uniform Crime Reporting (UCR) data by geographic area. Below in sections 3.4.1 and 3.4.2 we discuss these data in detail.

[37] See more information here Schools, 2016

3.4.1 School Administrative Data

Our primary data sources is comprised of annual administrative school report card data which was requested from The Oklahoma Department of education's Office of Education Quality & Accountability. We make use of all available data, with the exception of excluding Tulsa School District, from school year 2005-2006 to school year 2018-2019. The data is observed at the school level where we have roughly 1761 observed in each year from 514 school districts with an average enrollment of 645,845 students per year.

Tulsa School District is excluded because the district introduced its own discipline reform policies which took place a year after OKCSD in 2017. Where OKCSD's policies were federally required and in direct response to the OCR's complaint, Tulsa implemented reforms voluntarily, which made identifying formal documents outlining policy changes and implementation difficult. Due to this, our study focuses on the effect of OKCSD's policies and we exclude Tulsa from the analysis.

It's important to note that OKCSD stands out as an outlier in the Oklahoman context in several ways. It's the largest district in Oklahoma in terms of student enrollment, where in 2015 OKCSD accounted for 6% of the entire public school student population in Oklahoma.[38] Additionally, OKCSD stands out not only for its size, with 86 schools making it the largest district in terms of facilities, but also for its demographic composition. In 2015, African American students made up 36% of OKCSD's student body, marking one of the highest proportions of black students among Oklahoma school districts.[39]

In defining our outcomes, we articulate total suspensions as the aggregate average number of incidents of suspensions within a given school. The Oklahoma data categorizes school suspensions in two categories based on duration with short-term suspensions, which last fewer than 10 days, and long-term suspensions, which exceed 10 days. We analyse long suspensions, short suspensions, and their combined total as outcomes using these variables. Observations for long-term suspensions are relatively sparse as they make up only 8% of all suspensions. This is because long suspensions are typically reserved for severe disciplinary issues and often follow a short-term suspension for initial infractions. For absences, we use the department's definition in our outcome, which is the average days absent per student.

For our analysis on academic performance, the OKCSD report card data presents school level academic scores as a percentage of the student body that can be considered as having at least satisfactory test scores in the given grade and subject. We use

[38] In 2015, OKCSD had a total student enrollment of 43,190. The 6% proportion of total state enrollment is comparable to other years.

[39] OKCSD also has the 20^{th} lowest proportion of white students in Oklahoma, where only 20% of the student body is white.

Table 3.1:Descriptive Statistics

Variable Name	Median	Mean	Max
School Data			
Absences	9.1	9.4	59
Suspensions	8	29.2	1729
Grade 3 Reading	69	69	100
Grade 3 Math	68	67	100
Grade 4 Reading	76	71	100
Grade 4 Math	77	70	100
Adolescent Arrests	9	81	5064
Covariates (% of enrollment)			
School Enrollment	300	363	6496
Free School Lunch Eligible (%)	63	62	100
African American		8	100
Asian		2	40
Hispanic		10	98
Native American		20	100
White		58	100
UCR Crime Data			
Full Sample (n = 4001)		95.01	
K-means Sample (n = 341)		209.00	

Note: This table summarises descriptive statistics for the OKCSD dataset. The Outcomes Section covers academic performance, absences, and disciplinary actions, including suspensions (in-school and out-of-school) and adolescent arrests. The UCR Crime Data Section provides statistics on arrests based on different samples: the full sample and the reduced K-means sample. The Covariates Section includes school characteristics such as enrollment, proportion of students which are eligible for free lunches, and demographic composition (percentages). Statistics are based on the median, mean, and maximum values across all OKCSD schools during the study period, sourced from annual report cards.

this percentage based score as our primary outcome for grade level test results. It's important to note that for 6th grade test scores, data is only available until 2016, as these scores are marked as "data not shown" to adhere to data protection and privacy laws. Therefore we can only estimate the impact of the policy for the first post-treatment period of 2016 for 6th grade academic outcomes. In our analysis, we also include several key covariates to control for school-level dynamics and demographics. These include total school enrollment, ethnicity background information at the school level, and the proportion of students who are eligible for free lunch programs (see table 3.1).

In referencing the key policy changes in section 3.3, we do not have access to targeted survey data on teacher discipline strategies or student experiences, which limits our ability to directly assess the adoption of Positive Behavioral Interventions and

Supports (PBIS) strategies and staff retraining. However, we provide results showing that the district implemented these reforms by analyzing changes in staffing levels, specifically the number of school administrators and counselors. These reforms are captured using discrete variables that track the quantities of administrators and school counselors, serving as measurable indicators of the district's commitment to implementing these changes.

Next, we discuss the data we use to analyse the effect of OKCSD's discipline reform policies on adolescent crime.

3.4.2 Adolescent Crime Data

To analyse the impact of the OKCPSD's policy changes on adolescent crime, we utilize data from the Federal Bureau of Investigation's (FBI) Uniform Crime Reporting (UCR) program. The UCR compiles crime data on a national scale, gathering voluntary information from a broad spectrum of law enforcement agencies at the city, county, state and federal levels. The core of our analysis leverages the UCR's annual Age, Sex, and Race (ASR) release, a subset of the UCR. The ASR is particularly well suited to research the effect of OKCSD's policy changes on adolescent crime because of its granularity, which delineates crime statistics by age groups, with good coverage of instances of juvenile crime. The data also includes detailed geographic identifiers at the law enforcement agency (LEA) level.

LEAs are federal or state level entities that are authorized by law or government agencies to engage in prevention, detection, investigation, and prosecution of any law violations (Nigatu et al., n.d.). Examples of LEA's include city, state, county, and federal police and sheriff's offices. While participation in the UCR is voluntary, the dataset stands as the most comprehensive publicly available data on crime and coverage is representative of 97.4% of the population (DoJ, 2011). In our dataset, which includes juvenile records, we observe over 4,000 agencies across the United States from 2005 to 2019. The ASR records the count of specific crimes that occur in each age group with ages starting as young as 10. These include all types of crimes for children and adolescents aged 10 to 17. Since many of the jurisdictions that release juvenile crime data are subject to strict juvenile privacy requirements, detailed information about the type of crime committed is often redacted, this makes analyzing the effect of the policy on particular types of crime difficult.

Our analysis narrows its focus on to one LEA, the Oklahoma City Police Department (OKCPD). This allows us to examine the specific impact of OKCSD's policy changes within a defined geographic and administrative boundary. It is imperative to acknowledge that the jurisdiction of OKCPD extends beyond OKCSD's boundary, covering areas which are zoned to other school district as well. However, OKCSD is by far the largest school district in Oklahoma City and was the only district which saw descriptive trend shifts in suspensions. Further, placebo test on our school level outcomes such as suspensions and test scores show no effect for other school districts in Oklahoma city.

3.5 Empirical Method

In this section, we focus on our identification strategy to assess the causal impact of OKCSD's discipline reform policies. Our analysis utilizes a natural experiment arising from the landmark agreement between the OCR and OKCSD just before the 2016-2017 academic year, which sought to foster inclusive school environments and limit the use of exclusionary discipline practices.[40] To ascertain the impact of these policy changes, our analysis employs a Synthetic Difference-in-Differences (SDID) approach, comparing trends in OKCSD outcomes against those in a constructed counterfactual of similar districts that did not enlist reforms. Section 3.5.1 below presents the Synthetic Difference-in-Differences (SDID) approach in detail, including the rationale for its choice and its implementation.

3.5.1 Synthetic Difference-in-Differences

We apply the Synthetic Difference-in-Differences (SDID) model, outlined in Arkhangelsky et al. (2021), which integrates the conventional Difference-in-Differences (DID) framework with the flexibility and adaptability of synthetic control (SC) methods, proposed in Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010), to create a reliable counterfactual comparison for causal inference (Porreca, 2022). This is accomplished by re-weighting and matching on pre-treatment trends, resulting in estimates that are significantly more robust to the parallel trends assumption compared to traditional DID methods (Clarke et al., 2023). As with SC methods, the SDID approach is particularly well suited for policy settings in which there are few (or singular) treated units and many control units Arkhangelsky et al. (2021). SDID also has advantages over SC methods where instead of weighting a set of control units that minimize the average differences in levels between the synthetic and treated unit, in SDID it selects a weighted set of control units that minimizes the difference in trends in the pre-period. In effect, this allows the SDID to estimate the effect of policies in cases where the trends of treated and untreated units do not overlap (Dench, Pineda-Torres, and Myers, 2023).

The SDID method compares the pre- and post-treatment changes in outcomes between treated and untreated groups to infer causal effects by incorporating the synthetic control aspect that constructs a weighted combination of control units. This combination creates an 'artificial' control group that closely mirrors the trends of the treated group before the intervention. This helps address one of the key limitations of DID, the reliance on finding a naturally occurring control group that may not perfectly match the treated group in all pre-treatment characteristics. The SDID method

[40] For a detailed outline of the policies refer to section 3.3.3.

improves causal identification by combining unit and time weighting, allowing for better adjustment for confounding trends and unit-level heterogeneity.

We estimate the average treatment effect on the treated (ATT) as follows:

$$(\hat{\tau}^{sdid}, \hat{\mu}, \hat{\alpha}, \hat{\beta}) = \underset{\tau,\mu,\alpha,\beta}{\operatorname{argmin}} \Big\{ \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \hat{w}_i^{sdid} \hat{\lambda}_t^{sdid} \Big\} \tag{2}$$

Equation 2 above describes our optimization problem where we find the set of parameters $(\hat{\tau}^{sdid}, \hat{\mu}, \hat{\alpha}, \hat{\beta})$ that minimizes the weighted sum of squared differences between the observed outcomes Y_{it} and their predicted values. Here α_i and β_t represent unit and time fixed effects respectively, where W_{it} is a binary treatment indicator, and the term $W_{it}\tau$ captures the effect of treatment and is the main coefficient of interest. Specifically, τ represents the average treatment effect on the treated (ATT) for the treated units over the entire post-treatment period.

In our model specified for estimating the ATT, the synthetic element can be understood as the integration of unit weights (w_i^{sdid}) and time weights (λ_t^{sdid}). These weights both play important roles in constructing our synthetic control group.

Here, the unit weights (w_i^{sdid}) are determined in a way that minimizes the average squared difference in the trend between the observed treatment and a pool of potential donor units. In doing so, each control unit is assigned a weight to reflect its relevance and similarity to the treated units in terms of their outcome trajectories in the pre-treatment period. These weights subject to a regularization parameter, to prevent overfitting and to increase dispersion and uniqueness in weights, to prevent model overfitting.

The time weights (λ_t^{sdid}) are optimized to minimize the sum of squared differences between the time-weighted pre-treatment outcomes of the donor pool and the treated unit(s). These weights ensure that the pre-treatment periods with the most similar outcomes to the treatment group are given greater emphasis, allowing the synthetic control group to closely replicate the temporal patterns of the treated group before the intervention. This time-weighting component is a key innovation that distinguishes our SDID specification from traditional synthetic control methods.

As outlined in Arkhangelsky et al. (2021) there are several variance estimation methods which can be used to inform our estimated ATT. Following, Clarke et al. (2023) we rely on the placebo inference method to statically validate our ATT estimates.[41] The placebo-induced variance estimator, similar to traditional placebo tests in SC setups, offers a robust measure of treatment effect variability by simulating multiple scenarios in which control units are hypothetically treated. In our analysis, these

[41] Arkhangelsky et al. (2021) proposes three methods for statistical inference to be used under different assumptions: block bootstrap, block jackknife and block placebo methods. We justify the use of block placebo inference methods because of its suitability for policy settings with small number of treated units. This is reflected in Clarke et al. (2023), where they explain that the placebo inference approach should be used in instance with a small number of treated units.

placebo simulations were conducted 1,000 times. This approach ensures that our confidence intervals accurately reflect the variability of our outcomes based on our treatment. The placebo induced variance estimator, akin to traditional placebo tests in SC setups, assigns placebo treatment to untreated units. For each estimate, this method performs 1,000 placebo iterations, collects the these estimates, and uses their distribution to construct confidence intervals that accurately reflect the variability in outcomes attributable to treatment.

For school related outcomes we incorporate covariates into our model by using the optimization procedure outlined in Kranz (2022). We chose this method over the default "projected" method outlined in Clarke et al. (2023) because it has been shown to provide more precise causal estimates in a number of settings by using Monte-Carlo simulations. In Appendix Section B.5, we include results from an alternative specification that excludes covariates and show that these controls do not influence our findings in any significant way.

SDID results are relatively easy to assess for goodness of fit. Here, standard evaluation entails a visual inspection of the SDID trend with the treatment group trend. These plots help us visually assess the pre-treatment fit between treated and control units, which is especially important for Difference-in-Differences (DID) and synthetic control methods. In DID, the key identifying assumption is that treated and control units would have followed parallel trends in the absence of treatment. While SDID enhances this framework by re-weighting units and time periods to better match pre-treatment trends, it's important to recognize that this match is enforced mechanically and does not itself validate the identifying assumption. The core identifying assumption in SDID remains untestable and pertains to what would have happened in the post-treatment period in the absence of treatment. A separate critical assumption for SC methods, including SDID, is the necessity of a perfectly balanced panel. By integrating traditional Difference-in-Differences with synthetic control methods, SDID constructs a more accurate counterfactual by re-weighting and matching pre-treatment trends. However, this precision depends on achieving a perfect balance in the panel, where control units are weighted to closely match the treated units' pre-treatment trends. This meticulous matching is crucial for reliable policy effect estimations, making the balanced panel assumption central to SDID.

3.5.2 K-means Clustering for Donor Pool Selection

Recent computer science literature has addressed some limitations inherent to SC and SDID methods, namely in settings with high-dimensionality in the data, where synthetic control methods are prone to overfitting (Agarwal, Shah, and Shen, 2020; Bayani, 2021; Kinn, 2018). This complexity is not merely a function of high-dimensionality in terms of variables, but rather the difficulty in managing and meaningfully interpreting the vast array of untreated units compared to a single or few treated units.

Kinn (2018) notes that in cases with several hundred control units and small numbers of treated units, pre-estimation donor selection procedures should be used before using SC methods like the ones outlined in section 3.5.1 above. Here, our analysis on school level outcomes such as suspensions and test-scores, which have a much more reasonable ratio of treated to control units[42] does not necessitate the need for donor selection in the pre-treatment period. Therefore this pre-estimation data refinement procedure is only used in our analysis of the FBI's UCR data. In this specification, without donor selection, we have over 4,000 control units to be weighted to reflect one treated unit, OKC. Here, in addressing the problem of overfitting, we must carefully choose a method which is robust to under-fitted biased estimates. In light of these considerations, our approach incorporates K-means clustering as a pre-estimation donor selection procedure. Recent works described above have proposed K-means clustering as a suitable way to reduce data dimensionality (Agarwal, Shah, and Shen, 2020; Bayani, 2021). Papers which do not explicitly recommend K-means clustering, make the use of other similar machine learning processes (Kinn, 2018).

Following the methodology proposed in Hartigan (1979) and utilizing insights from Agarwal, Shah, and Shen (2020), we implement K-means clustering to refine the selection of our control units. Here because the included demographic data is not sufficient, we perform our K-means analysis using supplementary demographic information from control unit geographies from the American Community Survey (ACS).[43] We utilize K-means clustering over alternatives due to its relative simplicity and its widespread application in similar contexts, as it's arguably the simplest and most widely used unsupervised machine learning algorithm. The basic premise behind the method is that it defines data in terms of clusters so that the total intra-cluster variation (or total within-cluster variation) is minimized. In essence, the method partitions observations into k distinct clusters based on their attributes. The objective is to minimize the total intra-cluster variation, or the total within cluster sum of squares (TWSS), ensuring that the control units within each cluster are as similar as possible. By refining the synthetic donor control pool to only include control units which share the same k cluster as the treatment group, we can improve the performance of our SDID model.[44]

The total within-cluster sum of squares for a given single cluster C_k is defined as the sum of squared Euclidean distances (i.e. the"straight-line" distance between two points in space) between the data points x_i within the cluster and the cluster's centroid μ_k. This can be mathematically represented as:

[42] For OKCSD school data, we have 86 treatment units which are matched to 1,762 control units.

[43] The ACS was chosen for clustering due to its extensive size and comprehensive coverage of relevant demographic, social, and economic topics. As the largest annual household survey in the United States, the ACS samples approximately 3.5 million addresses annually, providing robust and representative data for such analyses.

[44] Characteristics we used to inform our cluster included a wide array of variables which were chosen based on model performance

$$W(C_k) = \sum_{x_{ik}} (x_i - \mu_k)^2 \tag{3}$$

Where: X_i represents an observation belonging to cluster C_K, μ_K is the centroid of cluster, calculated as the mean of all points assigned to C_k.

The objective function that K-means seeks to minimize across all clusters is the total within-cluster sum of squares (TWSS), which aggregates the intra-cluster variations C_k for all clusters can be represented as:

$$tot.withinss = \sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{4}$$

Here$W(C_k$ represents the within-cluster sum of squares for a single cluster, C_k and *tot.withinss* measures the total intra-cluster variation of all K clusters. The goal of K-means is to minimize this total sum, where on aggregate, points are as close to their respective cluster centroids μ_k as possible. It's important to note that in K-means algorithm, the number of clusters is user-defined and should be tuned. The standard solution for this is to use the elbow method, which is a plot of the within group sum of squares of distances from the centers, versus the cluster numbers and then choose the"elbow" of the curve. Elbow plots for our K-means results can be found in the Appendix Section B in Figure B.4.

We apply the algorithm outlined in Eq. 4 by utilizing data from American Community Survey (ACS) to identify clusters based on provided demographic information. We consider 14 variables to categorize these geographic towns and cities into groups. These variables include the percentages of population identifying as African American, Asian, Hispanic, or Native American, along with population densities, labor force participation rates, the percentage of the population lacking medical insurance, Siegel's Occupational Prestige Scores, the percentage of children attending public schools, average commute times, and the average socioeconomic index.[45] By applying clustering to identify relevant data for our SDID model, we significantly reduced our dataset from 4001 units to a more manageable 341. This approach not only enhances the efficiency of our model by focusing on the most comparable control units but also mitigates the risk of overfitting—a common challenge in high-dimensional data settings. By meticulously selecting control units that fall within the same clusters as the treatment group, we refine our synthetic control pool, thereby improving the fidelity and interpretability of our SDID model's findings.

[45] Socioeconomic Index follows the ACS's Hauser and Warren specification.

3.6 Results & Discussion

3.6.1 Descriptive Analysis

Before presenting the Synthetic Difference-in-Differences estimates, we provide some preliminary descriptive context for our analysis. Table 3.2 presents median trends in the pre- and post-treatment period outcomes by treatment group status. Here, the "treatment" column describes OKCSD's trends in outcomes with reference to the control units and excluding Tulsa. At a descriptive level, OKCSD's policies appear to have influenced the occurrence of school suspensions, as the median decreases by around 13% compared to an increase in suspensions of around 14% for the rest of the state. These figures also speak to the wide disparities between OKCSD and the state median, which was around 140% lower. Absences increased for both groups, with a slightly steeper trend observed for OKCSD.

Table 3.2Outcome Medians by Treatment Status in the Pre- and Post-Treatment Periods

Outcome Category	Treated			Untreated			Simple DID
	Pre	Post	(% Δ)	Pre	Post	(% Δ)	
Suspensions & Absences							
Short Suspensions	38	33	(-13%)	7	8	(+14%)	-6
Total Suspensions	39	34	(-13%)	8	9	(+14%)	-6
Absences	9.94	10.5	(+5%)	8.90	9.09	(+2%)	+0.37
Grades 6 to 8 Outcomes							
Grade 6 Math	63	24	(-61%)	78	43	(-44%)	-0.5
Grade 6 Reading	62	38	(-38%)	75.5	52	(-31%)	-0.5
Grade 7 Math	65	29	(-56%)	76	44	(-42%)	-4
Grade 7 Reading	69	36.5	(-47%)	72.4	46.9	(-36%)	-7
Grade 8 Math	52	19	(-64%)	69	30	(-57%)	5.5
Grade 8 Reading	68	39	(-42%)	80	50	(-37%)	1
Grade 8 Science	37	26	(-30%)	59	50	(-15%)	-2
Counselors	0.4	0.6	(50%)	0.58	0.57	(-1%)	+0.21
Arrests	4022	2682	(-33%)	50	21	(-58%)	-1131

Note: Pre-treatment and post-treatment medians are reported for both treated and untreated groups. Percentage change (% Δ) is calculated relative to the pre-treatment period. The Simple DID column represents the Difference-in-Differences estimate for each variable. Long Suspensions are excluded because the median value for both treatment and control units is zero. Arrest results are based off of the clustered sample.

We observe that median proficiency rates declined during this period for both the treatment and control groups. This coincides with the 2017 introduction of new state-level test assessments in Oklahoma, which were designed to adhere to national assessment standards (Felder, 2017)^{46}. While the reform was implemented statewide, its effects may not have been entirely uniform across districts. In particular, shifts in the mapping between observed test scores and underlying ability could lead to varying impacts depending on the distribution of student performance. For example, districts with a higher proportion of students near the new cutoff may have experienced more noticeable declines in measured proficiency.

With this in mind, OKCSD exhibited more pronounced shifts in school test scores during this period compared to untreated districts. Table 3.2 shows that the magnitude of changes varied by subject and grade level. For instance, grade 6 scores declined at a similar rate as those in the rest of Oklahoma, while declines in 7th and 8th grades were somewhat more pronounced in OKCSD. The median number of counselors per school was less than one in both groups; however, OKCSD's pre-treatment median was slightly lower but rose to match the post-treatment level.

The descriptive trends in arrests are also noteworthy and indicate a substantial decline in the total number of youth and adolescent arrests relative to the control. More specifically, arrests decreased from 4,022 to 2,682, representing a 33% reduction. While these patterns are striking, they warrant deeper investigation to assess whether they are attributable to the intervention or reflect broader trends or unobserved factors. We next conduct a more rigorous analysis to disentangle these effects and better understand the mechanisms at play.

3.6.2 The Effect on Suspensions & Absences

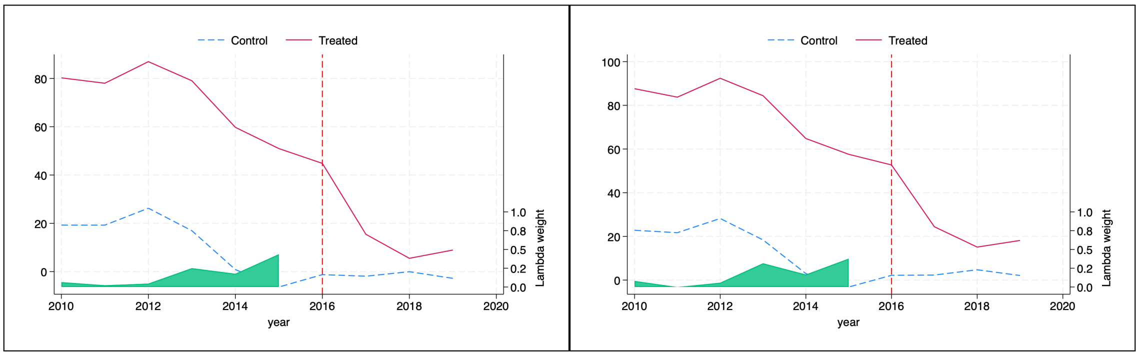
Suspensions. Figure 3.7 and Table 3.3 present our main findings on school-level outcomes from applying the SDID method outlined in Equation 7. This method weights pre-treatment control unit observations to mirror pre-treatment trends for OKCSD. The synthetic comparison serves as a counterfactual, estimating what the outcomes for OKCSD would have been in the absence of the policy reform. In the graphical output, the solid line represents the observed trend of the outcome for OKCSD, while the dotted line represents the synthetic counterfactual, composed of the weighted set of control districts throughout Oklahoma. The shaded area below these trends indicates the time weights,[47] and the vertical line marks the first treatment period, 2016. The results show that suspensions trended similarly in OKCSD and the weighted set of

[46] In 2016–2017, Oklahoma adopted new academic standards and replaced its previous state assessments with more rigorous tests aligned to college- and career-readiness benchmarks. This included the elimination of End-of-Instruction (EOI) exams and the introduction of new statewide assessments designed to emphasize critical thinking and problem-solving. While implemented statewide, the impact of these changes likely varied across districts depending on student performance distributions and instructional alignment.

[47] See Section 3.5.1 for more details.

Oklahoma school districts during the pre-treatment periods leading up to 2016, with time weights concentrated on pre-treatment periods near 2016.[48]

Figure 3.7: SDID Total and Short Suspensions



Note: Figure 3.7 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSD's discipline policies on total suspensions and short suspensions. Here panel A. outlines the impact on the combined effect of long and short suspensions, while panel b. presents short suspensions alone. Long suspension results can be found in figure B.1 in section B.2
[49]

Taken together, these results indicate a substantial decline in suspensions following OKCSD's discipline policy reforms, with the most pronounced change occurring in short suspensions, which fell by approximately 50%. Given the relatively low baseline rate of long suspensions across the sample, trends in total suspensions closely mirror those of short suspensions. Long suspensions also declined meaningfully, with an estimated effect of –2.37 relative to a pre-treatment mean of 1.8 (standard error 0.41).

While the estimated effects are large, it is worth noting that the structure of the outcome variable, as a count bounded at zero, may introduce challenges when treated and control units begin from very different baseline levels. In particular, treated units like OKCSD, starting from higher suspension rates, may have more scope to decline relative to synthetic controls, potentially contributing to post-treatment divergence. Although SDID matches on pre-treatment trends rather than levels, differences in available "distance to floor" could influence the estimated effect size. This possibility

[48] The time weights in the Synthetic Difference-in-Differences (SDID) analysis are concentrated around 2014, aligning with the increasing discourse surrounding the Office for Civil Rights (OCR) investigation that gained momentum in 2015. This pattern suggests that the SDID method effectively identifies and weights pre-treatment periods closely resembling the post-treatment context.

[49] In figure 3.7 the plot in panel A corresponds with an estimated coefficient of -39.06 significant at the 0.99 confidence level with a pre-treatment mean of 79.3. Panel B, sho rt suspension, corresponds with an estimated ATT of -38.37, with a p-value lower than 0.01 and a pre-treatment mean of 77.49.

highlights the importance of interpreting these results in the context of both statistical design and policy implementation.

Even so, the magnitude and consistency of the observed reductions provide compelling evidence that substantive changes to disciplinary practice did occur. While I cannot formally identify which components of the reforms were most impactful, the observed decline in suspensions serves as a first-stage validation that substantive changes to disciplinary practices occurred. The decline likely reflects a combination of intentional interventions designed to reshape the district's disciplinary framework. Key among these were direct steps to limit suspensions, such as re-defining the disciplinary codebook to narrow the scope of behaviors warranting exclusionary measures, and the adoption of Positive Behavioral Interventions and Supports (PBIS) strategies (Gage et al., 2020), which emphasize proactive approaches to improving school climate and reducing misbehavior. Along with this, OKCSD augmented these policy changes with structural support, including the hiring of additional school counselors. The expansion of counseling services may have supported these efforts by addressing the root causes of student behavior and fostering environments that reduce the need for punitive measures, but this is purely speculative. Existing literature supports the efficacy of these approaches, with PBIS strategies consistently associated with reductions in suspension rates (Gage et al., 2020), and substantial evidence highlighting the role of counselors in driving positive behavioral and academic outcomes (Carrell and Hoekstra, 2014; Osodo et al., 2016).

These findings underscore the comprehensive nature of the district's efforts, illustrating how coordinated policy and structural changes can yield meaningful improvements. Importantly, as shown in Section 3.6.4, the benefits of these reforms extended beyond the school setting, contributing to a measurable decline in adolescent arrests.

Absences. Figure 3.8 below plots the graphical SDID output for absences. Here, just as in the plots depicting suspensions, the synthetic counterfactual trends closely with OKCSD's actual absences in the pre-treatment period. We again see that time weights focus on 2014, just before OKCSD's trend slightly increases in 2015. With an estimated ATT of 0.43 and a pre-treatment mean of 11.4 absences, we find that the average number of days absent per student rose by around 4%.[50] As discussed previously, these effects could potentially be explained by Becker's theory on crime and incentives (Becker, 1968). The paper suggests that individuals weigh on the costs and benefits of their actions, with punishment serving as a deterrent to misbehavior. In the context of OKCSD, the adoption of more lenient disciplinary policies could have reduced the perceived consequences of absenteeism. This could effectively lower the cost of absenteeism, leading to an increase. Although data constraints make this theory difficult to test, the concept can generally be applied to all forms of misbehavior.

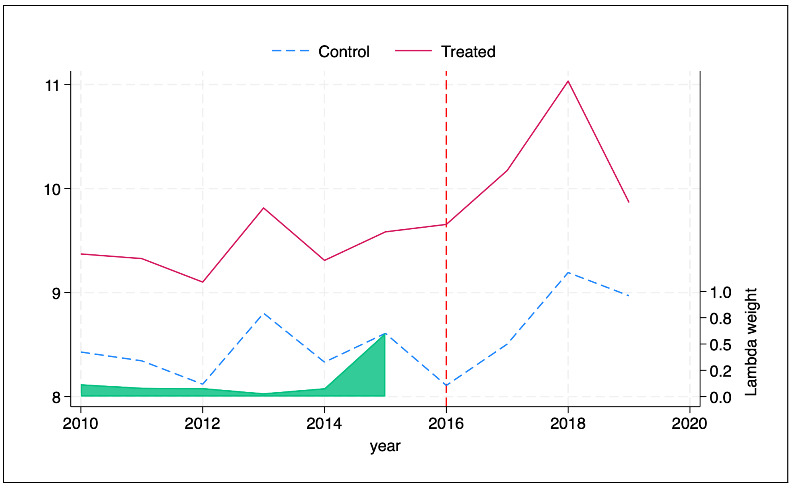
[50] See table 3.3

Table 3.3: **Estimated Treatment Effects on School-Related Outcomes and Arrests**

Outcome	**Estimate**	**Standard Error**	**P-value**
Suspensions & Absences			
Short Suspensions	-38.64***	(2.80)	¡ 0.01
Long Suspensions	-2.37***	(0.41)	¡ 0.01
Total Suspensions	-39.06***	(3.42)	¡ 0.01
Absences	0.43**	(0.20)	0.029
Grades 6 to 8			
Grade 6 Math	2.38	(3.88)	0.482
Grade 6 Reading	-6.59***	(2.73)	0.015
Grade 7 Math	7.24***	(2.89)	0.008
Grade 7 Reading	-1.85	(3.27)	0.580
Grade 8 Math	6.87*	(3.51)	0.044
Grade 8 Reading	0.23	(2.37)	0.973
Grade 8 Science	7.01***	(1.66)	¡0.01
Mechanisms			
Counselors	0.18***	(0.03)	¡ 0.01
Arrests			
Youth Arrests	-593.25***	(20.88)	¡ 0.01

Note: Standard errors are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table presents results from the main model. The data for academic and behavioral outcomes, including test scores and suspensions, are sourced from the Oklahoma School Report Card, which provides detailed school-level administrative records.

Figure 3.8: SDID Absences



Note: Figure 3.8 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSDs discipline policies on the number total number of absences at the school level. Data is from the Oklahoma School Report Card data. Covariates are included.

Conversely, this effect could stem from statistical artifacts rather than actual behavioral changes. In this context, the same students who would have been suspended under a no-reform policy environment might now be recorded as absent instead. Essentially, both suspension and absenteeism can be viewed as forms of student disengagement. This theme is thoroughly explored in Anderson (2021), a descriptive study that examines how students with disabilities respond to varying levels of classroom inclusivity and highlights a connection between absenteeism and suspension.[51]

In conclusion, while the district's move away from exclusionary discipline policies has successfully reduced suspension rates, the accompanying rise in absenteeism highlights the need for balanced strategies that address the root causes of student disengagement to mitigate unintended consequences. However, when evaluating the magnitude of both effects, the increase in absenteeism can be considered modest.

3.6.3 Adolescent Student Achievement

In this section, we outline the effects of OKCSD's policies on academic performance. Figures 3.9 and 3.10, along with Table 3.3, present our results on academic outcomes. Overall, we observe mixed effects on academic performance following the OKCSD discipline policy reforms.

Grade 6 reading scores declined significantly, with an estimated effect of -6.51 points, equivalent to an 11% decrease (or -0.276 standard deviations).[52] The effects on grade 7 and 8 reading scores is insignificant, with the point estimate being positive for 7th grade and negative for 8th grade. Statistically significant results for 7th grade math and 8th science outcomes suggest that students improved in these subjects. Grade 7 math competency rose significantly by 7.31 points (0.261 standard deviations), a 52% increase compared to a mean outcome of 14.26.[53] Additionally, Grade 8 science competency rates demonstrated the largest relative gain, with an increase of 7.01 points (0.320 standard deviations), corresponding to a 61% improvement, relative to the pre-treatment mean of 11.56. Although only Grade 6 reading scores decline significantly, the overall pattern in reading is mixed, with estimates varying across grades. This contrasts with the more consistently positive point estimates observed in math and science. This divergence aligns with the broader pattern in our results, suggesting that more structured subjects like math and science may have benefited more from the reforms. As discussed in Section 3.4, academic outcomes for 6th grade students were not reported in 2017 to adhere to data protection and privacy laws. As a result, the 6th-grade reading analysis includes only a single year of post-treatment data. This limits our ability to capture dynamic treatment effects that may unfold over multiple years. For example, if improvements in classroom behavior compound over

[51] See Gottfried (2014) for a discussion of the negative effects associated with chronic absenteeism.

[52] This result is significant at the 95% confidence level.

[53] While not conventionally significant, the increase in Grade 8 math scores by 6.87 points (0.285 standard deviations) would be significant at the 90% confidence level, representing a 50% improvement.

time, allowing teachers to build on prior-year gains, the effect observed in 6th-grade reading may understate the full impact of the policy relative to subjects and grades with longer post-treatment windows.

These findings may suggest that while the reforms introduced some challenges in areas such as reading, they were more conducive to supporting outcomes in structured subjects like math and science. The differential impacts underscore the complexity of the policy's effects on academic performance across various subjects. Importantly, the district's Code of Conduct specifies that students receiving out-of-school suspensions (OSS) are still expected to keep up with coursework and participate in standardized assessments, which helps mitigate concerns that changes in suspension policy may have altered the composition of test takers (Schools, 2016).

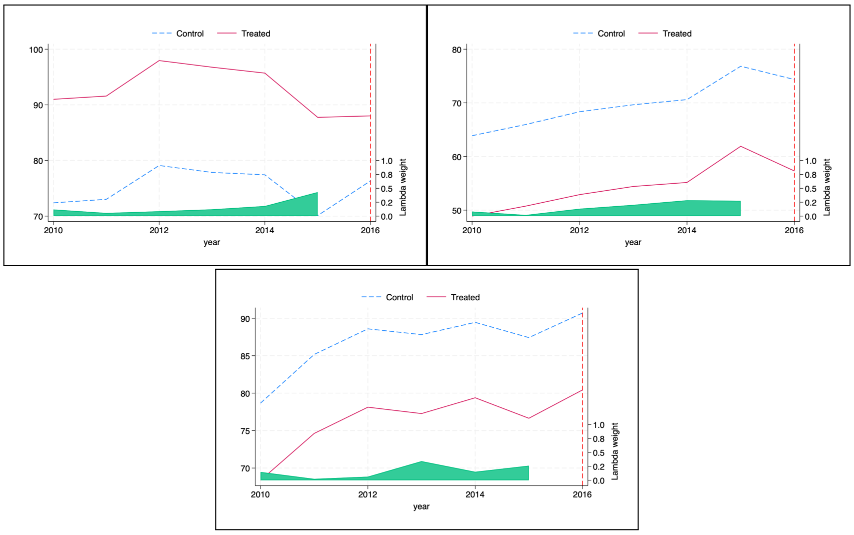
The variation in effects across subject areas, with reading showing mixed results and math showing more consistent gains, calls for closer examination. One possible explanation, though not directly testable with the data, relates to how OKCSD's policies may have interacted with gender dynamics in the classroom.

While we do not observe gender in our data, prior research has consistently shown that boys are more likely to be suspended and experience school discipline(Kothari et al., 2018a,b; NCES, 2019). In targeting the reduction of suspensions, it is possible that it may have provided immediate benefits to boys, such as keeping them in the classroom and addressing their behavioral challenges.

These changes could have interacted with subject-specific classroom dynamics—for example, potentially supporting more structured subjects like math while presenting different challenges in less structured environments like reading. However, given the mixed significance of the estimates, this interpretation remains tentative. This theory aligns closely with literature discussing the gender gap in math scores (Fryer Jr and Levitt, 2010; Guiso et al., 2008), which has been attributed to social structures that advantage boys (Breda, Jouini, and Napp, 2018; García-Echalar, Poblete, and Rau, 2024). A particularly relevant paper, (Goulas, Megalokonomou, and Zhang, 2023), exploits the random assignment of students to classrooms in Greece high schools and concludes that classrooms with a higher share of girls are associated with less disruptive behavior for boys and improved engagement from girls.[54] In OKCSD the opposite effect could have taken place, where classrooms with a higher proportion of boys, including those prone to disruptive behavior, could have reduced engagement among girls and contributed to lower reading test scores. While the decline in reading scores is not uniform across grades, it may reflect, in part, subtle shifts in classroom dynamics that differentially affect less structured subjects. Some literature has suggested that girls may have a comparative advantage in reading-oriented environments (Breda,

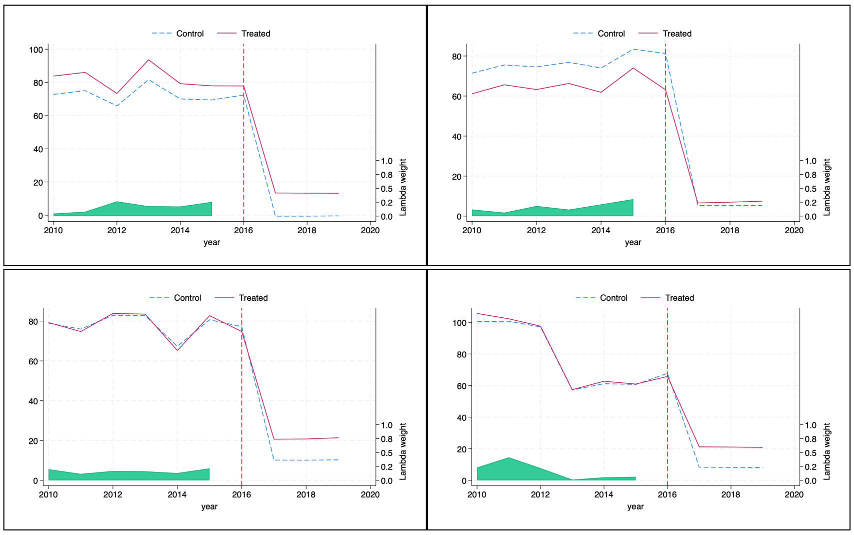
[54] The paper assumes aggregate randomness in classroom gender composition but potentially overlooks structural differences in sorting mechanisms across schools or cohorts, which could mask localised randomness and bias the results see Dieterle et al. (2015)

Figure 3.9: SDID Reading Scores



Note: Figure 3.9 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSDs discipline policies on reading proficiency for 6th, 7th, and 8th grade students begining with 6th grade reading in the upper left and ending with 8th grade reading results on the bottom.

Figure 3.10: SDID Math and Science



Note: Figure 3.10 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSDs discipline policies on math and science proficiency. Here panel A. outlines the impact on the combined effect on grade 6 math and panel b. presents results on grade 7 math, panel c, presents 8th grade math results, and panel d. presents 8th grade science scores. Data is from the Oklahoma School Report Card data. Covariates are included.

Jouini, and Napp, 2018).[55] This is further supported by García-Echalar, Poblete, and Rau (2024), a recent paper that uses administrative data from Chile to model teacher value added and concluded that teachers account for up to 18% of the math score variation by gender.

In the context of our findings, teacher-driven variation could also potentially explain the observed gains in math performance where boys may have disproportionately benefited from the reform's focus on classroom inclusion. For example, a study by Alan et al. (2021) looks at the impact of teacher re-training programs in Turkey designed to enhance social cohesion and reduce conflict within schools. The idea here is that in the case of OKCSD, it is possible that boys were better able to capitalize on the positive aspects of school inclusion and structured approaches to responding to disruption, thereby rising test scores in areas where they have a comparative advantage[56] in subjects like math and science.

It's worth mentioning that previous literature on exclusionary discipline policies has suggested that peer effects might lead to worse academic performance (Carrell and Hoekstra, 2010; Pope and Zuo, 2023), but in our context, some of this effect might be mitigated by changes in attendance patterns. While reduced suspensions increased the overall number of students in the classroom, it is possible that the most disruptive students, were more likely to be absent and drive the increase in absenteeism. This dynamic could lessen the negative peer effects often associated with reintegrating disruptive students.

Our results point to mixed academic impacts following the OKCSD policy reform, with some evidence of gains in structured subjects like math and science, and more variable outcomes in reading. While the analysis highlights potential mechanisms, such as changes in classroom dynamics and inclusivity, these explanations cannot be explicitly tested with the available data. This underscores the need for further research to better understand the factors driving these heterogeneous effects.

3.6.4 Adolescent Arrests

Figure 3.11 presents the SDID graph on the estimated impact of OKCSDs discipline reform policies on incidents of youth arrests. This outcome includes all youth crime and observations for children and adolescents aged 10 through 17. The results correspond with a significant 22.8% decline in youth arrests, where the estimated coefficient is -593 corresponding with a pre-treatment mean of 2602 arrests.[57] Our synthetic OKC youth arrest line tracks relatively well with the observed OKC data, but we do see a small spike in arrests for the OKC in 2015 leading up to the introduction of the policy.

[55] This interpretation remains speculative, given mixed statistical evidence across grades and the lack of gender-disaggregated data.

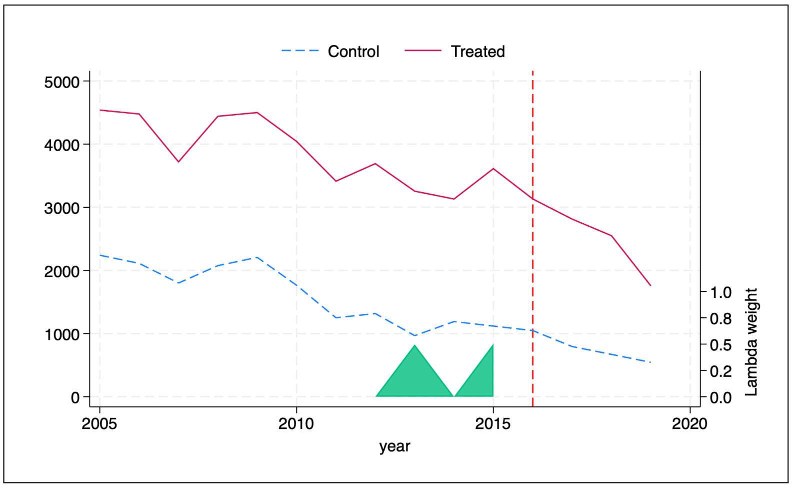
[56] See Dossi et al. (2021) for some evidence of why this advantage may exist

[57] See table 3.3

This story is reflected in our time weights, which focus near the treatment, but with a 1 period lag.

In Section 3.7, we present event study results for adolescent arrests, identifying some instances of poor matching in the pre-treatment period. Still, the findings provide strong evidence supporting that the policy led to a reduction in suspensions. Although the synthetic control tracks observed data closely in most periods, the pre-treatment trends in 2007 and 2014 show some degree of misalignment, suggesting imperfect matching in these periods. These discrepancies likely reflect inherent challenges in constructing a synthetic control group that perfectly mirrors the treated unit's dynamics in all pre-treatment periods. However, the substantial decline in arrests following the policy's introduction remains robust due to SDID's weighting mechanism, which minimizes the influence of periods with poor pre-treatment fit. While these results offer strong support for the policy's effectiveness, the presence of some pre-treatment mismatch underscores the need for cautious interpretation, especially regarding the estimated magnitude of the effects.

Figure 3.11: Adolescent Arrests



Note: Figure 3.11 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSDs discipline policies on adolescent arrests.

While previous studies have emphasized a link between school suspensions and an increased likelihood of long-term involvement in crime(Bacher-Hicks, Billings, and Deming, 2019; Wald and Losen, 2003),[58] we found no research demonstrating the converse effect, where a reduction in suspensions is associated with a decline in crime. Against this backdrop, our findings provide evidence that changes in student disciplinary policies can reduce adolescent arrest rates.

[58] Bacher-Hicks, Billings, and Deming (2019) found that suspended students are up to 20% more likely to face arrest as adults.

When taken together, the results in this section contribute to the existing literature by providing empirical evidence highlighting both the benefits and trade-offs of inclusive disciplinary policies. While these policies can foster improved school cohesion, decrease violence, and reduce adolescent crime rates (Craig and Martin, 2023; Gallego, Oreopoulos, and Spencer, 2023; Perry and Morris, 2014), they may also pose challenges, such as declines in academic performance (Carrell and Hoekstra, 2010; Pope and Zuo, 2023). This dual perspective underscores the importance of understanding how disciplinary reforms impact both immediate educational outcomes and broader societal issues.

The literature has established that certain student outcomes, including academic performance (Carrell and Hoekstra, 2010; Pope and Zuo, 2023), may be adversely affected by more permissive disciplinary policies.[59] While the full spectrum of unintended consequences remains to be thoroughly explored, the findings presented in this study suggest wide-ranging implications extending beyond school-level outcomes. By shifting away from exclusionary disciplinary practices that alienate students from the school environment, OKCSD's policies appear to create conditions that mitigate the likelihood of interactions with the criminal justice system.[60]

In doing so, OKCSD is taking proactive measures that may contribute to addressing the "school-to-prison pipeline." This recognition of the limitations inherent in traditional disciplinary approaches, coupled with the trade-offs associated with more inclusive school environments, underscores the need for comprehensive investigation into how schools can effectively balance discipline with student support. This study offers a significant contribution to this discourse by presenting a nuanced examination of the short-term effects of school inclusion policies. In addition to addressing immediate educational outcomes, our findings underscore the substantial impact these policies can have on student behavior beyond the classroom, particularly in relation to our results on adolescent crime rates.

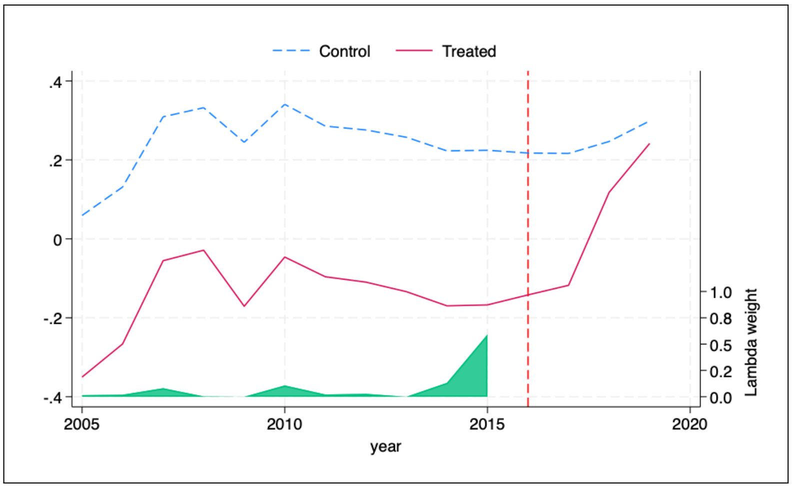
3.6.5 Mechanisms

Previously, we introduced the impact that OKCSD's policy changes have had on students' lives inside and outside the classroom. Here, we outline mechanisms that explain why this policy was effective in reducing suspensions and adolescent arrests. In OKCSD's resolution agreement, the district committed to ensuring that students are adequately supported by assessing the staffing levels of trained counselors and school administration. In our analysis on staff-level outcomes, we find that the policy led to a moderate and statistically significant increase in the number of school counselors and administrative staff. Here, in the years following the resolution agreement, OKCSD increased its of full-time counseling staff by 17%.

[59] This is evident in our findings of declining reading scores; see Section 3.6.3

[60] See Jacobs et al. (2020) for a systematic review of factors contributing to juvenile reoffending, which provides additional context for this issue.

Figure 3.12: SDID OKCSD Counseling Staff



Note: Figure 3.12 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSDs policy changes on the hiring of full-time counseling staff.

As shown in Figure 3.12, the synthetic control closely tracks OKCSD's pre-treatment trend, lending credibility to the parallel trends assumption for total school counselors. In this analysis, our outcomes represent the number of staff in each category recorded at the school level, with treated units belonging to OKCSD.

There is an existing body of evidence in the economics literature showing that additional student support staff and school counselors can positively impact various cognitive and non-cognitive outcomes for students. Robinson and Roksa (2016) found that school counselors can help mitigate inequalities by playing a significant role in predicting successful college applications. Furthermore, Carrell and Hoekstra (2014) found that hiring additional school counseling staff can be a cost-effective way to support students, even when compared to hiring additional teachers. They found that a one-unit increase in the number of counseling staff in a given school increases boys' reading and math achievement and reduces misbehavior for both girls and boys by 20% and 29%, respectively. Additionally, the presence of school counselors has been shown to impact the perceptions of other school staff, including teachers, particularly as a result of improved student behavior associated with school counselors (Carrell and Hoekstra, 2014).

These results strengthen the evidence of the significance of OKCSD's policies. By hiring additional counseling staff, OKCSD took meaningful steps to create a more supportive and inclusive learning environment, going beyond merely redefining the school code of conduct to address student over-suspension. Furthermore, the increase in school counseling staff could also serve as a mechanism through which the policy impacts students, contributing to a reduction in school suspensions and adolescent arrests outside the classroom. Below are a few plausible explanations for this effect.

The hiring of additional counseling could have facilitated early intervention and provided students with individualized attention. This increase in resources enabled counselors to deliver more targeted and timely support, addressing the unique challenges faced by troubled students. These interventions are critical for mitigating behavioral issues before they escalate into more serious disciplinary actions or academic setbacks. Research demonstrates that early intervention strategies can significantly reduce negative outcomes for at-risk youth (Carrell and Hoekstra, 2014).

Additionally, improvements in student behavior may have positively influenced the overall school climate. With teachers better able to focus on effective instruction, which could fostering a more conducive learning environment. Enhanced teacher confidence and satisfaction in their instructional capacity could further contribute to improved educational outcomes for all students. This aligns with findings in the literature that emphasize the relationship between improved student behavior, a supportive school climate, and higher academic achievement.[61] Where we find mixed effects on academic outcomes, it suggests that the relationship between disciplinary reforms and academic performance may be mediated by other factors, such as variations in implementation fidelity or differing baseline conditions across schools.

3.7 Robustness

3.7.1 Event Study Output

In addition to a visual inspection of the SDID output presented in Section 3.6, we also provide dynamic event study-style output for our main results on suspensions and youth crime. The event study-style output, proposed by Clarke et al. (2023), extends the static SDID estimates by focusing on the temporal evolution of treatment effects rather than their overall aggregate impact. This is achieved by comparing outcomes between treated units and their synthetic controls for each time period relative to a baseline pre-treatment period. The baseline is constructed using optimally weighted averages of pre-treatment outcomes, as determined by the SDID framework.

This approach allows us to trace changes in the treatment effect over time. By plotting these differences across time periods, we can assess both the immediate and longer-term impacts of the policy. For instance, we can determine whether the policy's effects were delayed, took time to build, or diminished after an initial impact. This contrasts with standard SDID plots, which provide a single point estimate summarizing the treatment effect across all periods without illustrating how these effects vary over time.

[61] See Carrell and Hoekstra (2014) for a discussion on the impact of school climate on educational outcomes.

The dynamic event study output also includes confidence intervals for these time-specific estimates, enabling a statistical evaluation of whether observed changes are significant. This helps clarify whether variations in treatment effects are genuine or attributable to noise. Together, these features offer a more nuanced visualization of the policy's trajectory, illuminating both the timing and magnitude of its impacts. This temporal focus complements the static SDID results by revealing not only whether the policy had an effect but also how and when those effects materialized.

In Figure B.2 in Section B.2, the SDID results for suspensions are plotted. Here, we find some evidence of isolated anticipation effects associated with the policy. Recall that the OCR issued a formal complaint in February 2014, and local news outlets began covering the policy in mid to late 2015. This initiated a sequence of events that preceded the reforms discussed in this paper. The timeline suggests that some schools may have begun adjusting their behaviors in response to the impending investigation and anticipated policy changes, leading to observable effects on suspension rates even before the official implementation of the reforms. We provide evidence in Appendix Section B.3 to show that in re-defining the treatment to 2015, some estimates are still significant, but with estimated effects sizes that are considerably smaller. Further, when we re-define the treatment to 2014, these effects vanish completely. This suggests that OKCSD teachers and students may have slightly altered their behavior in light of the significant news coverage on the topic.[62] As an additional check, in B.2 we provide dynamic event-study style output proposed by Clarke et al. (2023).

In Figure B.3, we can see that several periods leading up to treatment showed trends that were somewhat misaligned between the treated and control groups, particularly in 2007 and 2014. While this suggests that there may be some degree of imperfect matching in these pre-treatment periods, the substantial decline observed following the treatment remains significant due to the SDID method's weighting mechanism. This approach minimizes the influence of poorly matched periods by assigning greater weight to periods where the treated and control groups demonstrate more comparable trends. Nevertheless, the presence of these discrepancies in the pre-treatment trends suggests some caution is warranted when interpreting the magnitude of the estimated effects.

3.7.2 In Time Placebo

We conduct a series of in time placebo checks, shown in section B.3 where we artificially assign the treatment to a period prior to the actual policy implementation. This allows us to test whether the observed effects in our primary analysis could be due to random fluctuations or factors unrelated to the policy. By applying the same model specification to this placebo period, we assess whether significant associations emerge in the absence of the policy.

[62] See (Ellis, 2015) for an early example of news coverage on the topic

Our placebo analysis examines three pre-treatment periods: 2013, 2014, and 2015. In 2014, while awareness of a pending investigation into OKCSD's discipline policies existed, no formal adjustments had yet been implemented. By 2015, as the investigation progressed, schools may have begun anticipating mandated changes. The 2013 and 2014 results show no significant effects across all outcomes, providing strong evidence for the parallel trends assumption and confirming that the observed effects in our primary analysis are unlikely to be driven by pre-existing trends or random fluctuations.

In contrast, the 2015 placebo results reveal small but significant effects on some outcomes, particularly suspensions and select academic measures. These effects are consistent with the policy timeline, where schools, aware of the investigation's progression, may have begun to make modest adjustments to discipline practices before the formal policy implementation in 2016. However, the magnitude of these effects is minimal, and they are largely absent for key metrics such as youth arrests. This pattern reinforces the credibility of our main findings, suggesting that the observed treatment effects are primarily attributable to the policy's implementation rather than confounding factors or temporal noise in the data.

3.7.3 Placebo Treatment

We conduct a series of treatment placebo checks, shown in section B.4, where we exclude all treated units and randomly assign the treatment to a subset of control units that were never actually treated. This approach tests the robustness of our findings by evaluating whether the observed effects in our primary analysis could be artifacts of random chance or driven by unobserved confounding factors unrelated to the treatment. By applying the same SDID model specification to these placebo groups, we assess whether significant effects arise when no true treatment occurred. The absence of significant results in these tests supports the validity of our methodology and strengthens confidence in the causal interpretation of our findings.

3.8 Conclusion

In this study, we analyse the impact of school reform policies implemented in 2016 in the Oklahoma City School District (OKCSD). These reforms included a redrafting of the district's code of conduct to discourage exclusionary discipline practices, staff retraining to improve school climate, and the implementation of PBIS policies across OKCSD schools. These efforts align with national initiatives, such as the U.S. Department of Education's "Guiding Principles for Improving School Climate and Discipline," aimed at reducing disparities in school disciplinary measures. We address significant gaps in the literature by examining the short-term effects of inclusive discipline policies on both academic outcomes and adolescent arrests. While existing studies often focus on the long-term consequences of exclusionary discipline, such as adult criminal behavior or graduation rates, our study uniquely combines quasi-experimental methods to explore immediate impacts. By analyzing differential academic effects such as improved math and science scores but declining reading performance, we shed light on the nuanced trade-offs of discipline reform policies. Furthermore, our work contributes to the limited evidence on how discipline reforms can influence youth crime, bridging a critical gap in understanding the broader societal implications of fostering more inclusive and supportive school environments.

Our findings show mixed results. Using a SDID approach and K-means clustering to focus our model on suitable comparison units, we find that the policies led to a substantial reduction in school suspensions (approximately 50%), serving as evidence that OKCSD took significant steps in implementing these policies. However, the academic outcomes were varied. We observe declines in some subjects, such as Grade 6 reading scores (-6.51 points, or 11%), which may reflect challenges in adapting to changes in classroom dynamics. Conversely, there were notable improvements in Grade 7 math (ATT of 7.24) and Grade 8 science (7.08 points). Other subjects, such as Grade 6 math and Grade 7 and 8 reading, show minimal or statistically insignificant changes. While point estimates for math scores were consistently positive, the reading results were more mixed—shifting from a significant decline in Grade 6 to a small, statistically insignificant gain in Grade 8. These patterns, however, are not uniformly significant and should be interpreted cautiously. These findings suggest that structured subjects may have benefited from the reforms, while less structured subjects like reading may have been more sensitive to disruptions in classroom management during the transition. These results underscore the complexity of implementing inclusive policies and their differential impacts across academic domains.

Beyond academic outcomes, our analysis extends to the broader societal impacts of these reforms. We find a significant 22.8% decline in youth arrests following the policy changes, suggesting a potential reduction in interactions with the criminal justice system. While these findings contribute to the understanding of the relationship between school discipline policies and adolescent behavior, it is important to interpret them cautiously, as the mechanisms underlying these changes remain unclear. This study has several limitations that should be considered when interpreting the results. First, the analysis does not account for heterogeneity in effects, such as differences by gender, which may mask important subgroup variations because the school administrative data did not include any information on gender. Second, the study does not explore other potential mechanisms through which the reforms may have influenced outcomes, such as changes in teacher behavior, student-teacher interactions, or family engagement, due to the lack of direct data to evaluate these. Lastly, the policy's stated goals included improvements in school climate and equity in discipline practices, which were not directly measured in this analysis, for the same reasons as above.

In summary, this study provides evidence of the nuanced impacts of school discipline reforms on student outcomes and adolescent crime. While the results highlight promising reductions in suspensions and arrests, the mixed academic outcomes underscore the challenges of balancing inclusivity with maintaining engagement across diverse subjects. Future research could address the limitations identified here by incorporating analyses of heterogeneity and additional mechanisms to deepen our understanding of how these policies affect students and broader societal outcomes.

Chapter 4

Supervised Consumption Sites in Toronto: Did Harm Reduction Work?

4.1 Introduction

Opioid-related harms have become a critical public health challenge across many high-income countries, with surges in overdose incidents, opioid-related deaths, and the proliferation of highly potent synthetic opioids such as fentanyl. In the United States, the opioid crisis has claimed over 500,000 lives in the past two decades, with the Centers for Disease Control and Prevention reporting more than 100,000 overdose deaths in 2021 alone (Ahmad, Rossen, and Sutton, 2021). The scope of the problem is also significant in Canada, as the country battles similar challenges with the U.S. (Belzak and Halverson, 2018; Fischer, 2023; Toronto, 2024). Here, in 2022, opioid related deaths reached a staggering number of 7,537, which corresponds to an increase of nearly 300% when compared to the 1,931 incidents of mortality attributed to car accidents (Belzak and Halverson, 2018; Canada, 2024a; Canada, 2024b). In the same year, opioid mortality stands as the leading cause of accidental death, with an average of 21 deaths per day. While opioids play an important role in pain management in clinical settings or when used as prescribed by medical professionals, it's clear that the misuse prescription and illicitly obtained opioids, and synthetic opioids such fentanyl, has led to a variety of serious health and societal harms.[63]

In response to the ongoing crisis, policy makers have targeted harm reduction interventions as a potential solution. This study focuses on one distinct application of harm reduction policies, namely, the implementation of supervised consumption sites (SCS) in Toronto. SCS are medical facilities which are designed to prevent incidents of overdose and overdose mortality by providing a designated space for people who use drugs (PWUD) to consume pre-obtained substances under the safety and support of trained medical professionals (Hayle, 2018).

The evidence on the effectiveness of this form of harm reduction is mixed. Several descriptive studies suggest that SCS mitigate opioid related harms by reducing overdose mortality, lowering the likelihood of injection-related skin infections, and encouraging PWUD to access primary healthcare (Behrends et al., 2019; Greenwald et al., 2023; Lloyd-Smith et al., 2010; Rammohan et al., 2024a). Studies in economics have linked SCS with negative spillovers, where SCS have been shown to have a negative effect on house prices in the immediate vicinity of the sites (Liang and Alexeev, 2023a; Schaefer and Panagiotoglou, 2024). A separate strand of the literature focuses on other

[63] See section 4.2 along with Figures 4.13 and 4.14 for more context.

forms of opioid harm reduction, where studies have shown that harm reduction can lead to increased or riskier drug use (Erfanian, Grossman, and Collins, 2019a,b). These studies argue that this effect is driven by the moral hazard problem, where PWUD are driven to increase their drug use due to the perceived lower cost (risk) of consumption. Further, some have suggested that SCS might lead to crime, where substance abuse has been linked with criminal activity, as PWUD seek to fund their addictions (Bennett, Holloway, and Farrington, 2008; Brownstein, 2015).

Overall, SCS remain a point of contention with local governments and residents over the fear that these sites are associated with rising crime rates and anti-social behavior. Thus, while SCS show promise in reducing certain harms, the overall impact on the surrounding communities remains unclear. From an economic perspective, the opportunity costs associated with investing in SCS such as potential alternative uses of public funds for law enforcement, prevention, or treatment, further complicate evaluations of their overall societal benefit.

In this paper, we aim to examine the multifaceted causal effects of SCS by analyzing the staggered rollout of sites in Toronto. Here, over a five-year period from 2017 to 2021, health Canada approved the opening of 11 sites. Toronto provides an ideal setting to study the impact of SCS as it is Canada's largest urban center and has a diverse population with varied socio-economic backgrounds, which enhances the external validity of this study.[64] This study is the first to rigorously and causally estimate the impact of SCS on local demand for emergency services, drug-use-related crimes, and the prevalence of severe mental breakdowns. Existing descriptive studies tend to report reductions in overdose-related harms, but these often rely on pre-post comparisons or simple trends and may overstate the benefits by not accounting for underlying changes or selection into treatment. By using a more credible identification strategy, this study provides a clearer picture of the actual effects and a useful benchmark for interpreting earlier findings. By analyzing these effects in the context of Toronto, we assess the effectiveness of SCS in mitigating harm and address concerns about potential negative spillovers often cited by previous descriptive studies. From an economic perspective, this analysis contributes to the literature by quantifying the direct benefits of harm reduction interventions. While we acknowledge the potential for externalities, such as changes in public resource allocation, neighborhood desirability, and social capital, a thorough examination of these effects is beyond the scope of the current analysis. However, this represents a promising area for future research that can build upon our findings.

To evaluate Toronto's experience, our primary findings rely on a difference-in-differences setup, employing the model recently developed by Callaway and Sant'Anna (2021), which accounts for potential heterogeneity in treatment timing due to the staggered rollout of the policy. Our analysis is performed at the neighborhood level, neigh-

[64] Find a more detailed discussion on Toronto in section 4.2.

borhoods are considered treated if they fall within the "catchment area" of an SCS. Thus, our approach takes advantage of both the differential timing and geographic variation of the staggered adoption of SCS across Toronto. As in any difference model, our analysis relies on the parallel trends assumption, implying that in the absence of the SCS, treated and untreated neighborhoods would have followed similar trends.[65] We study the impact of SCS on emergency overdose callouts, incidents of assault and breaking and entering, and incidents of mental health apprehensions.

Our results are robust against a series of checks and provide well-identified evidence that, in the context of Toronto, we find no evidence that the implementation of SCS led to a significant increase or reduction in opioid-related emergency service callouts. These findings contribute to the ongoing debate about the efficacy of harm reduction policies and suggest that while SCS do not exacerbate riskier drug use or escalate harmful behavior, their ability to reduce the burden on emergency services was limited. However, we find a moderate increase in break-ins near SCS locations, indicating that concerns about localised property crime may have some validity. This underscores potential community-level trade-offs and highlights the necessity of integrating SCS within a broader public health framework that includes targeted crime prevention and support services to mitigate any negative spillovers. Finally, our study also finds no discernible effect on the incidence of serious mental health crises, specifically in extreme cases leading to mental health apprehensions under the Mental Health Act. This is significant because a core argument for SCS has been their potential role in stabilizing the health of people who use drugs (PWUD), which could, in theory, reduce the frequency of severe mental health episodes. However, the data suggest that the presence of SCS does not translate into measurable improvements in mental health outcomes at the community level. It is important to note that these are intent-to-treat estimates averaged across the neighborhood, and the proportion of residents directly affected by SCS may be quite small. As a result, even meaningful improvements for PWUD could be diluted when measured at the aggregate level.

Overall, while SCS are positioned as a promising intervention to mitigate the health-related harms associated with opioid use, there remains limited empirical evidence on their broader community impacts. This study fills this gap by offering a rigorous analysis on sites in Toronto. We provide strong evidence that the effects of these sites on opioid related overdoses and mental health incidents are marginal, and that, in the medium term, financially incentivized opportunistic crimes rise. The findings suggest that while SCS may not exacerbate negative behaviors often cited by critics, their potential to significantly alleviate community-level burdens like emergency services or crime appears limited. This paper adds to the ongoing debate by providing robust evidence that, despite the potential promise of SCS, their community-wide impact

[65] For all of the results discussed, we present event study analysis to diagnose the validity of the parallel trends in all periods leading up to SCS implementation.

includes nuanced outcomes, with marginal benefits and some increase in property crime.

The structure of the paper is as follows: Section 4.2 provides context on the opioid crisis and harm reduction efforts in Canada. Section 4.3 provides some background literature. Section 4.4 describes the data used, including opioid-related callout records and crime statistics. Section 4.5 details the spatial difference-in-differences approach used to identify the causal effects of interest. Section 4.6 presents the main findings, and Section 4.7 offers robustness checks to validate our results. Section 4.9 discusses the implications of these findings for public health policy.

4.2 Institutional Background

The opioid crisis has rapidly escalated into one of the most pressing public health emergencies in Canada, where opioid-related deaths far surpassed other leading causes of accidental fatalities since 2016 (Fischer, 2023). To speak to the context for these claims, between January 2016 and March 2024, there were 47,162 opioid-related deaths in Canada (Canada, 2024a), with incidents continuing to rise since the government began tracking in 2016, as shown in Figure 4.13. Specifically in Toronto, annual opioid-related deaths surged by 283% between 2015 and 2023, rising from 137 deaths in 2015 to 524 in 2023 (Toronto, 2024) (see Figure 4.14 below).

While raw figures in the incidents of opioid mortality provide stark statistics that highlight the urgency of the problem, the impact of the opioid crisis extends far beyond these numbers. In 2017, the White House's Council of Economic Advisors found that the opioid crisis costs the US's healthcare system an estimated $596 Billion each year Advisors (2017), a figure which can be seen as conservative relative to the CDC's estimate cost of 1.7 Trillion in the same year (Florence, Luo, and Rice, 2021). The crisis has been linked to increasing emergency department visits, rising pharmaceutical costs, and increasing overall healthcare costs in the U.S. (Maclean et al., 2022). Although the economic burden in Canada has not been as extensively studied, this underlines the substantial strain that opioid-related harms impose on healthcare systems and the economy at large.

The sharp rise in opioid-related deaths during the late 2000s highlighted the shortcomings of traditional punitive strategies in addressing the complexities of opioid misuse, sparking a broader discussion on effective policy approaches. This debate, much like earlier discussions in the economics of crime regarding "deterrence versus rehabilitation" (Becker and Murphy, 1988; Nagin, 2013), has divided the literature on drug policy. The central contention revolves around the contrasting paradigms of zero tolerance and harm reduction (Houborg, Frank, and Bjerge, 2014; Marlatt, 1996; Sung, 2003). In light of these debates and the persistent challenges posed by opioid misuse, public health officials have increasingly turned to harm reduction strategies as a potential solution. In effect, harm reduction policies are those that move away

Figure 4.13: Canada - Overall Count of Deaths and Hospitalizations

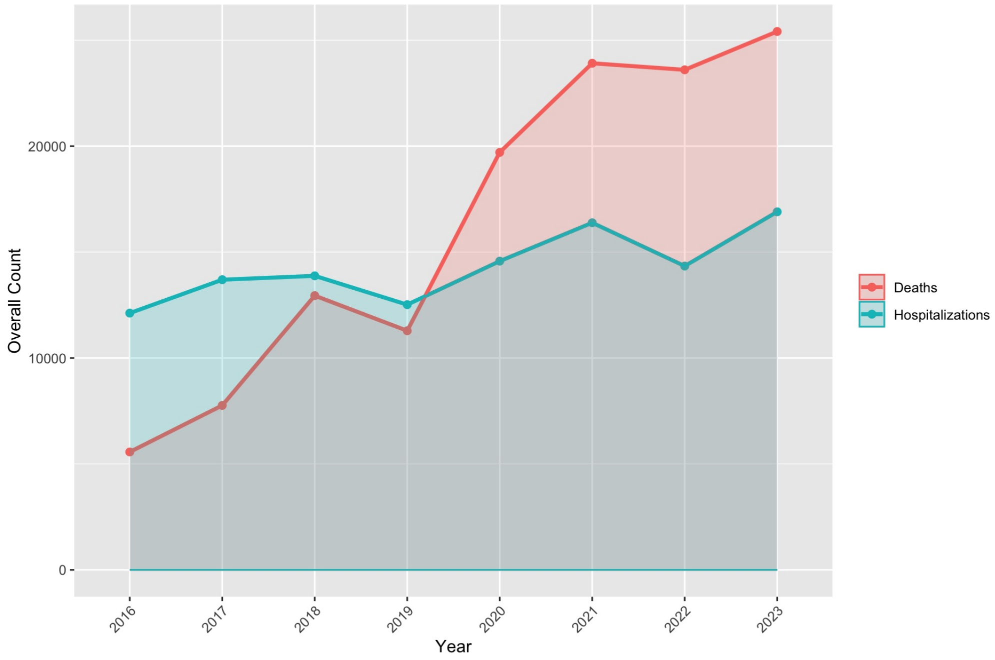


Figure 4.13 above plots data from the Government of Canada's Opioid and Stimulant Related Harm data and plots overall counts of deaths and hospitalizations in Canada. This data is available online from `https://health-infobase.canada.ca/substance-related-harms/opioids-stimulants/`.

Figure 4.14: Toronto - Overall Count of Deaths

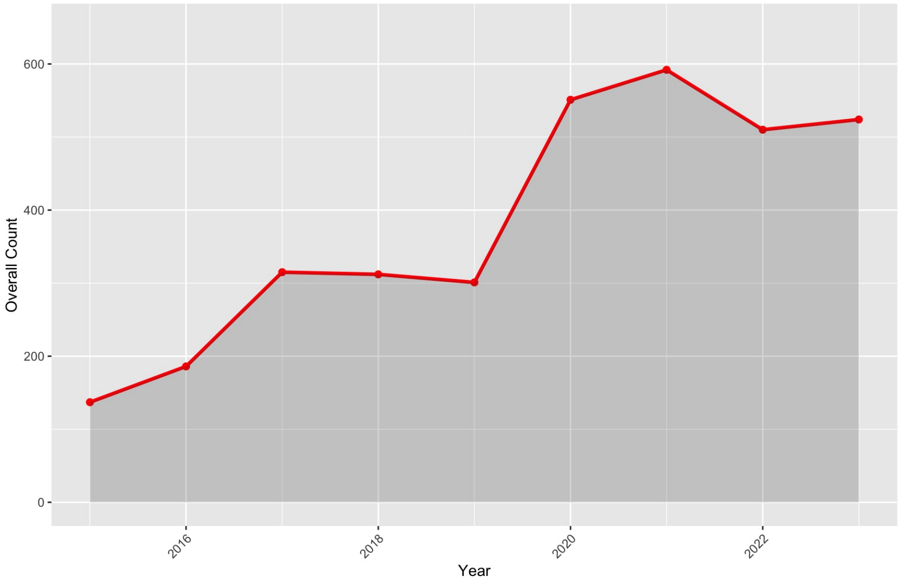


Figure 4.14 above plots the overall count of opioid related deaths in Toronto from 2015 to 2023. This data is available from Government of Canada's Opioid and Stimulant Related Harm Data `https://public.tableau.com/app/profile/tphseu/viz/TOISDashboard_Final/ParamedicResponse`.

from the "zero tolerance approach" to drug use in an embrace of harm mitigation. These policies operate under the assumption that some drug users are unable to cease drug use behaviors and prioritize short-term and realizable goals (Single, 1995). In the context of the opioid crisis, notable harm reduction approaches include syringe exchanges, increasing access to Naloxone and Narcan (overdose reversal drugs), and buprenorphine/ methadone administration ("safer" substitutes) (Cawley and Dragone, 2023; Doleac and Mukherjee, 2018; Goyer, Castillon, and Moride, 2022; Maclean et al., 2022). One form of harm strategy, and the focus of this study, that has gained increasing attention is the implementation of supervised consumption sites (SCS).

SCS, sometimes also referred to as "supervised injection sites", are medical facilities which are designed to prevent incidents of overdose and overdose mortality by providing a designated space for people who use drugs (PWUD) to consume pre-obtained substances under the safety and support of trained medical professionals. Canada's first SCS, Vancouver's Insite, opened in 2003 and was largely seen as a success, with several descriptive studies concluding that Insite was a cost effective way to reduce incidents of opioid related harms such as mortality, overdose, and HIV transmission (Young and Fairbairn, 2018). Despite this evidence, the Canadian government from 2006 to 2015 staunchly opposed the opening of new sites which required an exemption from Canada's *Controlled Drugs and Substances Act* (section 56) (Government, 2013). In practice, this exemption was exceedingly difficult to obtain, with Insite remaining the only site in operation, despite high demand in other cities and towns across Canada. Despite the success of Insite and the growing body of evidence supporting SCS, these facilities were subject to intense scrutiny and opposition.[66] Critics have long suggested that SCS, in enabling PWUD, could lead to increased drug use, and a host of localised negative spillovers associated with the sites.[67]

While Insite remained the only SCS in Canada for nearly a decade, the landscape changed dramatically with the introduction of the newly elected Liberal party leadership's Bill C-37 (of Canada, 2016). C-37 was included in the Controlled Drugs and Substances Act (CDSA) and vastly simplified the eligibility criteria and streamlined the application process for opening and operating a legally licensed SCS in line with federal standards. For instance, prior to the passage of Bill C-37, potential sites were required to meet 26 criteria in order to be considered for an exemption to gain permission to operate. After the bill, these requirements were significantly reduced to 5 clear criteria. Following C-37's implementation, the number of SCS in Canada ballooned from 1 to 51 over the course of four years, with 17 opening between 2016 and 2017 and at least 35 still in operation across Canada today.[68]

This paper focuses on the opening of SCS that are distinct from the more common grassroots Overdose Prevention Sites (OPS), which have and continue to be operated

[66] See Boyd (2013) for an overview of the evidence on Insite, and how this evidence ultimately led to the passage of C-37.

[67] Section 4.3 provides an overview of literature concerning harm reduction, and supervised consumption policies.

[68] As of September 24, 2024.

illicitly in many cities across Canada, including Toronto, since the early 2000s. OPS sites have been criticized for operating under unsanitary conditions, irregular hours, and untrained staff which often would not meet the eligibility to gain federal SCS status. Further, the sporadic opening and closure of OPS makes robust analysis of these types of premises difficult. It is also possible that some OPS continued operating alongside newly opened SCS, which may have muted the measurable impact of the SCS on outcomes like overdose callouts. As such, the estimates presented in this paper may reflect a lower bound on the true effects of formal SCS, particularly if individuals were already accessing some form of unsanctioned support prior to official site openings.

While the rapid expansion of SCS in Canada following the implementation of Bill-35 provides ample opportunities for analyzing the impact of these sites across Canada, this analysis focuses specifically on Toronto. Toronto's significance as Canada's largest metropolitan area, coupled with the availability of detailed open data from the Toronto Police Service, makes it an ideal case for this study. The accessibility and comprehensiveness of Toronto's data allow for a more precise and feasible analysis of the impact of SCS, particularly in understanding the localised effects on incidents of overdoses, crime, and mental health apprehensions. By concentrating on Toronto, we can leverage unique neighborhood level data to conduct a thorough evaluation of the effectiveness of harm reduction policies. Beginning in late 2017 with The Works, Toronto's first SCS, over the course of 5 years the Canadian government approved 11 sites in Toronto with the bulk of them (6) opening in 2018.[69]

4.3 Literature

In this section, we aim to meet two aims. First, we contextualize the primary findings of this study, on the impact of SCS on overdose callouts, within the existing literature on opioid related harm reduction strategies and supervised consumption sites. We follow this by contextualizing our paper within the economic framework of "moral hazard" to explore mechanisms which may be driving the results. In doing so, we are able to examine whether the presence of SCS has an influence on the perceived risk of drug use.

4.3.1 Opioid and Drug Related Harm Reduction

This study addresses an important gap in the literature concerning the effectiveness of harm reduction strategies as a potential solution to the opioid crisis. While there are a number of papers which look into the efficacy of various harm reduction strategies, like needle exchange programs (Franco and Koster, 2024; Strathdee and Vlahov, 2001; Wodak and Cooney, 2006), the provision of housing (Strathdee et al., 2006), and

[69] See map in Section 4.4 below for a map of Toronto's sites.

various approaches to spur cigarette smoking (Friedman, 2015; Friedman and Pesko, 2022), few studies have examined SCS, and no study has provided robust causal evidence on health and crime effects.

A number of descriptive papers have looked into the effect of SCS on health outcomes and generally point to reductions in overdoses and overdose mortality.[70] One example, a recent descriptive study in the public health literature (Rammohan et al., 2024b) uses publicly available geographic data to look at the effect of Toronto's SCS on overdose mortality and finds that Toronto's sites led to a reduction in overdose mortality by around 2 overdoses per 100,000 people. While suggestive, the design lacks a formal counterfactual and may overstate the true effect, especially if sites were targeted to high-risk areas where overdose rates were already peaking or beginning to decline. Given these limitations, the estimated reduction may be somewhat too large. Our own study focuses on emergency overdose callouts, which serve as an intermediate measure of overdose risk. If SCS were driving large reductions in overdose mortality at the population level, one might expect parallel declines in callouts. However, we find no clear evidence of such a pattern, which helps to contextualize the descriptive findings and suggests a more cautious interpretation of their magnitude.

Another recent study (Franco and Koster, 2024) analyses the impact of Dutch "drug consumption room" (DCR) and finds that the closure of these sites led to an increase of drug use by around 13 percentage points. They also look at the impact on crime and housing prices and provide some evidence that crime decreased by around 24% and lead to marginal increases in housing prices. While they use a similar identification strategy to ours, they measure the effect of site closure on drug use by leveraging self reported survey responses to the Longitudinal Internet studies for the Social Sciences (LISS). While this dataset has many advantages, including an abundant arrangement of relevant socioeconomic covariates, previous literature has discussed limitations implicit in using multi-purpose longitudinal studies in substance abuse research, where key demographic groups are commonly under-reported or include unreliable responses (Evans et al., 2010; Farabee, Hawken, and Griffith, 2011). Additionally, the paper does not include event study plots to examine how treatment effects evolve over time. The inclusion of event study plots is generally standard practice in difference-in-differences setups because it can help give readers an idea of pre-treatment trends and assess the plausibility of the parallel trends assumption. Finally, their setting focuses primarily on SCS closures, which may be less informative for policymakers since closures often reflect shifting political or neighborhood pressures rather than deliberate policy design. In contrast, openings provide a more relevant context for evaluating whether introducing SCS can effectively address public health and safety concerns, which is the central focus of most policy discussions.

[70] See Levengood et al. (2021) for an overview of the existing literature prior to 2021.

Two other papers also look at the impact of SCS on housing prices (Liang and Alexeev, 2023b; Schaefer and Panagiotoglou, 2024) with the former using a spatial DID method and the latter using a hedonic price model. These papers are among the few to look into the spillover effects of supervised consumption sites. Both papers point towards moderate and short-term adverse effects on house prices immediately following the opening of the sites. Aside from Franco and Koster (2024) referenced above, no other economic studies that we are aware of look into the effect of supervised consumption sites. Thus, while existing literature provides some valuable insights into the effects of harm reduction strategies, particularly on house prices, there remains a significant gap in in the literature which evaluates the causal effects of these sites, particularly with respect to their intended effect on health outcomes and other important spillovers such as crime and mental health.

4.3.2 Moral Hazard, Crime, and Mental Health

This study contributes to several strands of economic literature, particularly in how our findings relate to the theory of moral hazard, a concept first popularized in Peltzman (1975). This seminal study suggests traffic laws which target safety, like the introduction of mandatory seatbelt laws, could lead to an unintended increase in risky driving as drivers compensate for the lower perceived cost of risky driving. This conceptual framework has also been applied to harm reduction policies. For instance Packham (2019) applies this framework to examine the impact of syringe exchange programs (SEPs) in North America, which were designed to reduce HIV transmission among PWUD by providing sterile syringes and facilitating safe disposal after use. They find that while SEPs decreased the spread of HIV, these benefits were at least partially offset by a significant rise in opioid use, opioid overdoses, and opioid poisoning related mortality. Deiana and Giua (2021) makes a similar claim, and associates naloxone distribution programs, with rising opioid related ER visits and crime.

On the contrary, a separate strand of literature focusing on Naloxone training programs and naloxone co-dispensing programs find no evidence of moral hazard (Binswanger et al., 2022; Colledge-Frisby et al., 2023; Rees et al., 2017). With this background in mind, this study aims to contribute to the economic literature by providing evidence on a link between "moral hazard" and introduction of SCS.[71]

Despite widespread discussion and news coverage, there is limited evidence regarding the association between SCS and crime. One study looks into trends in crime before and after the opening of the first SCS in Australia, and found no significant effects (Freeman et al., 2005). Franco and Koster (2024) also links Dutch drug consumption rooms with lower levels of crime. Some previous studies have examined the

[71] Franco and Koster (2024), as outlined earlier, discusses the effect of supervised consumption sites in the context of a "moral hazard" mechanism. The study provides a comprehensive overview of effects of SCS on health outcomes, with results pointing to absence of a moral hazard mechanism. Nonetheless, the authors noted that the estimates should not be interpreted as causal evidence.

effect of harm reduction programs on crime, with some indicating that programs may increase crime by encouraging risky behavior (Colledge-Frisby et al., 2023; Deiana and Giua, 2021), and others concluding that these programs might lead to a reduction in crime (Bondurant, Lindo, and Swensen, 2018; Freeman et al., 2005).

We contribute to the ongoing debate on harm reduction policies and concerns about moral hazard by offering new evidence on the impact of SCS, leveraging geolocated data on health and crime in Toronto. While descriptive studies have generally pointed to reductions in overdose mortality and improvements in drug-related outcomes following the introduction of SCS, these findings are often based on simple pre-post comparisons or trend analyses that lack credible control groups. As such, they may overstate the benefits of SCS by failing to account for broader secular trends or the possibility of regression to the mean in areas selected for treatment. For example, sites are typically located in areas already experiencing high overdose rates, making it difficult to disentangle true policy effects from underlying trends without a valid counterfactual.

Our study addresses this gap by using a more rigorous identification strategy to estimate the causal impact of SCS on key outcomes, including emergency service use, drug-related crime, and mental health crises. This allows us to test whether the more optimistic conclusions of prior descriptive work hold up under causal scrutiny, and to provide a clearer benchmark for policymakers weighing the tradeoffs of SCS implementation.

4.4 Data

4.4.1 Primary SCS Data

To assess the impact of SCS in Toronto, we first needed to find data on exact location and opening date of each site. As of fall of 2024, Public Health Canada does not provide historical data on the opening and closing of SCS. Thus, in early 2023, we collected historical data on all SCS in Canada from various sources.

Our primary source is the web archived Canadian Health Supervised Consumption Site approvals page which lists, as of the last webpage update, which sites Health Canada had approved. A secondary source of information on the opening and closing of the sites was archived Google news articles.[72] To ensure the identification of all approved sites, we employed a systematic approach, beginning with a broad search at the provincial level and progressively narrowing the focus to major cities, smaller municipalities, and towns. Using Google News archives, we applied Boolean search terms related to supervised consumption sites, filtering results within quarterly date

[72] We were able to locate news articles covering the opening of every SCS approval listed on the Canadian Health website. At the time, the opening of these sites were subject to intense debate, and news coverage was plentiful

windows from 2016 to 2021.[73] An example, where only a single news source was available, we conducted follow-up phone calls to the sites to verify details. This step was also essential for confirming dates and addresses when sites had relocated. In two cases where sites had closed and we could not verify closure dates through news articles or phone interviews, we defaulted to the dates provided on the web-archived CHI website.[74]

Using the methods listed above, we were able to confirm the opening of 52 Canada Health approved supervised consumption sites beginning in January of 2016 and ending in July of 2022. Over this period 15 sites either closed or moved. We believe this dataset to be the only comprehensive source of historical Canadian Health Institute approved SCS openings and closings in existence.

Figure 4.15: Toronto Supervised Consumption Sites Map

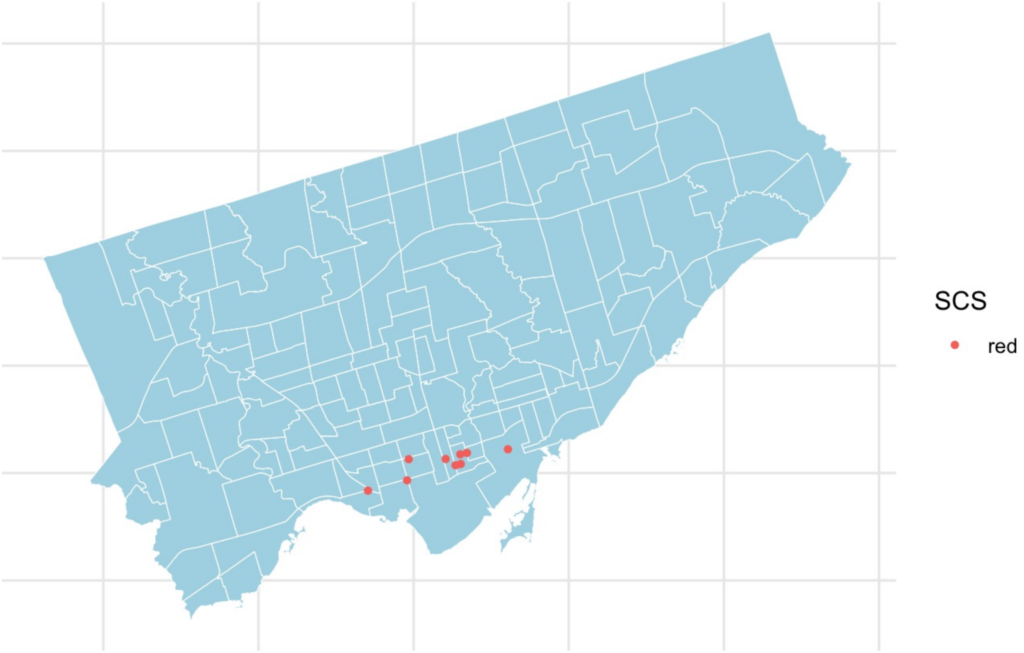


Figure 4.15 above displays all neighborhoods within the city of Toronto, with the locations of 11 supervised consumption sites (SCS) marked in red. The map serves as a reference for understanding the geographic relationship between SCS and their surrounding areas, which is essential for spatial analyses of their impacts on public health and emergency services.

[73] There were no SCS prior to the opening of The Works in November of 2017.

[74] Before 2017, the Canadian Health institute did not publish SCS approvals on their website. Therefore, for pre 2017 sites, we are unable to corroborate our findings from news sources with official CHI documentation.

This analysis makes use of a subset of this database which focuses on Toronto site openings from November of 2017[75] to March of 2019. In this period, no sites in Toronto moved or closed. Table 4.4 presents the historical rollout of Supervised Consumptions Sites in Toronto.

Table 4.4: Supervised Consumption Sites and Opening Dates

Consumption Site Name	Opening Date
The Works	11/8/17
South Riverdale Community Health Centre (keepSIX)	11/27/17
Fred Victor Centre	2/21/18
Parkdale Queen West Community Health Centre	3/16/18
Kensington Market	4/27/18
Regent Park Bevel Up CTS Site	5/1/18
Moss Park	6/1/18
Street Health	6/27/18
Parkdale Supervised Consumption Site	3/29/19
Casey House	8/1/21

Note: Opening dates reflect the actual opening date of the SCS and not the legal approval date of the C-37 exemption. Our analysis excludes Casey House, as its opening falls after COVID-19.

4.4.2 Toronto Police Open Data

Our outcome variables come from the Toronto Police Service Public Safety Data Portal (TPS). The portal provides public access to all data collected or maintained by the Toronto Police Service, excluding any identifiable personal information including detailed individual characteristics. To outline the impact of SCS on emergency overdose callouts, drug related crime, and mental health crises, we use distinct TPS subsets which focus on each topic, namely the Person in Crisis (PIC) Calls for Service dataset, TPS Neighborhood Crime data, and the Mental Health Act (MHA) Apprehensions subset. A key feature of the TPS data is that many of the subsets, including the ones used in this analysis, provide neighborhood level geographic information on observations. One drawback of the TPS open data, and the use of Toronto neighborhood level data in general, is a lack of information on time varying covariates such as neighborhood demographics and migration. Demographic data does exist, but is only updated in census years (every 5 years), therefore we can't use them, as we leverage quarterly frequency in our analysis.

While the TPS makes all data from January 2011 to present available, we restrict the analysis period to late 2016 through 2019 to provide a clear view of the impact of supervised consumption sites (SCS) on relevant outcomes, avoiding the confounding effects introduced by COVID-19. During the pandemic, many SCS locations in

[75] Toronto's first site opening date.

Toronto either closed temporarily or operated at limited capacity, leading to inconsistent service availability that would obscure the effects of these sites on public health and safety.[76] By focusing on the period just before the pandemic, we ensure that observed impacts are attributable to the regular operation of SCS and not influenced by the significant disruptions and behavioral changes related to COVID-19. This restricted timeframe captures the immediate pre- and post-implementation periods in a stable context, enabling a more reliable assessment of SCS outcomes. In our robustness section we include results which use all pre-treatment periods, from Q2 of 2014, to show that our results are robust.

While the three TPS subsets used generally follow the same structure, below, we briefly discuss each of them separately. Our primary results focus on the impact of SCS on emergency overdose callouts uses the Person in Crisis (PIC) Calls for Service dataset which includes all emergency calls associated with the following event types: attempted suicide, person in crisis, overdose, and threatened suicide. Here, each row denotes an incident occurrence. Important variables include the event date and hour, event type description, and the neighborhood in which the event occurred. We clean the data to only include overdose related incidents. See Figure 4.17 below for trends in callouts for each treatment group[77] and Figure 4.16 for a heat map of overdose callouts in Toronto in the last pre-treatment year, 2016. In looking at Figure 4.17, we cannot confidently assert that the treated groups follow parallel trends with the never-treated groups in the pre-treatment period. It is important to briefly address these descriptive differences. While treated sites show some deviations from the untreated sites pre-treatment, there does not appear to be a systematic pattern, nor does treatment timing seem to align consistently with peaks or drops in the outcome variable. Some of these trends could be driven by the large number of control neighborhoods with little or no overdose activity, which pulls down the average. When focusing on higher-need areas within the control group, the levels of callouts prior to treatment are more comparable to those in treated neighborhoods.[78] The difference in levels may make trends appear non-parallel, but once we condition on high-risk areas or examine group-specific dynamics, the concern is less clear-cut. In addition, the event study and group-time ATT framework mitigate these concerns by leveraging untreated observations in untreated periods, while allowing for variation in timing and levels across groups.Additionally, our estimator conducts analysis by group, and the event study will provide some insight into potential violations of the parallel trends assumption.

It is also worth noting that overdoses occurring within SCS facilities are typically managed by on-site staff and do not result in emergency callouts. As such, any reductions in overdose-related incidents captured in this dataset may, in part, reflect

[76] See Cassie et al. (2022).

[77] For a discussion on treatment assignment, see section 4.4.3. Find a similar plot for crime outcomes in Figure C.1 and in Figure C.2.

[78] Find further discussion in 4.6.1.

improved management rather than a direct decline in the total number of overdoses. However, this distinction does not diminish the relevance of our findings. From the perspective of the healthcare system and the broader community, preventing the need for ambulance dispatch is a meaningful outcome. It signals a reduction in the most acute, resource-intensive overdose events and reflects one of the key aims of SCS, to reduce pressure on emergency services by handling overdoses in a safer, controlled setting. Therefore, even if some portion of the observed decline is mechanical, it is still consistent with the public health goals that underpin SCS implementation.

Figure 4.16: Log Overdose Emergency Responses in Toronto

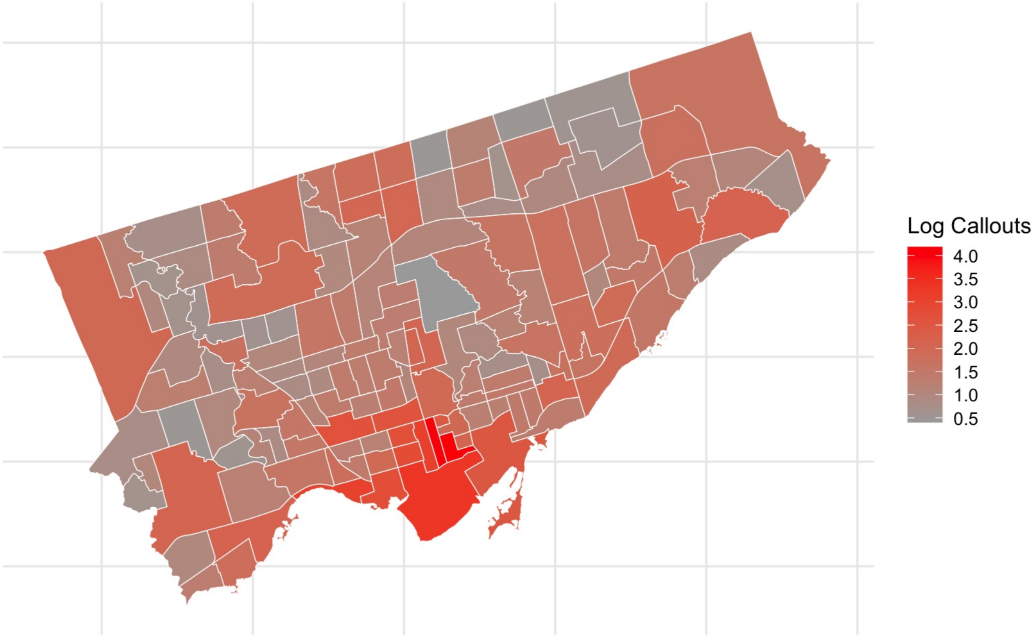


Figure 4.16 visualizes the log-transformed number of overdose emergency callouts across Toronto neighborhoods, based on data from the Toronto Data Service. Darker red areas indicate neighborhoods with higher incidents of overdose-related emergency responses. Neighborhood boundaries are plotted to highlight geographic variations in emergency service demand.

To evaluate neighborhood level crime, we use the Neighborhood Crime dataset. Similar to the PIC data above, the dataset is relatively concise and includes row observations on crime reports by neighborhood. Key variables include the date of the incident, spatial neighborhood data, and a 4 category MCI code. This includes codes for assault, incidents of breaking and entering and robberies and theft over CA$ 5000. We focus on breaking and entering and assault due to these crimes having a longstanding association with drug use (Doleac and Mukherjee, 2022; Maclean et al., 2022; Sim, 2023). See figure 4.18 below for a heat map of the geospatial distribution of assault callouts in Toronto in 2016.

Figure 4.17: Callouts per 1000 People By Treatment Group

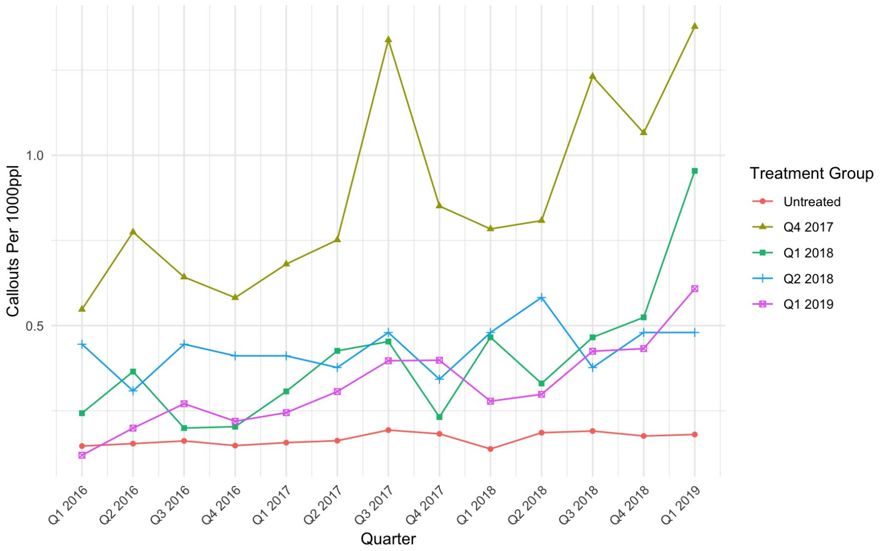


Figure 4.17 plots the descriptive trends in callouts per 1000 people for each treatment group and control group. Treatment groups are denoted with different shapes based on their quarter of first treamtnet.

Figure 4.18: Geospatial Distribution of Assault Callouts per 1000 People

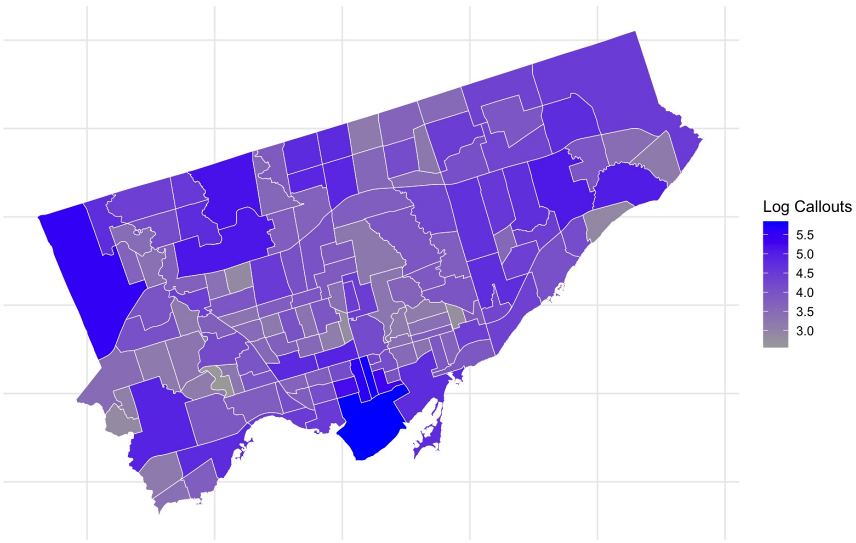


Figure 4.18 illustrates the geospatial distribution of assault incidents per 1,000 people across Toronto neighborhoods, based on data from the Toronto Data Service. Darker shaded areas represent neighborhoods with higher rates of assault-related emergency callouts. Neighborhood boundaries are plotted to emphasize spatial patterns and disparities in crime distribution.

Finally, we make use of the Mental Health Act (MHA) Apprehension data to test if SCS lead to increases or decreases in extreme cases of mental breakdown which require police apprehensions. In Canada, Mental Health Act (MHA) apprehensions refer to instances where law enforcement or medical professionals detain individuals under the Mental Health Act due to concerns about their safety or the safety of others, enabling them to be taken to a healthcare facility for psychiatric assessment and care. This dataset is recorded in a similar manner to the above datasets.

To analyse trends over time, we convert the row-wise incident observations into quarterly neighborhood-level counts. In this setup, our unit of analysis is defined at the neighborhood level at a quarterly frequency. Our primary outcomes for emergency overdose calls, crime, and MHAs use simple incident counts per 1000 by neighborhood and quarter. We use the simple incident per 1000 variable for our primary results in light of recent literature which outlines how the misinterpretation and misuse of near zero transformed outcome variables can lead to bias (Chen and Roth, 2024). Still, to ensure that our results are robust to the distributional properties of the data, we apply alternative transformations to our outcome variable. We do this in two ways, first we transformed the outcome variable using log(1+x), where x represents the raw count of overdose callouts. This transformation helps mitigate skewness in the data and helps account for observed zero values. Second, we applied the inverse hyperbolic sine (IHS) transformation defined as $arsinh(x) = log(x + \sqrt{x^2 + 1})$. The IHS transformation behaves similar to $log(Y)$ but is well defined at zero. In Figures C.11 and C.12 we show that our results are robust to these alternative transformations.

4.4.3 Treatment Assignment

Over the course of 5 years, 11 SCS have been opened at varying times in Toronto beginning in November of 2017 with the most recent site having opened in January of 2021. In this setting, where distinct neighborhoods in Toronto are the unit of analysis, we define treatment status relative to the distance to the nearest newly opened SCS.

Here, we employ the DID framework to estimate the average treatment effect on the treated (ATT). Due to the staggered rollout of SCS, where sites have sporadically opened over time and there are no incidents of treatment reversal (*i.e.* a site closure) we use the Callaway and Sant'Anna (2021) (CS) method where treatment status is defined by geographic distance from a supervised consumption site. We compile callouts into quarterly aggregates where if a SCS is opened within a given quarter, observations associated with that quarter and every following quarter are considered to be in the post-treatment period.

Our primary identification strategy defines neighborhoods as being treated if the centroid of a given geographic neighborhood falls within a 1,000 meter radius of a supervised consumption site. Once a neighborhood is assigned to a treatment group (*i.e.* it falls within the catchment area of an SCS), it will remain in that treatment group. For example, if a neighborhood centroid is within the radius of the first SCS opening in the post-treatment period, even if a subsequent SCS opens which is closer in distance to the neighborhood, the treatment status will remain unchanged. This treatment assignment strategy can be expressed mathematically, following equations 7 below.

Let:

$$S_n = \{s \in S | d(n,s) \leq 1000 \, meters\} \tag{5}$$

Here, S represents the set of all supervised consumption sites (SCS), and $d(n,s)$ measures the distance between the centroid (geographic center) of a neighborhood n (where n belongs to the set of neighborhoods N) and a specific supervised consumption site s (where s belongs to S). To show treatment assignment relates to timing equation 6 below provides additional context.

Following this, let:

$$t_n^{\text{first}} = \min_{s \in S_n} t_s \tag{6}$$

Where t_s is the opening time of SCS s and t_n^{first} is the earliest opening time of any SCS within 1,000 meters of neighborhood n's centroid.

This approach assumes that the initial exposure to an SCS has a lasting impact on the neighborhood, and that subsequent openings do not meaningfully change the treatment status. To address concerns about a neighborhood being exposed to an additional site in later periods, we also test an alternative treatment definition that excludes "double" treated observations, or the cases where a neighborhood is exposed to a second SCS within a 1,000 meter radius. Results are consistent with the results in our main specification and can be found in the Appendix. We excluded double-treated neighborhoods in this robustness check primarily to address concerns about potential overcrowding at SCS sites. If overdoses were underreported in areas where initial sites were at capacity (leading to the opening of additional nearby sites), our baseline estimates might conflate the effects of treatment saturation with the independent impact of a single site. By isolating neighborhoods exposed to only one SCS, we ensure a cleaner comparison and mitigate bias from such dynamics. A dosage measure (e.g., 1 vs. 2+ sites) could offer additional insights, but we prioritized simplicity here for two reasons: First, the primary policy question revolves around the impact of access to any SCS, rather than incremental effects of additional sites. Second, without more robust data on the factors influencing site selection, we prioritized a more transparent and conservative comparison by focusing on single exposures. Future work could explore dosage effects if clearer data on site-selection criteria become available.

We choose a baseline distance of 1,000 meters in line with survey results stating that this was the maximum distance PWUD are willing to travel Hyshka et al. (2016). We also provide a series of robustness checks to show that our results are not sensitive to decreases or increases in this threshold. Further, we also test an alternative treatment definition where we consider units treated based on if the geographic boundaries of a neighborhood, as apposed to if a neighborhood centroid, falls within the "catchment radius" of a supervised consumptions site. This alternative treatment definition yields similar results to our primary results and will be discussed in section 4.6 and displayed in the Appendix in figure C.8.

Figure 4.19: Treatment Group Map

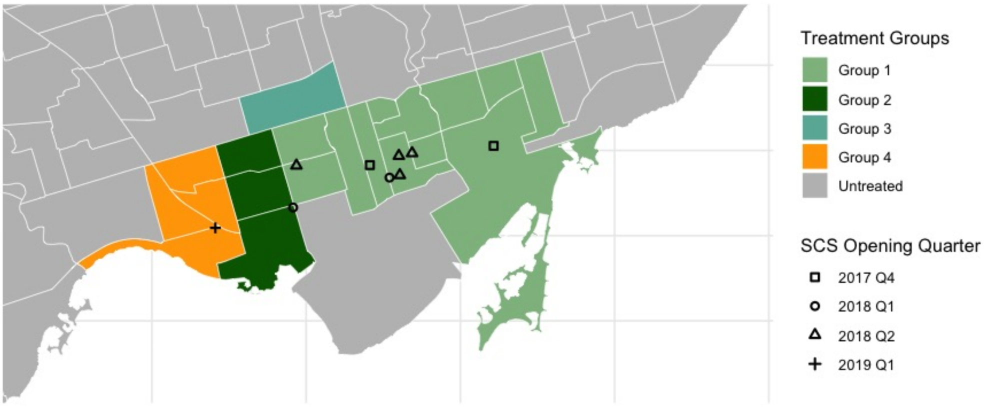


Figure 4.19 above displays supervised consumption sites (SCS) across Toronto using distinct shapes to represent individual locations. Neighborhoods are color-coded based on their assigned treatment groups, reflecting the timing of SCS implementation. The visualization provides a clear geographic overview of treatment group boundaries and SCS distribution.

Finally, to capture the staggered nature of SCS location openings where treatment is assigned over time we can refer to Figure 4.19 which plots each neighborhood treatment group in different colors.

Here, we can see that there is substantial variation in the number of treatment units in each treatment group, and that the first few supervised consumption sites contribute to a larger number of neighborhoods moving from untreated to treated. This occurs because as additional sites are rolled out, they often overlap with areas already exposed to earlier sites due to their proximity. As a result of this, 11 neighborhoods are attributed to group 1, with all other groups representing less than 5 neighborhoods.

In Toronto, SCS site selection was formally guided by a combination of public health criteria (e.g., existing harm reduction services, visible public drug use) and community/political considerations (e.g., municipal zoning, local stakeholder support). While overdose hot spots were a factor in broad prioritization, the final placement of sites depended heavily on operational feasibility, including building availability and willingness of host organizations, rather than short-term changes in overdose trends.

For our research design, this timing and location process helps mitigate concerns about endogenous placement: because sites required months of planning (including community consultations and federal approvals), their openings were unlikely to be direct reactions to contemporaneous neighborhood-level overdose spikes.

That said, we acknowledge that unobserved neighborhood factors could correlate with both SCS placement and overdose risk. We address this in two ways: (1) neighborhood fixed effects absorb time-invariant confounders, (2) our robustness checks exclude 'double-treated' areas where placement may reflect saturation effects. Future work could more rigorously investigate site-selection mechanisms with institutional data.

4.5 Empirical Strategy

Our empirical estimation strategy makes use of the generalized difference-in-differences (DID) approach to estimate the effect of the opening of supervised consumption sites in Toronto on a number of outcomes, including the prevalence of overdose emergency callouts, crime, and mental health apprehensions. In our specification using the TPS data, we exploit neighborhood and time variation in SCS adoption.

In our setting, SCS are implemented in different geographic areas at different times, therefore we adopt the CS method, that allows for a staggered design with multiple treatment groups and periods, and it is robust to potential heterogeneity in treatment timing. Historically, in similar research settings it was common to model the estimated effect of the policy by employing the standard two-way fixed effects (TWFE) model to estimate the average treatment effect on the treated (ATT), controlling for group and time fixed effects. While modified TWFE estimators can be used in a setup similar to ours by adding a series of dummies to account for time heterogeneity as described in Wooldridge (2021), the canonical TWFE has received valid scrutiny in recent years beginning with Goodman-Bacon (2021)

Previous literature outlines several limitations of this approach including additional stringent assumptions implied in TWFE estimation that are easily violated by setups with differential treatment timing (Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021). One main issue with using the TWFE model with multiple treatment timing is that it compares all treatment groups regardless of treatment timing of each group. In our specification, we are ideally interested in comparing treated groups to groups that are not yet treated, or will never be treated. TWFE, by design, estimates the ATT by including comparisons between treated groups, and those already treated ('late treated' with 'early treated'). This comparison is generally a bad comparison which leads to biased estimates because outcomes for the already treated group include treatment effect dynamics. In effect, already treated groups are not a good representation of the counterfactual for the treatment group, had the treatment not occurred (Callaway and Sant'Anna, 2021). With this in mind, we apply the alternative DID estimator proposed by Callaway and Sant'Anna (2021) which is robust to setups with differential treatment timing and treatment effect heterogeneity.

The estimator proposed by Callaway and Sant'Anna (2021) requires our specification to meet certain assumptions. We first denote $\{Y_{n1}, Y_{n2}, ..., Y_{nj}, X_n, D_{n1}, D_{n2}, ..., D_{nj}{}_{n=1}^{n}\}$ where Y_n denotes an outcome of interest (overdose callouts, ect...) for any neighborhood where D_n marks the treatment indicator. The CS estimator also requires *treatment irreversibility*, meaning that treatment status must be permanent. We maintain the irreversible treatment assumption as once neighborhoods are first treated, they remain treated in all observed periods. Treated groups in the post-treatment period are denoted by G, with not-yet-treated groups expressed as g. Treatment groups, $G_g \in \{0, 1\}$ is a binary variable which indicates that a neighborhood within group G is first treated in period g. We can adequately express this difference-in-differences framework through potential outcome notation using equation 7 below:

$$\mathrm{Y}_{nt} = \mathrm{Y}_{nt}(0) + \sum_{g=2}^{T} \left(\mathrm{Y}_{nt}(D_n = g) - \mathrm{Y}_{nt}(0)\right) \cdot \mathrm{G}_{ng} \tag{7}$$

Where $Y_{nt}(0)$ stands as the counterfactual and $Y_{nt}(D_n = g)$ denotes the potential outcome that a neighborhood n would experience at time t if they were to be treated first at the time period g. Equation 7 provides the intuition behind the DID estimator where treatment effect is isolated by finding the difference between treated outcome, $Y_{nt}(g)$ and the counterfactual $Y_{nt}(0)$.

The CS estimator effectively solves the issues that plagues TWFE by calculating a series of simple 2 X 2 DD for each group, at each time period. This approach essentially isolates specific group-time average effects, ATT(g,t), unique to each individual group, g in each period, t. Here, each g is defined as a cohort of units which are treated in the same time period. We estimate the group-time average treatment effects with the following specification, where we use never treated and not yet treated groups as our control units:

$$\mathrm{ATT}_{g,t} = \mathrm{E}\left[\mathrm{Y}_t - \mathrm{Y}_{g-1}|\mathrm{G}_g = 1\right] - \mathrm{E}\left[\mathrm{Y}_t - \mathrm{Y}_{g-1}|\mathrm{D}_t = 0, \mathrm{G}_g = 0\, or\, \mathrm{G}_g \neq 1\right] \quad (8)$$

Here, treatment groups, G, can contain more than one neighborhood when their implementation dates coincide. Equation 8 estimates the ATT for outcome variable Y at time t for treated units first treated in time g. This simple 2X2 calculates the difference in expected change in Y for treated units to the change in units that are eventually treated or never treated but are yet to be treated at time t (expressed by $D_t = 0$). Using this specification, we obtain a vector of ATT's for each group in each period. This approach gives the CS estimator a significant advantage over other staggered DID estimators[79] by allowing us to observe group-time average treatment effects rather than pooling across units. In our context, this is especially valuable for identifying how different cohorts respond following treatment and for detecting possible spillover effects from previously established SCS sites in nearby area

Crucially, the CS estimator also relies on the parallel trends assumption to hold. The parallel trends assumption in this setting posits that, in absence of treatment, the outcomes of treated, not-yet-treated and never treated groups would have followed similar trends over time. This means that the trajectory of outcomes in the post-treatment period for the treated groups should mirror the outcomes of untreated groups if they had not received treatment or will never be treated. Formally, we assume that:

$$E[Y_{nt}(0) - Y_{nt-1}(0)|G_g = 1] = E[Y_{nt}(0) - Y_{nt-1}(0)|G_g = 0] \quad (9)$$

Here, Y_{nt} represents the potential outcome in the absence of a treatment for a neighborhood n at time t. This assumption implies that any observed difference in outcomes pot-treatment are attributable to the treatment itself rather than pre-existing trends or unobserved factors. While it is not possible to test the validity of the parallel trends for the counterfactual outcome, the scenario in which treated units did not become treated, we can infer whether the parallel trends hold in the pre-treatment period by using event study estimates to confirm that the pre-treated trends for the treated group do not differ substantially from those of the never-treated groups.

In equation 9 above, the first term represents the expected change in the outcome for treated units, $G_g = 1$ from the pre-treatment period $t - 1$ to the post-treatment period t,and the second term represents the expected change in the outcome for the not-yet-treated or never treated units over the same time.

[79] This refers to the DID estimators proposed in Callaway and Sant'Anna (2021) and Sun and Abraham (2021).

4.6 Results

As a starting point, we display our results which examine whether the SCS rollout in Toronto had any impact on emergency response overdose callouts in the surrounding neighborhoods. Next, we analyse the effect of SCS introduction on important community spillovers, namely, crime and incidents of mental health apprehensions. For each outcome, we present both regression summaries and event study results. Event studies will be used to analyse treatment heterogeneity over time and to check for evidence of parallel trends assumption violations.

4.6.1 Overdose Callouts

Descriptive Analysis. We first examine the effect of the SCS rollout on the prevalence of overdose callouts. We begin by briefly discussing the simple descriptive trends in the incidents of overdose callouts in Toronto which we report in Table **??**. The table shows the average number of callouts (standardized per 1,000 people) for both the treatment and control group totals. For better readability, the post-treatment period averages, regardless of treatment group, include all observations which occur after quarter 4 of 2017, or the opening of the first site.

Table 4.5: Descriptive Averages for Pre- and Post-Treatment Periods

Variable	Pre-Treatment Mean	Post-Treatment Mean
Callouts per 1000		
Control Group	0.152	0.210
Treatment Group	0.489	1.37
Assult Reports per 1000		
Control Group	1.54	1.64
Treatment Group	2.87	3.13
Break & Enter Reports per 1000		
Control Group	0.63	0.67
Treatment Group	0.96	1.36
Mental Health Apprehensions per 1000		
Control Group	0.63	0.72
Treatment Group	1.22	1.42

Note: The table shows pre- and post-treatment means for both control and treatment groups. Pre-treatment and post-treatment periods reflect averages before and after exposure to treatment. Here, the post-treatment period observations include observations from all treatment groups pooled after Q4 2017. The data in this table comes from Toronto Police Service Public Data Portal described in 4.4.

In the section on Callouts per 1,000, the pre-treatment mean levels differ substantially between the control and treatment groups. Specifically, the control group averages 0.152 callouts per 1,000 people, while the treatment group exhibits higher levels prior to the intervention. This indicates that neighborhoods where supervised consumption sites (SCS) were eventually established generally experienced a greater number of callouts in the periods leading up to treatment. In the post-treatment period, both groups show an increase in these averages, a pattern consistent with trends observed across much of Canada (Canada, 2024a). This increase is more pronounced in the treatment neighborhoods, where the average number of callouts per 1,000 rises by over 150%. While this alone might suggest that the establishment of SCS is associated with a greater increase in overdose callouts compared to untreated neighborhoods, we present evidence below that challenges this interpretation.

It is important to note that these means do not capture the full story. For instance, the control group includes many more neighborhoods, with 119 untreated neighborhoods compared to 21 which are treated. In a landscape where overdoses tend to be concentrated in specific areas, as shown in figure 4.16, whether treated or untreated, we expect the large number of control group observations to be underemphasizing the effect of the most severely impacted neighborhoods within that group. For reference, when isolating control group neighborhoods within the top 10% of callouts, the pre-treatment average number of callouts is around 0.686, a figure much more in line with the treatment group average. Although baseline levels differ to some extent between treated and untreated neighborhoods, the decision to include the full set of controls is guided by the need to maintain statistical power. Many untreated areas still exhibit meaningful levels of overdose activity, and excluding lower-risk neighborhoods would lead to substantially wider confidence intervals and limit the ability to detect treatment effects. Comparisons are made only in periods when both treated and comparison units are untreated, and identification relies on variation in treatment timing across groups. This structure limits potential contamination from already treated units and helps mitigate bias from differences in baseline levels or outcome dynamics.

While the untreated group includes many more neighborhoods, their inclusion is supported by the broader spatial distribution of overdose activity across Toronto. As shown in Figure 4.16, several untreated areas exhibit high baseline levels of overdose callouts and share characteristics with treated neighborhoods. The estimator leverages both never-treated and not-yet-treated areas in untreated periods, which helps ensure that the comparison is driven by meaningful variation across time and space rather than simple differences in levels. This design, along with the event study, provides a credible framework despite variation in baseline intensity.

Estimated Effects. Table 4.6 displays a summary of our results on overdose callouts, where the coefficient can be interpreted as the effect of an SCS on neighborhoods within a 1,000 meter radius of a site opening by estimating equation 8. First, in reference to the overall ATT summary, which displays the aggregated effect on all treatment groups, we find no significant effect, with the confidence intervals nearly symmetrically overlapping zero. Keeping in mind that the pre-treatment mean in the callouts per 1000 is 0.42, the relative magnitude of the estimated effect can be considered as relatively small and close to zero. It is also important to note that, although some point estimates are close to zero, the confidence intervals in several cases are relatively wide, particularly for the overall ATT, meaning we cannot rule out moderately sized effects, including those that approach half the magnitude of the pre-treatment mean.

Table 4.6: Aggregated ATT and Group Effects: Callouts Per 1000

Group	ATT	SE	95% CI	Pre-Mean
Overall ATT				
	-0.0188	0.11	[-0.2344, 0.1968]	–
Group Effects				
Group 1	-0.2048	0.0864	[-0.3859, -0.0237]*	0.689
Group 2	0.2913	0.2187	[-0.1672, 0.7497]	0.286
Group 3	-0.0483	0.0119	[-0.0732, -0.0235]*	0.356
Group 4	0.3142	0.3094	[-0.3343, 0.9627]	0.289
Pre-Treatment Mean (Sample) = 0.42, SD = 1.01, $N = 2240$; Group Avg = 0.405				

Note: Significance levels are indicated by confidence intervals or bands not covering 0 (*).

For group specific effects, we find either insignificant effects or very moderate magnitude significant effects which show marginal declines in overdose callouts.[80] [81] In reference to Figure 4.19, and our short discussion on the variation in group sizes, we think it is reasonable to place more emphasis on group 1 relative to other treatment groups, this group includes 11 neighborhoods making it the largest treatment group, with at least double the number of units as other groups. The smaller number of neighborhoods in later groups (each with only 1 to 3 sites) means that these groups are more susceptible to variability, which is reflected in group 2 and group 4 relatively big standard errors. Another consideration is that group 1 is completely isolated from anticipation effects and outlines the treatment's initial impact, as there were no sites in Toronto prior to quarter 4 or 2017.

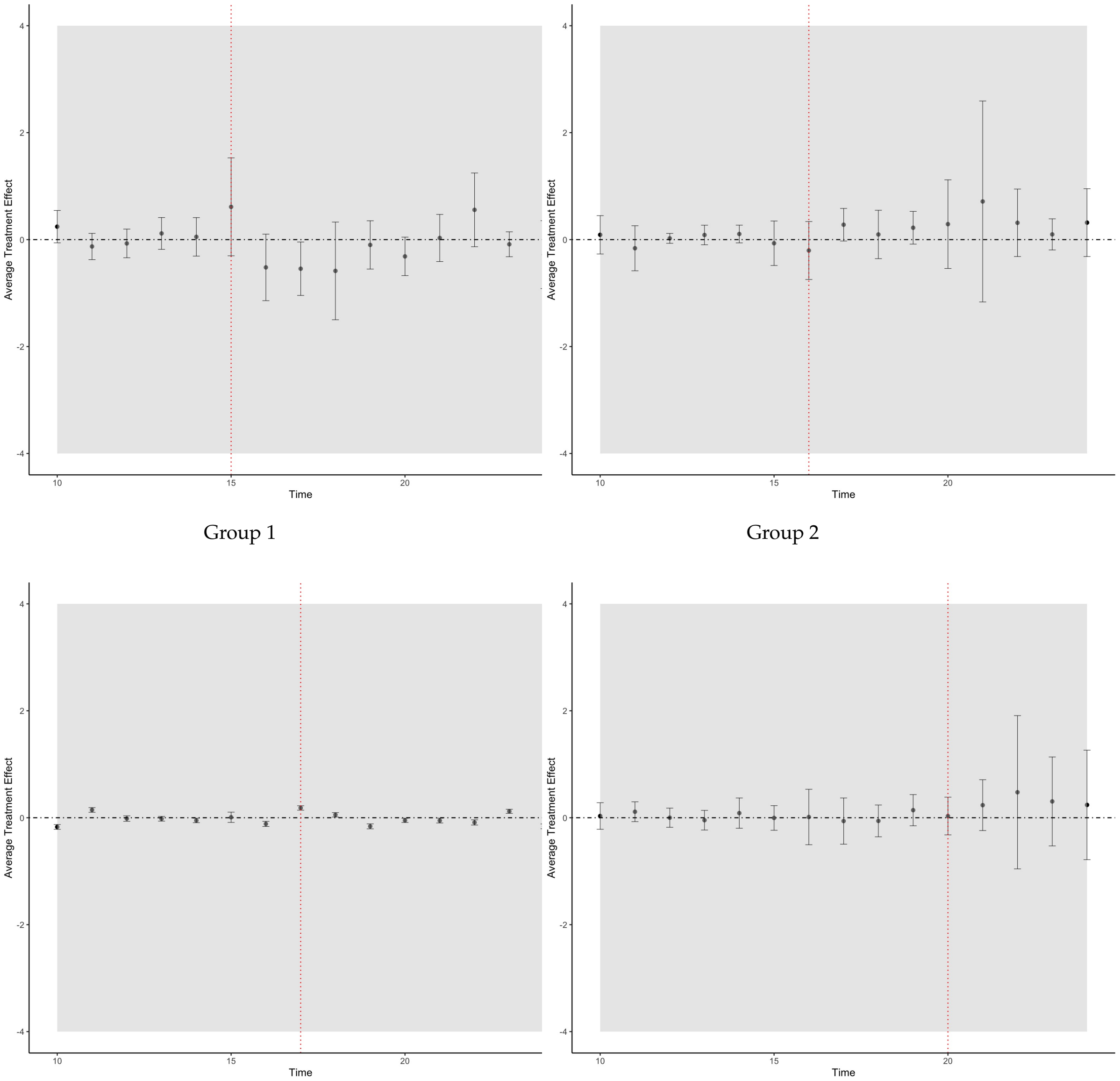
With this in mind, Group 1 and Group 3 show statistically significant effects as their confidence intervals do not include zero. However, the effects are small in magnitude or only marginally significant, suggesting only minimal changes in callouts. Meanwhile, Group 2 and Group 4 show no significant effects, with confidence intervals that include zero, indicating no clear impact for these groups. The event study in the Appendix provides additional context, showing some initial declines in callouts; however, the long-term trends return to null effects, suggesting that the declines for these groups were likely short lived.

Figure 4.21 presents the results from table 4.6 as an event study plot. Here, we check the pre-treatment period observations for evidence of parallel trend violations,

[80] Significant effects can be seen as marginally significant as they are only significant at the 10% level.

[81] Our inference accounts for serial correlation within neighborhoods over time by clustering standard errors at the neighborhood level. This approach allows for arbitrary within-neighborhood correlation across periods (e.g., if overdose callouts are serially correlated).

Figure 4.21: Overdose Callouts near SCS - Callouts per 1000



Group 1

Group 2

Group 3

Group 4

Figure 4.21 The figures above illustrate the event study results, presenting the estimated average treatment effects (ATEs) for overdose callouts near supervised consumption sites (SCS) in Toronto, measured in callouts per 1,000 people. Each panel corresponds to a specific group categorized based on the timing of SCS openings, capturing variations in treatment effects across different implementation periods.

or places where the trends of treatment groups significantly diverge from zero. In all cases, for groups 1,2, & 4, there are no instances of significant pre-trends. Two periods associated with group 3, Q3 of 2016 and Q4 of 2017, have confidence intervals which do not overlap with zero, which could indicate parallel trend violations specific to this group. Group 3, the smallest group, is solely made up of one neighborhood. Since this group's effects rely on one unit, we should consider that the DID is highly susceptible to unit specific variations therefore, these divergences may reflect neighborhood-level and highly localised trends rather than systematic treatment effects.

Figure 4.21 helps address concerns about parallel trends while showing that SCS had little to no effect on emergency overdose callouts. For Group 1, the Figure 4.21 suggests that the opening of SCS might have slightly reduced overdose callouts for about three quarters, after which trends returned to baseline levels. As noted earlier, the confidence intervals for Group 1, which has the most statistical power, suggest that large positive effects are unlikely. Groups 2 and 4 show similar post-treatment patterns over time, with many confidence intervals centered around zero and some leaning toward small positive trends in overdose callouts.

In summary, while considering both regression summary output and event study plots we observe that SCS rollout in Toronto had no discernible impact on Emergency SCS overdose callouts in nearby neighborhoods.

4.6.2 Crime and Mental Health

In this section, we outline our results on crime and mental health. For crime, we study the effect of SCS openings on reported incidents of assault and breaking and entering. As discussed previously, a common worry surrounding the opening of SCS is the idea that in attracting PWUD, crime rates around the sites might increase. We also examine the potential impact of supervised consumption sites (SCS) on mental health apprehensions (MHA), or instances where police intervene in severe mental health incidents. A common concern is that the presence of SCS might lead to an increase in severe mental health incidents, particularly involving individuals with severe substance use disorders, in areas surrounding SCS. We begin by outlining the descriptive trends based on outcome averages for crime and mental health and then proceed to presenting our estimates.

Descriptive Analysis. We again refer to table 4.5 for descriptive pre- and post- treatment averages. For crime outcomes, assault callouts and incidents of breaking and entering, we can see that average number of incidents of both forms of crime were trending upwards in from Q2 2016 to Q4 2020. Treated neighborhoods witness steeper sloping trends in crime compared to control neighborhoods. Interestingly, when focusing on control neighborhoods with in the 90th percentile of assault callouts, pre-and post treatment averages surpass that of the control units with 3.45 in the pre-treatment period and 3.74 in the post treatment period. For mental health apprehensions, which are incidents of severe medical distress requiring police seizure, the trends in both treated and control areas were similar, with both experiencing an increase at comparable rates. We note that these trends are different in their levels, where averages associated with the treated areas are notably higher in both periods. We proceed to present the estimated effects below.

Assault Callouts. Beginning with table 4.7 which summarises the estimated effect on assault callouts per 1,000 residents, the aggregated ATT estimate for all groups is not statistically significant. This suggests that the introduction of SCS in Toronto did not lead to any significant increase or decrease in reported incidents of assault in the short run. In terms of magnitude, the estimated effect size is small compared to the pre-treatment mean of 1.9.

Table 4.7 Aggregated ATT and Group Effects: Assault Callouts Per 1000

Group	ATT	SE	95% CI	Pre-Mean
Overall ATT				
	0.0471	0.1967	[-0.3385, 0.4327]	–
Group Effects				
Group 1	-0.0640	0.2652	[-0.6225, 0.4945]	4.000
Group 2	0.1478	0.6435	[-1.2076, 1.5033]	2.610
Group 3	0.2366	0.2419	[-0.2729, 0.7462]	2.140
Group 4	0.2644	0.2707	[-0.3058, 0.8346]	1.960
Pre-Treatment Mean (Sample) = 1.905, SD = 1.4, $N = 1906$				

Note: Significance levels are indicated by confidence bands not covering 0.
Estimation Method: Doubly Robust; * Denotes significance at the 10% level, ** at the 5% level, and *** at the 1% level.

For group specific effects, Group 1, the largest treatment cohort in the analysis also displays null and statistically insignificant effects. The smaller groups also have insignificant estimated effects, with positive coefficients. The standard error for the aggregated and group-specific results reflects some level variability, indicating that while large impacts can be ruled out, smaller effects could exist within the bounds of the estimates. While the point estimates suggest no systematic effect of SCS on assault callouts, we acknowledge that the wide confidence intervals in Table 4.7 mean we cannot rule out moderate-sized effects in either direction. Nevertheless, the consistency of null effects across multiple specifications supports the interpretation that any undetected effects are unlikely to be substantively large.

The event study results for assault callouts displayed in Appendix A Figure 4.18 show no evidence of pre-trends leading up to the treatment, with the exception of group 3. For other groups, the plotted residuals also show no discernible shift, either significant or insignificant, over time. The stability in estimates further supports the conclusion that there is no strong causal link between the introduction of SCS and variations in assault callouts.

Breaking and Entering. To discuss results on the effect of SCS on reports of incidents of breaking and entering we refer to table 4.8. Here, the aggregate results are significant at the 10% level. While these results indeed suggest that the introduction of SCS in Toronto could have led to an increase in incidents, the overall effect size is considerable compared to the pre-treatment mean of 0.79.[82] These increases are consistent across groups, but particularly in groups 1 and 3. Unlike the previous outcomes we examined, point estimates for all groups are consistently positive.

Table 4.8 Aggregated ATT and Group Effects: Breaking and Entering Per 1000

Group	ATT	SE	95% CI	Pre-Mean
Overall ATT				
	0.309	0.1015	[0.1102, 0.5079]*	–
Group Effects				
Group 1	0.3957	0.1684	[0.0265, 0.7649]*	1.260
Group 2	0.1767	0.1236	[-0.0941, 0.4475]	0.804
Group 3	0.3285	0.0285	[0.2660, 0.3909]*	0.900
Group 4	0.0819	0.1756	[-0.3030, 0.4668]	1.080
Pre-Treatment Mean (Sample)= 0.7917, SD = 0.6545, $N = 2240$; Group Avg = 0.8445				

Note: Significance levels are indicated by confidence bands not covering 0.
Estimation Method: Doubly Robust; * Denotes significance at the 10% level, ** at the 5% level, and *** at the 1% level.

While these increases are statistically significant at the 10% level, the magnitude of the effect is considerable, with the observed rise in breaking and entering reports associated with SCS representing a meaningful proportion of pre-treatment levels. Further research is needed to investigate potential mediating factors or external influences, including changes in local policing strategies or shifts in community dynamics, that may have contributed to these results.

The event study plots depicted in Appendix Figure C.4 are in line with these findings. Overall, there are no significant signs of pre-trends in either plot. For most groups, including Groups 1, 2, and 4, the earlier post-treatment periods show no effect, while the trend in breaking and entering increases in the later periods. Interestingly, while the event study plot for Group 1 suggests a rise in incidents of breaking and entering in the later periods, none of the individual period-specific estimates are significant. This is not the case for Group 2, where the last observed period shows a significant increase in breaking and entering, and Group 4 reflects a similar trend. The discrepancy between the largely insignificant period-specific estimates in the event study and the significant aggregate results arises because the aggregate estimates cap-

[82] The standardized effect size of the estimate, given by dividing the ATT by the pre-treatment standard deviation correspond to 0.47 of a standard deviation.

ture the average effect over the entire post-treatment period, while the event study examines effects at specific time points. Period-specific estimates may lack sufficient power to detect significance when effects emerge gradually or are dispersed over a longer horizon. Overall, the main effects of SCS on incidents of breaking and entering appear to manifest in the medium to long term. Given our quarterly data, this suggests that neighborhoods near SCS may experience an increase in breaking and entering incidents over time following their opening.

Mental Health Apprehensions. Here, we introduce the potential impact of supervised consumption sites (SCS) on mental health apprehensions (MHA). In reference to our estimated effects displayed in Table 4.9, we find null effects in aggregate and with the exception of group 3, no significant effects of SCS opening on the prevalence of MHA callouts. For aggregate effects, the magnitude of the estimate, when compared to the pre-treatment mean is small. Apart from group 3, all of the group specific coefficients are insignificant with positive point estimates. Group 3 is the only group with significant negative effects. As discussed previously, these results should be interpreted with caution, as this group contains only one treated unit, making it the smallest group in the analysis. With such a limited sample size, the ability to detect true effects and avoid false positives is lower compared to other groups, which include at least three and up to twelve treated units. Consequently, the findings for Group 3 are more susceptible to random variation.

Turning to our the event study plot displayed in Figure 4.22 and with a particular focus on the pre-treatment period trends for each group, we find some evidence of pre-trends for group 3, which aligns with our suggestion to interpret the significant findings for this group with caution. For all other groups, there is no evidence of pre-trends. When taken as a whole these results do not suggest that opening SCS in Toronto led to significant increases or decreases in incidents of severe mental health apprehensions. We can argue that SCS had a null effect on the prevalence of MHA.

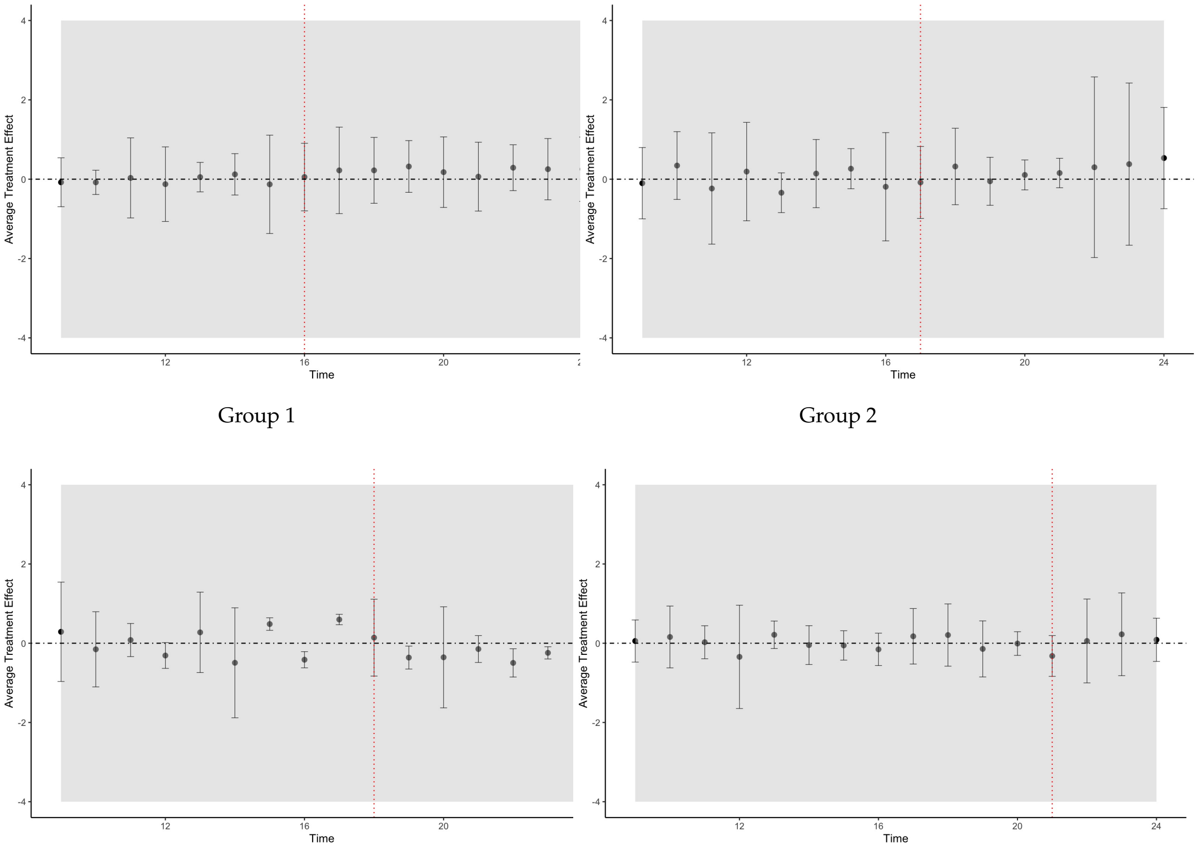
4.7 Robustness Checks

4.7.1 Changing SCS "Catchment" Distances

As outlined in section 4.4.3, our primary specification assigns treatment status to neighborhoods based on if the geographic centroid falls within a 1,000 meter radius of a SCS. A potential concern with this approach is that the 1,000 meter distance could been seen as arbitrary.[83] For instance, if significant changes to our findings are observed when changing the radius distance, it could indicate that the treatment effect is highly sensitive to the definition of the boundary itself. To address this concern, in section C.4 we estimate our model using various SCS catchment radiuses. Using a smaller radius, such as 500 meters, result in fewer treated units, adding noise to

[83] See (Hyshka et al., 2016) for evidence on PWUD willingness to travel to SCS.

Figure 4.22: Mental Health Apprehensions Near SCS in Toronto - Calls per 1000 People



Group 1

Group 2

Group 3

Group 4

Note: This figure illustrates the estimated effects of supervised consumption site (SCS) openings on mental health apprehensions, measured as calls per 1,000 people. Each panel corresponds to a specific treatment group, reflecting staggered exposure to SCS over time. The data is sourced from Toronto Police records and highlights trends in mental health-related incidents in neighborhoods near SCS. Confidence intervals provide insight into the statistical significance of observed changes, allowing for an assessment of how SCS implementation may have influenced mental health emergencies in the surrounding areas.

Table 4.9 Aggregated ATT and Group Effects: Mental Health Apprehensions Per 1000

Group	ATT	SE	95% CI	Pre-Mean
Overall ATT				
	0.1334	0.1019	[-0.0662, 0.3331]	–
Group Effects				
Group 1	0.2062	0.1467	[-0.0812, 0.4937]	1.420
Group 2	0.2079	0.2025	[-0.1890, 0.6048]	0.935
Group 3	-0.2324	0.0902	[-0.4091, -0.0557]*	0.891
Group 4	0.0118	0.1190	[-0.2214, 0.2450]	1.060
Pre-Treatment Mean(Sample) = 0.95, SD = 0.519, $N = 2375$; Group Avg = 1.077				

Note: Significance levels are indicated by confidence bands not covering 0.
Control Group: Not Yet Treated, Anticipation Periods: 0
Estimation Method: Doubly Robust; * Denotes significance at the 10% level, ** at the 5% level, and *** at the 1% level.

our estimates, and lead to wider confidence intervals. The reverse is true when using a larger radius. However, we do not find different results using these different measures, which indicates that our findings are robust regardless of the radius used.

4.7.2 Using Neighborhood Boundaries Instead of Centroids

A critical aspect of our primary specification is that we define treatment status based on SCS proximity to a neighborhood's geographic centroid. In other words, in order to be considered treated, the SCS radius must encapsulate the neighborhoods centroid. While this approach offers simplicity and consistency, it may not fully capture neighborhood access to a site. For instance, larger neighborhoods with centroids just outside the defined treatment radius might have significant portions of their area exposed to an SCS, which may downplay the true effect of SCS. To address this concern, we conduct a robustness check by redefining treatment status based on whether any part of the neighborhood boundary falls within the specified catchment area, rather than solely relying on the centroid location. This approach accounts for partial exposure and provides a more nuanced assessment of how SCS impacts neighborhoods with varying levels of proximity. In C.6 we show that this approach has little impact on our main results focused on overdose callouts. In reference to table C.5, the overall effect remains insignificant. Similarly, there is no evidence of pre-trends, with most post-treatment period point estimates being insignificant.[84]

4.7.3 Dropping Neighborhood Observations Which are Treated Twice

The treatment assignment method we used assigns neighborhoods to a group based on their first exposure to an SCS. This necessitates that once a neighborhood is classi-

[84] See Figure C.8.

fied as treated, its status remains unchanged, even if a new SCS opens closer to it at a later time. This raises a potential concern that neighborhoods experiencing subsequent exposures to additional SCS could introduce bias or complicate the interpretation of treatment effects, particularly if these later exposures exerts a stronger effect, for example if sites suffer from congestion, users may take advantage of having access to multiple sites. To address these concerns, we ran the analysis with a subset of the data that excludes observations that could be considered as "treated twice" (*i.e.* when a subsequent SCS opened after initial treatment nearby a given neighborhood). This approach allows us to focus solely on the impact of the initial SCS exposure and asses whether follow-up exposures could bias or alter our main findings. As shown in C.4 this approach has no bearing on our overall findings.

4.7.4 Placebo Treatment

We conduct a placebo exercise as shown in Section C.4, where we run the analysis using an artificial dataset that excludes all treated units and randomly assigns treatment to a subset of control neighborhoods that were not exposed to SCS. In constructing the placebo dataset, we maintain the same number of treatment groups, the same group sizes, and assign treatment at the same relative time periods as in the original analysis. This ensures that the placebo structure mimics the temporal and group dimensions of the actual data. However, we do not constrain the placebo assignments to replicate the geographic clustering observed in the real treatment rollout. This design choice allows us to examine whether features of the research design itself—such as the small number of units per group, staggered treatment timing, or the structure of the estimator could produce spurious effects even when no true treatment has occurred. By applying the same model specification to this placebo dataset, we observe that the estimated effects are consistently null across groups. This outcome reinforces the robustness of our findings, suggesting that the treatment effects identified in the primary analysis are unlikely to be driven by random variation, mechanical features of the model, or the specific grouping structure of the data.

4.7.5 Alternative Estimator: Sun and Abraham (2020)

When analyzing staggered rollout treatment effects, there are a number of potential estimators to choose from. The estimator proposed in Callaway and Sant'Anna (2021), that we use in our primary specification is appropriate for assessing heterogenous treatment effects over time and across groups. This is particularly useful for our setting, where the disaggregated group and time estimates allow us to capture heterogenous effects by group. For instance, our Group 1 estimates help us isolate the impact of the initial opening of sites in Toronto. Testing the results using an alternative estimator like the one proposed in Sun and Abraham (2021) can provide insights into how robust our findings are to differences in methodological assumptions. The Sun and Abraham (2021) estimator offers similar robustness properties to the estimator proposed in Callaway and Sant'Anna (2021), especially in addressing potential bias from using already treated units as comparisons inherent in TWFE. However, the estimator defaults to presenting aggregate effects, which can be helpful in understanding the overall impact across all treated cohorts in a single event study plot but it is less informative on group-specific effects than CS.[85]

In our analysis, we applied the Sun and Abraham (2021) estimator to test the robustness of our primary findings. The results confirmed that the null effects observed in our primary findings on overdose callouts are consistent with the estimated effects found by using the Sun and Abraham (2021) estimator. See Figure C.13 in C.4 for the event study plot and event study.

As a final robustness check, we also provide results that include all available pretreatment periods in Figure C.14. This check is included to alleviate concerns related to selection bias.

4.8 Discussion

The results presented in section 4.6 provides robust evidence that the opening of SCS in Toronto was not associated with large changes in the frequency of emergency overdose response callouts, [86], increases or decreases in the frequency of assaults, or police administered mental health apprehensions. We do find some evidence to show that incidents of breaking and entering tend to trend increase in neighborhoods within a 1,000 meter radius of a SCS. These findings prompt a deeper exploration into the mechanisms behind the observed null effects and potential factors driving the increase in property crime, as well as implications for public health and community safety.

Our primary null findings on overdose callouts touch on two important intended effects of SCS. Overdose callouts serve as a downstream measure of the severity of opioid-related harm, in capturing instances that require emergency response (Li et al., 2019), and simultaneously act as a proxy for emergency service demand. A considerable incentive for the opening of SCS is the potential to prevent the prevalence of overdoses. Thus, the argument of the moral hazard comes into play. This argument proposes that, PWUD, in seeking to to maximize their utility from opioid consumption, may choose to engage in riskier or more frequent drug use due to the lower perceived costs associated with the safety net that SCS provide. If the moral hazard idea sufficiently explained how PWUD might respond to the opening of sites, we would generally expect to observe a corresponding rise in emergency callouts due to the need for external medical interventions. The absence of such an increase suggests that any

[85] See (Callaway and Sant'Anna, 2021) for details on the advantages of this method over other similar methods.

[86] While we cannot rule out modest or localized effects due to the variability in our data, the consistency of null results across specifications suggests that any undetected impacts are likely limited in magnitude

potential moral hazard effect may be limited or effectively managed within the SCS environment.

This begs the question, why do we not see an increase in overdose callouts? One potential explanation is that PWUD might be indifferent to any perceived reduction in risk associated with SCS. In other words, the model may fundamentally overlook the reality that for many, opioid use is not driven by a calculated response to changes in perceived risk but by an overwhelming need that supersedes considerations of safety or consequence.[87] This story is reflected in the wider literature which associates substance use and addiction with risky behavior on both a mental and physiological level (Crews and Boettiger, 2009; Rømer Thomsen et al., 2018). Still, while SCS may not have driven a rise in riskier drug use, they also had no discernible impact on reducing overdose callouts in surrounding areas. These findings underscore the importance of viewing SCS as one component of a broader public health strategy. To achieve a more substantial reduction in opioid-related harm at the community level, SCS must be integrated with comprehensive measures that address the underlying social, psychological, and economic factors contributing to addiction.

Our results on crime, which show that SCS have no significant effect on assault callouts and show significant, medium run, impacts on breaking and entering, outline potential negative spillovers which have been associated with SCS. It is important to recognize that PWUD are linked to two distinct types of crimes, those associated with aggressive or violent behaviors and those driven by financial needs or economic desperation (Bondurant, Lindo, and Swensen, 2018). Given these distinct motivations, it is reasonable to expect that SCS might influence these types of crimes differently.

However, when considering the results on breaking and entering, we acknowledge that in the medium-term, concerns about increasing crime rates near SCS are valid, as the communities surrounding sites in Toronto were subject to a rising trend in opportunistic breaking and entering crimes. Still, it is important to contextualize these findings within the broader understanding of crime patterns and spatial dynamic. Research has shown that crime tends to cluster in small, specific areas, attempts to displace crime often face limitations (Wang, Liu, and Eck, 2014) and that hotspot policing strategies are effective in reducing net crime (Mohler et al., 2015; Ratcliffe et al., 2011) even in areas outside of these hotspots (Weisburd et al., 2006). Therefore it is reasonable to suggest that in concentrating crime near SCS, the frequency of arrests or reported incidents may rise, driven by more effective policing in these focused areas. This increase may be a matter of statistics, where these crimes might still have taken place in the counterfactual situation, but they would have occurred elsewhere, with less effective police enforcement. In a sense, this could mean that while SCS might contribute to a localised increase in crime visibility and concentration the aggregate impact on overall crime levels may be minimal.

[87] See Connors (1992) for an ethnographic study on the risk of HIV and the decision to share needles.

A potential limitation of this study is the focus on medium-term trends without capturing longer-term outcomes. While we observe a moderate rise in localised breaking and entering incidents near SCS, it remains unclear how these trends might evolve over an extended period. Literature on hotspot policing suggests that effective, targeted enforcement can lead to long-term reductions in crime (Nagin, 2013). Therefore, without long-term data, we cannot fully determine whether the observed increase represents a temporary spike due to heightened concentration and policing or if it will translate into sustained or reduced crime rates over time. Future research should explore these long-term trends to better understand the enduring impact of SCS on localised and citywide crime patterns.

Finally, concerning our null results on the impact of SCS on police-administered mental health apprehensions, these findings suggest that SCS alone may not be sufficient to influence broader mental health-related incidents in their surrounding areas. While SCS provides crucial support for PWUD by mitigating immediate health risks, their role in addressing the complex mental health needs of the community may be limited. highlights the importance of integrating SCS with comprehensive mental health services and community resources to better support individuals facing mental health crises and enhance overall community well-being.

One key limitation is that the data is constrained in terms of other important outcomes, particularly those relating to community health such as non-emergency overdose incidents or public health changes not captured by callout statistics. Additionally, due to the granularity of the data and the difficulty in identifying unit level covariates, it might be difficult to isolate the specific impact of SCS from other simultaneous social and policy changes within the community (*i.e.,* economic shifts or grassroots community level overdose prevention sites). While our study gives a good indication of the effects of SCS in the short and medium term trends, we are unable to evaluate the potential long-term impacts of SCS on both health related outcomes or community spillovers.

Additionally, this study is limited by the inability to effectively quantify the underlying mechanisms driving the observed outcomes. While we can identify changes in overdose callouts and crime rates, pinpointing the precise behavioral or social processes responsible for these trends remains challenging. This limitation means that, although we observe the end effects, such as the lack of significant changes in overdose callouts or the moderate rise in breaking and entering, understanding the mechanisms—such as changes in user behavior, economic conditions, or community interactions—requires further investigation. This gap highlights the need for more in depth research that can delve deeper into the pathways through which SCS influence public health and community dynamics.

4.9 Conclusion

This paper provides the first comprehensive analyses using causal inference techniques to evaluate the impact of supervised consumption sites (SCS) on opioid-related emergency callouts, crime, and mental health crises in Toronto. Using a spatial variation of the Callaway and Sant'Anna (2021) DID estimator we leverage neighborhood-level Toronto Police Open Data to analyse the staggered rollout of 10 Health Canada-approved SCS between 2017 and 2021. We find no significant changes in emergency overdose callouts, incidents of assault, or mental health apprehensions.

However, our analysis points to an increase in opportunistic break-ins near SCS. These findings challenge the moral hazard concern, suggesting that PWUD may not engage in significantly riskier behavior due to perceived safety nets provided by SCS. Instead, it aligns with literature emphasizing the complex, often non-rational drivers behind substance use. The observed increase in localised property crime near SCS highlights potential community-level trade-offs that merit further investigation. Crime clustering and heightened policing efforts near SCS may contribute to increased visibility of such incidents, suggesting that while the sites themselves do not directly exacerbate citywide crime, their localised impacts may be more pronounced. This underscores the importance of contextualizing SCS as part of a broader public health approach that includes targeted crime prevention and support services to mitigate any potential negative spillovers.

Our findings speak to the complexity of harm reduction approaches, and show that while SCS may not lead to significant and large negative spillovers as feared by critics, their ability to significantly reduce emergency service burdens or improve mental health outcomes is limited.

Chapter 5

Thesis Conclusion

This thesis examines how public policy decisions in the areas of fiscal policy, education, and health have affected a wide range of economic, social, and public health outcomes. By analyzing austerity measures, school discipline reforms, and harm reduction initiatives, the findings shed light on how such interventions reshape individual behaviors, familial dynamics, and community-level outcomes. While these policies often target specific goals, their broader consequences reveal important lessons about the interplay between policy design and social structures, offering insights for crafting more effective and equitable approaches.

The analysis of UK austerity policies illustrate how fiscal reforms can produce unintended consequences for households. These measures restructured labor market participation and income dynamics within families, which potentially disproportionately burdening mothers while sparing fathers from comparable financial pressures. Beyond economic effects, austerity reshaped relational and household dynamics, influencing divorce rates, new relationships, and parental engagement with children. These findings underscore the far-reaching ripple effects of fiscal tightening, emphasizing the need for gender-sensitive policy frameworks that balance economic imperatives with family well-being.

The studies on educational reform demonstrate how thoughtfully designed policies can lead to meaningful improvements in disciplinary outcomes while also revealing the challenges of balancing social and academic objectives. The Oklahoma City School District's reforms successfully shifted disciplinary practices away from exclusionary measures, resulting in a 38% reduction in suspensions and a notable decrease in adolescent arrests. These outcomes highlight the potential of inclusive approaches, such as Positive Behavioral Interventions and Supports (PBIS), to foster safer, more supportive school environments. However, the reforms resulted in subject-specific academic impacts, with mixed results on reading performance and some identified gains in structured subjects like math and science. These findings highlight the complexity of educational interventions, emphasizing the need to carefully consider and address potential trade-offs to maximize the benefits of reforms across all subjects and student groups.

The evaluation of supervised consumption sites (SCS) in Toronto addresses the ongoing debate surrounding harm reduction strategies and their effectiveness in addressing the opioid crisis. While SCS were designed to reduce overdose-related callouts and enhance public safety, the study found no statistically significant changes in emergency services use or most crimes; however, there is some evidence of increased break-ins in neighborhoods near SCS locations. Additionally, the presence of SCS did not seem to exacerbate mental health crises, suggesting that they do not impose additional strain on community resources. These findings raise important questions about the role of SCS in harm reduction efforts, suggesting that while they do not significantly impact most aspects of public safety, their association with increased break-ins highlights potential negative consequences that require attention, and their capacity to alleviate broader systemic burdens may be limited.

Together, these studies demonstrate the importance of applying evidence-based approaches to policymaking, with this thesis making several key contributions. First, it quantifies the effects of benefit reduction programs on household dynamics, highlighting the far-reaching consequences of austerity measures on labor markets, gendered income disparities, and family relationships. Second, it examines the effects of school discipline reforms on juvenile arrests, providing new insights into how inclusive policies influence both academic outcomes and social behaviors. By integrating computer science techniques such as K-means clustering and synthetic difference-in-differences, this work also advances methodological approaches to policy evaluation. Finally, it provides the first robust analysis of the impact of supervised consumption sites on public safety and health outcomes, offering a nuanced perspective on harm reduction strategies. Collectively, these contributions offer valuable guidance for policymakers seeking to address complex social and economic challenges through more effective and equitable interventions.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program". In: *Journal of the American statistical Association* 105.490, pp. 493–505.

Abadie, Alberto and Javier Gardeazabal (2003). "The economic costs of conflict: A case study of the Basque Country". In: *American Economic Review* 93.1, pp. 113–132.

Abdi, Hervé and Lynne J Williams (2010). "Principal component analysis". In: *Wiley interdisciplinary reviews: computational statistics* 2.4, pp. 433–459.

Adam, Stuart and James Browne (2013). *Do the UK Government's welfare reforms make work pay*. Report. IFS Working Papers.

Advisors, White House Council of Economic (2017). "The underestimated cost of opioid crisis". In: *Retrieved from*.

Agarwal, Anish, Devavrat Shah, and Dennis Shen (2020). "On model identification and out-of-sample prediction of principal component regression: Applications to synthetic controls". In: *arXiv preprint arXiv:2010.14449*.

Agarwal, Bina (1997). ""Bargaining"and gender relations: Within and beyond the household". In: *Feminist economics* 3.1, pp. 1–51.

Agostinelli, Francesco and Giuseppe Sorrenti (2021). "Money vs. time: family income, maternal labor supply, and child development". In: *University of Zurich, Department of Economics, Working Paper* 273.

Ahmad, Farida B, Lauren M Rossen, and Paul Sutton (2021). "Provisional drug overdose death counts". In: *National Center for Health Statistics* 12.

Aizer, Anna and Joseph J Doyle Jr (2015). "Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges". In: *The Quarterly Journal of Economics* 130.2, pp. 759–803.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney (2016). "The long-run impact of cash transfers to poor families". In: *American Economic Review* 106.4, pp. 935–71.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay (2021). "Building social cohesion in ethnically mixed schools: An intervention on perspective taking". In: *The Quarterly Journal of Economics* 136.4, pp. 2147–2194.

Alesina, Alberto, Carlo Favero, and Francesco Giavazzi (2019). "Effects of austerity: Expenditure-and tax-based approaches". In: *Journal of Economic Perspectives* 33.2, pp. 141–62.

— (2020). *Austerity: When it Works and when it Doesn't*. Princeton University Press.

Anderson, Kaitlin P (2021). "The relationship between inclusion, absenteeism, and disciplinary outcomes for students with disabilities". In: *Educational Evaluation and Policy Analysis* 43.1, pp. 32–59.

Anderson, Kaitlin P, Gary W Ritter, and Gema Zamarro (2019). "Understanding a vicious cycle: The relationship between student discipline and student academic outcomes". In: *Educational Researcher* 48.5, pp. 251–262.

Angrist, Joshua D (2014). "The perils of peer effects". In: *Labour Economics* 30, pp. 98–108.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager (2021). "Synthetic difference-in-differences". In: *American Economic Review* 111.12, pp. 4088–4118.

Ashton, Sam (2014). *A New Poll Tax?: The Impact of the Abolishment of the Council Tax Benefit in London*. Report. Child Poverty Action Group.

Austin, Peter C (2011). "An introduction to propensity score methods for reducing the effects of confounding in observational studies". In: *Multivariate behavioral research* 46.3, pp. 399–424.

Bacher-Hicks, Andrew, Stephen B Billings, and David J Deming (2019). *The school to prison pipeline: Long-run impacts of school suspensions on adult crime*. Report 26257. National Bureau of Economic Research.

Bastiaansen, Coco, Emmie Verspeek, and Hedwig van Bakel (2021). "Gender differences in the mitigating effect of co-parenting on parental burnout: The gender dimensión applied to COVID-19 restrictions and parental burnout levels". In: *Social Sciences* 10.4, p. 127.

Basu, Bharati and Pushkar Maitra (2020). “Intra-household bargaining power and household expenditure allocation: Evidence from Iran”. In: *Review of Development Economics* 24.2, pp. 606–627.

Bayani, Mani (2021). “Robust PCA Synthetic Control”. In: *arXiv preprint arXiv:2108.12542*.

Beaman, Robyn, Kevin Wheldall, and Coral Kemp (2007). “Recent research on troublesome classroom behaviour: A review”. In: *Australasian Journal of Special Education* 31.1, pp. 45–60.

Becker, Gary S (1968). “Crime and Punishment: An Economic Approach”. In: *The Economic Dimensions of Crime/Springer*.

Becker, Gary S and Kevin M Murphy (1988). “A theory of rational addiction”. In: *Journal of political Economy* 96.4, pp. 675–700.

Behrends, Czarina N, Denise Paone, Michelle L Nolan, Ellenie Tuazon, Sean M Murphy, Shashi N Kapadia, Philip J Jeng, Ahmed M Bayoumi, Hillary V Kunins, and Bruce R Schackman (2019). “Estimated impact of supervised injection facilities on overdose fatalities and healthcare costs in New York City”. In: *Journal of substance abuse treatment* 106, pp. 79–88.

Belzak, Lisa and Jessica Halverson (2018). “Evidence synthesis-The opioid crisis in Canada: a national perspective”. In: *Health promotion and chronic disease prevention in Canada: research, policy and practice* 38.6, p. 224.

Bennett, Fidel, Dante Contreras, and Matías Morales Cerda (2022). “The consequences of exclusionary discipline on school dropout: Evidence from Chile”. In: *International Journal of Educational Development* 95, p. 102671.

Bennett, Trevor, Katy Holloway, and David Farrington (2008). “The statistical association between drug misuse and crime: A meta-analysis”. In: *Aggression and Violent Behavior* 13.2, pp. 107–118.

Bertrand, Marianne and Jessica Pan (2013). “The trouble with boys: Social influences and the gender gap in disruptive behavior”. In: *American economic journal: applied economics* 5.1, pp. 32–64.

Binswanger, Ingrid A, Deborah Rinehart, Shane R Mueller, Komal J Narwaney, Melanie Stowell, Nicole Wagner, Stan Xu, Rebecca Hanratty, Josh Blum, and Kevin McVaney (2022). “Naloxone co-dispensing with opioids: a cluster randomized pragmatic trial”. In: *Journal of general internal medicine* 37.11, pp. 2624–2633.

Bitler, Marianne P, Jonah B Gelbach, Hilary W Hoynes, and Madeline Zavodny (2004). “The impact of welfare reform on marriage and divorce”. In: *Demography* 41.2, pp. 213–236.

Bondurant, Samuel R, Jason M Lindo, and Isaac D Swensen (2018). “Substance abuse treatment centers and local crime”. In: *Journal of Urban Economics* 104, pp. 124–133.

Boyd, Neil (2013). “Lessons from INSITE, Vancouver’s supervised injection facility: 2003–2012”. In: *Drugs: Education, Prevention and Policy* 20.3, pp. 234–240.

Breda, Thomas, Elyès Jouini, and Clotilde Napp (2018). “Societal inequalities amplify gender gaps in math”. In: *Science* 359.6381, pp. 1219–1220.

Brewer, Mike, Thang Dang, and Emma Tominey (2024). “Universal Credit: Welfare Reform and Mental Health”. In: *Journal of Health Economics* 98, p. 102940.

Browning, Martin and Pierre-André Chiappori (1998). “Efficient intra-household allocations: A general characterization and empirical tests”. In: *Econometrica*, pp. 1241–1278.

Brownstein, Henry H (2015). “Drugs and violent crime”. In: *The Handbook of Drugs and Society*, pp. 369–386.

Callaway, Brantly and Pedro HC Sant’Anna (2021). “Difference-in-differences with multiple time periods”. In: *Journal of econometrics* 225.2, pp. 200–230.

Canada, Government of (2016). *Government of Canada announces new comprehensive drug strategy supported by proposed legislative changes*. Press Release. URL: `https://www.canada.ca/en/health-canada/news/2016/12/government-canada-announces-new-comprehensive-drug-strategy-supported-proposed-legislative-changes.html`.

Canada, Ottowa Public Health Agency of (2024a). *Opioid- and Stimulant-related Harms in Canada*. Government Document. URL: `https://health-infobase.canada.ca/substance-related-harms/opioids-stimulants/`.

Canada, Transport (2024b). *Canaidan Motor Vehicle Traffic Collision Statistics: 2022*. Aggregated Database. URL: `https://tc.canada.ca/en/road-transportation/statistics-data/canadian-motor-vehicle-traffic-collision-statistics-2022#`.

Carlson, D, R Petts, and J Pepin (2021). “Changes in US Parents’ Domestic Labor During the Early Days of the COVID-19 Pandemic”. In: *Sociological Inquiry*.

Carneiro, Pedro Manuel, Italo Lopez Garcia, Kjell G. Salvanes, and Emma Tominey (2015). “Intergenerational mobility and the timing of parental income”. In: *NHH Dept. of Economics Discussion Paper* 23.

Carrell, Scott E and Mark Hoekstra (2014). “Are school counselors an effective education input?” In: *Economics Letters* 125.1, pp. 66–69.

Carrell, Scott E and Mark L Hoekstra (2010). "Externalities in the classroom: How children exposed to domestic violence affect everyone's kids". In: *American Economic Journal: Applied Economics* 2.1, pp. 211–228.

Cassie, Rachel, Kanna Hayashi, Kora DeBeck, M-J Milloy, Zishan Cui, Carol Strike, Jeff West, and Mary Clare Kennedy (2022). "Difficulty accessing supervised consumption services during the COVID-19 pandemic among people who use drugs in Vancouver, Canada". In: *Harm reduction journal* 19.1, p. 126.

Cawley, John and Davide Dragone (2023). *Harm reduction: When does it improve health, and when does it backfire?* Report. National Bureau of Economic Research.

Chen, Jiafeng and Jonathan Roth (2024). "Logs with zeros? Some problems and solutions". In: *The Quarterly Journal of Economics* 139.2, pp. 891–936.

Chen, Natalie, Paola Conconi, and Carlo Perroni (2007). "Women's Earning Power and the Double Burden of Market and Household Work". In: 20.

Chen, Xiao, Bihong Huang, and Dezhu Ye (2020). "Gender gap in peer-to-peer lending: Evidence from China". In: *Journal of Banking Finance* 112, p. 105633.

Choe, Chung and Seunghee Yu (2022). "The effect of child abuse and neglect on trajectories of depressive symptoms and aggression in Korean adolescents: exploring gender differences". In: *International journal of environmental research and public health* 19.10, p. 6160.

Clarke, Damian, Daniel Pailanir, Susan Athey, and Guido Imbens (2023). "Synthetic difference in differences estimation". In: *arXiv preprint arXiv:2301.11859*.

Clarke, Damian and Kathya Tapia-Schythe (2021). "Implementing the panel event study". In: *The Stata Journal* 21.4, pp. 853–884.

Colledge-Frisby, Samantha, Kasun Rathnayake, Suzanne Nielsen, Mark Stoove, Lisa Maher, Paul A Agius, Peter Higgs, and Paul Dietze (2023). "Injection drug use frequency before and after take-home naloxone training". In: *JAMA network open* 6.8, e2327319–e2327319.

Connors, Margaret M (1992). "Risk perception, risk taking and risk management among intravenous drug users: Implications for AIDS prevention". In: *Social Science Medicine* 34.6, pp. 591–601.

Craig, Ashley C and David Martin (2023). *Discipline reform, school culture, and student achievement*. Report 15906. IZA Discussion Papers.

Crandall-Hollick, Margot L. (2018). "The earned income tax credit (EITC): A brief legislative history". In: *Congressional Research Service Report* 44825, p. 2018.

Crews, Fulton Timm and Charlotte Ann Boettiger (2009). “Impulsivity, frontal lobes and risk for addiction”. In: *Pharmacology Biochemistry and Behavior* 93.3, pp. 237–247.

Cunha, Flavio and James Heckman (2007). “The technology of skill formation”. In: *American economic review* 97.2, pp. 31–47.

Cunha, Flavio and James J. Heckman (2008). “Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation”. In: *Journal of human resources* 43.4, pp. 738–782.

De Brey, Cristobal, Lauren Musu, Joel McFarland, Sidney Wilkinson-Flicker, Melissa Diliberti, Anlan Zhang, Claire Branstetter, and Xiaolei Wang (2019). “Status and Trends in the Education of Racial and Ethnic Groups 2018. NCES 2019-038”. In: *National Center for Education Statistics*.

De Chaisemartin, Clément and Xavier D’Haultfoeuille (2022). *Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey*. Report. National Bureau of Economic Research.

Deiana, Claudio and Ludovica Giua (2021). “The intended and unintended effects of opioid policies on prescription opioids and crime”. In: *The BE Journal of Economic Analysis Policy* 21.2, pp. 751–792.

Dench, Daniel, Mayra Pineda-Torres, and Caitlin Knowles Myers (2023). “The effects of the Dobbs decision on fertility”. In: 16608.

Diedrick, Patricia (1991). “Gender differences in divorce adjustment”. In: *Journal of Divorce Remarriage* 14.3-4, pp. 33–46.

Dieterle, Steven, Cassandra M Guarino, Mark D Reckase, and Jeffrey M Wooldridge (2015). “How do principals assign students to teachers? Finding evidence in administrative data and the implications for value added”. In: *Journal of Policy Analysis and Management* 34.1, pp. 32–58.

Doepke, Matthias, Giuseppe Sorrenti, and Fabrizio Zilibotti (2019). “The economics of parenting”. In: *Annual Review of Economics* 11, pp. 55–84.

Doepke, Matthias and Fabrizio Zilibotti (2017). “Parenting with style: Altruism and paternalism in intergenerational preference transmission”. In: *Econometrica* 85.5, pp. 1331–1371.

DoJ, US (2011). “About the Uniform Crime Reporting (UCR) Program”. In: URL: https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/aboutucrmain.pdf.

Doleac, Jennifer L and Anita Mukherjee (2018). “The moral hazard of lifesaving innovations: naloxone access, opioid abuse, and crime”. In.

— (2022). “The effects of naloxone access laws on opioid abuse, mortality, and crime”. In: *The Journal of Law and Economics* 65.2, pp. 211–238.

Dong, Xinwei (2022). “Intrahousehold property ownership, women’s bargaining power, and family structure”. In: *Labour Economics* 78, p. 102239.

Dossi, Gaia, David Figlio, Paola Giuliano, and Paola Sapienza (2021). “Born in the family: Preferences for boys and the gender gap in math”. In: *Journal of Economic Behavior Organization* 183, pp. 175–188.

Eissa, Nada and Hilary W Hoynes (2006). “Behavioral responses to taxes: Lessons from the EITC and labor supply”. In: *Tax policy and the economy* 20, pp. 73–110.

Ellis, Rnady (2015). *Mental health center files civil rights complaint against OKC school district*. Newspaper Article. URL: https://www.oklahoman.com/story/news/columns/2015/11/28/mental-health-center-files-civil-rights-complaint-against-okc-school-district/60707258007/.

Entwisle, Doris R, Karl L Alexander, and Linda S Olson (2007). “Early schooling: The handicap of being poor and male”. In: *Sociology of education* 80.2, pp. 114–138.

Erfanian, Elham, Daniel Grossman, and Alan R Collins (2019a). “The impact of naloxone access laws on opioid overdose deaths in the US”. In: *Review of Regional Studies* 49.1, pp. 45–72.

— (2019b). “The impact of naloxone access laws on opioid overdose deaths in the US”. In: *Review of Regional Studies* 49.1, pp. 45–72.

Evans, Elizabeth, Christine E Grella, Debra A Murphy, and Yih-Ing Hser (2010). “Using administrative data for longitudinal substance abuse research”. In: *The journal of behavioral health services & research* 37, pp. 252–271.

Farabee, David, Angela Hawken, and Peter Griffith (2011). “Tracking and incentivizing substance abusers in longitudinal research: results of a survey of National Institute on Drug Abuse-funded investigators”. In: *Journal of addiction medicine* 5.2, pp. 87–91.

Farber, Henry S, Jesse Rothstein, and Robert G Valletta (2015). “The effect of extended unemployment insurance benefits: Evidence from the 2012–2013 phase-out”. In: *American Economic Review* 105.5, pp. 171–176.

Felder, Ben (Oct. 2017). *Oklahoma changes test score rates, prepares for ‘shock to the system’*. Newspaper Article. URL: https://www.oklahoman.com/story/news/

education/2017/10/09/oklahoma-changes-test-score-rates-prepares-for-shock-to-the-system/60569794007/.

Fetzer, Thiemo (2019). "Did austerity cause Brexit?" In: *American Economic Review* 109.11, pp. 3849–86.

Fischer, Benedikt (2023). "The continuous opioid death crisis in Canada: changing characteristics and implications for path options forward". In: *The Lancet Regional Health–Americas* 19.

Fisher, Hayley and Anna Zhu (2019). "The effect of changing financial incentives on repartnering". In: *The Economic Journal* 129.623, pp. 2833–2866.

Florence, Curtis, Feijun Luo, and Ketra Rice (2021). "The economic burden of opioid use disorder and fatal opioid overdose in the United States, 2017". In: *Drug and alcohol dependence* 218, p. 108350.

Franco, Sofia and Hans Koster (2024). *Drug-related harm reduction and local communities: Evidence from Dutch drug consumption rooms*. Report. CEPR Discussion Papers.

Freeman, Karen, Craig GA Jones, Don J Weatherburn, Scott Rutter, Catherine J Spooner, and Neil Donnelly (2005). "The impact of the Sydney medically supervised injecting centre (MSIC) on crime". In: *Drug and Alcohol Review* 24.2, pp. 173–184.

Freyaldenhoven, Simon, Christian Hansen, Jorge Pérez Pérez, and Jesse M. Shapiro (2021 2021). *Visualization, identification, and estimation in the linear panel event-study design*. Report. National Bureau of Economic Research.

Friedman, Abigail S (2015). "How does electronic cigarette access affect adolescent smoking?" In: *Journal of health economics* 44, pp. 300–308.

Friedman, Abigail S and Michael F Pesko (2022). "Young adult responses to taxes on cigarettes and electronic nicotine delivery systems". In: *Addiction* 117.12, pp. 3121–3128.

Fryer Jr, Roland G and Steven D Levitt (2010). "An empirical analysis of the gender gap in mathematics". In: *American Economic Journal: Applied Economics* 2.2, pp. 210–240.

Gage, Nicholas A, Lydia Beahm, Rachel Kaplan, Ashley S MacSuga-Gage, and Ahhyun Lee (2020). "Using positive behavioral interventions and supports to reduce school suspensions". In: *Beyond Behavior* 29.3, pp. 132–140.

Gallego, Francisco, Philip Oreopoulos, and Noah Spencer (2023). *The importance of a helping hand in education and in life*. Report 31706. National Bureau of Economic Research.

García-Echalar, Andrés, Sebastián Poblete, and Tomás Rau (2024). "Teacher Value-Added and the Test Score Gender Gap". In: *Labour Economics* 89.102588.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". In: *Journal of econometrics* 225.2, pp. 254–277.

Gopalan, Maithreyi and Ashlyn Aiko Nelson (2019). "Understanding the racial discipline gap in schools". In: *Aera Open* 5.2, p. 2332858419844613.

Gornick, Janet C. and Timothy M. Smeeding (2018). "Redistributional policy in rich countries: Institutions and impacts in nonelderly households". In: *Annual Review of Sociology* 44, p. 441.

Gottfried, Michael A (2014). "Chronic absenteeism and its effects on students' academic and socioemotional outcomes". In: *Journal of Education for Students Placed at Risk (JESPAR)* 19.2, pp. 53–75.

Goulas, Sofoklis, Rigissa Megalokonomou, and Yi Zhang (2023). "Female Classmates, Disruption, and STEM Outcomes in Disadvantaged Schools: Evidence from a Randomized Natural Experiment". In: 16689.

Government, Canadian (2013). *Harper Govenremnt respects community concerns with new legislation for Supervised Drug Consumption Sites*. Press Release. URL: `https://www.canada.ca/en/news/archive/2013/06/harper-government-respects-community-concerns-new-legislation-supervised-drug-consumption-sites.html`.

Goyer, Camille, Genaro Castillon, and Yola Moride (2022). "Implementation of interventions and policies on opioids and awareness of opioid-related harms in Canada: a multistage mixed methods descriptive study". In: *International journal of environmental research and public health* 19.9, p. 5122.

Greenwald, Zoë R, Zachary Bouck, Elizabeth McLean, Kate Mason, Bernadette Lettner, Jennifer Broad, Zoë Dodd, Tanner Nassau, Ayden I Scheim, and Dan Werb (2023). "Integrated supervised consumption services and hepatitis C testing and treatment among people who inject drugs in Toronto, Canada: A cross-sectional analysis". In: *Journal of viral hepatitis* 30.2, pp. 160–171.

Gregory, Anne, Russell J Skiba, and Pedro A Noguera (2010). "The achievement gap and the discipline gap: Two sides of the same coin?" In: *Educational researcher* 39.1, pp. 59–68.

Guiso, Luigi, Ferdinando Monte, Paola Sapienza, and Luigi Zingales (2008). "Culture, gender, and math". In: *Science* 320.5880, pp. 1164–1165.

Hartigan, JA (1979). "A k-means clustering algorithm". In: *JR Stat. Soc. Ser. C-Appl. Stat.* 28, pp. 100–108.

Hayle, Steven (2018). "A tale of two Canadian cities: Comparing supervised consumption site (SCS) policy making in Toronto and Vancouver". In: *Drugs: Education, Prevention and Policy* 25.5, pp. 397–407.

He, Fang (2016). *An Economic Analysis of the Relationship between Household Income and Fertility*. Report. Institute for Economics Studies, Keio University.

Heckman, James J. (2005). "Lessons from the technology of skill formation". In: *NBER Working Paper Series* 11142.

Heckman, James J. and Stefano Mosso (2014). "The economics of human development and social mobility". In: *Annu. Rev. Econ.* 6.1, pp. 689–733.

Henney, Susan M. (2016). "The relationship between personality and parental confidence in mothers of school-aged children". In: *SAGE Open* 6.3, p. 2158244016659317.

Horning, Melissa L, Robin Schow, Sarah E Friend, Katie Loth, Dianne Neumark-Sztainer, and Jayne A Fulkerson (2017). "Family dinner frequency interacts with dinnertime context in associations with child and parent BMI outcomes". In: *Journal of Family Psychology* 31.7, p. 945.

Houborg, Esben, Vibeke Asmussen Frank, and Bagga Bjerge (2014). "From zero tolerance to non-enforcement: Creating a new space for drug policing in Copenhagen, Denmark". In: *Contemporary Drug Problems* 41.2, pp. 261–291.

Hyshka, Elaine, Jalene Anderson, Zing-Wae Wong, and T Cameron Wild (2016). *Risk behaviours and service needs of marginalized people who use drugs in Edmonton's inner city*. Generic.

Jacobs, Leah A, Laura Ellen Ashcraft, Craig JR Sewall, Barbara L Folb, and Christina Mair (2020). "Ecologies of juvenile reoffending: A systematic review of risk factors". In: *Journal of Criminal justice* 66, p. 101638.

Jones, Blake L (2018). "Making time for family meals: Parental influences, home eating environments, barriers and protective factors". In: *Physiology behavior* 193, pp. 248–251.

Karamessini, Maria and Jill Rubery (2013). *Women and austerity*. Taylor Francis.

Kinn, Daniel (2018). "Synthetic control methods and big data". In: *arXiv preprint arXiv:1803.00096*.

Kothari, Brianne H, Bethany Godlewski, Bowen McBeath, Marjorie McGee, Jeff Waid, Shannon Lipscomb, and Lew Bank (2018a). "A longitudinal analysis of school discipline events among youth in foster care". In: *Children and Youth Services Review* 93, pp. 117–125.

— (2018b). "A longitudinal analysis of school discipline events among youth in foster care". In: *Children and Youth Services Review* 93, pp. 117–125.

Kranz, Sebastian (2022). "Synthetic difference-in-differences with time-varying covariates". In: *ArXiv Preprint ArXiv:2202.02903*. URL: `https://github.com/skranz/xsynthdid/blob/main/paper/synthdid_with_covariates.pdf`.

Lenhart, Otto (2019). "Higher wages, less gym time? The effects of minimum wages on time use". In: *Southern Economic Journal* 86.1, pp. 253–270.

Levengood, Timothy W, Grace H Yoon, Melissa J Davoust, Shannon N Ogden, Brandon DL Marshall, Sean R Cahill, and Angela R Bazzi (2021). "Supervised injection facilities as harm reduction: a systematic review". In: *American journal of preventive medicine* 61.5, pp. 738–749.

Li, Zehang Richard, Evaline Xie, Forrest W Crawford, Joshua L Warren, Kathryn McConnell, J Tyler Copple, Tyler Johnson, and Gregg S Gonsalves (2019). "Suspected heroin-related overdoses incidents in Cincinnati, Ohio: A spatiotemporal analysis". In: *PLoS medicine* 16.11, e1002956.

Liang, Jian and Sergey Alexeev (2023a). "Harm reduction or amplification? The adverse impact of a supervised injection room on housing prices". In: *Regional Science and Urban Economics* 98, p. 103856.

— (2023b). "Harm reduction or amplification? The adverse impact of a supervised injection room on housing prices". In: *Regional Science and Urban Economics* 98, p. 103856.

Lincove, Jane Arnold, Catherine Mata, and Kalena Cortes (2024). *The Effects of a Statewide Ban on School Suspensions*. Report 33086. National Bureau of Economic Research.

Lloyd-Smith, Elisa, Evan Wood, Ruth Zhang, Mark W Tyndall, Sam Sheps, Julio SG Montaner, and Thomas Kerr (2010). "Determinants of hospitalization for a cutaneous injection-related infection among injection drug users: a cohort study". In: *BMC Public Health* 10, pp. 1–7.

Losen, Daniel J, Cheri L Hodson, Michael A Keith II, Katrina Morrison, and Shakti Belway (2015). "Are we closing the school discipline gap?" In.

Maclean, Johanna Catherine, Justine Mallatt, Christopher J Ruhm, and Kosali Simon (2022). "The opioid crisis, health, healthcare, and crime: A review of quasi-experimental economic studies". In: *The ANNALS of the American Academy of Political and Social Science* 703.1, pp. 15–49.

Makowiecki, Barbara (2014). "Great Britain's welfare reform act of 2012-Implementation overview, preliminary impact, and future implications". In: *Int'l Law.* 48, p. 243.

Marlatt, G Alan (1996). "Harm reduction: Come as you are". In: *Addictive behaviors* 21.6, pp. 779–788.

Milligan, Kevin and Mark Stabile (2011). "Do child tax benefits affect the well-being of children? Evidence from Canadian child benefit expansions". In: *American Economic Journal: Economic Policy* 3.3, pp. 175–205.

Mishra, Vaibhav, Golnoush Seyedzenouzi, Ahmad Almohtadi, Tasnim Chowdhury, Arwa Khashkhusha, Ariana Axiaq, Wing Yan Elizabeth Wong, and Amer Harky (2021). "Health inequalities during COVID-19 and their effects on morbidity and mortality". In: *Journal of healthcare leadership* 13, p. 19.

Mohler, George O, Martin B Short, Sean Malinowski, Mark Johnson, George E Tita, Andrea L Bertozzi, and P Jeffrey Brantingham (2015). "Randomized controlled field trials of predictive policing". In: *Journal of the American statistical association* 110.512, pp. 1399–1411.

Molina, José Alberto, Jorge Velilla, and Helena Ibarra (2023). "Intrahousehold bargaining power in Spain: An empirical test of the collective model". In: *Journal of Family and Economic Issues* 44.1, pp. 84–97.

Mowen, Thomas and John Brent (2016). "School discipline as a turning point: The cumulative effect of suspension on arrest". In: *Journal of research in crime and delinquency* 53.5, pp. 628–653.

Nagin, Daniel S (2013). "Deterrence: A review of the evidence by a criminologist for economists". In: *Annu. Rev. Econ.* 5.1, pp. 83–105.

NCES (2019). *Indicator 15: Retention, Suspension, and Expulsion*. Web Page. URL: https://nces.ed.gov/programs/raceindicators/indicator_rda.asp#:~:text=More%20than%20twice%20as%20many,for%20all%20racial%2Fethnic%20groups..

Nigatu, Hellina Hailu, Lisa Pickoff-White, John Canny, and Sarah Chasins (n.d.). "Co-Designing for Transparency: Lessons from Building a Document Organization Tool in the Criminal Justice Domain". In: *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pp. 1463–1478.

No, Sung Chul, Donald Andrews, and Ashagre Yigletu (2007). "Dynamic Analysis of Income and Independence Effect of African American Female Labor Force Participation on Divorce". In: *Atlantic Economic Journal* 35, pp. 159–171.

Norman, Helen and Mark Elliot (2015). "Measuring paternal involvement in childcare and housework". In: *Sociological Research Online* 20.2, pp. 40–57.

Nowicki, Jacqueline M (2018). "K-12 Education: Discipline Disparities for Black Students, Boys, and Students with Disabilities. Report to Congressional Requesters. GAO-18-258". In: *US Government Accountability Office*.

Nunley, John M and Joachim Zietz (2012). "The long-run impact of age demographics on the US divorce rate". In: *The American Economist* 57.1, pp. 65–77.

Olmos, Antonio and Priyalatha Govindasamy (2015). "Propensity scores: a practical introduction using R". In: *Journal of MultiDisciplinary Evaluation* 11.25, pp. 68–88.

Osodo, John Mark, Joseph Osodo, Jane Wagumba Mito, Pamela Raburu, and Peter Aloka (2016). "The Role of Peer Counselors in the Promotion of Student Discipline in Ugunja Sub-County, Kenya". In: *Asian Journal of Education and Training* 2.2, pp. 63–69.

Packham, Analisa (2019). *Are syringe exchange programs helpful or harmful? New evidence in the wake of the opioid epidemic*. Report. National Bureau of Economic Research.

PBIS.org (2020). "Classroom PBIS". In: URL: `https://www.pbis.org/classroom-pbis`.

Peltzman, Sam (1975). "The effects of automobile safety regulation". In: *Journal of political Economy* 83.4, pp. 677–725.

Périvier, Hélène (2018a). "Recession, austerity and gender: A comparison of eight European labour markets". In: *International Labour Review* 157.1, pp. 1–37.

— (2018b). "Recession, austerity and gender: A comparison of eight European labour markets". In: *International Labour Review* 157.1, pp. 1–37.

Perry, Brea L and Edward W Morris (2014). "Suspending progress: Collateral consequences of exclusionary punishment in public schools". In: *American Sociological Review* 79.6, pp. 1067–1087.

Perugini, Cristiano, Jelena Žarković Rakić, and Marko Vladisavljević (2019). "Austerity and gender inequalities in Europe in times of crisis". In: *Cambridge Journal of Economics* 43.3, pp. 733–767.

Pieters, Janneke and Samantha Rawlings (2020). "Parental unemployment and child health in China". In: *Review of Economics of the Household* 18.1, pp. 207–237.

Pilkauskas, Natasha, Katherine Michelmore, Nicole Kovski, and H Luke Shaefer (2022). *The effects of income on the economic wellbeing of families with low incomes: Evidence from the 2021 expanded Child Tax Credit*. Report. National Bureau of Economic Research.

Pope, Nolan G and George W Zuo (2023). "Suspending suspensions: The education production consequences of school suspension policies". In: *The Economic Journal* 133.653, pp. 2025–2054.

Porreca, Zachary (2022). "Synthetic difference-in-differences estimation with staggered treatment timing". In: *Economics Letters* 220, p. 110874.

Price, Joseph, Luke P Rodgers, and Jocelyn S Wikle (2021). “Dinner timing and human capital investments in children”. In: *Review of Economics of the Household* 19.4, pp. 1047–1075.

Rambachan, Ashesh and Jonathan Roth (2020). “Design-Based Uncertainty for Quasi-Experiments”. In: *arXiv preprint arXiv:2008.00602*.

— (2023). “A more credible approach to parallel trends”. In: *Review of Economic Studies* 90.5, pp. 2555–2591.

Rammohan, Indhu, Tommi Gaines, Ayden Scheim, Ahmed Bayoumi, and Dan Werb (2024a). “Overdose mortality incidence and supervised consumption services in Toronto, Canada: an ecological study and spatial analysis”. In: *The Lancet Public Health* 9.2, e79–e87.

— (2024b). “Overdose mortality incidence and supervised consumption services in Toronto, Canada: an ecological study and spatial analysis”. In: *The Lancet Public Health* 9.2, e79–e87.

Ratcliffe, Jerry H, Travis Taniguchi, Elizabeth R Groff, and Jennifer D Wood (2011). “The Philadelphia foot patrol experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots”. In: *Criminology* 49.3, pp. 795–831.

Rees, Daniel I, Joseph J Sabia, Laura M Argys, Joshua Latshaw, and Dhaval Dave (2017). *With a little help from my friends: the effects of naloxone access and good samaritan laws on opioid-related deaths*. Report. national Bureau of economic research.

Robinson, Karen Jeong and Josipa Roksa (2016). “Counselors, information, and high school college-going culture: Inequalities in the college application process”. In: *Research in Higher Education* 57, pp. 845–868.

Rømer Thomsen, Kristine, Mette Buhl Callesen, Morten Hesse, Timo Lehmann Kvamme, Michael Mulbjerg Pedersen, Mads Uffe Pedersen, and Valerie Voon (2018). “Impulsivity traits and addiction-related behaviors in youth”. In: *Journal of behavioral addictions* 7.2, pp. 317–330.

Romiti, Agnese (2018). “The effects of immigration on household services, labour supply and fertility”. In: *Oxford Bulletin of Economics and Statistics* 80.4, pp. 843–869.

Schaefer, Maximilian and Dimitra Panagiotoglou (2024). “Evaluating the effects of supervised consumption sites on housing prices in Montreal, Canada using interrupted time series and hedonic price models”. In: *Drug and Alcohol Dependence Reports* 11, p. 100242.

Schmidheiny, Kurt and Sebastian Siegloch (2019). “On event study designs and distributed-lag models: Equivalence, generalization and practical implications”. In: *IZA Discussion Paper Series* 12079.

Schools, Oklahoma City Public (2016). “Student Rights, Responsibilities, and Code of Conduct”. In: URL: `https://www.okcps.org/cms/lib/OK01913268/Centricity/Domain/911/OKCPS%20Code%20of%20Conduct%202015-2016.pdf`.

Siennick, Sonja E and Alex O Widdowson (2020). “Juvenile arrest and later economic attainment: Strength and mechanisms of the relationship”. In: *Journal of Quantitative Criminology*, pp. 1–28.

Sim, Yongbo (2023). “The effect of opioids on crime: Evidence from the introduction of OxyContin”. In: *International Review of Law and Economics* 74, p. 106136.

Single, Eric (1995). “Defining harm reduction”. In: *Drug and Alcohol Review* 14.3, pp. 287–290.

Skiba, Russell J, Mariella I Arredondo, and Natasha T Williams (2014). “More than a metaphor: The contribution of exclusionary discipline to a school-to-prison pipeline”. In: *Equity Excellence in Education* 47.4, pp. 546–564.

Ştefan, Catrinel A and Julia Avram (2017). “Investigating direct and indirect effects of attachment on internalizing and externalizing problems through emotion regulation in a cross-sectional study”. In: *Journal of Child and Family Studies* 26.8, pp. 2311–2323.

Strathdee, Steffanie A, Erin P Ricketts, Steven Huettner, Lee Cornelius, David Bishai, Jennifer R Havens, Peter Beilenson, Charles Rapp, Jacqueline J Lloyd, and Carl A Latkin (2006). “Facilitating entry into drug treatment among injection drug users referred from a needle exchange program: Results from a community-based behavioral intervention trial”. In: *Drug and alcohol dependence* 83.3, pp. 225–232.

Strathdee, Stephanie A and David Vlahov (2001). “The effectiveness of needle exchange programs: a review of the science and policy”. In: *AIDScience* 1.16, pp. 1–33.

Sun, Liyang and Sarah Abraham (2021). “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects”. In: *Journal of econometrics* 225.2, pp. 175–199.

Sung, Hung-En (2003). “Differential impact of deterrence vs. rehabilitation as drug interventions on recidivism after 36 months”. In: *Journal of Offender Rehabilitation* 37.3-4, pp. 95–108.

Toronto, Public Health (2024). *Toronto Overdose Information System*. Dataset. URL: `https://public.tableau.com/app/profile/tphseu/viz/TOISDashboard_Final/ParamedicResponse`.

Verdugo, Gregory and Guillaume Allègre (2020). “Labour force participation and job polarization: Evidence from Europe during the Great Recession”. In: *Labour Economics* 66, p. 101881.

Wald, Johanna and Daniel J Losen (2003). “Defining and redirecting a school-to-prison pipeline”. In: *New directions for youth development* 2003.99, pp. 9–15.

Wang, Ninghua, Lin Liu, and John E Eck (2014). “Analyzing crime displacement with a simulation approach”. In: *Environment and Planning B: Planning and Design* 41.2, pp. 359–374.

Weisburd, David, Laura A Wyckoff, Justin Ready, John E Eck, Joshua C Hinkle, and Frank Gajewski (2006). “Does crime just move around the corner? A controlled study of spatial displacement and diffusion of crime control benefits”. In: *Criminology* 44.3, pp. 549–592.

Wendler, Emily (2015). *High Suspension Rates at Oklahoma City Public Schools*. Newspaper Article.

White, Gregory (2016). “The impact of welfare reform on the social services workforce”. In: 32.

Willert, Tim (2015). *Oklahoma City school District begins discipline intervention training*. Web Page. URL: https://eu.oklahoman.com/story/news/local/oklahoma-city/2015/08/16/oklahoma-city-school-district-begins-discipline-intervention-training/60729304007/.

Wodak, Alex and Annie Cooney (2006). “Do needle syringe programs reduce HIV infection among injecting drug users: a comprehensive review of the international evidence”. In: *Substance use misuse* 41.6-7, pp. 777–813.

Wooldridge, Jeffrey M (2021). “Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators”. In: *Available at SSRN 3906345*.

Young, Samantha and Nadia Fairbairn (2018). “Expanding supervised injection facilities across Canada: lessons from the Vancouver experience”. In: *Canadian Journal of Public Health* 109, pp. 227–230.

A Appendix A

A.1 Summary Stats

Descriptive Statistics: Parents

	Control (SD)	**Treatment (SD)**	**Difference (p-value)**
Demographics			
Male	0.47 (0.499)	0.312 (0.463)	-0.158 (0.000***)
Female	0.53 (0.499)	0.688 (0.463)	0.158 (0.000***)
White	0.727 (0.446)	0.657 (0.475)	-0.07 (0.001***)
Black	0.0114 (0.103)	0.0166 (0.127)	0.0052 (0.424)
Asian (	0.0341 (0.181)	0.0322 (0.176)	-0.0019 (0.910)
Marital Status			
Single	0.214 (0.411)	0.382 (0.486)	0.168 (0.000***)
Married	0.781 (0.414)	0.602 (0.49)	-0.179 (0.000***)
Divorced	0.0373 (0.189)	0.139 (0.346)	0.1017 (0.000***)
Widow	0.00496 (0.071)	0.0137 (0.116)	0.00874 (0.072*)
Employment			
Average Job Hours	33.4 (10.2)	27.4 (12.3)	-6.0 (0.000***)
Paid Employment	0.766 (0.424)	0.37 (0.483)	-0.396 (0.000***)
Unemployed	0.0548 (0.229)	0.328 (0.47)	0.2732 (0.000***)
Income			
Median Total Monthly Income	1577.0 (500.0)	1297.0 (600.0)	-280.0 (0.000***)
Median Monthly Labour Income	1250.0 (800.0)	0.0 (0.0)	-1250.0 (0.000***)
Household			
Household Hours	10.6 (4.2)	16.1 (5.5)	5.5 (0.000***)
Dinner with Family	3.33 (1.2)	3.44 (1.3)	0.11 (0.575)
GHQ			
GHQ Score (Total)	10.9 (3.4)	13.3 (4.1)	2.4 (0.019**)

Significance levels: * p¡0.1, ** p¡0.05, *** p¡0.01

Notes: This table summarises descriptive statistics for individual-level variables in the adult file of the UKHLS. The statistics are split into two groups: individuals exposed to cuts and the whole sample. All variables are expressed as proportions or monetary values, and income is reported in GBP. GHQ Scores include the sum of all GHQ categories, where a higher score indicates higher levels of dissatisfaction and distress.
Data Source: UKHLS Understanding Society dataset.

Descriptive Statistics: Youth File

Individual Level Variables	**Exposed to Cuts**	**Whole Sample**
Age	12.51	12.53
Gender		
Male	0.49	0.49
Female	0.50	0.50
Outcomes		
Dinner with Mother (p/w)	3.433	3.448
Dinner with Father (p/w)	3.293	3.274
SDQ Score (Total)	11.72	10.65
N Obs	8265	36 188

Notes: This table summarises descriptive statistics for individual-level variables in the youth portion of the UKHLS. The statistics are split into two groups: individuals exposed to cuts and the whole sample. SDQ Scores include the sum of all SDQ categories, where a higher score indicates higher levels of dissatisfaction and distress. *Data Source:* UKHLS Understanding Society dataset.

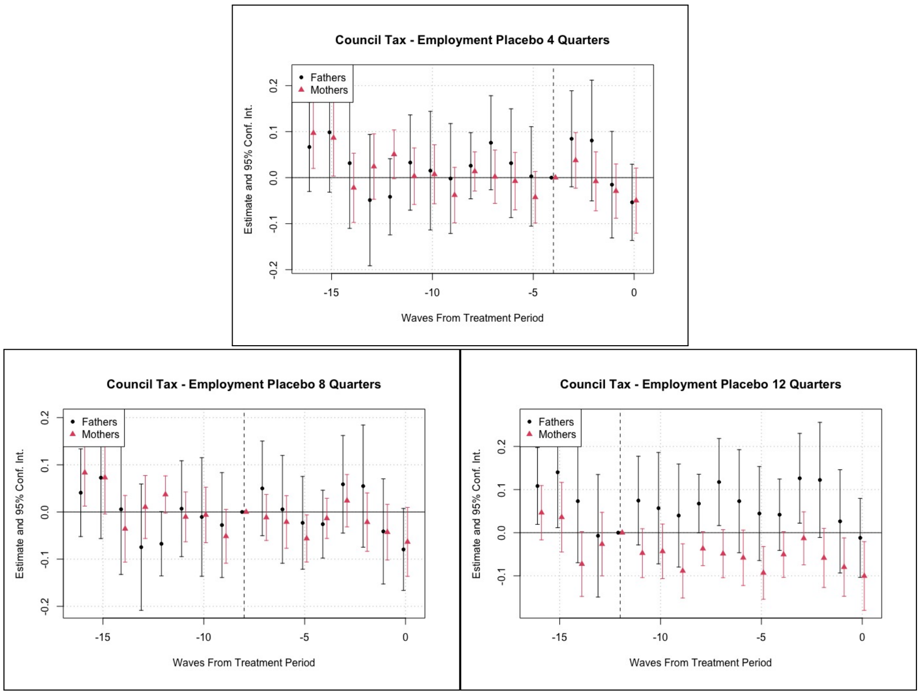
Descriptive Statistics: Treatment Cross-tab

Council Tax Benefit	**Bedroom Tax Benefit (0**	**Bedroom Tax Benefit (1)**
0	3820	3
1	424	832

Notes: This table shows a cross-tabulation of treatment status in 2013 for those exposed to measures in the WRA. The majority of units are either untreated or jointly treated. Very few observations fall into the partially treated groups. *Data Source:* UKHLS Understanding Society dataset.

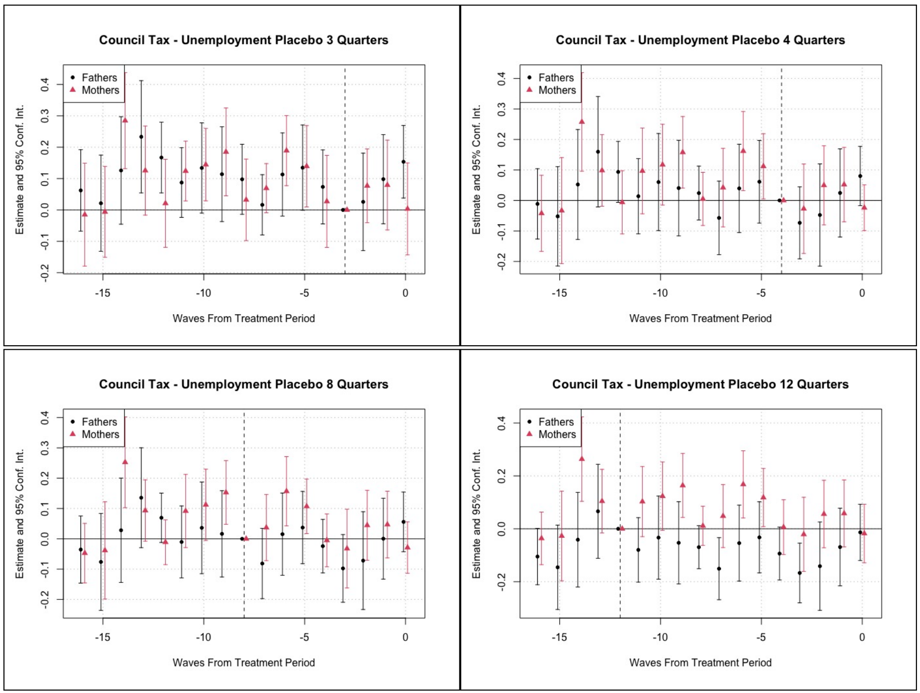
A.2 Placebo Tests

Placebo Test Paid Employment



Note: This table presents results from placebo treatment tests on the employment outcome, where the treatment assignment is shifted to 4, 8, and 12 quarters before the actual treatment period. These tests assess the validity of the parallel trends assumption by examining whether significant effects are observed in pre-treatment periods.

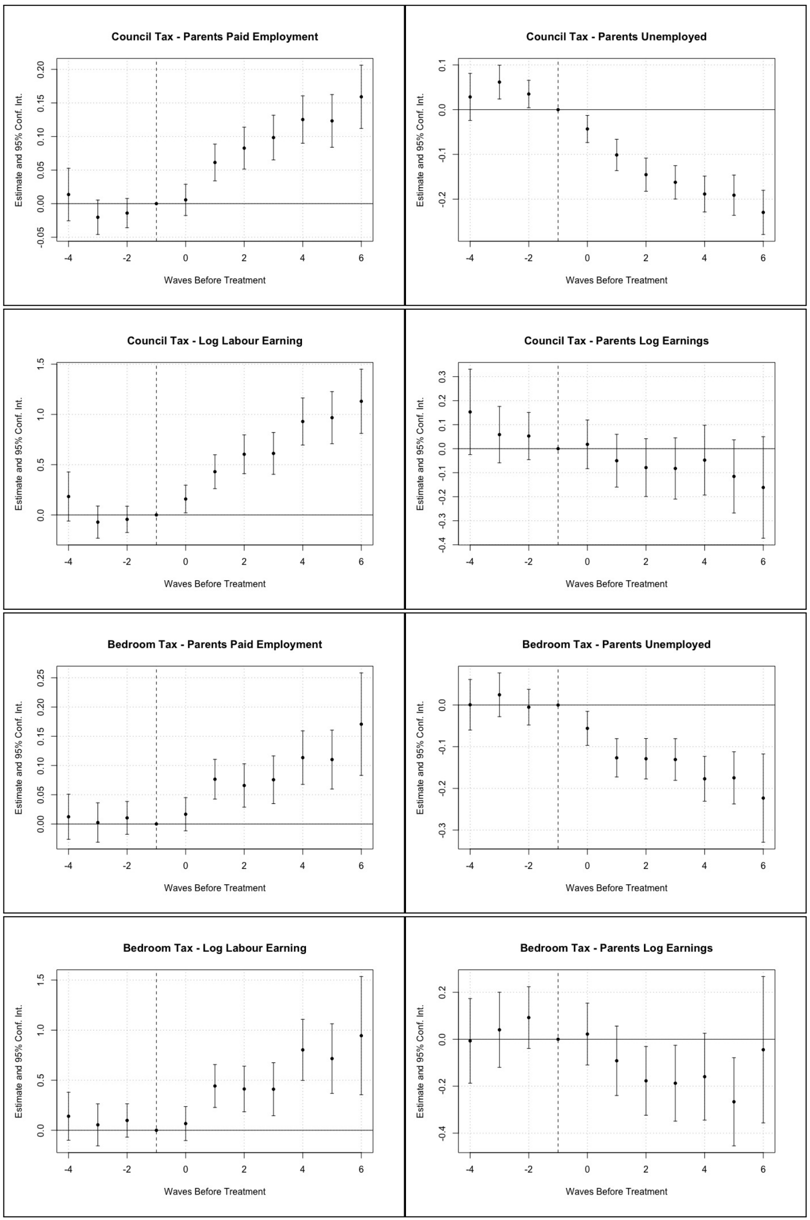
Placebo Test Unemployment



Note: This table presents results from placebo treatment tests on the unemployment outcome, where the treatment assignment is shifted to 2, 4, 8, and 12 quarters before the actual treatment period. These tests assess the validity of the parallel trends assumption by examining whether significant effects are observed in pre-treatment periods.

A.3 Outcome Heterogeneity

A.4 Full Sample of Mothers and Fathers



Note: Figure A.4 sections A and B show the effect of austerity policies on being in paid employment (panel a), being unemployed (panel b), earnings from labour (panel c), and total income (Panel d). Pre-treated trends appear before the vertical dotted line indicating the reference period, $t = -1$. Figure A.4 is broken into two sections with section A displaying the impact of the council tax benefit abolishment and section B displaying the effect of the "spare bedroom tax." Mother and Father results are plotted together, with the coefficients associated with mothers shown in red and the results associated with fathers shown in black.

A.5 Rambachan & Roth (2023) Test

Council Tax: Mothers Paid Employment

Lower Bound (lb)	**Upper Bound (ub)**	**Method**	**Delta Type**	**M-bar**
0.0560	0.223	C-LF	DeltaRM	0.5
0.0533	0.227	C-LF	DeltaRM	1.0
0.0398	0.239	C-LF	DeltaRM	1.5
0.0222	0.254	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. Higher *M-bar* values allow for greater deviations, relaxing the assumption of exact parallel trends. These results provide insight into the sensitivity of the effects of council tax on mothers' paid employment.

Council Tax: Unemployment Effects

Lower Bound (lb)	**Upper Bound (ub)**	**Method**	**Delta Type**	**M-bar**
-0.435	-0.156	C-LF	DeltaRM	0.5
-0.444	-0.143	C-LF	DeltaRM	1.0
-0.465	-0.103	C-LF	DeltaRM	1.5
-0.502	-0.0624	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. Higher *M-bar* values allow for greater deviations, relaxing the assumption of exact parallel trends. These results provide insight into the sensitivity of the effects of council tax on unemployment.

Council Tax: Labor Earnings Effects

Lower Bound (lb)	**Upper Bound (ub)**	**Method**	**Delta Type**	**M-bar**
0.182	1.48	C-LF	DeltaRM	0.5
0.191	1.45	C-LF	DeltaRM	1.0
0.0399	1.60	C-LF	DeltaRM	1.5
-0.129	1.78	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. Higher *M-bar* values allow for greater deviations, relaxing the assumption of exact parallel trends. These results provide insight into the sensitivity of the effects of council tax on labor earnings.

Council Tax: Total Income Effects

Lower Bound (lb)	**Upper Bound (ub)**	**Method**	**Delta Type**	**M-bar**
-0.667	0.154	C-LF	DeltaRM	0.5
-0.656	0.160	C-LF	DeltaRM	1.0
-0.749	0.224	C-LF	DeltaRM	1.5
-0.860	0.329	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. Higher *M-bar* values allow for greater deviations, relaxing the assumption of exact parallel trends. These results suggest that the estimated effects on total income are highly sensitive to even modest departures from parallel trends, as the confidence intervals often include zero, indicating a lack of robustness in the estimated treatment effect.

Council Tax: Mothers Being Married

Lower Bound (lb)	**Upper Bound (ub)**	**Method**	**Delta Type**	**M-bar**
0.182	1.48	C-LF	DeltaRM	0.5
0.191	1.45	C-LF	DeltaRM	1.0
0.0399	1.60	C-LF	DeltaRM	1.5
-0.129	1.78	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. Higher *M-bar* values allow for greater deviations, relaxing the assumption of exact parallel trends. These results suggest that the estimated effects on mothers being married are generally robust to modest parallel trend violations, though sensitivity increases with larger deviations, as seen with *M-bar* values approaching 2.

Council Tax: Father Paid Employment

Lower Bound (lb)	**Upper Bound (ub)**	**Method**	**Delta Type**	**M-bar**
-0.0382	0.276	C-LF	DeltaRM	0.5
-0.0295	0.276	C-LF	DeltaRM	1.0
-0.0666	0.309	C-LF	DeltaRM	1.5
-0.106	0.348	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. These results suggest that the estimated effects on father paid employment are moderately sensitive to parallel trend violations, as confidence intervals widen with increasing *M-bar* values, and some intervals include zero.

Council Tax: Father Unemployment

Lower Bound (lb)	Upper Bound (ub)	Method	Delta Type	M-bar
-0.345	-0.0901	C-LF	DeltaRM	0.5
-0.348	-0.0715	C-LF	DeltaRM	1.0
-0.367	-0.0362	C-LF	DeltaRM	1.5
-0.399	0.000929	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. These results suggest that the estimated effects on father unemployment are relatively robust to small parallel trend violations, but sensitivity increases with larger *M-bar* values, as confidence intervals approach and eventually include zero.

Council Tax: Father Labor Earnings

Lower Bound (lb)	Upper Bound (ub)	Method	Delta Type	M-bar
0.0822	2.05	C-LF	DeltaRM	0.5
0.142	2.05	C-LF	DeltaRM	1.0
-0.00747	2.17	C-LF	DeltaRM	1.5
-0.202	2.37	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. These results suggest that the estimated effects on father labor earnings are robust to smaller deviations, but sensitivity increases as *M-bar* grows, with confidence intervals including zero for *M-bar* values of 1.5 and above.

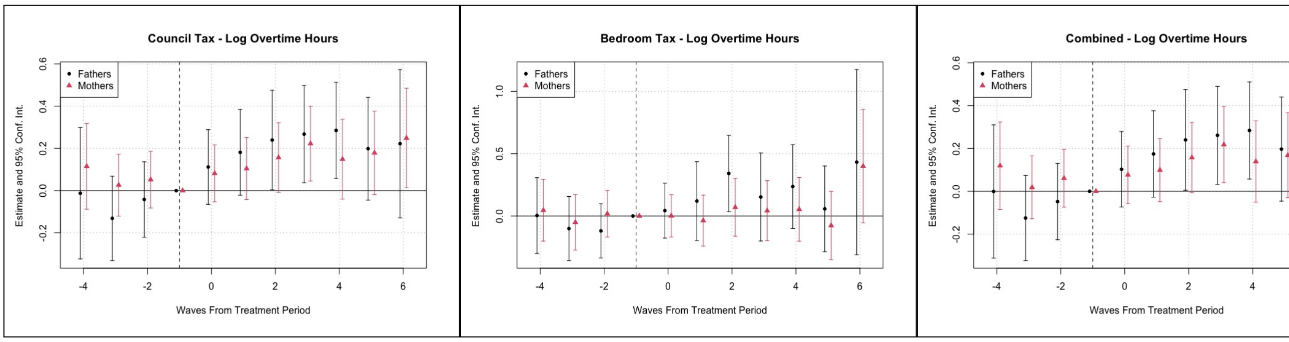
Council Tax: Father Total Income

Lower Bound (lb)	Upper Bound (ub)	Method	Delta Type	M-bar
-0.372	0.835	C-LF	DeltaRM	0.5
-0.478	0.922	C-LF	DeltaRM	1.0
-0.662	1.12	C-LF	DeltaRM	1.5
-0.874	1.32	C-LF	DeltaRM	2.0

Note: This table presents results from the Rambachan and Roth sensitivity analysis framework, assessing the robustness of estimated treatment effects to potential violations of the parallel trends assumption. The lower bound (lb) and upper bound (ub) indicate the confidence interval for treatment effects under the imposed restrictions. The parameter *M-bar* specifies the maximum allowable violation of parallel trends, relative to the largest pre-treatment deviation. These results suggest that the estimated effects on father total income are sensitive to parallel trend violations, as the confidence intervals widen significantly with increasing *M-bar* values and include both negative and positive values.

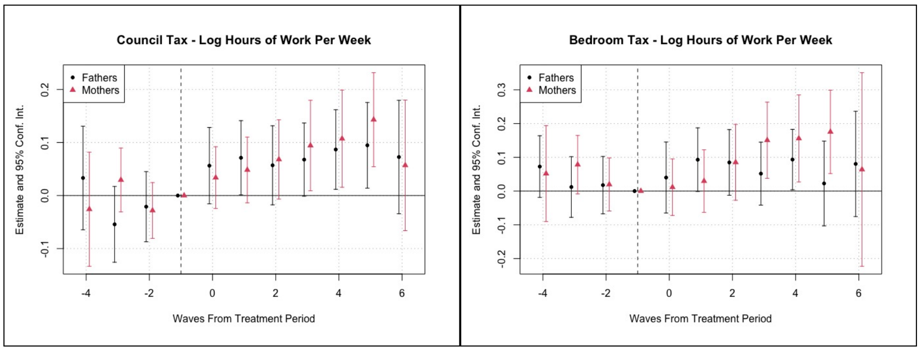
A.6 Supplementary Results

Log Overtime Hours



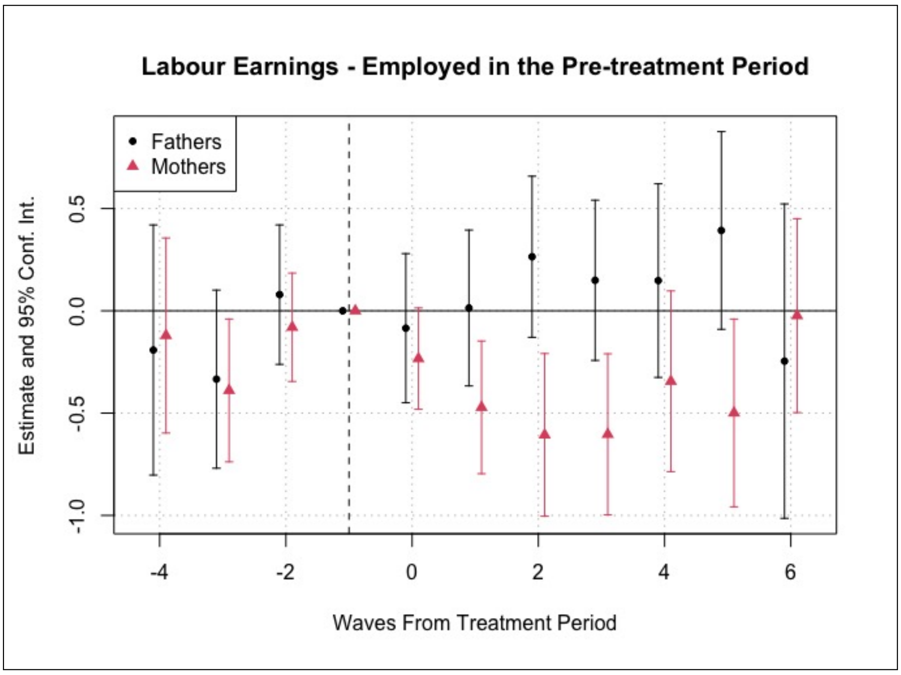
Note: This table presents supplementary results for "overtime hours" using data from the UK Household Longitudinal Study (UKHLS). These results provide additional insights into the relationship between the treatment and changes in overtime work.

Father and Mother Labour Hours



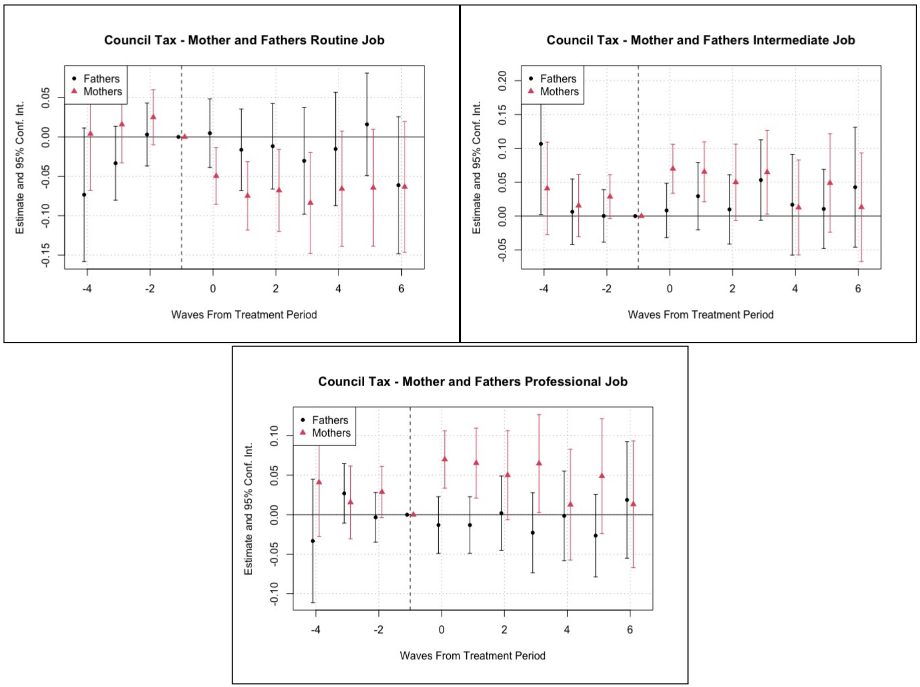
Note: This table presents supplementary results for "Labour Hours" using data from the UK Household Longitudinal Study (UKHLS). These results provide additional insights into the relationship between the treatment and changes in overtime work.

Mothers and Fathers- Labour Earnings - Subset of Treatment Group Who Were Employed



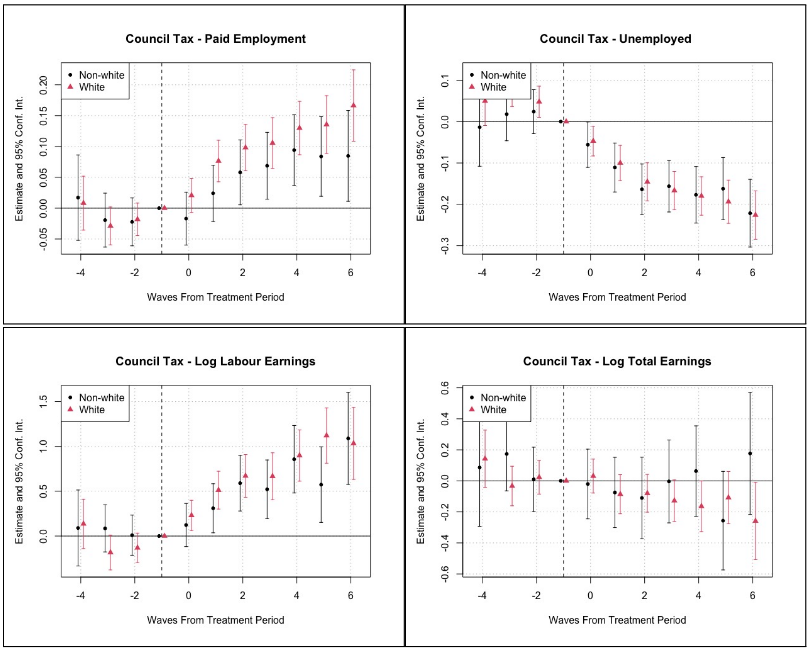
Note: Figure A.5 shows the impact of austerity on labour earnings for those who are both exposed to austerity and were employed in the pre-treatment period. I can see that the high coefficients are cancelled out when I exclude the unemployed.

Job Categories



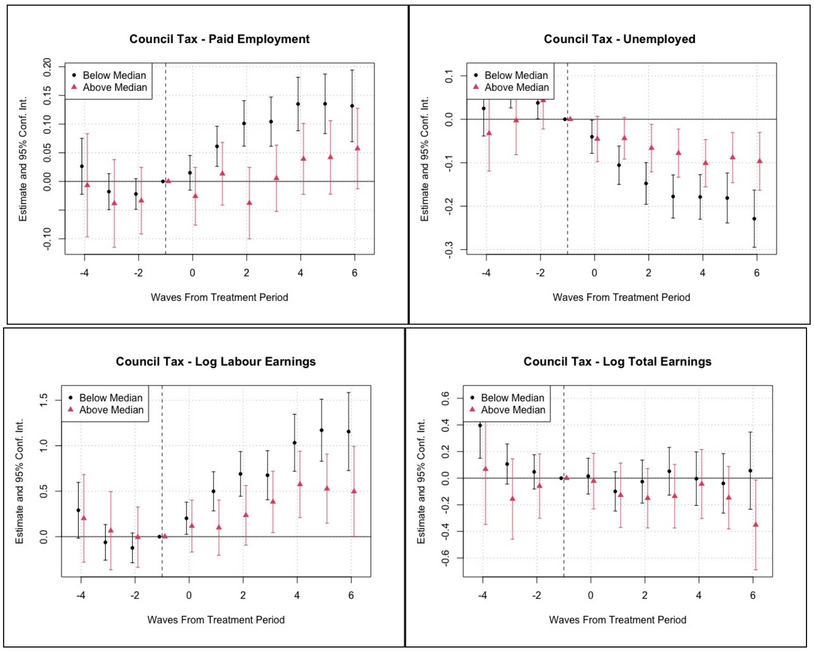
Note: Figure A.6 displays the estimated coefficients for the impact of the Welfare Reform Act (WRA) on employment trends across different job categories. Job categories are grouped based on their routine intensity, skill level, and sectoral characteristics to assess whether the reform had differential effects across occupational types. Positive coefficients represent an increase in employment within the respective category, while negative coefficients suggest a decline. Data comes from the UKHLS

Non-white and White Subsets



Note: This table presents results separately for the non-white and white subsets of the sample, using data from the UK Household Longitudinal Study (UKHLS). The results examine potential heterogeneity in the treatment effects across racial groups, providing insights into whether the observed impacts differ based on racial identity.

Above and Below Median Subsets


Council Tax - Paid Employment

Council Tax - Unemployed

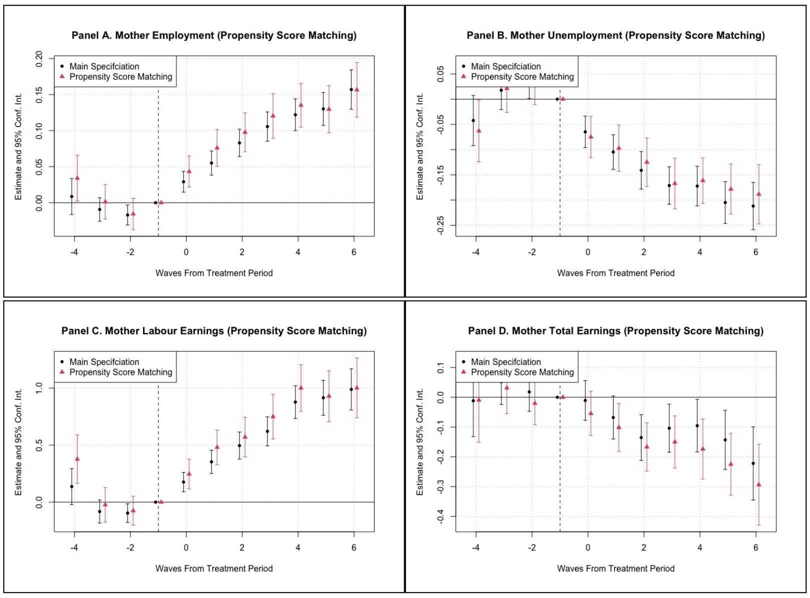
Council Tax - Log Labour Earnings

Council Tax - Log Total Earnings

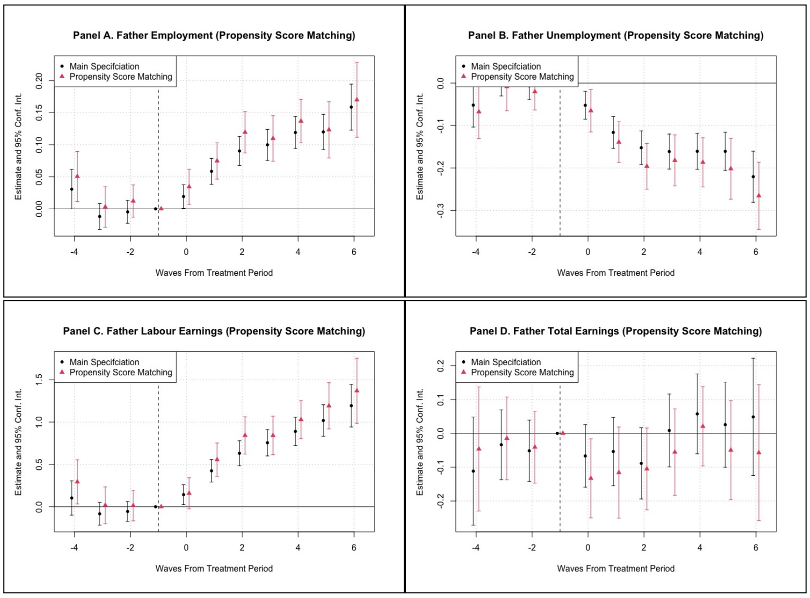
Note: Figure A.8 presents the estimated effects of the Welfare Reform Act (WRA) on employment, unemployment, wages, and total earnings for individuals in above- and below-median income subsets. Confidence intervals are included to indicate the precision of the effects, and standard errors are adjusted for matched sample dependency. Positive coefficients indicate increases in the respective outcome (e.g., employment or wages), while negative coefficients suggest declines. This disaggregated analysis highlights whether the WRA had differential impacts on individuals depending on their position in the income distribution, providing insight into the reform's equity effects.

Council Tax Results: Propensity Score Matching Method

Mother Results



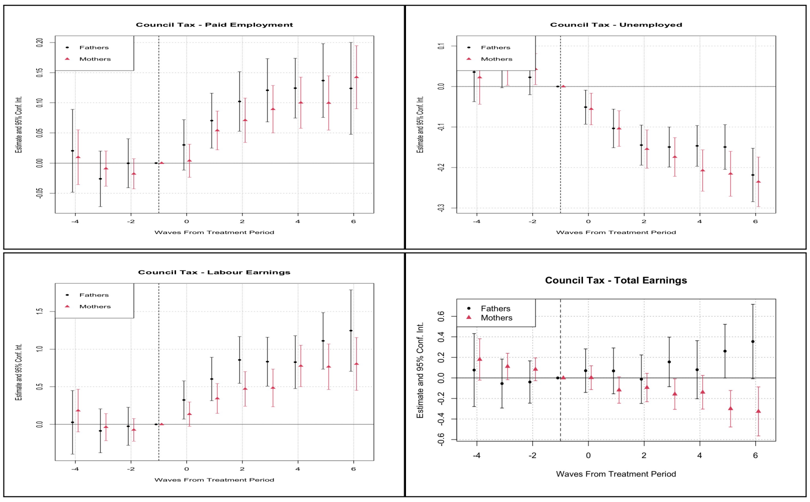
Father Results



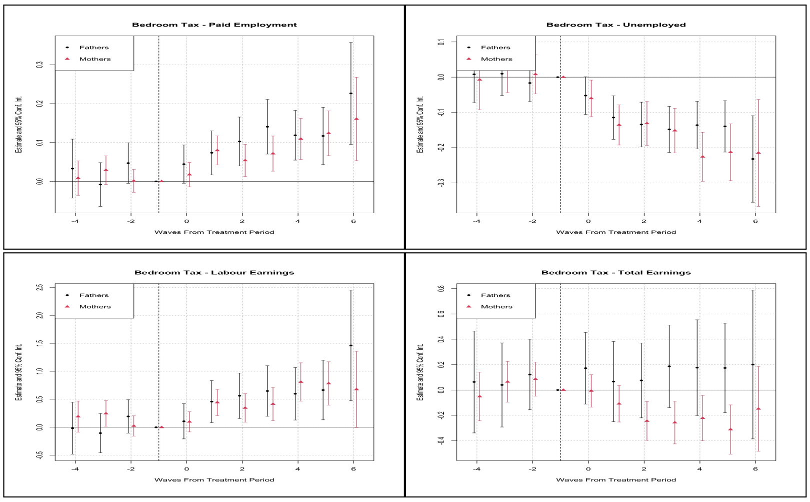
Note: Figure A.9 presents the estimated effects of the Welfare Reform Act (WRA) on employment, unemployment, wages, and total earnings. The analysis is conducted using the propensity score matching (PSM) method, which matches treated and control individuals based on their observable characteristics to estimate the causal impact of the reform. Each panel corresponds to one of the four labor market outcomes, with estimated treatment effects and confidence intervals displayed. Positive coefficients suggest an increase in employment or wages, while negative coefficients indicate a reduction.

Main Findings Excluding Kids Covariate

Council Tax Treatment

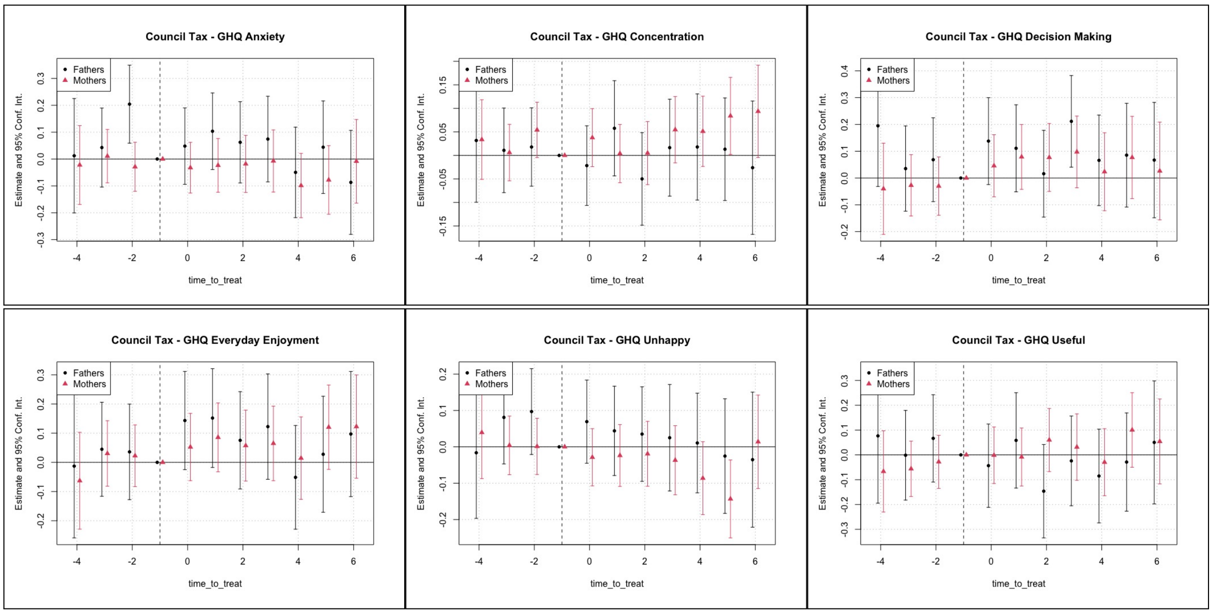


Bedroom Tax Treatment



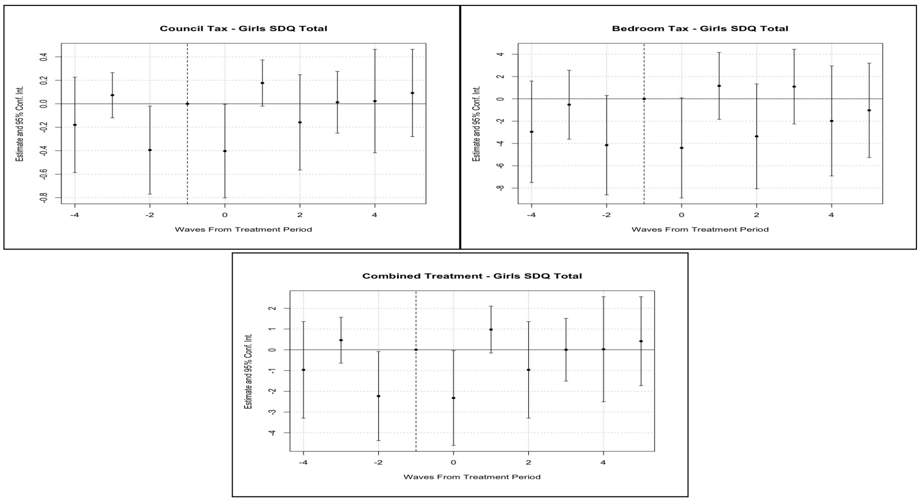
Note: Figure A.10 plots the estimated effect of austerity on employment, income, and labor market outcomes, excluding the covariate denoting the number of children in the household. The results are divided into sections (a) and (b), where section (a) presents the effect of changes to the Council Tax Benefit and section (b) refers to the ”bedroom tax.” Results are sub-divided by gender, with mothers in red and fathers in black.

Parent GHQ



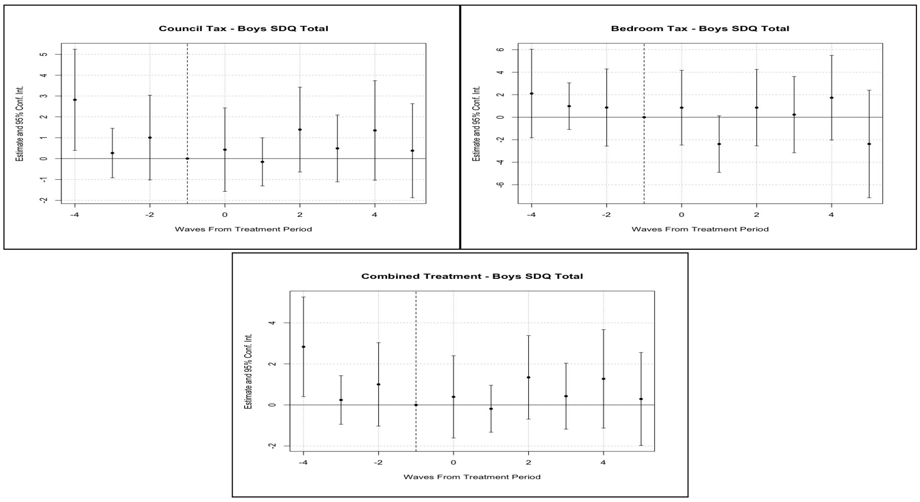
Note: Figure A.11 displays the estimated coefficients for the impact of the Welfare Reform Act (WRA) on all components of the General Health Questionnaire (GHQ) for both mothers and fathers. The analysis uses data from the UK Household Longitudinal Study (UKHLS). Each panel corresponds to a specific GHQ component, with separate coefficients plotted for mothers and fathers for comparison. Positive coefficients reflect worsening mental health outcomes, while negative coefficients indicate improvements.

Girls SDQ Outcomes



Note: Figure A.12 plots the estimated coefficients for the impact of the WRA on the overall SDQ composite for girls. The SDQ is a measure of behavioral and emotional difficulties in children. The analysis uses UKHLS data, and confidence intervals are displayed.

Boys SDQ Outcomes

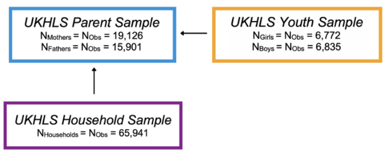


Note: Figure A.13 presents the estimated coefficients for the impact of the WRA on the overall Strengths and Difficulties Questionnaire (SDQ) composite score for boys.

A.7 Data Outline

I focus on the main units of analysis being the parents, the children, both of which are linked with household information. In the UKHLS each dimension of the household is contained in unique files, separated by waves. In order to research several dimensions of the household at once, (i.e., the parents and their household data) it is necessary to first match the household and parent information by wave, and secondly join the data waves together to create longitudinal data. Finally, a distinct youth dataset matches parental information from the adult file to the youth dataset. this allows us to include individual and household specific variation, in models which explore youth specific dependent variables. A brief outline of the data structure can be seen in the table below.

UKHLS Waves 1 to 9 (2009 – 2019): Families with Children


UKHLS Parent Sample
$N_{Mothers} = N_{Obs} = 19,126$
$N_{Fathers} = N_{Obs} = 15,901$
UKHLS Youth Sample
$N_{Girls} = N_{Obs} = 6,772$
$N_{Boys} = N_{Obs} = 6,835$
UKHLS Household Sample
$N_{Households} = N_{Obs} = 65,941$

In event study models unbiased estimates of post-treatment period effects relies intrinsically on the parallel, or common, trend assumption, which presumes that the counterfactual trend of the treated group, in absence of the event, would remain the same. By interpreting the pre-trend estimated coefficients, event study models help us assess the validity of this assumption in this unique specification (Clarke and Tapia-Schythe, 2021; Schmidheiny and Siegloch, 2019).

B Appendix B

Chapter 3 Appendix

B.1 Additional Descriptive Statistics

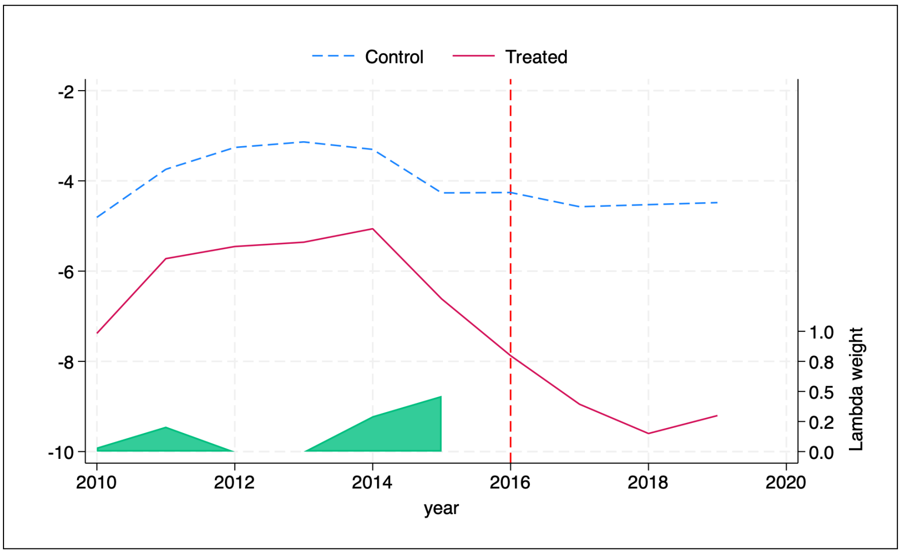
Oklahoma Department of Education Administrative Data

	OKCSD	**Control**	**Total (n obs)**
School Characteristics			
Schools (Count)	86	1,594	1,680
Avg. N Schools per District	na	3.11	na
Districts	1	512	513
School Enrollment*	510	370	6 342 132
Number of Teachers*	28	21	349 893
Teacher YOE($USD)^{x}	10	13	na
Teacher Salary^{x}	$49,965	$43 944	$44 816
School Counselors^{x}	1	1	13
Demographic Characteristics			
Free School Lunch Elig^{x}	80%	62%	65%
White^{x}	19%	63%	57%
Black^{x}	32%	4%	7%
Asian^{x}	2%	1%	1%
Hispanic^{x}	42%	11%	12%
Native^{x}	2%	19%	19%

Note: Table B.1 above presents school and district level staffing and demographic statistics. Note that rows denoted with superscript * correspond with statistics displayed as *school level average values*, rows denoted with superscript ^{x} display values as *district level average values*. YOE denotes years of education.

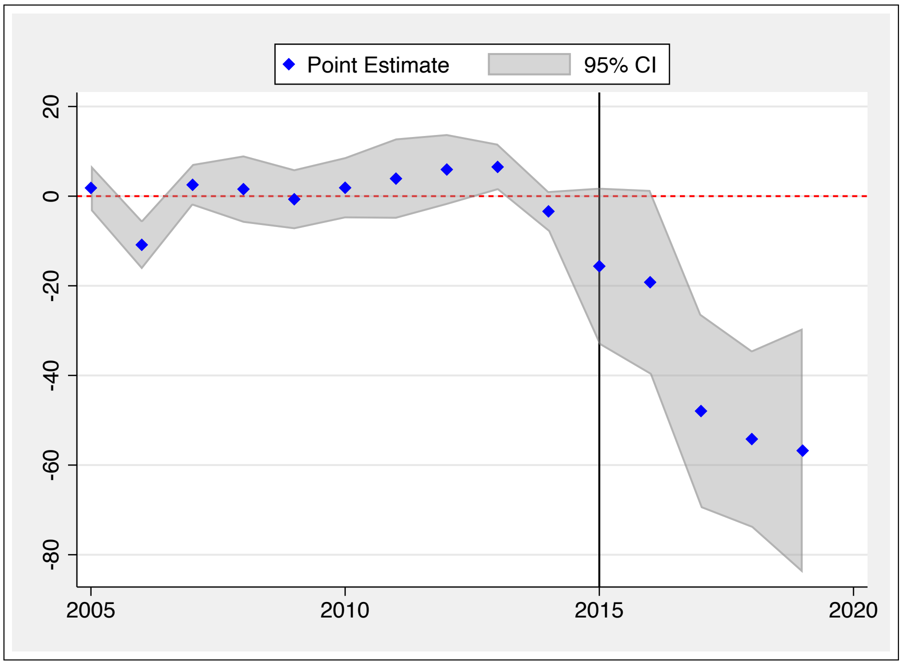
B.2 Additional Results

SDID: Long Suspensions



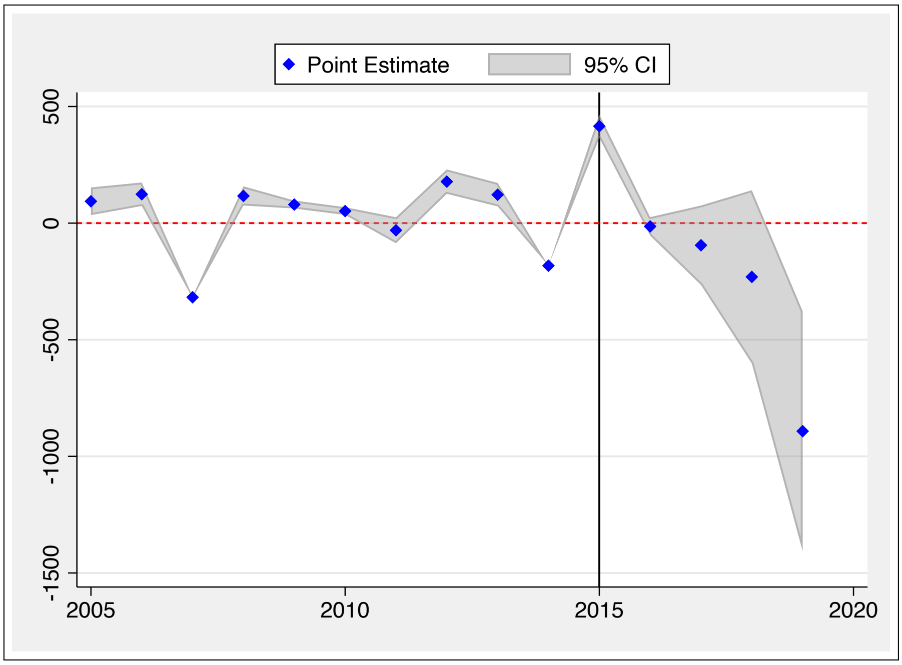
Note: Figure B.1 above presents the Synthetic Difference-in-Differences plot associated with the effect of OKCSDs discipline policies on long suspensions.

SDID: Event Study For Total Suspensions



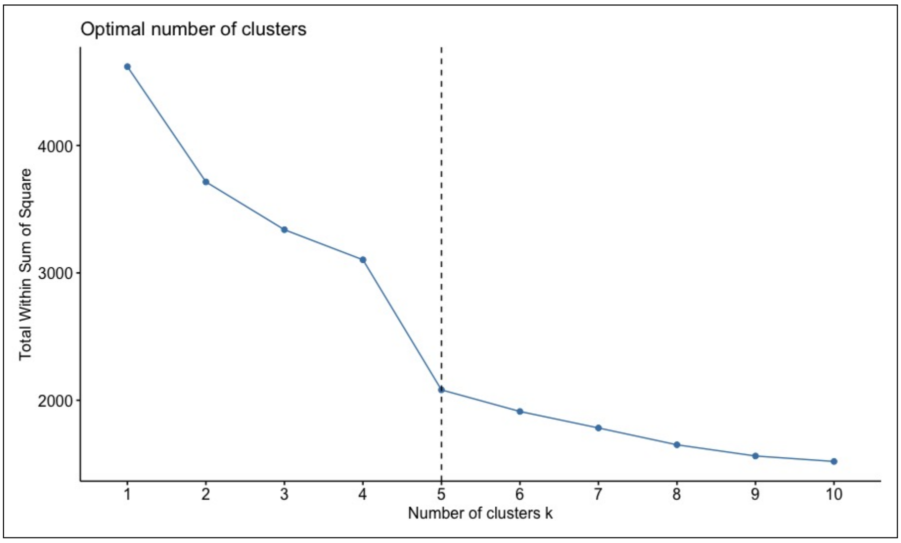
Note: Figure B.2 presents the dynamic event study results for suspensions, using the Oklahoma Department of Education administrative dataset. The plot compares suspension rates between OKCSD schools and their synthetic controls, relative to a pretreatment baseline constructed using the SDID framework.

SDID: Event Study For Adolescent Arrests



Note: Figure B.3 presents the dynamic event study results for adolescent arrests, using data from the Oklahoma Department of Education administrative dataset and the FBI's ASR crime dataset. The plot shows trends in arrests between OKCSD and synthetic controls, relative to a pre-treatment baseline.

K-Means Elbow



Note: Figure B.4 illustrates the elbow plot for the K-means clustering results, where the within-cluster sum of squares ($W(C_k)$) is plotted against the number of clusters. The goal of K-means is to minimize total intra-cluster variation (**tot.withinss**) by grouping points as closely as possible to their respective centroids (μ_k).

B.3 In Time Placebo Estimation

2015 In Time Placebo Estimated Treatment Effects

Outcome	**Estimate**	**Standard Error**	**P-value**
Suspensions			
Short Suspensions	-17.20***	(8.43e-12)	¡ 0.01
Long Suspensions	-1.96**	(0.01)	0.005
Grades 6 to 8			
Grade 6 Math	-5.86	(0.065)	0.065
Grade 6 Reading	-10.15***	(0.00003)	¡ 0.01
Grade 7 Math	2.92	(0.33)	0.333
Grade 7 Reading	1.09	(0.59)	0.587
Grade 8 Math	0.82	(0.84)	0.841
Grade 8 Reading	-2.45	(0.24)	0.245
Grade 8 Science	-3.39	(0.27)	0.273
Arrests			
Youth Arrests	-15.26	(2.65)	¡ 0.01

Notes: Standard errors are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table presents results from a placebo exercise with the post-treatment period beginning in 2015. The data for test scores and suspensions are sourced from the Oklahoma School Report Card data, which provides school-level administrative records. The crime statistics are derived from the FBI's ASR subset of crime data, focusing specifically on youth arrests from the ages of 10 to 12. The above specification drops years 2016 onwards.

2014 In Time Placebo Estimated Treatment Effects

Outcome	**Estimate**	**Standard Error**	**P-value**
Suspensions & Absenteeism			
Short Suspensions	-4.26	(0.14)	0.138
Long Suspensions	0.69	(0.17)	0.173
Total Suspensions	-2.21	(0.52)	0.522
Absences	0.30	(0.16)	0.072
Grades 6 to 8			
Grade 6 Math	0.11	(0.97)	0.971
Grade 6 Reading	-2.66	(0.32)	0.323
Grade 7 Math	2.31	(0.51)	0.511
Grade 7 Reading	2.28	(0.27)	0.268
Grade 8 Math	0.61	(0.91)	0.910
Grade 8 Reading	0.73	(0.80)	0.796
Grade 8 Science	-1.50	(0.67)	0.665
Arrests			
Youth Arrests	2.8381	(2.83)	0.822

Notes: Standard errors are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table presents results from a placebo exercise with the post-treatment period beginning in 2014. The data for test scores and suspensions are sourced from the Oklahoma School Report Card data, which provides school-level administrative records. The crime statistics are derived from the FBI's ASR subset of crime data, focusing specifically on youth arrests from the ages of 10 to 12. The above specification drops years 2015 onwards.

2013 In Time Placebo Estimated Treatment Effects

Outcome	**Estimate**	**Standard Error**	**P-value**
Suspensions			
Short Suspensions	0.83	(6.30)	0.637
Long Suspensions	0.06	(1.15)	0.908
Grades 6 to 8			
Grade 6 Math	-3.28	(8.25)	0.196
Grade 6 Reading	-2.46	(7.73)	0.361
Grade 7 Math	-1.53	(8.73)	0.678
Grade 7 Reading	-0.31	(6.09)	0.916
Grade 8 Math	3.93	(13.34)	0.413
Grade 8 Reading	-0.71	(5.52)	0.771
Grade 8 Science	-4.37	(11.49)	0.229
Youth Arrests	-0.34	(6.46)	0.913

Notes: Standard errors are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table presents results from a placebo exercise with the post-treatment period beginning in 2013. The data for test scores and suspensions are sourced from the Oklahoma School Report Card data, which provides school-level administrative records. The crime statistics are derived from the FBI's ASR subset of crime data, focusing specifically on youth arrests from the ages of 10 to 12. The above specification drops years 2015 onwards.

B.4 Placebo Treatment Estimated Treatment Effects

Placebo Treatement

Outcome	**Estimate**	**Standard Error**	**P-value**
Suspensions			
Short Suspensions	0.37	(9.12)	0.947
Long Suspensions	0.07	(1.55)	0.953
Grades 6 to 8			
Grade 6 Math	-2.26	(17.39)	0.546
Grade 6 Reading	1.54	(15.60)	0.463
Grade 7 Math	-1.21	(19.95)	0.730
Grade 7 Reading	-0.86	(14.57)	0.672
Grade 8 Math	-1.45	(13.13)	0.628
Grade 8 Reading	1.72	(14.51)	0.525
Grade 8 Science	2.77	(18.78)	0.377
Arrests			
Youth Arrests	-0.85	(353.24)	0.889

Notes: Standard errors are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table presents results from a placebo exercise with randomly selected treated units. The data for test scores and suspensions are sourced from the Oklahoma School Report Card data, which provides school-level administrative records. The crime statistics are derived from the FBI's ASR subset of crime data, focusing specifically on youth arrests from the ages of 10 to 12.

B.5 Results Without Covariates

SDID: School Outcomes Without Covariates

max width=

Outcome	**ATT**	**P-value**	**Lower CI**	**Upper CI**
School Metrics				
Absences	0.51***	0.02	0.09	0.93
Short Suspensions	-42.62***	0.00	-48.24	-36.99
Long Suspensions	-2.72***	0.00	-3.36	-2.08
Grades 6 to 8				
Grade 6 Math	2.77	0.45	-4.34	9.88
Grade 6 Reading	-5.68*	0.07	-11.72	0.36
Grade 7 Math	7.36**	0.02	1.39	13.33
Grade 7 Reading	-2.17	0.43	-7.58	3.24
Grade 8 Math	6.61**	0.03	0.72	12.49
Grade 8 Reading	0.63	0.79	-4.07	5.33
Grade 8 Science	8.05***	0.00	4.71	11.39
Counselors				
Counselors	0.15***	0.00	0.10	0.21

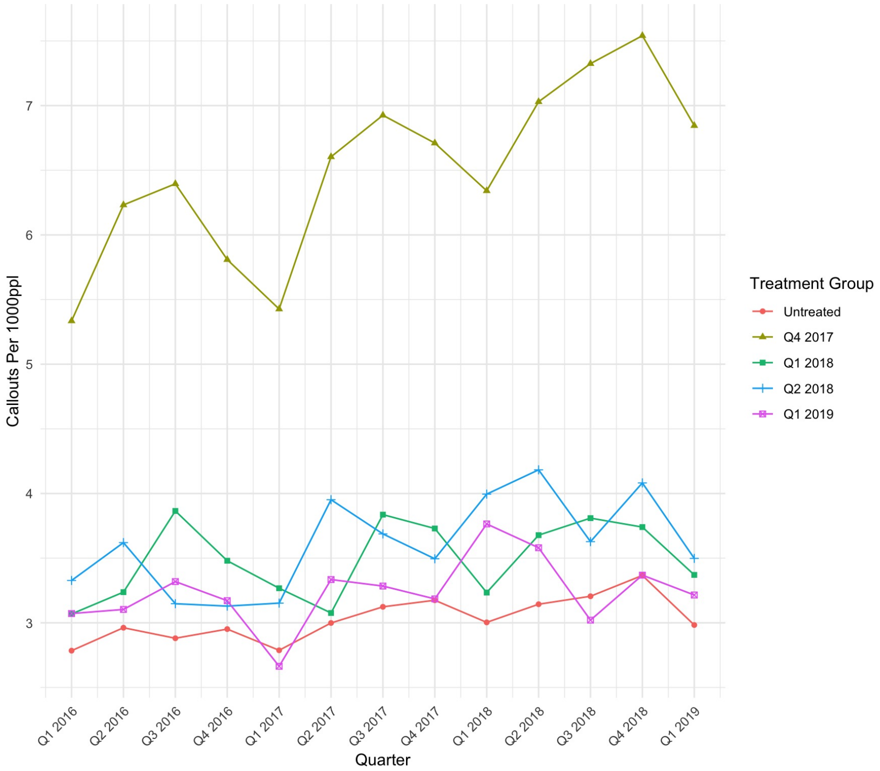
Notes: Confidence intervals (CI) are reported alongside ATT estimates. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. This table presents results from an alternative specification that does not include covariates. Data for test scores and suspensions are sourced from the Oklahoma School Report Card data, providing school-level administrative records.

C Appendix C

Chapter 4 Appendix

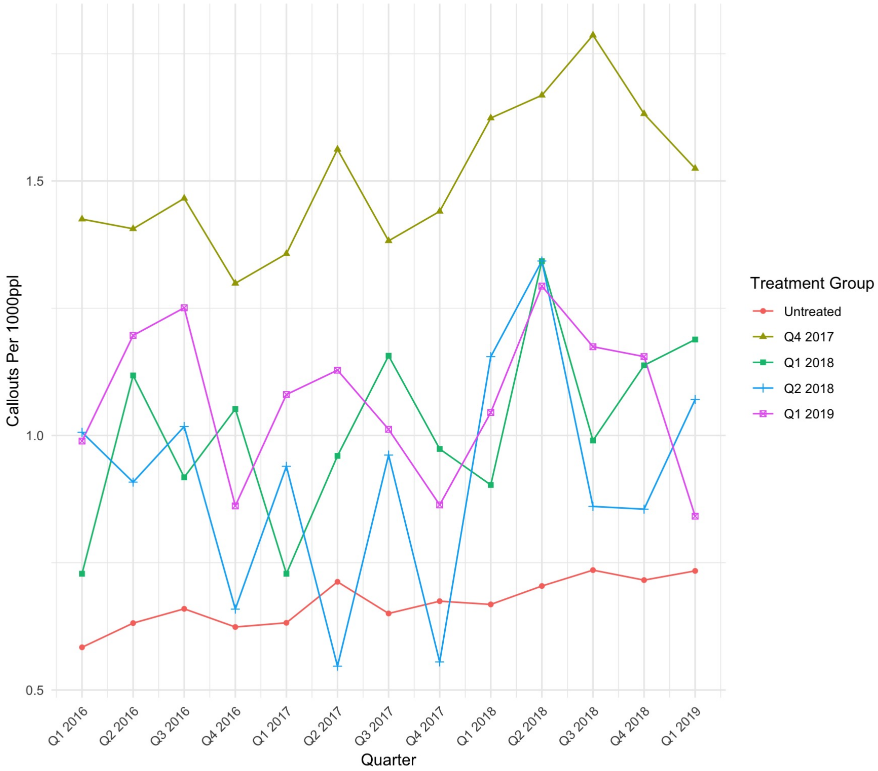
C.1 Descriptive Plots

Figure C.1: Crime Reports per 1000 People by Treatment Group



above plots the sum of counts for assault and break and entering by treatment group. Treatment groups are denoted by different shapes based on the quarter in which they enter the post-treatment period. Data come From the Toronto Police Open Data.

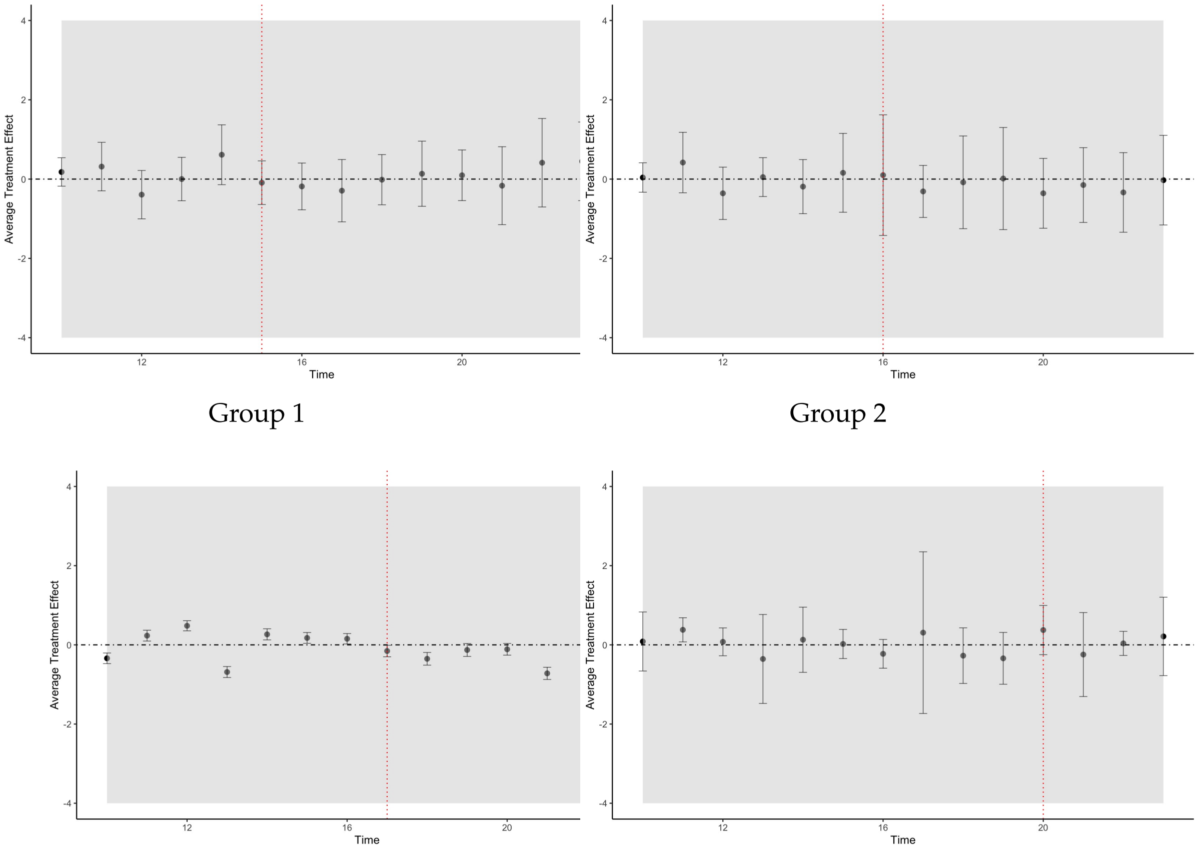
Figure C.2: Mental Health Apprehensions per 1000 People by Treatment Group



above plots the sum of counts for mental health apprehensions by treatment group. Treatment groups are denoted by different shapes based on the quarter in which they enter the post-treatment period. Data come From the Toronto Police Open Data.

C.2 Event Study Plots Crime

Figure C.3: Assault Emergency Callouts Near SCS in Toronto - Calls per 1000 People



Group 1

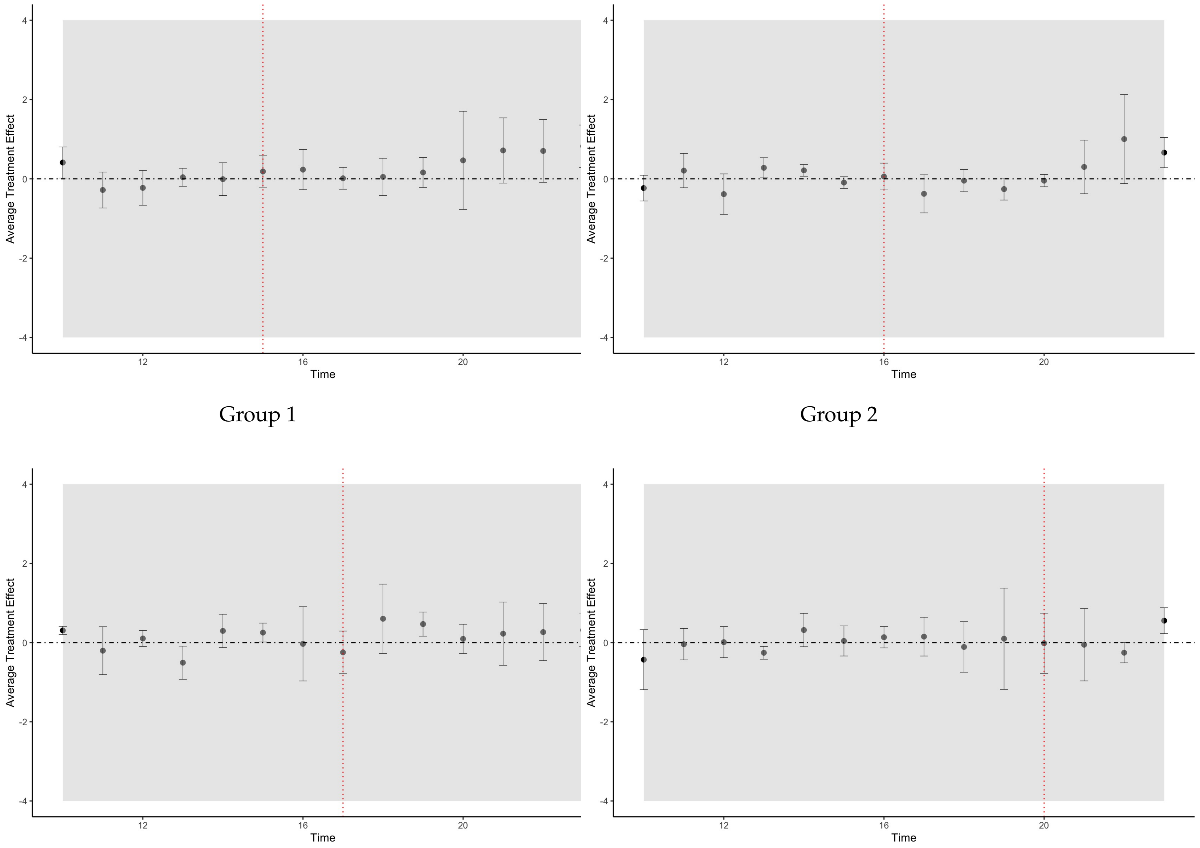
Group 2

Group 3

Group 4

Figure C.3 above illustrates emergency assault call trends near supervised consumption sites (SCS) in Toronto, measured per 1,000 people. Each panel represents a specific cohort group based on staggered treatment timings, capturing how call frequencies evolved before and after SCS implementation. The data, sourced from Toronto Police records, provides insights into potential community-level impacts of SCS on property-related emergencies. Confidence intervals indicate statistical significance, with overlapping trends across groups suggesting consistent patterns.

Figure C.4: Break & Enter Emergency Calls Near SCS in Toronto



Group 1

Group 2

Group 3

Group 4

Figure C.4 above illustrates break-and-enter emergency call trends near supervised consumption sites (SCS) in Toronto, measured per 1,000 people. Each panel represents a specific cohort group based on staggered treatment timings, capturing how call frequencies evolved before and after SCS implementation. The data, sourced from Toronto Police records, provides insights into potential community-level impacts of SCS on property-related emergencies. Confidence intervals indicate statistical significance, with overlapping trends across groups suggesting consistent patterns.

C.3 Event Studies

Event Study Results for Group 1 (Period 16 as Baseline)

Time (Period)	**Estimate (ATT)**	**Std. Error**	**[95% Conf. Interval]**
-6	0.3896	0.1639	[-0.0903, 0.8695]
-5	-0.0665	0.1475	[-0.4985, 0.3656]
-4	-0.2578	0.1674	[-0.7481, 0.2324]
-3	0.0743	0.1813	[-0.4567, 0.6054]
-2	0.0952	0.1485	[-0.3396, 0.5299]
-1	0.3248	0.1589	[-0.1404, 0.7900]
0	-0.4526	0.2021	[-1.0445, 0.1392]
1	-0.3714	0.1189	[-0.7195, -0.0232]
2	-0.4759	0.1775	[-0.9956, 0.0437]
3	-0.0636	0.1175	[-0.4077, 0.2805]
4	-0.3242	0.1387	[-0.7303, 0.0819]
5	0.1040	0.1503	[-0.3360, 0.5440]
6	0.2715	0.0961	[-0.0099, 0.5529]
7	0.0147	0.1104	[-0.3085, 0.3379]
8	-0.0395	0.1280	[-0.4144, 0.3354]

Note: This table presents the event study results for Group 1, with Period 16 serving as the baseline (Time 0). Estimates represent the average treatment effects on the treated (ATT) for each time period. These results provide insight into how supervised consumption sites (SCS) influenced overdose callouts for this group. Data comes from Toronto Police Open Data.

Event Study Results for Group 2 (Period 17 as Baseline)

Time (Period)	**Estimate (ATT)**	**Std. Error**	**[95% Conf. Interval]**
-7	0.0090	0.5464	[-1.5909, 1.6089]
-6	-0.1879	0.6357	[-2.0494, 1.6736]
-5	0.0440	0.1319	[-0.3421, 0.4301]
-4	0.3102	0.1899	[-0.2460, 0.8664]
-3	0.2246	0.1483	[-0.2096, 0.6588]
-2	-0.2051	0.2595	[-0.9649, 0.5547]
-1	-0.3320	0.3895	[-1.4724, 0.8084]
0	0.6715	0.2570	[-0.0810, 1.4240]
1	0.0916	0.3859	[-1.0383, 1.2216]
2	0.4755	0.2613	[-0.2898, 1.2407]
3	0.3248	0.5483	[-1.2806, 1.9302]
4	0.7519	0.6623	[-1.1874, 2.6911]
5	0.4600	0.5894	[-1.2660, 2.1859]
6	0.2022	0.2792	[-0.6155, 1.0199]
7	0.6512	0.3623	[-0.4096, 1.7120]

Note: This table presents the event study results for Group 2, with Period 17 serving as the baseline (Time 0). Estimates represent the average treatment effects on the treated (ATT) for each time period. These results provide insight into how supervised consumption sites (SCS) influenced overdose callouts for this group. Data comes from Toronto Police Open Data.baseline period.

Event Study Results for Group 3 (Period 18 as Baseline)

Time (Period)	Estimate (ATT)	Std. Error	[95% Conf. Interval]
-8	-0.3275	0.0666	[-0.5224, -0.1326]
-7	0.3351	0.0677	[0.1369, 0.5334]
-6	-0.0815	0.0696	[-0.2852, 0.1223]
-5	-0.0168	0.0710	[-0.2246, 0.1911]
-4	-0.1095	0.0553	[-0.2713, 0.0523]
-3	0.0796	0.0589	[-0.0927, 0.2520]
-2	-0.2416	0.0679	[-0.4406, -0.0427]
-1	0.4588	0.0665	[0.2640, 0.6535]
0	-0.0098	0.0628	[-0.1936, 0.1739]
1	-0.4192	0.0637	[-0.6058, -0.2326]
2	-0.1807	0.0619	[-0.3620, 0.0006]
3	-0.1779	0.0619	[-0.3591, 0.0034]
4	-0.2225	0.0618	[-0.4036, -0.0414]
5	0.1548	0.0607	[-0.0230, 0.3327]
6	-0.3759	0.0650	[-0.5663, -0.1854]

Note: This table presents the event study results for Group 3, with Period 18 serving as the baseline (Time 0). Estimates represent the average treatment effects on the treated (ATT) for each time period. These results provide insight into how supervised consumption sites (SCS) influenced overdose callouts for this group. Data comes from Toronto Police Open Data.

Event Study Results for Group 4 (Period 21 as Baseline)

Time (Period)	Estimate (ATT)	Std. Error	[95% Conf. Interval]
-11	0.3100	0.3772	[-0.7944, 1.4144]
-10	0.3747	0.2488	[-0.3539, 1.1033]
-9	-0.0793	0.2013	[-0.6686, 0.5100]
-8	-0.0635	0.1926	[-0.6275, 0.5006]
-7	0.1347	0.3176	[-0.7953, 1.0647]
-6	0.0835	0.1777	[-0.4369, 0.6038]
-5	0.0348	0.4402	[-1.2541, 1.3237]
-4	-0.0345	0.3894	[-1.1748, 1.1058]
-3	-0.4770	0.2819	[-1.3025, 0.3485]
-2	0.5879	0.4402	[-0.7011, 1.8769]
-1	-0.0833	0.3311	[-1.0529, 0.8863]
0	0.3401	0.2011	[-0.2487, 0.9290]
1	0.3247	0.4899	[-1.1099, 1.7593]
2	0.3180	0.2798	[-0.5014, 1.1373]
3	0.0809	0.5855	[-1.6336, 1.7953]

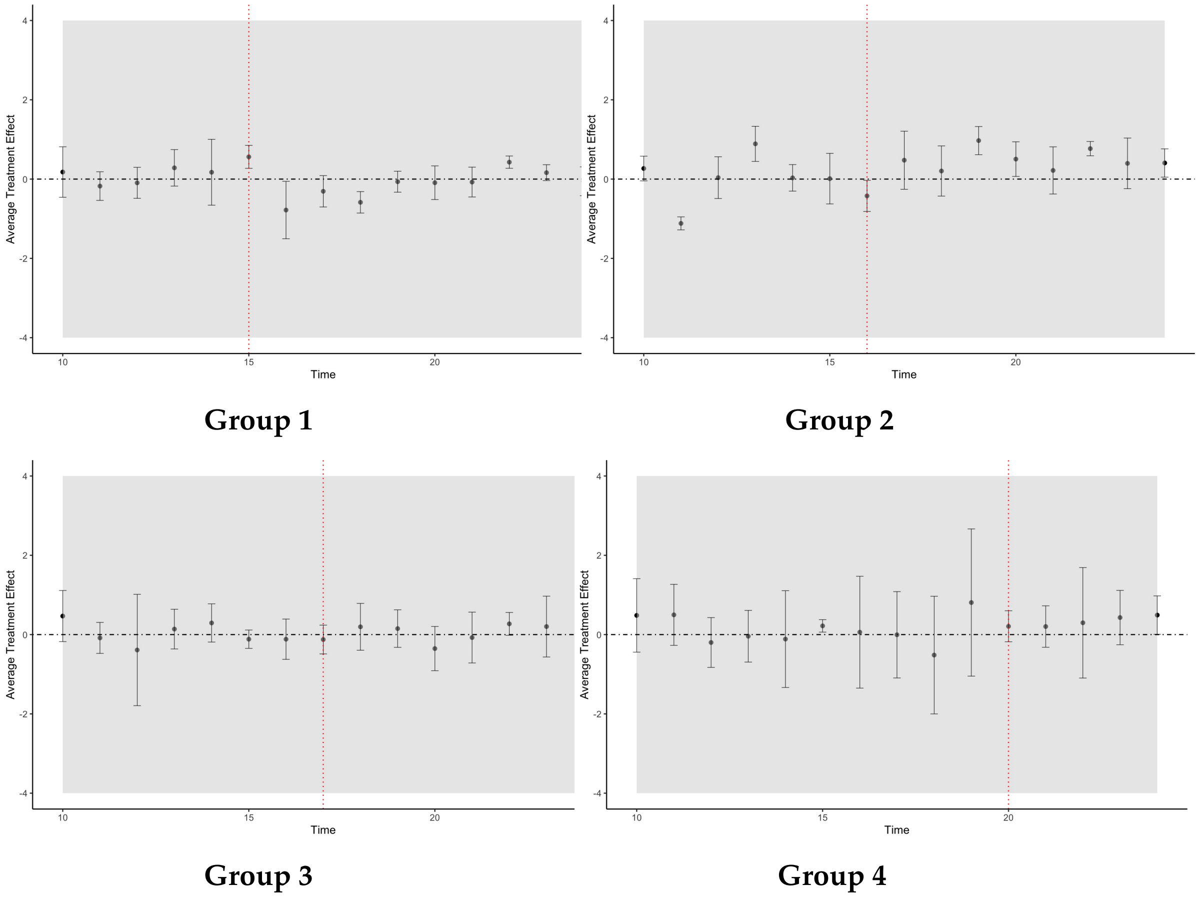
Note: This table presents the event study results for Group 4, with Period 21 serving as the baseline (Time 0). Estimates represent the average treatment effects on the treated (ATT) for each time period. These results provide insight into how supervised consumption sites (SCS) influenced overdose callouts for this group. Data comes from Toronto Police Open Data.

C.4 Robustness Checks

C.5 Decreasing and increasing SCS "Catchment areas"

In the Figures below we provide evidence to show that our results are not sensitive to adjusting the treatment definition to include smaller or larger geographic areas by decreasing and increasing the catchment areas of SCS substantially in each iteration.

Figure C.5: Overdose Callouts near SCS - Calls per 1000 People- Distance 500m



Group 1

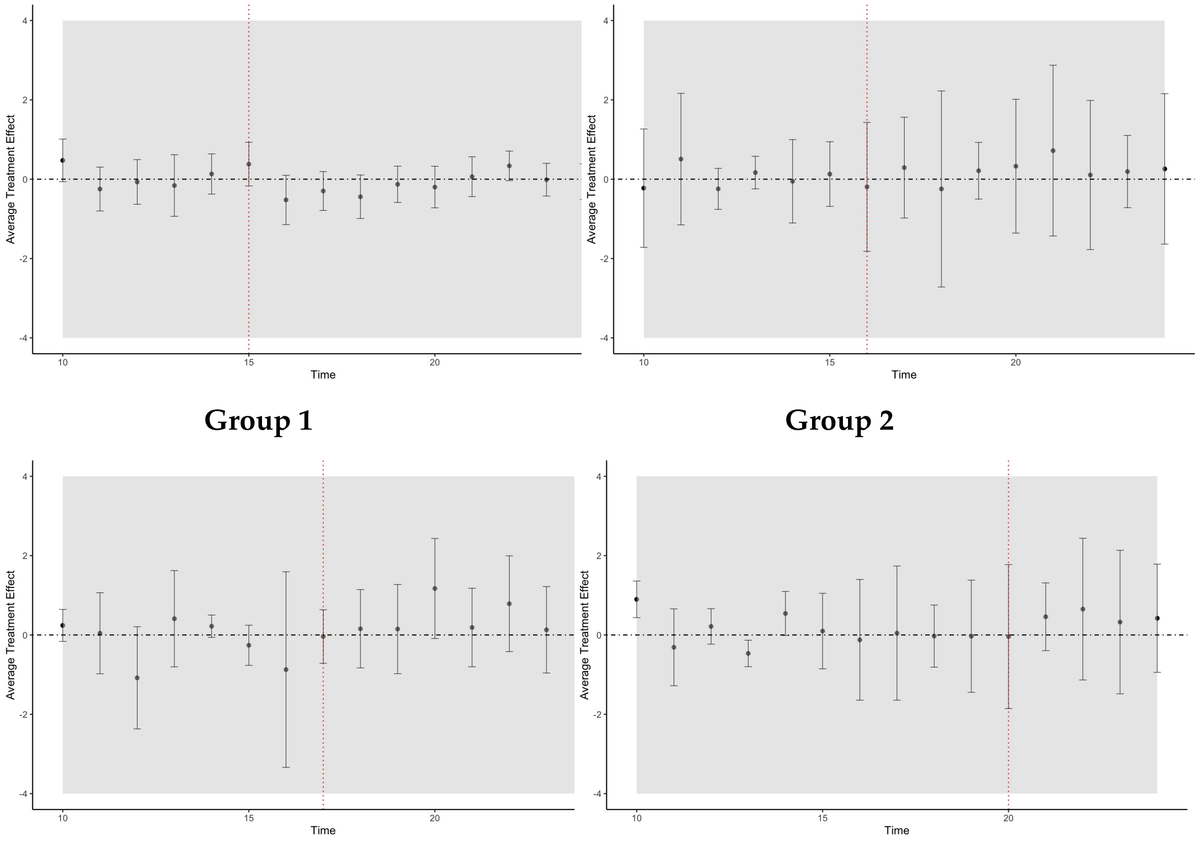
Group 2

Group 3

Group 4

Figure C.5 provides the estimated impact of SCS on overdose callouts when using an alternative treatment definition that defines treatment radius around each SCS as 500 meters. Data comes from Toronto Police Open Data.

Figure C.6: Overdose Callouts near SCS - Calls per 1000 People - Distance 1500m



Group 1

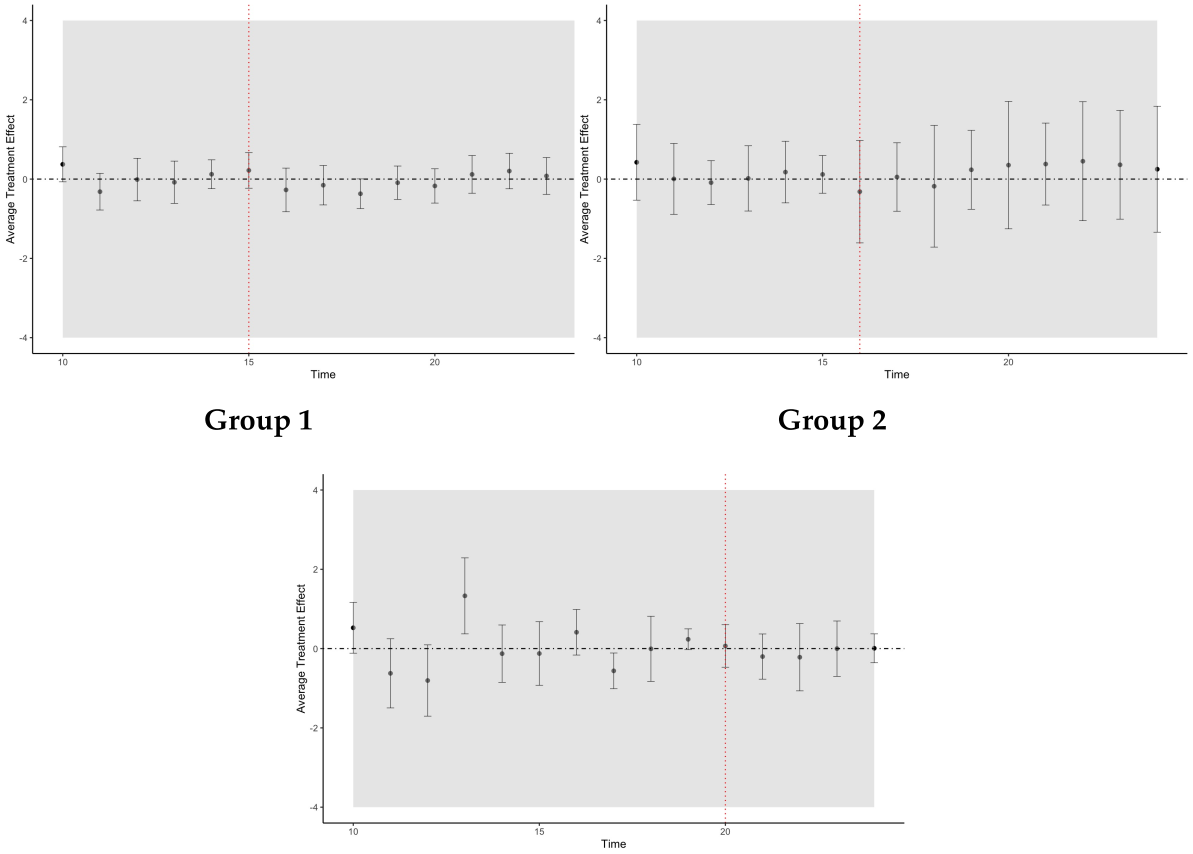
Group 2

Group 3

Group 4

Figure C.6 provides the estimated impact of SCS on overdose callouts when using an alternative treatment definition that defines treatment radius around each SCS as 1500 meters. Data comes from Toronto Police Open Data.

Figure C.7: Overdose Callouts near SCS - Calls per 1000 People - Distance 2000m



Group 1

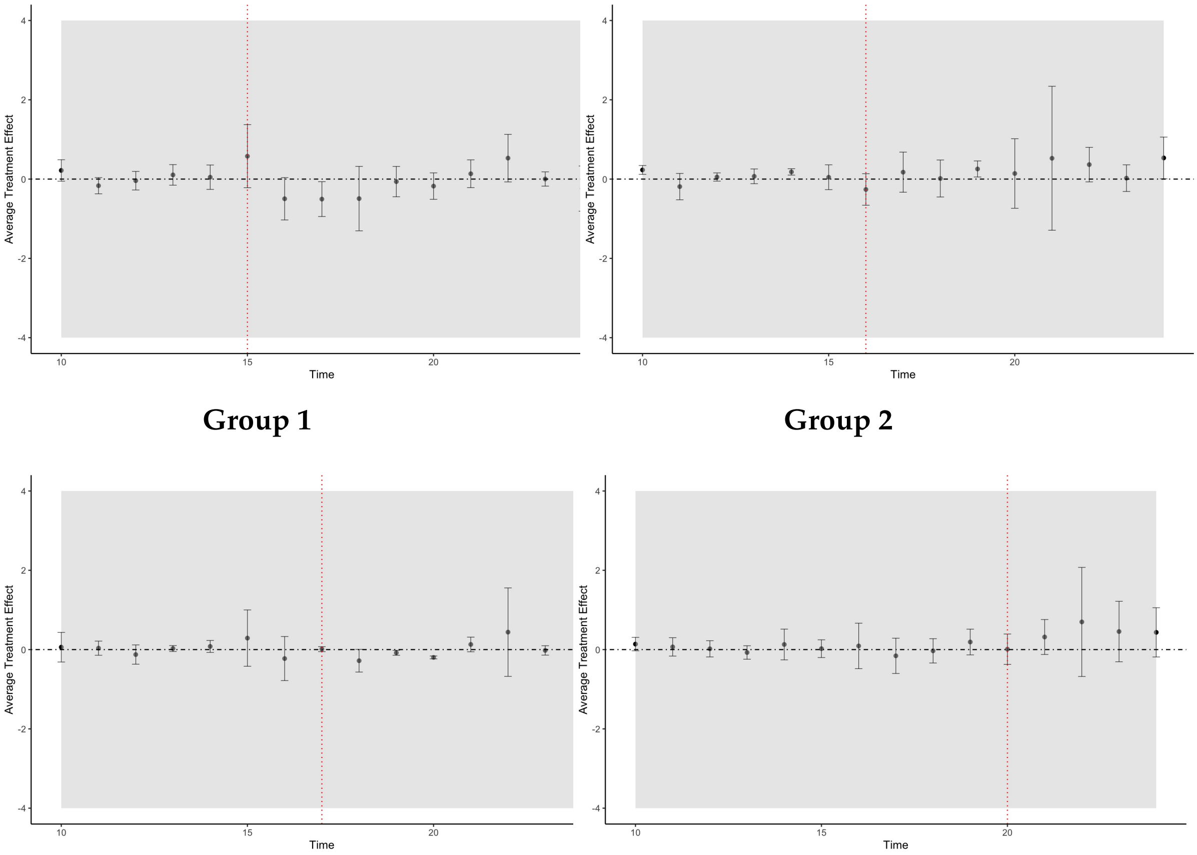
Group 2

Group 4

Figure C.7 provides the estimated impact of SCS on overdose callouts when using an alternative treatment definition that defines treatment radius around each SCS as 2000 meters. Sub-Figure C represents Group 4 and is centered below Sub-Figure A (Group 1) and B (Group 2). Data comes from Toronto Police Open Data.

C.6 Using Neighborhood Boundaries Instead of Neighborhood Centroids

Figure C.8: Overdose Callouts near SCS - Calls per 1000 People



Group 1 **Group 2**

Group 3 **Group 4**

Figure C.8 shows the impact of supervised consumption sites (SCS) on overdose callouts per 1,000 people, using a spatial setup where treatment is defined by **neighborhood boundaries** rather than centroids with the same 1000 meter radius used in our primary specification. Each panel corresponds to a specific group based on when treatment began. Data comes from the Toronto Police Open Data.

Table C.5: Aggregated ATT and Group Effects: Emergency Overdose Calloutss Per 1000

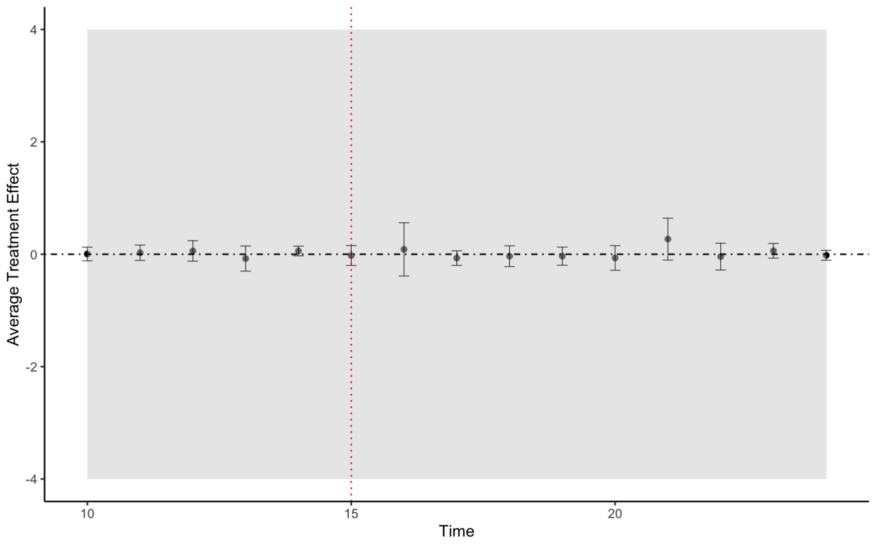
Cohort/Group	Estimate (ATT)	Std. Error	[95% Conf. Interval]
Overall ATT Summary			
—	0.0221	0.115	[-0.2033, 0.2475]
Group Effects			
Group 1	-0.1455	0.0870	[-0.3160, 0.0250]
Group 2	0.2541	0.2405	[-0.2172, 0.7254]
Group 3	-0.0019	0.0534	[-0.1066, 0.1028]
Group 4	0.4765	0.3347	[-0.1794, 1.1324]
Pre-Treatment Mean = 0.95, SD = 0.519, $N = 2375$			

Note: Significance levels are indicated by confidence bands not covering 0.
Estimation Method: Doubly Robust

* Denotes significance at the 10% level, ** at the 5% level, and *** at the 1% level.

Figure C.9aOverdose Callouts near SCS - Calls per 1000 People

Specification: Excludes Units Treated a Second Time After Initial Treatment



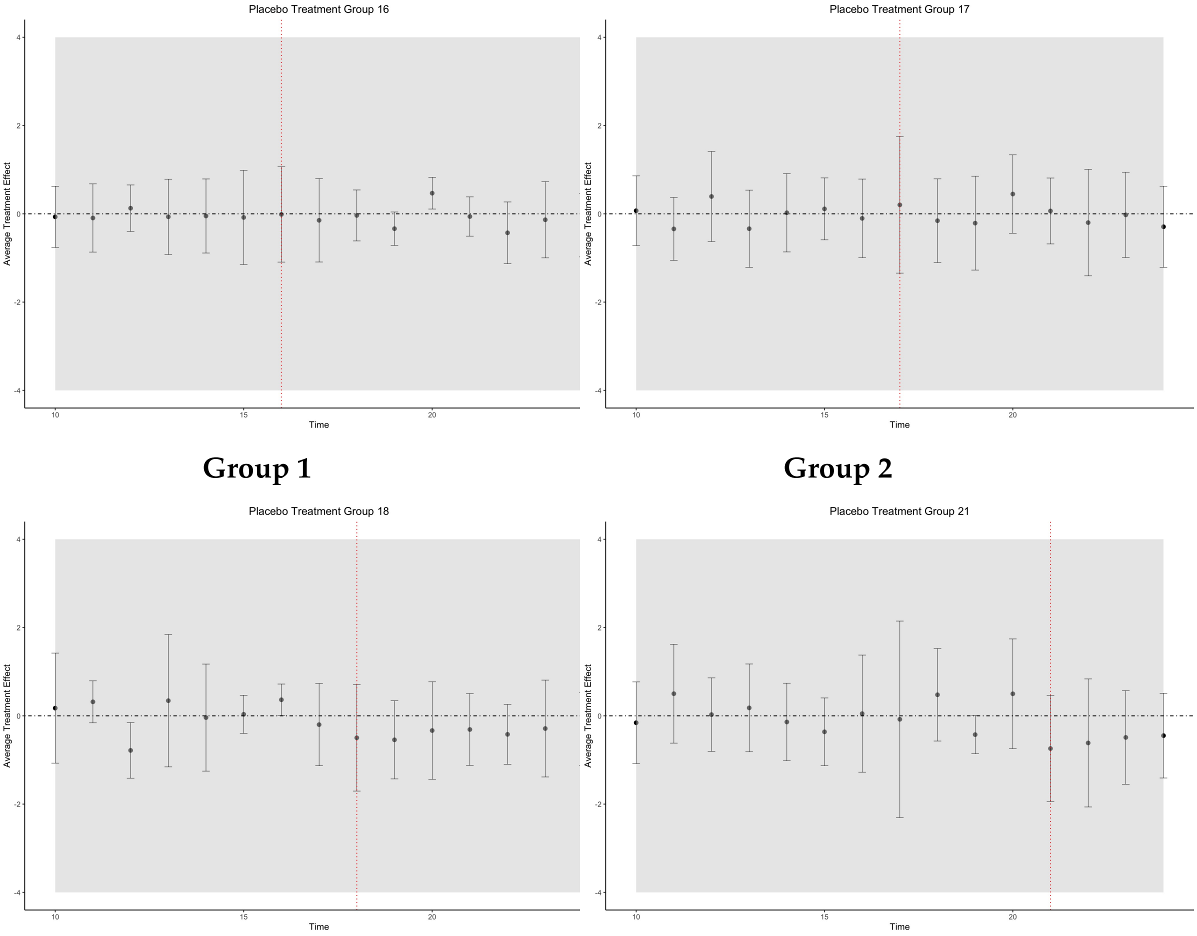
*

Note: This figure presents results from the Callaway and Sant'Anna (CS) estimator using Toronto Police data. The specification excludes all units that were treated a second time after their initial treatment (i.e., "second" treated units). The data is analysed to estimate the impact of supervised consumption sites (SCS) on overdose callouts per 1000 people in the affected areas.

C.7 Placebo Treatment

Here the treatment is randomly assigned to untreated neighborhoods (further than 1,000 meters away from a SCS opening at any time in the pre or post treatment periods.

Figure C.10: Placebo Treatment Overdose Callouts near SCS - Calls per 1000 People



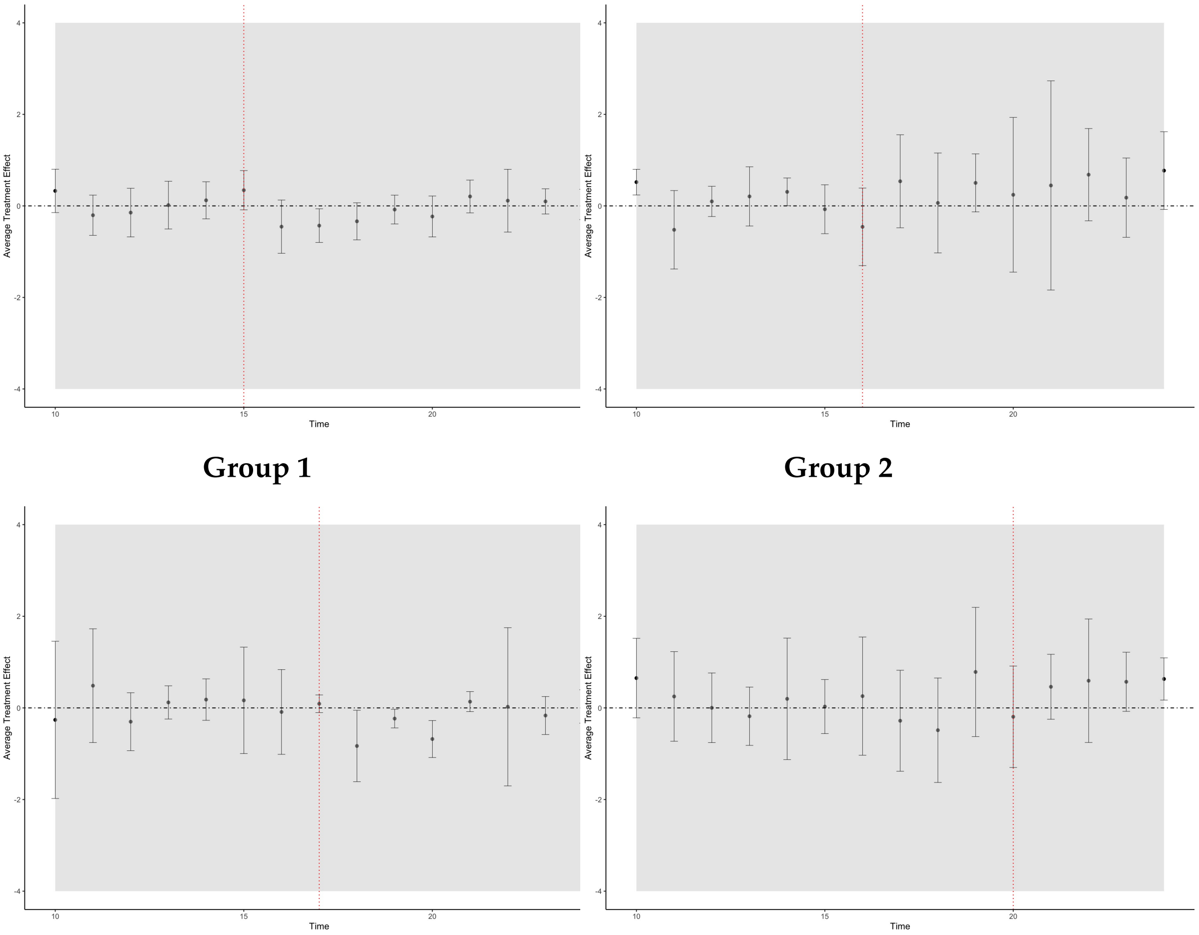
Group 1 **Group 2**

Group 3 **Group 4**

Figure C.10 presents results from a placebo test analyzing overdose callouts near supervised consumption sites (SCS). The data is sourced from Toronto Police records and uses the Callaway and Sant'Anna (CS) estimator. In this specification, placebo treatment periods are assigned to ensure that no actual treatment effects are captured. The purpose of this test is to verify that the observed trends are not driven by spurious correlations or pre-existing differences unrelated to the actual implementation of SCS. Confidence intervals for these placebo treatments provide a benchmark for assessing the robustness of the primary analysis.

C.8 Alternative Outcome Measurements

Figure C.11: Overdose Callouts near SCS - Log Callouts per 1000



Group 1

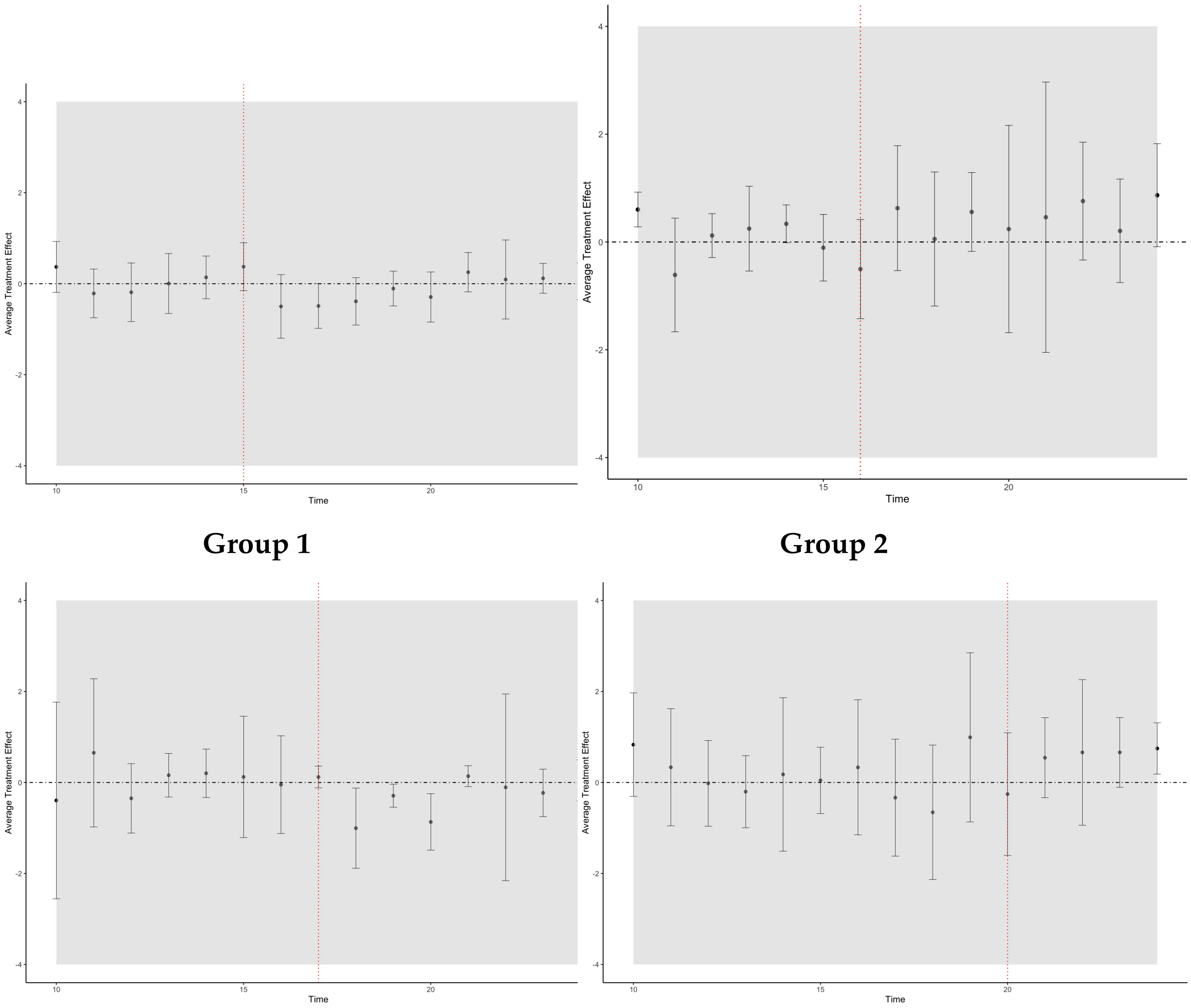
Group 2

Group 3

Group 4

Figure C.11 uses the 'log 1+x' transformation (logarithm of one plus the count) to model overdose callouts near supervised consumption sites (SCS) per 1000 people. This alternative specification is included to assess the robustness of the results to changes in the outcome variable's distribution. The data, sourced from Toronto Police records, is analysed using the Callaway and Sant'Anna (CS) estimator. The 'log 1+x' transformation helps to address skewness in callout data and provides a complementary perspective on the treatment effects.

Figure C.12: Overdose Callouts near SCS - Inverse Hyperbolic Sine Transformed Callouts per 1000



Group 1

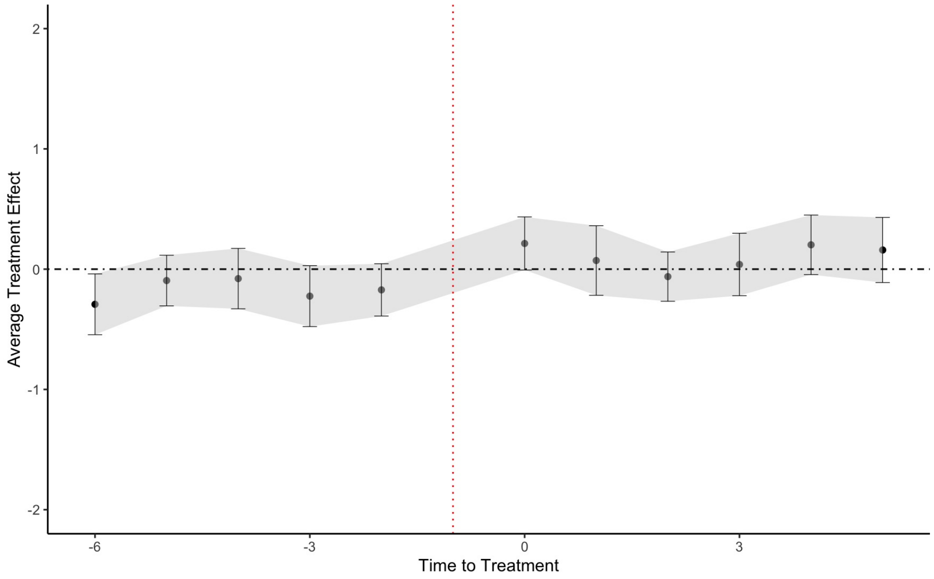
Group 2

Group 3

Group 4

Figure C.12 uses the inverse hyperbolic sine (IHS) transformation to model overdose callouts near supervised consumption sites (SCS) per 1000 people. The IHS transformation is applied as an alternative to traditional logarithmic transformations to handle zero and negative aA, making it particularly suited for this context. The data, sourced from Toronto Police records, is analysed using the Callaway and Sant'Anna (CS) estimator. This approach allows for a robustness check of the results by ensuring the findings are not sensitive to the choice of transformation for the dependent variable.

Figure C.13: Overdose Callouts near SCS - Callouts per 1,000: Sun and Abraham (2020)



Note: This figure presents results from the Sun and Abraham (2020) estimator for overdose callouts near supervised consumption sites (SCS), measured per 1,000 people. The Sun and Abraham method provides an alternative approach to estimating treatment effects in staggered adoption settings, addressing potential issues with heterogeneous treatment effects. The data, sourced from Toronto Police records, is analysed to compare results with the primary Callaway and Sant'Anna (CS) estimator. This figure demonstrates that the estimated effects remain consistent across different estimators, reinforcing the robustness of the findings.

Sun and Abraham Event Study: Overdose Callouts per 1000

Time (Period)	**Estimate (ATT)**	**Std. Error**	**[95% Conf. Interval]**
-6	-0.2922	0.1290	[-0.5451, -0.0393] *
-5	-0.0947	0.1074	[-0.3052, 0.1158]
-4	-0.0787	0.1281	[-0.3296, 0.1722]
-3	-0.2244	0.1292	[-0.4777, 0.0289] .
-2	-0.1723	0.1109	[-0.3897, 0.0451]
0	0.2137	0.1126	[-0.0070, 0.4344] .
1	0.0718	0.1476	[-0.2172, 0.3608]
2	-0.0620	0.1046	[-0.2669, 0.1429]
3	0.0390	0.1325	[-0.2203, 0.2983]
4	0.2024	0.1263	[-0.0451, 0.4490]
5	0.1594	0.1381	[-0.1104, 0.4292]

Note: Significance levels are indicated by confidence bands not covering 0. Group 1 results are based on LS estimation with fixed effects. Time 0 corresponds to the baseline period. Signif. codes: * 0.05, . 0.1

Figure C.14: Overdose Callouts near SCS - Calls per 1000 People

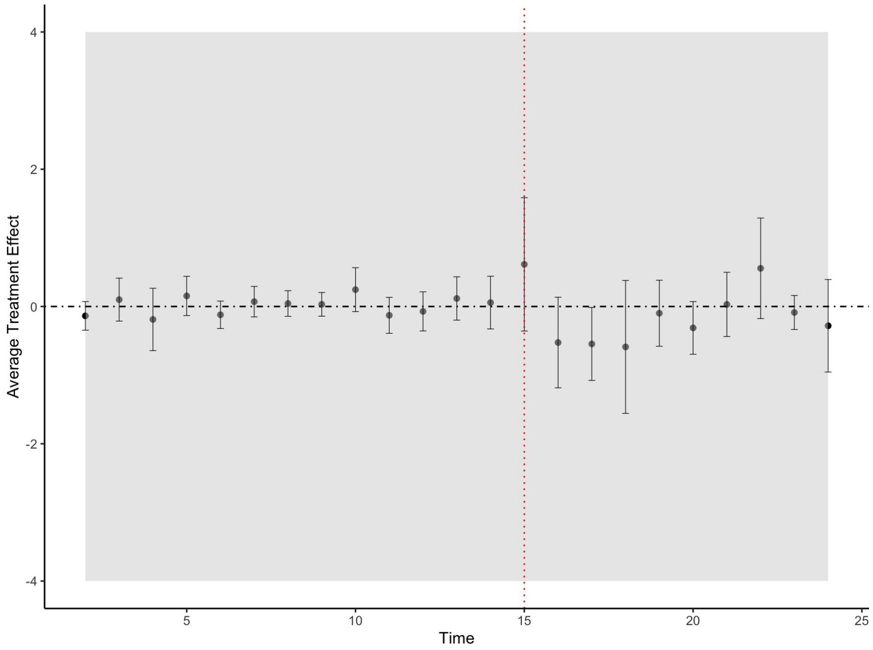


Figure C.14 includes all available pre-treatment periods in the analysis of overdose callouts near supervised consumption sites (SCS). By incorporating all pre-treatment periods, this specification provides a more comprehensive view of baseline trends before the implementation of SCS.