Life Unleaded: Three Essays on The Socio-Economic Effects of Lead Exposure

by

Anthony Christopher Higney

Submitted to the Economics Division in partial fulfilment of the requirements for the degree

Doctor of Philosophy

at the

University of Stirling

February 2023

Abstract

Chapter 1: Does lead pollution increase crime? We perform the first meta-analysis of the effect of lead on crime, pooling 542 estimates from 24 studies. The effect of lead is overstated in the literature due to publication bias. Our main estimates of the mean effect sizes are a partial correlation of 0.16, and an elasticity of 0.09. Our estimates suggest the abatement of lead pollution may be responsible for 7–28% of the fall in homicide in the US. Given the historically higher urban lead levels, reduced lead pollution accounted for 6–20% of the convergence in US urban and rural crime rates. Lead increases crime, but does not explain the majority of the fall in crime observed in some countries in the 20th century. Additional explanations are needed.

Chapter 2: How does lead pollution affect birth outcomes? Does a mother's lead exposure increase the risk of child death? Does it lower the infant's birthweight (a proxy for later health outcomes)? We test these hypotheses by examining an intervention in the Scottish water supply, which reduced water lead levels and blood lead levels in Scotland's largest two cities. In our main estimates, we use a staggered difference-in-differences design estimated with two-way Mundlak pooled OLS. We do not find the lead reduction interventions reduced birthweights. However, we do find that they may have had a large effect on infant deaths. Our main estimates suggest lead reduction accounts for 0.3-0.1 percentage points decrease in deaths in Glasgow and a 0.7-0.1 percentage points decrease in deaths in the Alnwickhill water plant supplied area of Edinburgh. However, these results are not robust to alternative specifications, and therefore can only be taken as weak evidence of an effect.

Chapter 3: Does lead pollution harm educational achievement? And are the marginal effects greater at low or high levels of lead? We use exogenous variation in lead pollution from water treatment in Glasgow, Scotland,

combined with within-household sibling differences, to estimate the effect of lead on education. We compare pre- and post-treatment sibling differences between treated and control areas with difference-in-differences estimation. We find a clear dose-response relationship. Treated areas with low prevalence of lead piping show no change compared to a control group. In contrast, high lead pipe prevalence areas show improvement in educational outcomes. Our findings indicate that countries and areas with very high levels of lead can expect large educational gains from even small amounts of lead abatement, while those with already low levels of lead can expect much lower marginal improvements.

Declaration

I wish to submit the thesis detailed above in accordance with the University of Stirling research degree regulations. I declare that the thesis embodies the results of my own research and was composed by me.

A modified version of Chapter 1 was published in Regional Science and Urban Economics in 2022. A modified version of Chapter 3 was made available as a working paper in 2022.

All three papers are co-authored. Chapter 1, 2 and 3 are co-authored with my PhD advisors Mirko Moro and Nick Hanley. Chapter 3 is also co-authored by Anne Gasteen, who obtained the education data. I confirm that I have contributed over 80% of the research and output for the paper.

Acknowledgements

Many people have contributed to this PhD thesis, in all sorts of ways and I would like to thank them.

Firstly, my PhD advisors, Mirko Moro and Nick Hanley. Mirko was brave enough to give me the ESRC scholarship and, spent several frustrating years getting access to the data for chapters 2 and 3, and has guided me throughout my academic journey. Nick has given me invaluable advice and help whenever I needed it, with both incisive comments and steady encouragement. I thank them both deeply.

Secondly, I would like to thank several others in academia. Tanya Wilson was also part of the interview team that allowed me to do this PhD, helped me get an invite to the Economics of Housing workshop at Royal Holloway, and provided comments on chapter 3. Hector Rufrancos also provided excellent comments on chapter 1 and 3, and was always willing to talk ideas. Ken Gibb taught me a lot about how academia works with policy, local and central government, and the wider community.

Thirdly, I would like to thank my family who have always been there for me. My mum and dad, and my brothers and sisters.

Finally, I would never have started this journey without Karen, my partner of 13 years. She encouraged me to go for a PhD when I was unsure if I would be able to do it, she helped me all the way through, and she always believed in me, and makes my life better every day (although she has a little help there from our dog, Pushkin).

Dedication

To Karen, my mum and dad, and Pushkin (my dog).

Contents

List of tables

List of Figures

Introduction

Pollutants often remain in use even after their harmful effects are known. It follows that they must either be providing benefit to some parties, or the cost of not using them is viewed as too great. Therefore, although pollution is a complex, multifaceted, and inter-disciplinary problem, economic methods offer value in quantifying and evaluating the extent of these costs and benefits. They can also assess the distributional effects of pollution. Pollution is not *just* an economic issue, but it *is* an economic issue.

Lead pollution has been known to be harmful, at least to some, since antiquity. Classical texts show the Greeks Hippocrates and Nicander, and the Romans Pliny and Vitruvius warning of the dangers of lead poisoning 2000 years ago (Waldron, 2012). In 1696 the Duke of Württemberg banned putting lead in wine on pain of death (Eisinger, 1982). As we shall see in chapters 1-3, there has been a further acceleration in knowledge of the harms of lead in the last hundred years. Yet lead is still found in clothing, food, spices, batteries, ceramics, electronics, water pipes, paint, and, although now phased out of road vehicle gasoline, is still used for light air fuel (UNICEF, 2020). As many as 800 million children globally may have elevated levels of lead (GBD, 2019).

There are a number of economic concepts that may explain why widespread lead pollution continues:

1. Imperfect information
2. Asymmetric information
3. Benefits outweigh costs, for some social welfare function
4. Externalities and Distributional Conflict

I explain the potential contribution of my thesis to each of these in turn.

1. Imperfect information

Vague knowledge that lead causes harm may not be enough to make decisions to undertake remediation measures. Lead abatement is costly. For example, the US environmental protection agency estimates removing each water pipeline made of lead could cost $1200-$12,300 per pipeline, for a total cost of $28-$47 billion (EPA, 2019). Estimates for removing all lead paint from US housing range from $32 billion to $442 billion (Ryan, 2013). It is, as yet, unknown how widespread the lead contamination of spices such as turmeric and saffron is in south Asia, or the costs of removal (GiveWell, 2021). There likely remains hundreds of thousands of homes in Scotland with lead piping (Robertson et al., 2020), but it is unknown which houses have them, and would be costly to find out, let alone replace.

Given these costs, combined with competing demands on fiscal and household budgets, especially in low- and middle-income countries, concrete quantification of lead pollution's damages may be required for action to occur. Uncertainty about these costs (and therefore the benefits of remediation) may contribute to a lack of action. My thesis increases the knowledge about the extent of harms caused by lead and helps quantify these costs.

2. Asymmetric information

Asymmetric information is when either a buyer or seller of a good or service has more information than the counterparty. For example, when leaded gasoline began to be produced and used the harms of lead were known not just to the companies making it, but to the US Surgeon General (The Nation, 2000). The FTC in 1936 issued an order that disallowed any criticism of leaded gasoline by competitors (The Nation, 2000). The public, despite the

attempts of some scientists, was thus unaware of the full costs while being told of the benefits**.**

Such information asymmetries still exist. Companies that produce lead tainted spices and clothing know they are doing so, yet it takes research and investigation to find out this is the case[1]. Many consumers may not know about the lead pollution in their goods. While water companies, both private and public, are well aware of the presence of lead piping, many consumers are not. A survey for Scottish Water found that more than 50% of Scottish Households were unaware that some old water service pipes may be made of lead, and 47% didn't know if there was any health risk from lead (Scottish Water, 2018). Research that contributes to the extent and presence of lead's harms, especially if well-disseminated, can help reduce these asymmetries. The papers contained in this dissertation contribute towards this.

3. Benefits outweigh costs, for some social welfare function

The environmental Kuznets curve (EKC) is firstly an empirical hypothesis that many types of pollution follow an inverted U shape with respect to income growth (Grossman and Krueger, 1995). That is, pollution is low when economic activity is low, then it increases as industrialisation and modern economic growth takes hold, until finally it decreases again as structural transformation, new technology, and income effects leading to calls for remediation happen. Secondly, it has been shown theoretically, under some assumptions, with some social welfare functions, the EKC can be optimal behaviour (e.g. Dinda, 2005, and Pasten, 2012).

[1] See, for example, this CBC investigation https://www.cbc.ca/news/business/marketplace-fast-fashion-chemicals-1.6193385

There is disagreement on the extent to which the EKC is empirically true for different pollutants, and, similarly, different theoretical models, with different assumptions, show it can be sub-optimal even if it exists (see the early criticism by Arrow et al., 1995 and also Carson, 2010 for a review of papers).

Nevertheless, under some assumed social welfare functions, with some measures of costs and benefits, and updating information, deciding when to expend marginal resources on pollution reduction can mean delaying pollution abatement. For example, given lead water piping is now a sunk cost with high removal costs, the best use of marginal resources at current income levels may be on other activities or investments (such as improving education or transfers to low-income families). Whether in a rich country with relatively low levels of lead, or a low- to middle-income country with higher levels of lead, lead remediation must compete with other worthy uses of productive capacity.

Quantification of lead pollution damages allows both individuals and governments to optimise when to take action. Although the optimal decision depends both on information, individual utility functions, and the social welfare function. My thesis contributes to a deeper understanding of these trade-offs by helping citizens and policy makers to make informed cost-benefit decisions by determining the optimal allocation of marginal resources for lead remediation efforts.

4. Externalities and Distributional Conflict

Finally, even if there were no information problems with lead pollution, and the social costs of abatement were less than the benefits, distributional conflict could mean lead is not abated. Pollution from lead could bring private benefits to producers but inflict costs on others. If those receiving the costs

are not politically powerful, or are unorganised, they may not be able to force abatement. Such issues are, unfortunately, beyond the scope of this thesis. However, evidence can aid both regulators and activists in raising awareness and gathering support for their cause[2].

Structure of the Thesis

The chapters of this thesis are organised as self-contained papers and appendices are collected at the end of the thesis.

Chapter 1 is the first meta-analysis of the effect of lead pollution on crime. The lead pollution theory of crime states that lead pollution exposure when young increases the propensity to commit crimes when older. The mechanisms being either due to the brain and nerve damage lowering self-control, and increasing aggressiveness, or that lead pollution lowers an individual's ability to invest in human capital and therefore reduces outside options. This theory has been held by some to account for 90% of the rise and fall in crime seen in the 20th century in some countries (Nevin 2000, 2007). While others have called it a statistical artifact (Lauritsen et al., 2016) responsible for none of the decrease. A variety of studies have quantitatively tested this hypothesis since the 1980s. For the first time, we gather all these studies and conduct a meta-analysis. We use cutting-edge publication bias techniques and find evidence for significant publication bias in favour of a positive link between lead and crime. We then carry out over a million different meta-regressions, with differing specifications, to investigate the effect of observable between-study heterogeneity on the lead-crime average effect. We find, after accounting for publication bias and between-study heterogeneity, that there is an effect of lead on crime on average, but this is

[2] Clair Cameron Patterson is widely credited with hastening the end of leaded gasoline: https://www.mentalfloss.com/article/94569/clair-patterson-scientist-who-determined-age-earth-and-then-saved-it

much smaller than is found in most studies alone. Our main estimates imply an elasticity of 0.06-0.12 between lead and crime (i.e. a 10% reduction in lead blood levels leads, on average, to a 0.6-1.2% decrease in crime levels). This further implies around 7-28% of the fall in US homicide was due to lead. Lead pollution does increase crime, but, contrary to the Nevin (2000, 2007) view, it is not responsible for the majority of the 20th century crime decline seen in some high-income countries.

Chapter 2 considers the impact of lead in drinking water on human health. It examines the case of water treatment in Scotland's two largest cities (Glasgow and Edinburgh) in the 1970s-1990s that lead to large falls in water lead levels in households. First in the 70s and 80s the water pH was raised, which made lead pipes less likely to leech their lead into the water supply. Then in the 80s and 90s, orthophosphate was added which again made lead less likely to leech from pipes into the household water out of the tap. Given lead exposure is correlated with income, race and education, some plausibly exogenous variation is needed to identify the effects of lead pollution. The sequence of water treatment events allows us to identify the effect of water-lead pollution on certain birth outcomes. The two outcomes we examine are birthweights and under-5 mortality. The various water plants in our sample were treated at different times. We therefore use a staggered difference-in-differences design to compare the separate treatment cohorts with each other and with the control group, which did not have treated water until much later. We do not find evidence for an effect of lead on birthweights. However, we do find evidence that lead pollution may have caused 23-186 infant deaths in Edinburgh and 216-867 in Glasgow, over the full 25 years of the sample, but this finding is not robust to several alternative specifications.

Chapter 3 examines the effect of lead pollution when born on education outcomes 15-16 years later in Glasgow. For this paper, due to data availability, we look at only the later 1989 water treatment in Glasgow, where the water department added orthophosphate in the water supply. This

orthophosphate treatment reduced water lead levels and is associated with reduced blood lead levels found in Glasgow. The richness of the data allows us to look at the difference between siblings born before and after this treatment, compared to the difference between siblings born before and after the same date in non-treated areas (i.e. the rest of Scotland). This allows us to control for the household fixed effects, including whether lead piping was in that house. We use a difference-in-difference design and find that younger siblings born in the treatment area, i.e. those exposed to lower levels of lead, performed relatively better compared to the control group of younger siblings born elsewhere in Scotland. Crucially, however, we find there is only a robust effect in areas with very high concentration of lead piping, i.e., areas with high levels of lead contamination before the intervention. Previous studies have suggested that even very low levels of lead could have a profound effect on education results. We do not find this to be the case in Glasgow. Lead pollution certainly lowered test scores, but only in areas with known high levels of lead piping. This suggests the worst marginal harms of lead may be at higher levels rather than at low levels of pollution concentration.

In summary, my thesis presents evidence that broad-based reductions of lead exposure in the 20th century have lowered the harms of lead. The extent of the damages may have been over-stated in previous literature (see Chapter 1), but are, nevertheless, real. Future research on lead exposure should explore potential mediating effects which lower the harms of lead (see discussion in chapter 2) and on mapping out the precise dose-response relationship for different outcomes (see discussion chapter 3). Finally, almost all the non-health related quantitative research on lead pollution is focused on rich countries – this needs remedied. For example, only one of the studies in the chapter 1 meta-analysis used data from Africa, while over 70% used data from North America. Given the vast majority of high blood lead levels

today are found in low- and middle-income countries, this is where new knowledge production should be prioritised.

References

Arrow K, Bolin B, Costanza R, Dasgupta P, Folke C, Holling CS, Jansson B-O, Levin S, Mäler K-G, Perrings C, et al. 1995. Economic Growth, Carrying Capacity, and the Environment. Science. 268(5210):520–521. doi:10.1126/science.268.5210.520.

Brookings, 2019. Lead in Minnesota Water. Arrow K, Bolin B, Costanza R, Dasgupta P, Folke C, Holling CS, Jansson B-O, Levin S, Mäler K-G, Perrings C, et al. 1995. Economic Growth, Carrying Capacity, and the Environment. Science. 268(5210):520–521. doi:10.1126/science.268.5210.520.

Carson RT. 2010. The Environmental Kuznets Curve: Seeking Empirical Regularity and Theoretical Structure. Review of Environmental Economics and Policy. 4(1):3–23. doi:10.1093/reep/rep021.

Dinda S. 2005. A theoretical basis for the environmental Kuznets curve. Ecological Economics. 53(3):403–413. doi:10.1016/j.ecolecon.2004.10.007.

Eisinger J. 1982. Lead and wine. Eberhard Gockel and the colica Pictonum. Med Hist. 26(3):279–302. doi:10.1017/S0025727300041508.

EPA (Environmental Protection Agency), 2019. Strategies to Achieve Full Lead Service Line Replacement. [accessed 2023 Jan 10]. https://www.epa.gov/ground-water-and-drinking-water/strategies-achieve-full-lead-service-line-replacement.

GBD, 2020. Global Burden of Disease (GBD 2019) | Institute for Health Metrics and Evaluation [WWW Document]. Glob. Burd. Dis. - IMHE. URL https://www.healthdata.org/gbd/2019 (accessed 1.26.22).

Givewell. 2021. Pure Earth — Support for Reducing Lead Exposure in Low- and Middle-Income Countries | GiveWell. [accessed 2023 Jan 10]. https://www.givewell.org/research/incubation-grants/Pure-Earth-lead-exposure-July-2021.

Grossman GM, Krueger AB. 1995. Economic Growth and the Environment. The Quarterly Journal of Economics. 110(2):353–377. doi:10.2307/2118443.

Lauritsen, J.L., Rezey, M.L., Heimer, K., 2016. When Choice of Data Matters: Analyses of U.S. Crime Trends, 1973–2012. J. Quant. Criminol. 32, 335–355. https://doi.org/10.1007/s10940-015-9277-2

Nevin, R., 2007. Understanding international crime trends: The legacy of preschool lead exposure. Environ. Res. 104, 315–336. https://doi.org/10.1016/j.envres.2007.02.008

Nevin, R., 2000. How lead exposure relates to temporal changes in IQ, violent crime, and unwed pregnancy. Environ. Res. 83, 1–22. https://doi.org/10.1006/enrs.1999.4045

Pasten R. 2012. The Environmental Kuznets Curve: A Survey of the Theoretical Literature. IRERE. 6(3):195–224. doi:10.1561/101.00000051.

Perman R. 2011. Natural resource and environmental economics. Fourth edition. Harlow, United Kingdom: Pearson Education.

Pure Earth — Support for Reducing Lead Exposure in Low- and Middle-Income Countries | GiveWell. [accessed 2023 Jan 10]. https://www.givewell.org/research/incubation-grants/Pure-Earth-lead-exposure-July-2021.

Robertson, G., Wilson, A., Papastathopoulos, Y., Graham, M., Eades, L., Chatzisymeon, E., 2020. Estimating the number of homes in Scotland with internal lead piping. [WWW Document]. https://www.dwqr.scot/media/1f1n3kv4/ueh-007-19-project-management-doc-048-final-project-report-20-november-2020.pdf (accessed 12.31.22)..

Ryan D. 2013. Calls for Removing All Lead Paint From US Housing Are Misguided. Am J Public Health. 103(5):e5. doi:10.2105/AJPH.2013.301262.

Scottish Water, 2018. SR21 Customer Engagement Phase 2 – The Removal of Lead in Water. [accessed 2023 Jan 10]. https://www.scottishwater.co.uk/-/media/ScottishWater/Document-Hub/Key-Publications/Strategic-Plan/Research-Projects/300120ResearchProject16LeadRemoval.pdf

Shahbaz M, Ozturk I, Afza T, Ali A. 2013. Revisiting the environmental Kuznets curve in a global economy. Renewable and Sustainable Energy Reviews. 25:494–502. doi:10.1016/j.rser.2013.05.021.

The Nation, 2000 Mar 2. The Secret History of Lead. [accessed 2023 Jan 10]. https://www.thenation.com/article/archive/secret-history-lead/.

UNICEF, 2020. The Toxic Truth: Children's exposure to lead pollution undermines a generation of potential | UNICEF [WWW Document]. URL https://www.unicef.org/reports/toxic-truth-childrens-exposure-to-lead-pollution-2020 (accessed 12.31.22).

Waldron HA. 1973. Lead Poisoning in the Ancient World. Medical History. 17(4):391–399. doi:10.1017/S0025727300019013.

Chapter 1

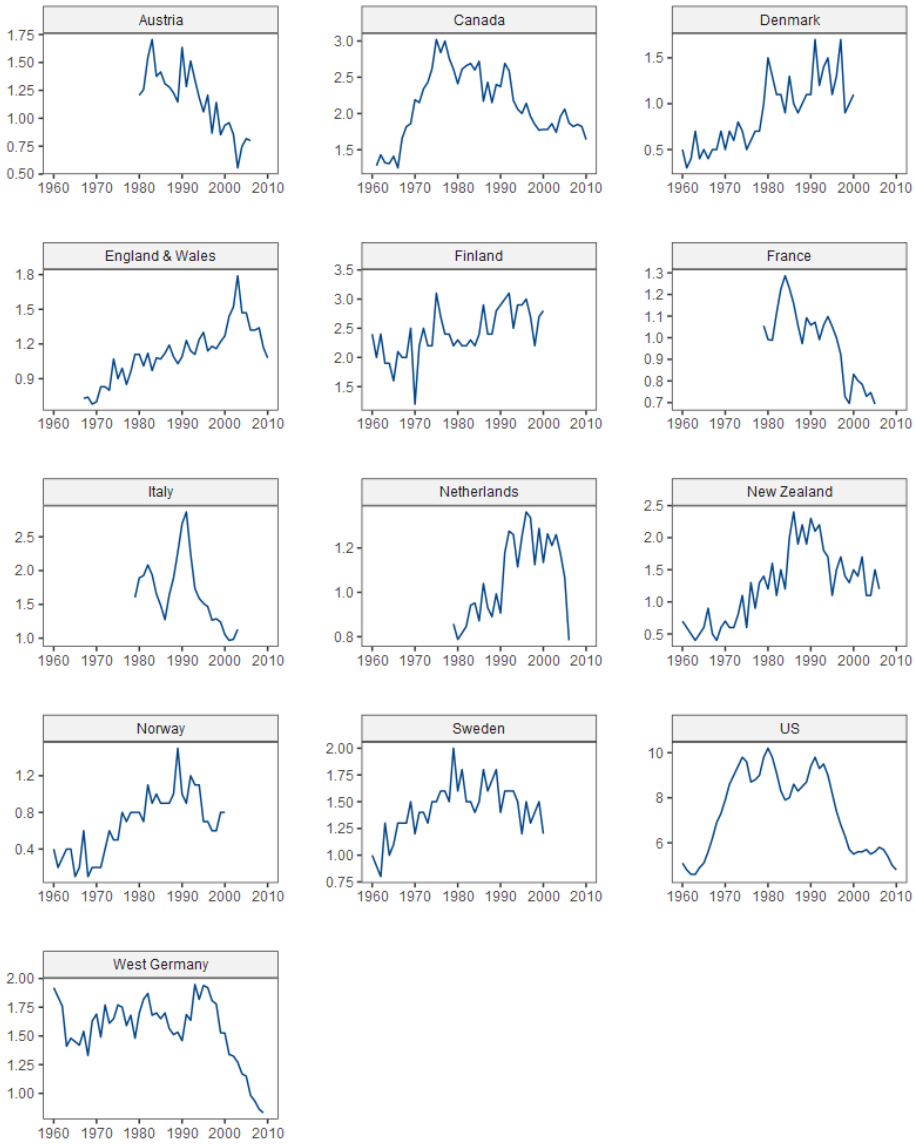
The Lead-Crime Hypothesis: A Meta-Analysis[a]

[a]This chapter is co-authored with Mirko Moro and Nick Hanley. I thank Paul Ferraro, Kevin Schnepel, Tom Stanley, Hector Gutierrez Rufrancos, Conny Wollbrant and Jonathan Norris, for their comments on this paper. We also thank participants at the European Association of Environmental and Resource Economists (EAERE) conference and the Meta-Analysis in Economics Research (MAER) colloquium. Any remaining errors are our own. We are also grateful for the financial support from the UK Economic and Social Research Council.

1.1. Introduction

Homicide rates spiked and then fell in a consistent pattern across many western countries in the 20th century (figure 1.1). In the US alone the homicide rate has halved since the 1980s, when it was as high as the road fatality rate is today. In other countries the falls are not so great in magnitude, but still amount to many lives saved. If the causes of this fall were known, many more deaths and trauma could be prevented.

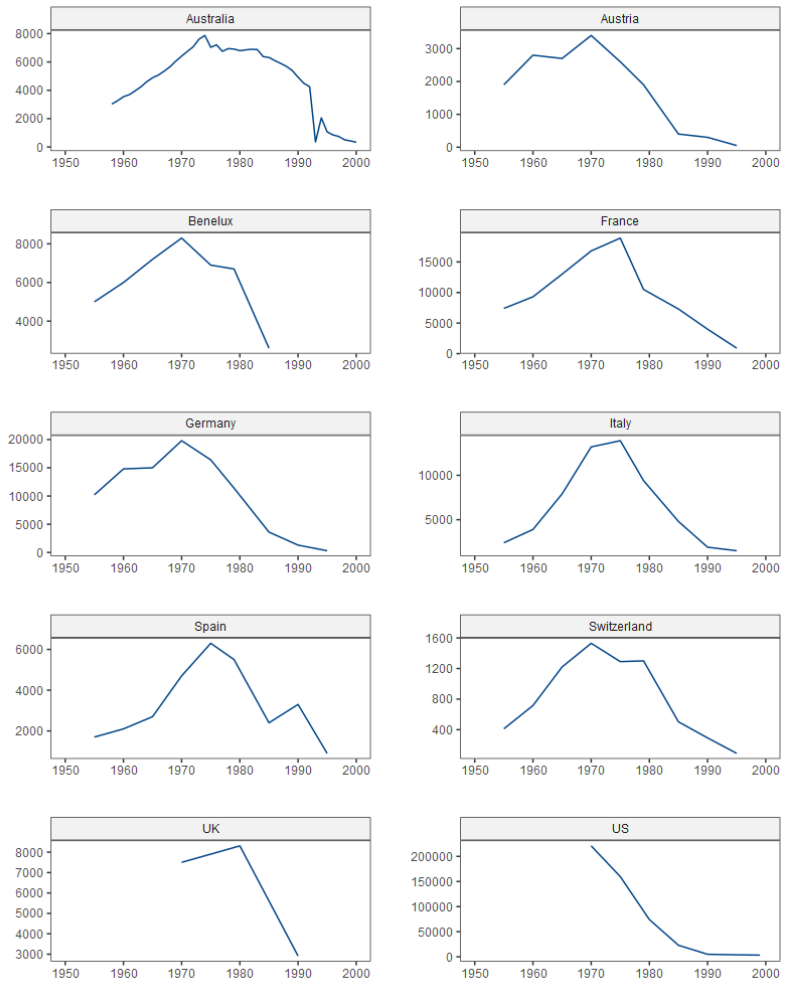
Figure 1.1 Homicide Rate per 100,000 by Country



Sources: New Zealand Police (2018); Buonanno *et al*. (2011), UK Home Office (2019); Uniform Crime Reports for the United States (2019); Falck, Von Hofer & Storgaard (2003); Statistics Canada (2019); Birkel and Dern (2012).

Is lead pollution responsible? Lead is a toxic metal linked to harmful health and behavioural outcomes (see section 1.2). Studies have pointed to falling lead levels in the environment as a cause of the falls in homicide, and as a factor in reducing crime rates in general. Some have claimed that lead emissions account for as much as 90% of the fall in violent crime (Nevin, 2000, 2007). The reduction in lead pollution over time is largely due to falling emissions from leaded gasoline (figure 1.2), but also due to less lead pollution from water pipes, paint, food, and soil.

Figure 1.2 Lead Emissions by Country (1000 kg Y^(-1))



Source: Dore *et al.* (2006), Schwikowski *et al.* (2004), Kristensen (2015), Statistical Abstract of the United States (2009).

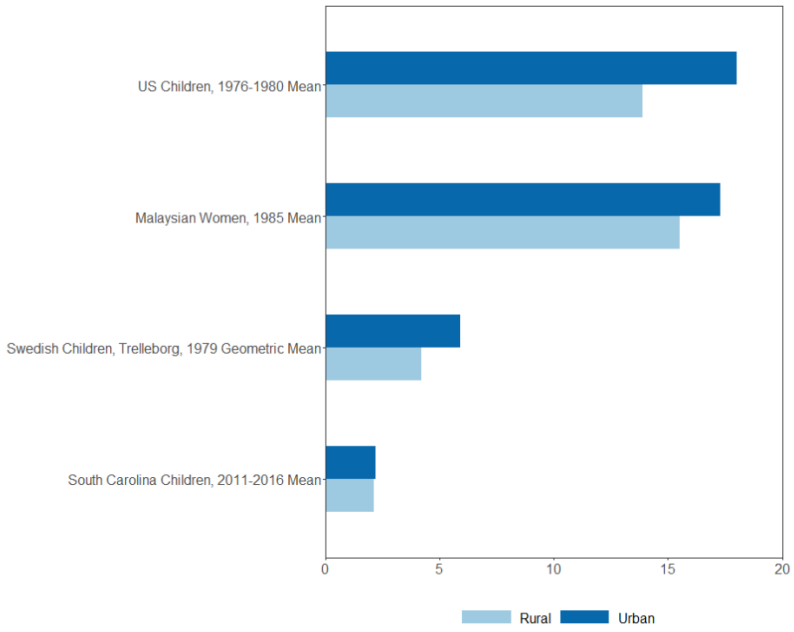
Crucially, this reduction in exposure to lead pollution over time has been spatially uneven. Pollution tends to be more concentrated within urban areas (Carrozzi and

Roth, 2020; Borck and Schrauth, 2021) and lead is no exception. The lead burden is likely to be higher in urban areas for several reasons (Levin et al., 2021, O'Flaherty and Sethi, 2015). Urban road traffic is higher, and urban residents often live closer to congested roads, a risk factor before the phase out of leaded gasoline in most countries. Urban dwellers tend to live in closer proximity to lead working sites. Urban areas also have less turnover in soil, which therefore accumulates a larger concentration of lead.

Figure 1.3 shows that blood lead levels were generally higher in urban areas than rural areas. Similarly, the left-hand chart in figure 1.4 shows that blood lead levels were elevated for children under 5 years of age in the US, at least in the period before the phase out of leaded gasoline. In the 1970s and 80s, blood lead levels in Metropolitan Statistical Areas (MSAs) with populations greater than 1 million were 15% higher than levels in other parts of the country. The chart also illustrates the swift convergence in blood lead levels across rural and urban areas in the 1990s. We see a similar pattern in the US crime trends in the right-hand chart of figure 1.4. The urban crime rate, as measured by the National Crime and Victimisation Survey (NCVS) was 70% higher than the rural rate in the early 1990s. There then followed a convergence in crime rates in the 21st century, although the urban rate remains somewhat higher.

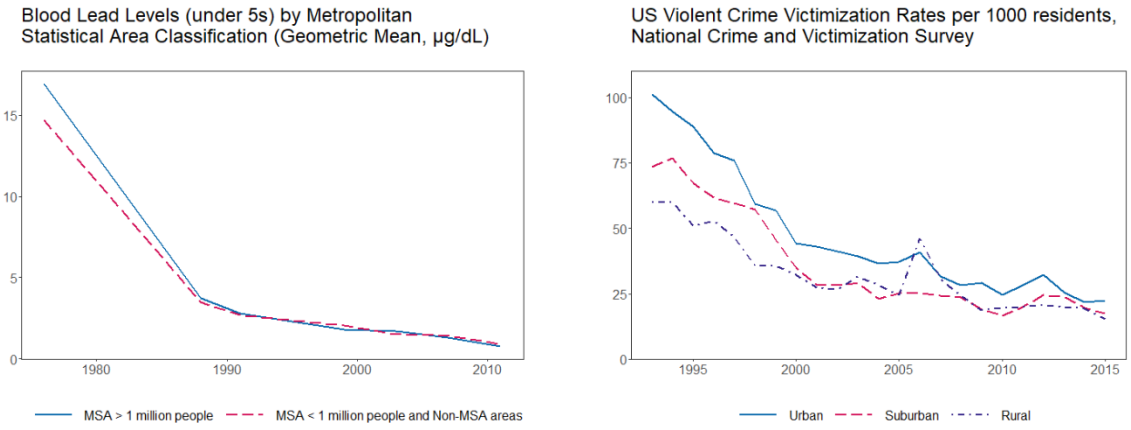
These trends indicate lead *could* explain a large part of the observed variation in crime, both over time and between urban and rural areas. However, the rise and fall pattern in figure 1.1 is by no means uniform. Furthermore, Buonanno *et al*. (2011) show that while total crime has behaved similarly to homicide in the US, it has not in Europe (figure 1.5). Similarly, outside the US, population density is associated with lower rather than higher crime rates. Ahlfeldt and Pietrostefani (2019), synthetising the literature on the economic effects of density, estimate that a log-point increase in density is associated with a decline in crime of 0.085 log-points. In the US they find the opposite, density is associated with higher crime.

Figure 1.3 Urban and Rural Blood Lead Levels (µg/dl)



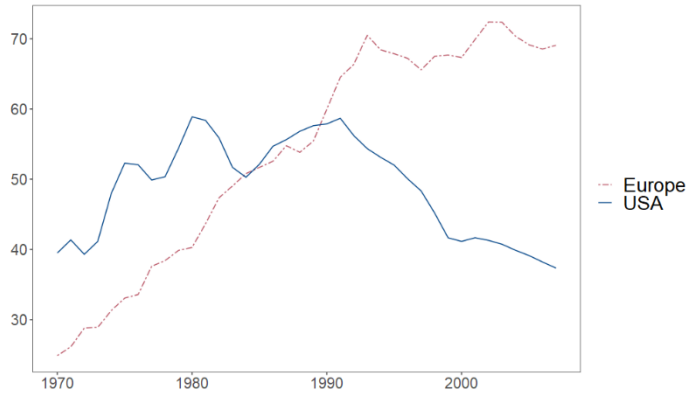
Source: Mahaffey *et al.* (1982), Lim *et al.* (1985), Strömberg Shütz and Skerfving (1995), and Aelion and Davis (2019).

Figure 1.4 Urban/Rural Blood Lead Levels and Violent Crime in the US



Source: Egan *et al.* (2021) and Bureau of Justice Statistics (2022)

Figure 1.5 Total Recorded Crime Rate per 100,000 in USA and Seven European Countries



Source: Buonanno *et al*. (2011). The countries are: Austria, France, Germany, Italy, The Netherlands, Spain, and the UK.

Alternative hypotheses for the observed fall in crime in some countries range from falling poverty levels (Rosenfeld and Fornango, 2007, and Messner, Raffalovich, and Mcmillan, 2001), to demographic transition, where an ageing population is less likely to be victimised by or engage in crime (Fox, 2005, chap. 9; Baumer, Rosenfeld, and Wolff, 2012), increased/better policing or incarceration (Levitt, 1996, 1997, 2004; Marvell and Moody, 1996; and Corman and Mocan, 2000), to more controversial hypothesis such as legalized abortion reducing the number of children born into "adverse home environments" (Donohue and Levitt, 2001, 2019; Buonanno et al., 2011). Tcherni-Buzzeo (2019) provides a recent summary of potential causes.

Against this background, our paper conducts the first meta-analysis of the effect of lead on crime. We systematically review the literature and construct a dataset containing 542 estimates from a total of 24 studies. We convert these estimates to comparable effect sizes. For this full sample we use partial correlation coefficients. We also convert estimates to elasticities, where it is possible to do so, and analyse this subsample of 312 estimates from 11 papers. Throughout the paper we account for the importance of the research design in identifying credible treatment effects by running separate analyses on the subsample of papers that address sorting or endogeneity bias of lead and crime explicitly. They do this by examining natural experiments where there is plausibly exogenous variation in lead exposure. These

studies estimate effects using research designs such as difference-in-difference or instrumental variables. For simplicity, we label this subsample as the "addressing endogeneity" sample. This subsample consists of 7 studies and 220 estimates when using partial correlation coefficients. The sample declines to 5 studies and 211 estimates when we compute elasticities instead.

We perform tests for publication bias and find that the effect of lead on crime is overstated in the literature due to this bias. Furthermore, we find substantial between-study heterogeneity in our sample. We therefore use meta-regression to estimate an average effect size accounting for both publication bias and the observable between-study heterogeneity. We take into account model uncertainty by estimating over 1 million meta-regression specifications, using every combination of our covariates on both the full sample, the elasticity subsample, and several subsamples which exhibit less between-study heterogeneity. We plot the distributions of the estimated average effect size of lead on crime and calculate its mean.

Our main finding is that the estimated mean effect size, evaluated at sample averages, is a partial correlation of 0.16 in the full sample, and an elasticity of 0.09 in the subsample. We also find there are differences between the average effect size when we use the full sample, and when we use only study designs that address endogeneity with quasi-experimental methods. The mean partial correlation coefficient for the "addressing endogeneity" sample is only 0.01, far smaller than the full sample estimate. However, when we use the smaller sample of studies which address endogeneity and have elasticity estimates the mean elasticity range is 0.05-0.17.

We also distinguish between studies in which estimates are based on regional or area-level (e.g., US states) from individual-level data. The sample of studies that use crime in an area as the focus of analysis have a larger mean effect size compared to those of studies which focus on individual behaviour. Conversely, we do not find evidence of differences for the effect of lead on different types of crime when we use homicide, violent, and non-violent crime samples.

Finally, we examine the share of the fall in crime in the late 20th century that lead pollution accounts for. Using the example of homicide in the US, our range of elasticity estimates suggests the fall in blood lead levels is responsible for 4-15-percentage points of the 54% fall in homicide from its peak, with our main estimate being 8. This would mean lead explains around 7-28%15% of the fall in crime, leaving 93-72% unaccounted for. When we estimate the share of US urban/rural volent crime convergence explained by falling lead levels, we obtain a figure of 6-20%, with our main estimate being 11% Our findings suggest that, while the effect of lead pollution on crime is positive, it is not responsible for the majority of the fall in crime observed in some countries in the 20th century, or the majority of the urban/rural crime convergence. Therefore, other explanations require further investigation.

1.2 Lead and Crime

Lead has long been part of the human environment. It was used in cosmetics, paint, and as coinage in ancient China (Schafer, 1956). Similar uses were recorded in ancient Egypt, India, and across the Bronze Age world (Needleman, 1992). The sweet taste of lead acetate meant that the Roman Empire, and later medieval Europe, used lead to sweeten wine, cider, and food (Lessler, 1988). The Romans had many other uses for lead, using it for cooking utensils, pottery, and water pipes (Hernberg, 2000). Indeed, Roman use of lead was prodigious, with estimates from Greenland artic ice cores putting the increase in atmospheric lead pollution at around 4000 metric tons a year at its peak 2000 years ago (Hong *et al*. 1994). This is equivalent to the UK's lead pollution emissions in the mid-1980s, when leaded gasoline had not yet been phased out.

Lead is a useful but toxic metal. At high levels of exposure even adults will experience lead poisoning. Acute lead poisoning is rare but can kill quickly. Chronic poisoning can still kill and is associated with abdominal pain, organ failure, tumours, and exhaustion, amongst other symptoms (WHO, 2010a). Although chronic lead poisoning in adults still happens, and appears to affect behaviour, it is primarily the long-term lead exposure of children that is thought to influence crime rates.

Children are especially vulnerable to lead pollution. Children not only absorb more lead per unit body weight than adults, but, as the brain and nervous system are still developing, lead has more harmful long-term effects even at low levels (WHO, 2010b). Lead is chemically similar to calcium. Calcium is important for cell growth, and synaptic functioning, as well as a myriad of other body processes (Sanders et al., 2010). Therefore, lead is particularly harmful to the developing brain and nervous system, and thus in the womb and early infancy are the worst time to be exposed to lead (WHO, 2010b).

The extent to which children have been exposed to lead pollution has varied substantially, both over time and spatially. As detailed in the introduction, for many OECD countries lead air pollution rose sharply in the mid-20th century before peaking in the 70s and 80s (figure 1.2). Children in urban areas tended to have higher blood lead levels during this period (figures 1.3 and 1.4). The highest average blood lead levels for children today are in low and middle income countries, with one estimate putting the share of children with elevated blood levels (above 5μg/dL) at one third (GBD, 2019).

Yet even today, in countries that have reduced blood lead levels, there remain pockets with higher pollution. Cities with low pH water supplies tend to have higher lead levels if they also have lead pipes, because the water reacts more strongly upon the lead piping. Feigenbaum and Muller (2016), using distance to a lead refinery as an instrument, find these cities to have higher homicide rates in the early 20th century. Aizer and Currie (2019) find that blood lead levels are higher for those living near a road, but this only applies in the period before the phasing out of leaded gasoline. Tanaka *et al*., (2022) show that pollution around lead-acid battery recycling plants in the US sharply reduced after an air-quality law was introduced in 2009, but this led to offshoring of lead battery recycling to Mexico. Infants living near the Mexican plants began to experience worse health outcomes as a result. Highly concentrated lead pollution, and higher blood lead levels, have also been found near airports (Zahran *et al*., 2017), lead smelters (Stromberg *et al*., 1995) and NASCAR racetracks (Hollingsworth and Rudik, 2021). This inequality in lead exposure means that any effect of lead on crime will also be spatially uneven.

The causal chain of lead to crime starts with the biological changes it induces at this young age. The mechanism for these changes is laid out in Sanders *et al*. (2010), and there is an array of evidence for lead's negative effects. These include impaired nerve conduction (Sindhu and Sutherling, 2015), damaged myelination in the nerve system (Brubaker *et al*., 2009), impeded brain development (Lanphear, 2015), and reduced brain matter (Cecil *et al*., 2008).

The next link in the chain is from biological change to behavioural change in later life. Meta-analyses have found that lead exposure is associated with aggressiveness and other conduct problems (Marcus, Fulton, & Clarke, 2010), lower IQ (Schwartz, 1994), and impaired cognitive functioning (Vlasak et al., 2019, and Seeber et al, 2002).

The final link is from behavioural changes to an increased propensity to commit crime. There are several possible mechanisms. Needleman pioneered research on lead exposure and aggressiveness (1996), suggesting it is linked to violent crime in particular. In contrast, Denno (1990) and Fergusson, Boden and Horwood (2008) argue that the link is through lower education outcomes, leading to worse life outcomes, which causes increased criminality. This mechanism is consistent with Becker's (1968) economic theory of crime, where lower opportunity cost makes crime relatively more attractive, and suggests lead would show a stronger link to property crime than violent crime. A third mechanism was proposed by Gottfredson and Hirschi (1990), where lack of self-control, combined with opportunity, causes higher crime rates. Lead has been associated with increases in impulsivity (Winter and Sampson, 2017), and so may cause an increase in crime through this process. If this mechanism were true we might expect increases in violent crime, non-violent crime, or both. Separating the different types of crime may help identify which, if any, mechanism lead acts through. However, whilst a range of mechanisms have been laid out linking lead in the environment to the propensity to commit crime, the strength of this link is a matter of empirical enquiry. The main objective of this paper is to quantify the strength of this link from the range of empirical work reported to date. To do this, we use meta-analysis.

1.3 Data

Meta-analysis data collection begins by specifying the criteria which studies must fulfil to be accepted into the analysis.

The criteria we chose were:

1. The explanatory variable must be some quantitative measure of lead exposure.
2. Outcome variable must measure crime in some way (i.e. not other types of behaviour such as aggressiveness or depression).
3. Must have original estimates, i.e. no review papers.
4. Must have estimates that can be combined into a meta-analysis.
5. Be published before December 2019.
6. Study must be available in English.

We then undertook a systematic literature review for papers on Web of Science, PubMed, and Google Scholar in 2019. We also searched on NBER and REPEC for working papers to include as much "grey" literature as possible. The keyword combinations used were:

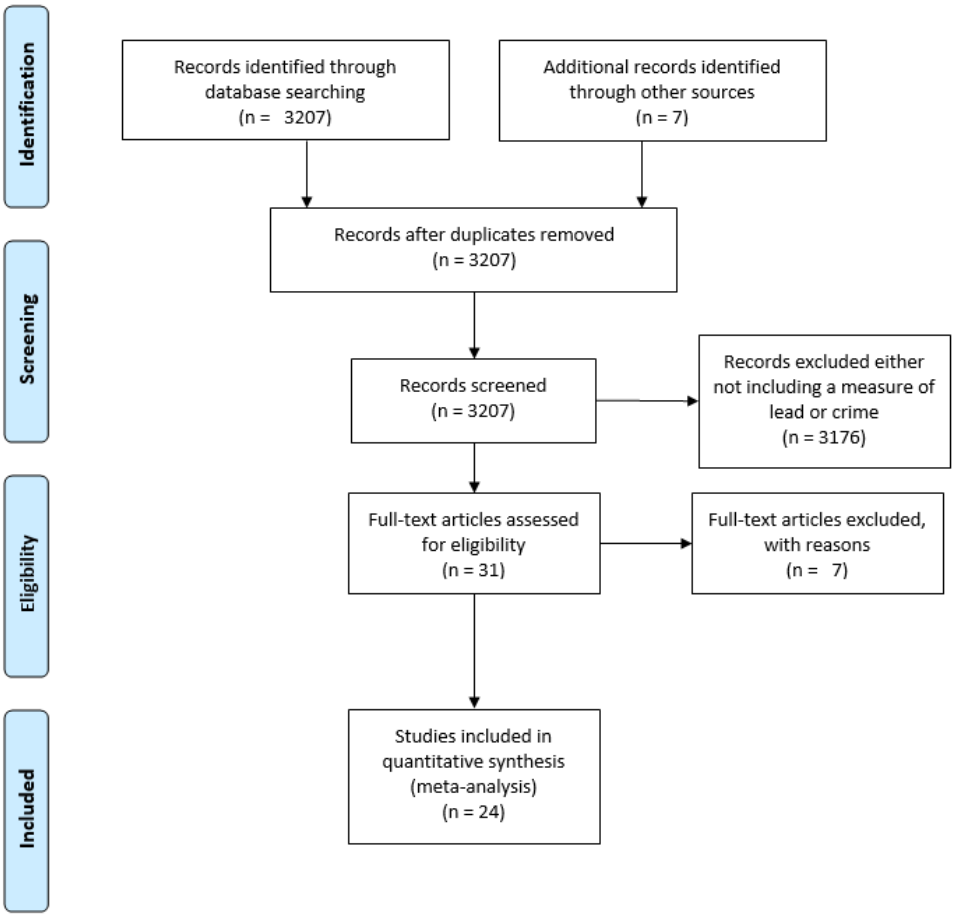
"lead", or "lead" AND "pollution", or "lead" AND "poisoning", or "lead" AND "exposure", or "lead" AND "blood", or "lead" AND "air", or "lead" AND "paint", or "lead" AND "water"

Combined with:

"crime" or "conviction" or "arrest" or "jail" or "prison"

After searching, papers were screened to see if they fulfilled the criteria, as laid out in the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram (figure 1.6). A review and description of the studies included is given in appendix A.

Figure 1.6 PRISMA Flow Diagram of Studies Selection Process



The vast majority of the studies identified in the literature review did not fulfil criteria one or two and therefore did not estimate the lead-crime relationship. These were then filtered out at the screening stage. 31 papers did estimate the lead-crime relationship, but 7 of these could not be converted into comparable effect sizes, failing criterion four. Criterion four is needed because estimates must be combined in a meta-analysis. Estimates are made comparable by converting into a common metric, such as the partial correlation coefficient (PCC), or an elasticity. Most regression coefficients and simple correlations can be converted into PCCs easily. Odds ratios and standardised mean differences can also be converted into PCCs. However, five papers used risk ratios (Boutwell *et al*., 2016; Boutwell *et al*., 2016; Haynes *et al*., 2011; Stretesky and Lynch, 2001; and Write *et al*., 2008). Risk ratios can be converted into odds ratios, which can then be converted to PCCs, but need a base rate risk to do so. It was not possible to infer a base rate risk from the data

available in the papers. Therefore, these papers were excluded at the eligibility stage. One other paper (Masters and Coplan, 1999) contained charts but not enough information to make PCCs and was excluded. Similarly, Denno (1990) did not have enough information to use the estimates. No papers were excluded based on criteria six, but search terms were only in English. This left 24 papers in the final meta-analysis dataset.

We organised accepted papers into a dataset following the guidelines for meta-analysis in economics in Havránek *et al*. (2020). Every paper gave multiple estimates for the effect of lead on crime. Meta-analyses tend to either select one estimate from each study as a "representative" estimate; or take all estimates and account for the potential clustering of estimates from the same study. Both are defensible. Taking all estimates means more information available for the meta-analysis. Representative estimates, on the other hand, may be less biased. For example, a researcher may show a simple OLS estimate before giving reasons for why it will be biased. They then go on to use their preferred method of estimation, which attenuates this bias. In most of our analysis we use all estimates from the studies, but as a robustness check we also test our results by using one representative estimate from each study in appendix E. The results are similar.

In the full sample, there are 542 estimates from the 24 studies. The dataset forms an unbalanced panel, with each estimate being an observation and observations grouped by study. The studies included span across a variety of disciplines including economics, sociology, medicine, epidemiology, and criminology.

Study effect sizes were then converted to the common effect size. Conversion is necessary because both lead and crime are measured in different ways in each paper, and therefore must be converted to be comparable. All studies in the full sample could be converted to PCCs. See appendix B for more details of how PCCs and the PCC standard errors are calculated.

PCCs measure the correlation between two variables holding other variables in the model constant. Their sizes are not intuitive. They have no unit and cannot be interpreted quantitively in a meta-analysis with varied measurements of outcome (Doucouliagos, 2011). However, as they are bounded from -1 to 1, they do offer a

sense of the magnitude and direction of an effect. In a survey of economic effect sizes Doucouliagos (2011) offers the following rough guidelines: 0.07-0.17 is a small effect size, 0.18-0.33 is a moderate one, and above 0.33 a large one. For most of the paper, we follow this taxonomy, but a small effect combined with a large absolute change in a variable can still mean it is significant for welfare.

We were also able to convert some study estimates into elasticities. The elasticities measure the percent change in some measure of crime, given a percent change in some measure of lead pollution. They provide a better measure of the real effect rather than the measure of statistical strength the PCCs provide. The trade-off is that the sample is smaller and therefore may be less representative of the literature. There are 11 studies and 312 estimations in what we label for simplicity the "elasticity subsample".

Table 1.1 presents the mean, median and weighted average PCC for each study (with weights being equal to the precision, 1/standard error of the PCC). It also includes some information on the characteristics of each study. We do the same for the elasticity sample in table 1.2.

Table 1.1 Partial Correlation Coefficients from the Studies Used in Full Sample Meta-analysis

Study & Year	Median	Mean	Weighted Average	Type of Crime	Individual or Area-level	Addresses Endogeneity
Aizer & Currie (2019)	0.027	0.019	0.019	Violent and non-violent	Individual	Yes
Barrett (2017)	0.556	0.556	0.589	Violent	Area	No
Beckley *et al.* (2018)	0.065	0.061	0.063	Violent and non-violent	Individual	No
Billings & Schnepel (2018)	0.122	0.113	0.103	Violent and non-violent	Individual	Yes
Curci & Masera (2018)	0.027	0.043	0.029	Violent	Area	Yes
Dills, Miron & Summers (2008)	0.022	0.021	0.021	Violent and non-violent	Area	No
Feigenbaum & Muller (2016)	0.054	0.056	0.053	Only Homicide	Area	Yes
Fergusson *et al*. (2008)	0.080	0.079	0.080	Violent and non-violent	Individual	No
Grönqvist, Nilsson and Robling (2019)	0.002	0.003	0.003	Violent and non-violent	Individual	Yes
Lauritsen *et al.* (2016)	0.740	0.495	0.742	Violent and non-violent	Area	No
Lersch & Hart (2014)	0.043	0.043	0.043	Violent and non-violent	Area	No
Manduca & Sampson (2019)	0.087	0.087	0.087	Violent and non-violent	Individual	No
Masters *et al.* (1998)	0.051	0.061	0.061	Violent and non-violent	Area	No
McCall & Land (2004)	-0.017	-0.017	-0.017	Only Homicide	Individual	No
Mielke & Zahran (2012)	0.526	0.497	0.515	Violent	Area	No
Needleman *et al.* (2002)	0.336	0.307	0.324	Non-violent	Individual	No
Nevin (2000)	0.914	0.912	0.937	Violent	Area	No
Nevin (2007)	0.808	0.710	0.874	Violent and non-violent	Area	No
Nkomo *et al.* (2017)	0.004	0.052	0.088	Violent	Individual	No
Reyes (2007)	0.059	0.053	0.053	Violent and non-violent	Area	Yes
Reyes (2015)	0.026	0.036	0.029	Violent and non-violent	Individual	Yes
Sampson and Winter (2018)	-0.065	-0.046	-0.046	Violent and non-violent	Individual	No
Stretesky & Lynch (2004)	0.396	0.352	0.331	Violent and non-violent	Area	No
Taylor *et al.* (2018)	0.371	0.377	0.429	Violent	Area	No

Notes. Table shows median and mean partial correlation coefficient (PCC) estimates from each study of the effect of lead on crime. These averages are computed from 542 estimates from 24 studies used for the full sample meta-analysis. It also shows an average where estimates are combined in a weighted average with the weights equal to one divided by the standard error. Table also shows what type of crime was used as dependent variable in each study, whether the study unit of interest was an individual or a geographic area, and whether any estimates in the study used a design that attempted to account for endogeneity. All coding is done at an estimate level, so a study may include both "addresses endogeneity" and "correlational" estimates, violent and non-violent estimates etc.

Table 1.2 Estimated Elasticities in Studies Used in Elasticity Subsample Meta-analysis

Study & Year	Median	Mean	Weighted Average	Type of Crime	Individual or Area-level	Addresses Endogeneity
Barrett (2017)	0.68	0.68	0.61	Violent	Area	No
Curci & Masera (2018)	0.20	0.22	0.12	Violent	Area	Yes
Feigenbaum & Muller (2016)	0.72	0.73	0.32	Only Homicide	Area	Yes
Fergusson et al. (2008)	2.45	2.14	0.94	Violent and non-violent	Individual	No
Grönqvist, Nilsson and Robling (2019)	0.04	0.06	0.07	Violent and non-violent	Individual	Yes
Mielke & Zahran (2012)	0.53	0.53	0.48	Violent	Area	No
Reyes (2007)	0.74	0.61	0.29	Violent and non-violent	Area	Yes
Reyes (2015)	0.50	0.64	0.40	Violent and non-violent	Individual	Yes
Sampson and Winter (2018)	-0.22	-0.29	-0.12	Violent and non-violent	Individual	No
Stretesky & Lynch (2004)	0.15	0.15	0.15	Violent and non-violent	Area	No
Taylor et al. (2018)	0.24	0.25	0.26	Violent	Area	No

Notes. Table shows median and mean elasticity estimates from each study of the effect of lead on crime. These averages are computed from 312 estimates from 11 studies used for the "elasticity" subsample. It also shows an average where estimates are combined in a weighted average with the weights equal to one divided by the standard error. Table also shows what type of crime was used as dependent variable in each study, whether the study unit of interest was an individual or a geographic area, and whether any estimates in the study used a design that attempted to account for endogeneity. All coding is done at an estimate level, so a study may include both "addresses endogeneity" and "correlational" estimates, violent and non-violent estimates etc.

1.4 Methods and Results

1.4.1 General Approach4

Let θ_j be an effect size of interest in study *j*. Study *j* uses some method to estimate θ_j and these we denote as $\hat{\theta}_{ij}$, for estimate *i* of study *j*. Researchers are often interested in both how close $\hat{\theta}_{ij}$ is to θ_j (internal validity), and in how useful θ_j would be in predicting results from a similar event or study. This can be interpreted as the degree of external validity of a study.

[4]This section owes much to the excellent expositions in Meager (2019), Rubin (1981), and Röver (2018). Much of their explanation deals with Bayesian methods but works equally well for non-Bayesian methods up to the point we arrive at.

If θ_j is a draw from some distribution with a likelihood function $\psi(\cdot\,|\Theta)$ such that $\theta_j \sim \psi(\cdot\,|\Theta)\ \forall\ j$, then there exists some parameter(s) Θ which can give information about a new draw θ_{j+1} from that distribution. It is the parameters contained in Θ that are estimated in a meta-analysis. There may be several parameters of interest, but in practice meta-analyses usually estimate two: θ , the mean of the distribution, and the variance τ^2. This is because meta-analyses tend to impose the assumption $\theta_j \sim N(\theta, \tau^2)\ \forall\ j$ in the interests of efficient estimation. Even if this is not the true shape of the distribution McCulloch and Neuhaus (2011) show, both in theory and simulation, that maximum likelihood estimates are robust to different distributions of θ_j around θ. If we also assume, as the individual studies themselves usually do, that $\hat{\theta}_{ij}$ follows a normal distribution with mean θ_j and variance σ_{ij}^2 , then this leads to the normal-normal hierarchical model of Rubin (1981):

(1) $$\theta_j \sim N(\theta, \tau^2)\ \forall\ j$$

(2) $$\hat{\theta}_{ij} \sim N\left(\theta_j, \sigma_{ij}^2\right)\ \forall\ i \text{ and } \forall\ j$$

(3) $$\hat{\theta}_{ij} \mid \theta, \sigma_{ij}^2, \tau^2 \sim N\left(\theta, \sigma_{ij}^2 + \tau^2\right)\ \forall\ i \text{ and } \forall\ j$$

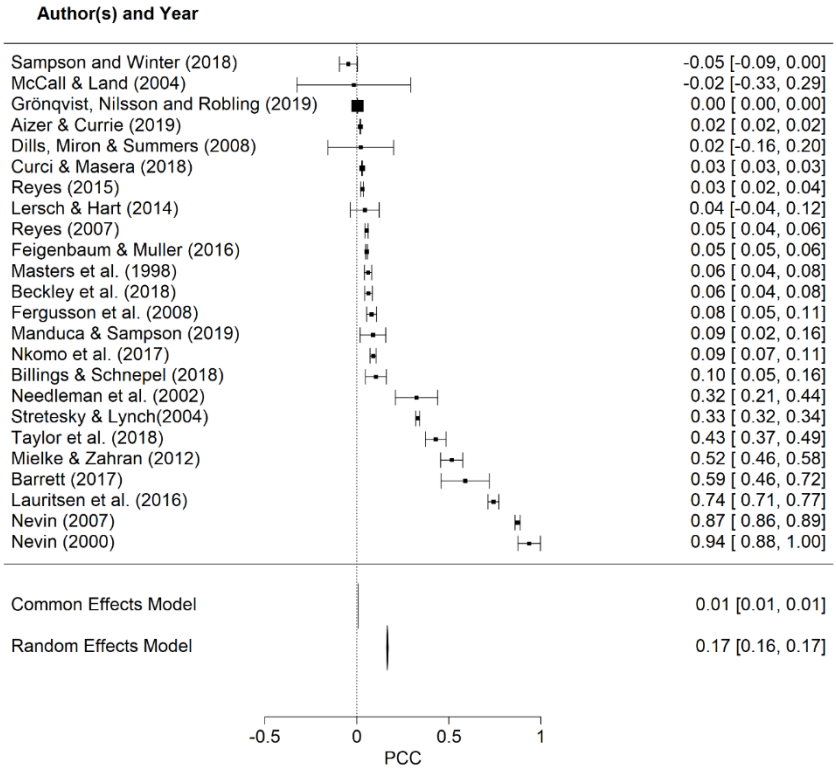
where the last expression follows from the previous two but is expressed in marginal form, as in Röver (2018). This marginal form can be further extended to be conditional on observable variables, common across the $\hat{\theta}_{ij}$'s, as we do in our meta-regression analysis.

The variance of the effect size distribution $\tau^2$is a crucial measure of how useful aggregation of estimates will be. If τ^2 is zero, then all studies are estimating the exact same effect and it is only the study variances that affect how well they can predict θ_{j+1}. This we call the common effect model following the Rice, Higgins, and Lumley (2018) terminology. As τ^2 grows larger, aggregation becomes less useful. $\tau^2 \to \infty$ represents an "apples and oranges" comparison where meta-analysis should never be undertaken.

1.4.2 Between-Study Heterogeneity

We begin investigating between-study heterogeneity in effect sizes by plotting each study's weighted average PCC along with their 95% confidence intervals in figure 1.7 and doing the same with the elasticities in figure 1.8.

Figure 1.7 Forest Plot, Partial Correlations

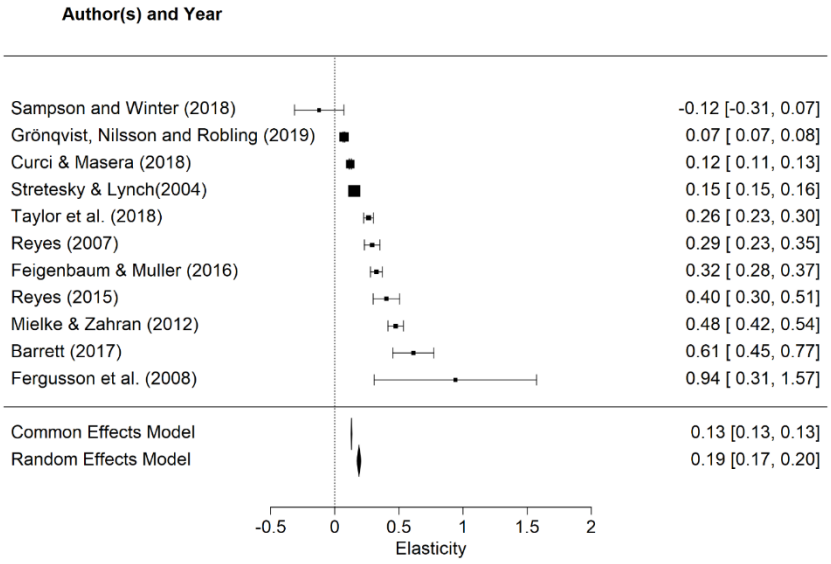


Notes. Chart shows weighted average partial correlation coefficients (PCCs) of each study's effect size along with corresponding 95% confidence intervals. The weighted averages are calculated by first normalizing the PCCs so that confidence intervals can be constructed, then the fixed effects average is calculated, finally the estimates are converted back to PCCs (see appendix B for details). Bottom of table shows common effects and random effects estimates for all studies combined (see appendix C for details). Numbers on right are the point estimates and the 95% confidence intervals.

We show the common and random effects estimates at the bottom of each figure. Both estimates are weighted averages, where more precise estimates get more weight. However, the random effects estimate will give more equal weight to each

study the larger the estimate of between-study heterogeneity. See appendix C for more details on the calculations. The PCC common effects point estimate is 0.01 and the random effects 0.17, while the elasticity common effects estimate is 0.13, and the random effects is 0.19. The difference between the common and random effects estimates indicates that between-study heterogeneity is important, as the lower the estimated heterogeneity between studies, the closer the random effects estimate will be to the common effects.

Figure 1.8 Forest Plot, Elasticities



Notes. Chart shows weighted average of each study's effect sizes converted to elasticities along with corresponding 95% confidence intervals. Bottom of table shows common effects and random effects estimates for all studies combined (see appendix C for details). Numbers on right are the point estimates and the 95% confidence intervals.

It is unlikely that the only source of this heterogeneity is the random, unobservable variances σ_{ij}^2 and τ^2. Distribution (3) can be extended to be conditional on a $1 \times K$ vector of variables $\boldsymbol{x}_{ij}$. In this case the study specific estimates θ_j are a function of this variation in $\boldsymbol{x}$ and we have the conditional distribution:

$$\hat{\theta}_{ij} \mid \sigma_{ij}^2, \tau^2, \boldsymbol{x_{ij}}, \boldsymbol{\beta} \sim N\left(\boldsymbol{x}_{ij}'\boldsymbol{\beta},\ \sigma_{ij}^2 + \tau^2\right) \forall\ i \text{ and } \forall\ j \tag{4}$$

If these variables are observable, we can include them in our estimation. To investigate sources of observable between-study heterogeneity, table 1.3 splits the data into further sub-samples, based on common characteristics. These

characteristics are also used as covariates in the meta-regression analysis and described fully in section 1.4.4. We then compare three measures of between-study heterogeneity for each sample, $\hat{\tau}^2$, $\hat{I}^2$, and $\widehat{H}^2$. For each of these measures, the higher they are, the higher the estimated between-study heterogeneity.

$\hat{\tau}^2$ is an estimate of the variance of the effect size distribution in (3) using the DerSimonian-Laird (1986) method. It is measured in the same units as the effect sizes, which is either PCCs or elasticities in our analysis. The larger is $\hat{\tau}^2$ then the greater the dispersion of the "true" effect sizes each study is attempting to estimate.

$\hat{I}^2$ is an estimate of the proportion of observed variance between effect sizes that is due to effect size heterogeneity, as opposed to sampling variation. It is a figure between 0% and 100%. If 100%, it means all the observed variation is due to between-study effect size heterogeneity. If 0% it means the effect being estimated is homogeneous between studies, and all observed variation is due to sampling error.

$\widehat{H}^2$ is more complicated to interpret. It is the residual standard deviation from regressing the t-statistic of each effect size on its precision. $\widehat{H}^2$ of 1 means that all studies are estimating the exact same effect. The larger $\widehat{H}^2$ is, the greater the between-study effect size variation.

$\widehat{H}^2$ and $\hat{I}^2$ are sensitive to the number of estimates and the variation in the standard error of those estimates. $\hat{I}^2$ tends to 100 as the number of estimates included increases. $\hat{\tau}^2$ is less sensitive to the number of studies used in the analysis compared to $\hat{I}^2$ and $\widehat{H}^2$, but it does not give a sense of how important between-study heterogeneity is compared to within-study sampling variation.

Table 1.3 Random Effects and Heterogeneity Estimates by Subsample

Sample	RE Estimate	SE	$\hat{\tau}^2$	$\hat{I}^2$	$\hat{H}^2$	Studies	Estimates (N)
Full Sample	0.166	0.002	0.002	99	108	24	542
Addressing Endogeneity	0.014	0.001	0.000	90	10	7	220
Correlational	0.505	0.014	0.059	99	159	20	322
Individual-level	0.008	0.001	0.000	95	20	11	125
Area-level	0.388	0.010	0.033	99	123	13	417
Homicide	0.172	0.012	0.010	94	18	8	103
Violent Crime	0.261	0.008	0.016	99	72	18	339
Non-Violent Crime	0.492	0.040	0.120	99	145	15	82
Total Crime	0.077	0.003	0.001	99	152	11	119
North America	0.217	0.006	0.011	98	58	19	386
Europe	0.069	0.003	0.001	100	201	2	85
Direct Lead Measure = TRUE	0.092	0.026	0.031	95	19	9	54
Direct Lead Measure = FALSE	0.171	0.002	0.002	99	118	15	488
Representative Estimate = TRUE	0.186	0.020	0.006	98	54	24	24
Representative Estimate = FALSE	0.167	0.002	0.002	99	111	24	518
Control Gender = TRUE	0.007	0.001	0.000	95	20	8	103
Control Gender = FALSE	0.355	0.007	0.017	99	123	18	439
Control Race = TRUE	0.084	0.008	0.005	97	29	13	114
Control Race = FALSE	0.190	0.003	0.002	99	128	14	428
Control Income = TRUE	0.028	0.002	0.000	97	31	13	174
Control Income = FALSE	0.399	0.008	0.016	99	139	16	368
Control Education = TRUE	0.006	0.001	0.000	95	19	11	106
Control Education = FALSE	0.345	0.007	0.015	99	124	17	436
Elasticity Sample*	0.189	0.008	0.010	91	12	11	312
Elasticity Sample (Addressing Endogeneity)*	0.198	0.012	0.016	88	8	5	211

Notes. RE Estimate is a random effects, meta-analysis estimate computed using DerSimonian-Laird (1986) method. All in PCCs except for the elasticity sample. SE is the standard error of the RE estimate. τ^2, $\hat{I}^2$, and $\widehat{H}^2$ are estimates of between-study heterogeneity. See section 1.4.2 for more details. *values are in elasticities, not PCCs.

Looking at table 1.3 we can see which variables seem important for heterogeneity and the different estimated average effect sizes. The subsample of studies which control for endogeneity has a lower estimated heterogeneity and a smaller effect size compared to the correlational sample. Endogeneity can arise from unobserved variables correlated with both crime and lead. These could bias upwards the estimate of the effect of lead on crime. We cannot rule out that these variables may cause individuals both to commit more crime and be more exposed to lead, rather than lead being the cause. Therefore, the difference between the "addressing

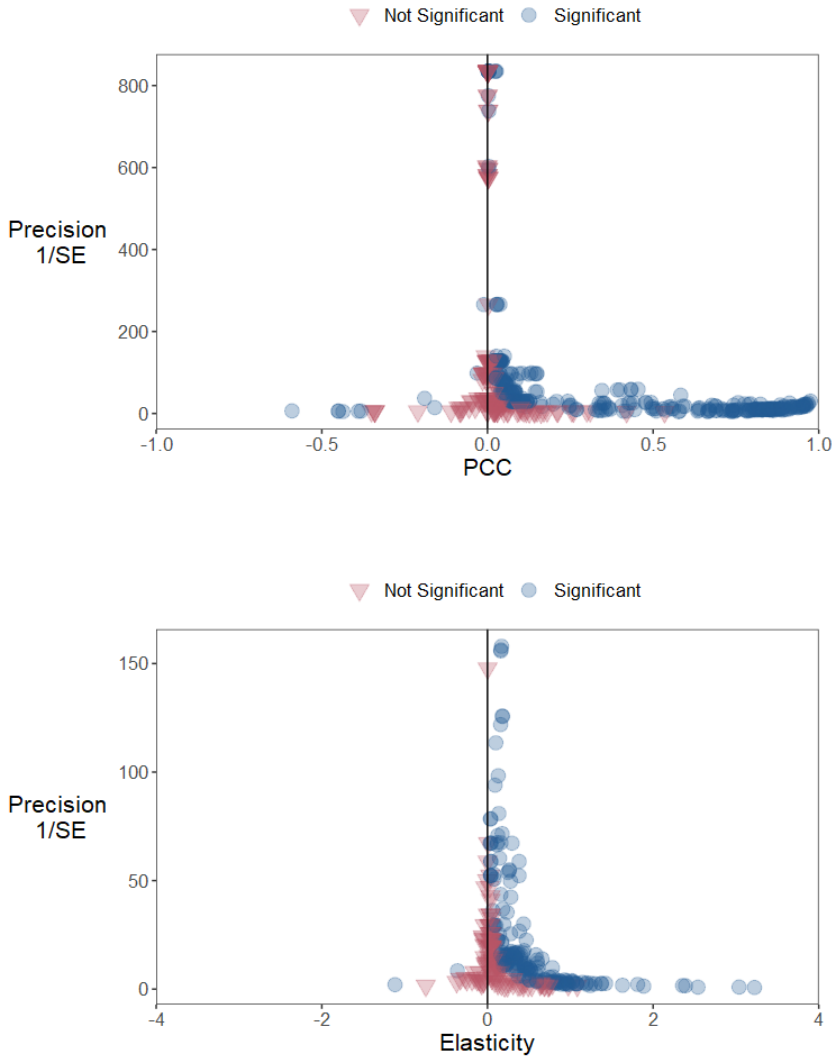
endogeneity" sample and the full sample could be related to these factors. The elasticity subsample also shows lower heterogeneity than the full sample.

Studies that look at individual-level data on the propensity to commit crime have lower estimated heterogeneity and estimated effect size compared to studies that look at crime committed within a geographic area. Studies which use homicide as the dependent variable appear to have less heterogeneity and find a smaller effect size. This reduction in heterogeneity may be due to lower measurement error in homicide data compared to other types of crime, combined with more similar classification of this crime across countries, and therefore less noise in the data. Finally, when race, gender, education, and income covariates are included in an estimation, these tend to lower the effect size. These subsamples also show less between-study heterogeneity than those which do not include these covariates. The estimated differences in effect size and heterogeneity between subsamples indicates observable variation is important and must be considered when we estimate an average effect. We incorporate the observable variation indicated in table 1.3 into our meta-regression analysis in section 1.4.4.
A further, and common source of heterogeneity in effect sizes in meta regression analysis comes from publication bias. We investigate this in the next section.

1.4.3 Publication bias

Publication bias is a well-known problem across disciplines (see for example: DeLong and Lang, 1992; Ioannidis, 2005; Ioannidis, Stanley and Doucouliagos, 2014; and Ferraro and Shukla, 2020). Papers which contain statistically significant effect sizes are more likely to be published than those which show no effects, or those which contain counter-intuitive results (also known as the bottom-drawer problem). It is standard practice to test for the presence of publication bias in meta-analysis.

Figure 1.9 Funnel Charts



Notes. PCC = Partial Correlation Coefficient. Precision is one divided by the standard error. "Significant" means statistically significant at the 95% confidence level using two-sided critical values of a normal distribution.

The first and most common step is to simply chart the data and visually inspect for bias, using a funnel plot. Figure 1.9 plots effect sizes against their precision. The upper funnel shows the PCCs for the full sample, and the lower the elasticities for that subsample. A funnel with no bias should be symmetrical around central tendencies. The estimates will tend to spread out as the precision decreases, but they should do so symmetrically if this is only due to sampling noise. Figure 1.9 shows a pronounced asymmetry in the estimates, suggesting there may be a positive bias. There appears to be less asymmetry in the elasticity panel. This suggests these studies may be more similar, and/or have less bias. Some of the studies with the largest effect sizes did not report enough information for

elasticities to be calculated, which may be the reason for this. Although there is asymmetry in both panels, suggesting publication bias, it is also possible this is due to heterogeneity within the sample. We explore this possibility in section 1.4.4.

More formal testing of publication bias is also possible. There are many tests for publication bias. We use seven methods, which we split into linear and non-linear methods. Linear tests involve regressions of a measure of sampling uncertainty on the estimated effect. A linear relationship between the estimate and its standard error, as figure 1.9 implies, would indicate the presence of publication bias (see appendix D). This naturally leads to the estimating equation (5).

$$(5) \qquad \hat{\theta}_{ij} = \theta + \beta_F \hat{\sigma}_{ij} + u_j + \epsilon_{ij}; \text{ where } \epsilon_{ij} \sim N\left(0, \sigma_{ij}^2\right) \text{ and } u_j \sim N(0, \tau^2)$$

This is the combined Funnel Asymmetry Test (FAT) and Precision Effect Test (PET). Here the FAT is β_F, and is an estimate of the size and sign of publication bias. It is a function of the inverse Mills' ratio. If positive then estimates that are positive are more likely to be published than negative ones. This test also gives an estimate of θ that takes into account this bias, called the PET. Equation (5) nests the common effects model where τ^2 is zero.

The test in (5) would be subject to heteroskedasticity, as can be observed from figure 1.9. We have estimates of the heteroskedasticity in $\hat{\sigma}_{ij}$. These can therefore be used to weight the regression and we estimate the test with weighted least squares following Stanley (2008).

$$(6) \qquad \hat{t}_{ij} = \theta \frac{1}{\hat{\sigma}_{ij}} + \beta_F + v_j + e_{ij}$$

Here the dependent variable $\hat{t}_{ij}$ is now the t-ratio, rather than the estimate alone. The intercept of the regression is the FAT and the coefficient on $\frac{1}{\hat{\sigma}_{ij}}$ is the PET.

We estimate four variations of linear publication bias tests. First with OLS and clustered standard errors by study, but no study fixed effects; second, a variation of this where we regress on the variance rather than the standard error (Stanley and Doucouligas, 2014); third a full hierarchical FAT-PET with study fixed effects.

We estimate this with restricted maximum likelihood (REML), as Monte Carlo simulations suggest REML performs well for unbalanced panels (Baltagi, Song and Jung, 2000). Finally, we use the square root of the sample size as an instrumental variable for the precision. This last method allows for the fact some estimation techniques may be less efficient but lead to unbiased estimates.

We also run three non-linear methods. The Weighted Average of Adequately Powered Estimates (WAAP) of Stanley, Doucouliagos, and Ioannidis (2017) estimates which studies are post-hoc "adequately powered" and only uses these to calculate an average effect size. The Trim and Fill (TF) method (Duval & Tweedie, 2000) adds imputed studies on the sparse side of the funnel before calculating an average effect. The Andrews and Kasy (2019) method reweights all observations by estimated relative publication probabilities and calculates an average effect size after reweighting. See appendix D for a full discussion of all methods.

Table 1.4 shows the results of all tests. We estimate the tests with four different samples. Panel A is the full sample using PCCs, panel B is all studies which address endogeneity (PCCs), Panel C is only studies with elasticity estimates available, and panel D is studies which both address endogeneity and have elasticities. Linear methods allow for not only an effect beyond bias estimate but an indication of the strength of bias in the FAT coefficient. In all four panels every estimate of publication bias is positive, indicating positive estimates are more likely to be published. Only the FAT-PEESE estimate in panel B, and the IV estimate in panel C, have 95% intervals that cover zero. In every panel, the effect beyond bias estimates are all smaller than the random effects estimate of table 1.3, indicating the effect size is overstated due to publication bias.

Table 1.4 Effect Beyond Bias and Publication Bias Estimates

	FAT-PET	FAT-PEESE	Multi-level FP	IV	WAAP	TF	AK
Panel A – Full Sample, PCCs							
Effect Beyond Bias	-0.003	0.005	0.006	-0.004	0.005	0.008	-0.773
	(0.002)	(0.002)	(0.004)	(0.002)	(0.002)	(0.018)	(0.438)
Publication bias	5.026	32.227	3.502	5.062	.	.	.
	(1.283)	(8.638)	(0.885)	(1.297)			
Groups	24	24	24	24	.	.	24
Observations	542	542	542	542	362	542	542
Panel B – Only Addressing Endogeneity Sample, PCCs							
Effect Beyond Bias	0.001	0.004	0.001	0.001	0.003	0.007	0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.002)	(0.002)
Publication bias	2.159	11.305	1.982	2.159	.	.	.
	(0.431)	(10.186)	(0.434)	(0.430)			
Groups	7	7	7	7	.	.	7
Observations	220	220	220	220	55	220	220
Panel C – Only Elasticity Sample*							
Effect Beyond Bias	0.110	0.128	0.107	-0.056	0.126	0.145	0.025
	(0.029)	(0.021)	(0.010)	(0.087)	(0.022)	(0.018)	(0.069)
Publication bias	1.202	3.355	1.966	4.579	.	.	.
	(0.545)	(0.805)	(0.681)	(2.935)			
Groups	11	11	11	11	.	.	11
Observations	312	312	312	312	122	312	312
Panel D –Only Elasticity and Addressing Endogeneity Sample*							
Effect Beyond Bias	0.040	0.084	0.084	0.013	0.116	0.081	0.018
	(0.007)	(0.019)	(0.016)	(0.014)	(0.028)	(0.015)	(0.021)
Publication bias	1.801	4.371	1.392	2.186	.	.	.
	(0.440)	(1.127)	(0.619)	(0.514)			
Groups	5	5	5	5	.	.	5
Observations	211	211	211	211	70	211	211

Notes. *Indicates effects are elasticities rather than PCCs. Estimates are presented with their standard errors in brackets. FAT-PET is Funnel Asymmetry test and Precision Effect Test (Stanley and Doucouliagos, 2014). FAT-PEESE is Funnel Asymmetry Test and Precision Effect Estimate with Standard Error. The multi-level FP is a FAT-PET multi-level model with fixed effects for each study. IV is a FAT-PET regression with square root of sample size used as an instrumental variable for the precision using two stage least squares. WAAP (Stanley, Doucouliagos, & Ioannidis, 2017) is the Weighted Average of Adequately Powered Estimates, where studies below a certain estimated power are removed before calculating the effect. TF is Trim and fill (Duval & Tweedie, 2000), which removes outlier studies and then adds imputed studies before calculation an average effect. AK is the Andrews-Kasy method (Andrews & Kasy, 2019), which is a step function selection model which reweights the observed sample with estimated publication probabilities. See Appendix D for full explanation of each method.

The estimates for the full sample and addressing endogeneity sample (panels A and B) are all close to zero, save the full sample Andrews-Kasy estimate which is -0.77. However, the 95% confidence interval covers zero, and this estimate is the outlier. For the elasticity sample (panel C) they vary from 0.15 to -0.06, but most estimates are around 0.11. For elasticity estimates that address endogeneity, the estimates range from 0.01-0.08. As a robustness check, we also estimate all methods using only representative estimates in appendix E and the results are similar.

All tests suggest publication bias is present in the sample. This should not be a surprise as Stanley and Doucouliagos (2013) show that bodies of literature with theoretically implausible signs or sizes tend to exhibit more publication bias. It is, of course, theoretically implausible that an increase in lead pollution would cause a decrease in crime, and therefore it may be researchers do not write up papers showing such findings. Nevertheless, we should expect negative estimates due to sampling noise. This may explain the finding of publication bias in all tests and the asymmetry in the funnel plots.

The tests also suggest the true mean effect size of lead on crime may be close to zero, but this could be due to the relatively small sample, or to characteristics of the studies. These characteristics can be investigated more thoroughly with meta-regression analysis.

1.4.4 Meta-Regression Analysis

Meta-regression analysis (MRA) follows from (4) where we include common observable variation in our estimation. Given all tests suggest the presence of publication bias we include the FAT in all regressions. We also weight all regression covariates by the standard errors as in (6). Therefore, the specification is the same as in (6) except we now also regress on a vector of observable covariates, $\boldsymbol{x_{ij}}$, weighted by the standard errors of the estimate. This includes the precision, and the coefficient on the precision is now only an estimate of the average effect size when all other covariates are set to zero. The meta-regression is shown in (7).

$$(7) \qquad \hat{t}_{ij} = \beta_F + \boldsymbol{z}'_{\boldsymbol{ij}}\boldsymbol{\beta} + v_j + e_{ij}$$

Where $\boldsymbol{z_{ij}}$ is a $1 \times K$ vector of weighted observable covariates.

The covariates included are based on common characteristics of the studies that are suggested by the literature. Their descriptive statistics are included in table 1.5. The majority are dummy variables indicating whether that characteristic is present for that estimate. All variables are coded at estimate level, not at study level. That is, different estimates from the same study may have different characteristics, and therefore have different values for the covariates. There is a dummy variable that equals one when an estimate comes from a quasi-experimental study design that attempts to deal with endogeneity concerns. There is a dummy variable which is one when an estimate is of crime in an area, and zero when it is at the individual level. There are four dummy variables which indicate whether specific controls were included in the estimation. Lead exposure is correlated with poverty (Baghurst e al. 1999) and race (Sampson and Winter, 2016), may have different effects on men and women (Denno, 1990), and may have a relationship with educational outcomes (Fergusson, Boden and Horwood, 2008). Therefore, when an estimation includes these variables we might expect it to influence the estimate. The interpretation of the effect of these variables depends on where they are in the causal chain. If these variables are confounders, causing changes in lead and changes in crime, then omitting them will tend to overstate the effect of lead on crime (given they change both in same direction). If they are mediators, changed by lead and then changing crime, then conditioning on them can lead to understating the effect of lead on crime. This is especially important when study designs do not use some method to deal with endogeneity issues. Of course, there are other variables that may be important controls, but these were not found to be common enough across studies to include.

Next there are three dummy variable that describe what type of crime was used as the dependent variable (homicide, violent, and non-violent), with a reference group of total crime. This allows us to test whether the different mechanisms proposed in section 1.2 matter. The violent crime category nests homicide within it. They are separate categories because homicide data is thought to be the best quality crime data, and thus less likely to suffer from bias (Fox and Zatz, 2000). We next have two dummy variables representing possible estimation effects. One for if simple OLS was

used, another for if maximum likelihood was used. The reference group is any other estimation such as GMM or mean differences. We have two dummy variables for further estimation effects. One for if panel data were used, and another for if the results are reported as odds ratios.

Table 1.5 Descriptive Statistics of Covariates used in the Meta-Regression Analysis

Variable	Mean	Median	Standard Deviation
Control_gender	0.19	0	0.39
Control_race	0.21	0	0.41
Control_income	0.32	0	0.47
Control_education	0.20	0	0.40
Homicide	0.19	0	0.39
Violent	0.63	1	0.48
Non_Violent	0.15	0	0.36
Both	0.22	0	0.41
Area	0.77	1	0.42
OLS	0.39	0	0.49
ML	0.13	0	0.34
Odds_Ratio	0.03	0	0.17
Panel	0.67	1	0.47
Addressing Endogeneity	0.41	0	0.49
North America	0.71	1	0.45
Europe	0.16	0	0.36
Direct Lead Measure	0.10	0	0.30
Publication Year*	2013	2015	6
Number of Covariates*†	445	13	802
Sample Size*	64478	901	186,709

*Notes:**Indicates variables have been standardised.

†Includes fixed effects for degrees of freedom adjustment

A further two dummy variables are geographic dummies that equal one when an estimate come from a either North America or Europe, with the rest of the world as the reference group. 70% of estimates use data from North America. The final dummy variable equals one when a direct measure of lead, from either blood, bone, or dentine samples, is used in the estimation and zero when a proxy measure or estimate, such as leaded gasoline use in an area, is used. This allows us to test

whether there is a systematic difference in effect sizes found when lead levels are taken directly from subjects, which we might expect to give a more accurate measure of the true effect, rather than proxied. The final three covariates are the publication year, sample size, and the number of covariates included in the estimation. These variables have been standardised to aid the restricted maximum likelihood convergence.

We estimate many specifications due to model uncertainty. Our sample is relatively small and coefficient estimation varies significantly in alternative specifications. The number of different covariate combinations is 2^K where K is the total number of covariates. It is common in the meta-analysis literature to employ some method of model averaging or shrinkage to deal with model uncertainty. However, with this many covariates and modern computational power it is possible to estimate all 2^K specifications[5]. In addition, table 1.3 showed that some subsamples have substantially less heterogeneity than the full sample. It may be that these sub-samples suit aggregation better than the full sample. For example, we might expect studies with individuals as the unit of analysis to share much more common information than those that have a geographic area as the unit of interest. We therefore also estimate all covariate specifications for these subsamples. It is not possible to estimate every combination as some dummy variables no longer have any variation in the subsamples, leading to collinearity. This can also lead to other variables being excluded as they become the new base case (for example if there are no studies from outside Europe or North America in a subsample, then Europe becomes the base case). A full list of the covariates included for each subsample is in table 1.6. We estimate every possible combination of covariates for the full sample and the subsamples. We include the FAT, the estimate of publication bias. We estimate with REML and include study fixed effects.

We do not interpret the coefficients on the covariates following best practice (see Westreich and Greenland, 2013 and Stevenson and Elwert, 2020), as they are not identified. Instead, we use the information from each meta-regression specification to construct a distribution of estimates of the average effect of lead on crime. We are

[5] As a robustness check we perform Bayesian Model Averaging in appendix F. The posterior mean PCC using the full sample and evaluated at the sample averages is 0.09, lower than the method we use here. The elasticity posterior mean is also lower at 0.07.

now estimating an average effect conditional on the observable heterogeneity in our specifications. In practice, meta-analysis tends to do this in two ways, either by using the sample averages or by taking some "ideal" specification. We do both. That is, for each specification we generate a predicted estimate of the effect of lead on crime, using both the sample averages, or by using an ideal specification, and not including the FAT in the predicted value (i.e. removing the publication bias).

Table 1.6 Variables Used in Combinations For Each Sample Estimation

Sample	**Variables Used**
Full Sample	Control gender, Control race, Control income, Control education, Homicide, Violent, Non-Violent, Area dummy, OLS, ML, Odds Ratio, Panel dummy, Addressing Endogeneity, North America, Europe, Direct Lead Measure, Publication Year, Covariates, Sample Size
Addressing Endogeneity Sample	Control gender, Control race, Control income, Homicide, Violent, Non-Violent, Area dummy, OLS, Panel dummy, Publication Year, Covariates, Sample Size
Correlational Sample	Control gender, Control race, Control income, Control education, Homicide, Violent, Non-Violent, Area dummy, OLS, ML, Odds Ratio, Panel dummy, North America, Direct Lead Measure, Publication Year, Covariates, Sample Size
Area-level Sample	Control race, Control income, Control education, Homicide, Violent, Non-Violent, OLS, ML, Panel dummy, Addressing Endogeneity, Direct Lead Measure, Publication Year, Covariates, Sample Size
Individual-level Sample	Control gender, Control race, Control income, Control education, Violent, Non-Violent, OLS, ML, Odds Ratio, Panel dummy, Addressing Endogeneity, Direct Lead Measure, Publication Year, Covariates, Sample Size
Homicide Sample	Control race, Control income, OLS, Panel dummy, Addressing Endogeneity, Publication Year, Covariates, Sample Size
Violent Crime Sample	Control gender, Control race, Control income, Control education, Area dummy, OLS, ML, Panel dummy, Addressing Endogeneity, North America, Direct Lead Measure, Publication Year, Covariates, Sample Size
Non-Violent Crime Sample	Control gender, Control race, Control income, Control education, Area dummy, OLS, ML, Odds ratio, Panel dummy, Addressing Endogeneity, North America, Direct Lead Measure, Publication Year, Covariates, Sample Size
Elasticity Sample	Control gender, Control race, Control income, Control education, Homicide, Violent, Non-Violent, Area dummy, OLS, ML, Panel dummy, Addressing Endogeneity, North America, Direct Lead Measure, Publication Year, Covariates, Sample Size
Elasticity and Addressing Endogeneity Sample	Control gender, Control race, Control income, Homicide, Violent, Non-Violent, Area dummy, Publication Year, Covariates, Sample Size

Notes. Table shows which covariates were included for each sub-sample estimation. Inclusion depended on whether there was variation in the covariate for that subsample.

The ideal specification we use is one that includes controls for race, education, income and gender, that uses individual data, directly measured lead levels, controls for endogeneity, uses panel data, is estimated without just using simple OLS or ML, uses total crime as the dependent variable, uses North American data (as most of

our sample is from there), and uses the sample averages for the publication year, sample size, and number of covariates. This ideal specification is chosen to represent a robust and high-quality estimation, and as such we would expect it be generally lower than the sample averages estimates.

The means, medians, and standard deviations of the full sample and subsample estimates are presented in table 1.7. The top panel shows the estimates effect sizes evaluated at the sample averages, while the bottom shows effect sizes evaluated at the "ideal" specification. The table also shows the number of specifications for each sample. The final column shows the how many of the estimates fell outside of the feasible interval of the PCC [-1,1]. This indicates whether there may be a misspecification issue with that particular sample estimation.

The distribution of coefficient sizes for the full sample estimation is in plotted in figure 1.10, panel A. The left figure shows effect sizes evaluated at the sample averages, while the right shows effect sizes evaluated at the "ideal" specification In each there is a distribution of 524,288 estimated effect sizes. The mean and median PCC for the sample averages distribution are 0.16 and 0.18 respectively, which is "moderately positive" according to the Doucouliagos (2011) taxonomy. The distribution appears to be bimodal with one peak close to zero and the other around 0.2. The distribution of the ideal specification is not bimodal and is roughly symmetrical. The mean and median are 0.13 and 0.09 respectively. As expected the ideal specification is lower than the sample averages.

We next restrict the sample to only the studies that estimate a causal effect with quasi-experimental methods rather than an association: our "addressing endogeneity" sub-sample. This consists of seven studies and 220 estimates. It is common in meta-analysis to exclude correlational studies altogether (e.g., Kraft, Blazer and Hogan, 2018). Although we have not excluded those studies in this meta-analysis, we now examine what a meta-analysis estimate with only causal studies would be. We saw in table 1.3 that the addressing endogeneity subsample has lower between-study heterogeneity than the full sample, so aggregation may yield comparatively more information.

We plot the sub-sample average specification and ideal specification in figure X, panel B (excluding those variables that cannot be included in the estimation, see table 1.6). The distribution of the sample average predicated values is tight around zero with a mean and median of 0.01, and a sample standard deviation of 0.01. The "ideal" specification also has a mean and median of 0.01. The results suggest there is a systematic difference between the "addressing endogeneity" studies and the rest of the sample.

In figure 1.11, we carry out the same exercise except only for those studies that have elasticity estimates available. The elasticity effect sizes in panel A, figure 1.11 are for the full elasticity sample. The mean and median effect size, evaluated at the sample averages, are both an elasticity of 0.09. Evaluated at the "ideal" specification they are 0.09 and 0.05 respectively. The "ideal" distribution shows more heterogeneity and is bimodal. The standard deviation is 0.2, much higher than the 0.03 evaluated at the sample averages.

Panel B of figure 1.10, the addressing endogeneity and elasticity subsample, is very similar to panel A. The mean is 0.10 and the median elasticity is 0.09 when evaluated at the sample averages. When evaluated at the "ideal" specification the addressing endogeneity, elasticity sample mean is 0.17 and the median is 0.16 These are considerably larger than when evaluated at the sample averages, or when looking at the "ideal" specification for the full elasticity sample. In both panel A and B the "ideal" specification distributions have a larger variance than the sample average distributions. 40% of the "ideal" specifications yield a negative elasticity when using the full elasticity sample, and 15% are negative when using the addressing endogeneity, elasticity sample. In contrast almost no estimates are negative when evaluated at the sample averages in panel A or B. This suggests the "ideal" specification is much more sensitive to model changes than when we evaluate at the sample averages.

Table 1.7 Meta-Analysis Average Estimates for The Full Sample and Each Subsample

Sample averages

Sample	Mean	Median	SD	N	% < −1 *or* > 1
Full Sample	0.16	0.18	0.07	524288	0%
Addressing Endogeneity Sample	0.01	0.01	0.01	4096	0%
Correlational Sample	0.29	0.29	0.10	131072	0%
Area-level Sample	0.25	0.26	0.06	16384	0%
Individual-level Sample	0.03	0.03	0.01	65536	0%
Homicide Sample	0.58	0.54	0.22	256	0%
Violent Crime Sample	0.39	0.39	0.22	16384	0%
Non-violent Crime Sample	0.75	0.71	0.24	32768	14%
Elasticity Sample*	0.09	0.09	0.03	131072	.
Elasticity and Addressing Endogeneity Sample*	0.10	0.09	0.04	1024	.

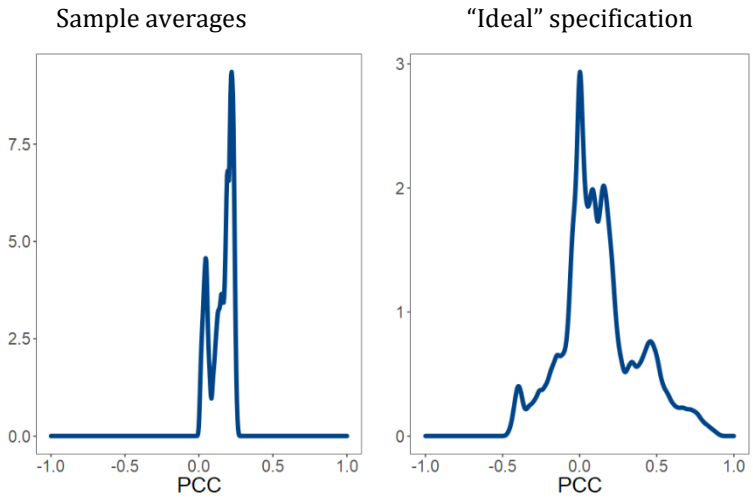
"Ideal" specification

Sample	Mean	Median	SD	N	% < −1 *or* > 1
Full Sample	0.13	0.09	0.25	524288	0%
Addressing Endogeneity Sample	0.01	0.01	0.02	4096	0%
Correlational Sample	0.49	0.37	0.6	131072	15%
Area-level Sample	0.23	0.20	0.22	16384	0%
Individual-level Sample	0.02	0.02	0.04	65536	0%
Homicide Sample	0.28	0.27	0.17	256	0%
Violent Crime Sample	0.57	0.14	1.29	16384	36%
Non-violent Crime Sample	1.26	0.58	3.50	32768	64%
Elasticity Sample*	0.09	0.05	0.20	131072	.
Elasticity and Addressing Endogeneity Sample*	0.17	0.16	0.16	1024	.

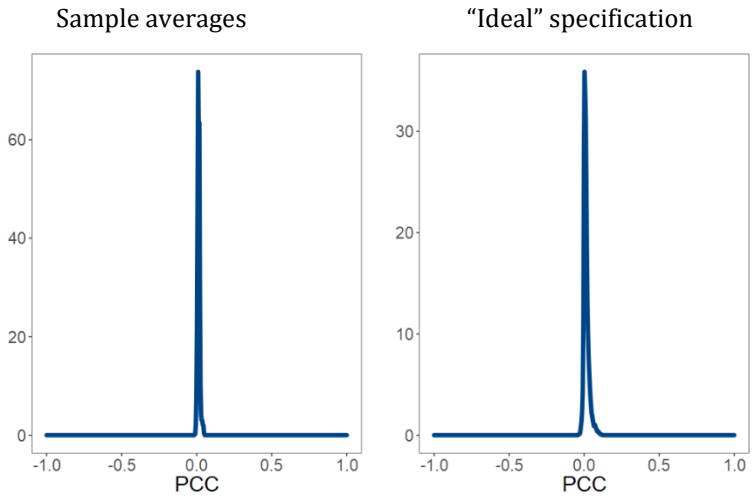
Notes. *Indicates values are elasticities rather than PCCs. Table shows results from combining multiple meta-regression estimates, each using different specifications. All regressions carried out by restricted maximum likelihood. This is done for the full sample and subsamples. N is the number of regressions carried out, each a different specification. The mean and median are the summary statistics of the average effect size from these regressions, given in Partial Correlation Coefficients (PCCs) or elasticities. PCCs are bounded between -1 and 1. The last column gives the percent of effects which fall outside this range.

Figure 1.10 Density of Meta-Analysis Average Effect Size Estimates from Full Sample

Panel A – Full Sample (PCCs)



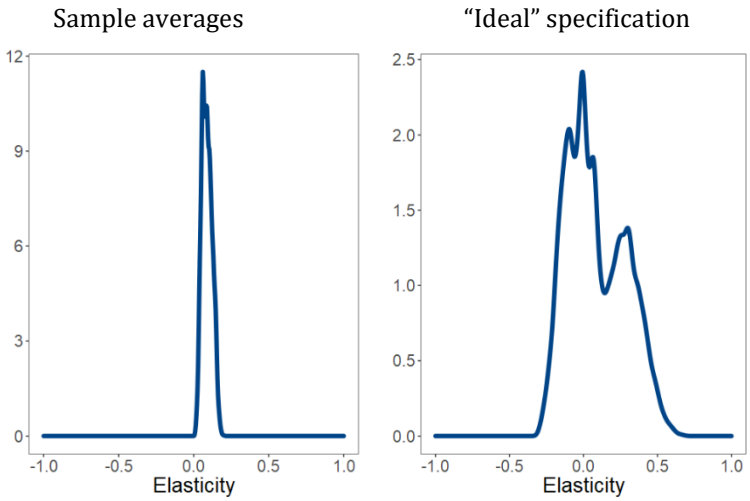
Panel B – Addressing Endogeneity Sample (PCCs)



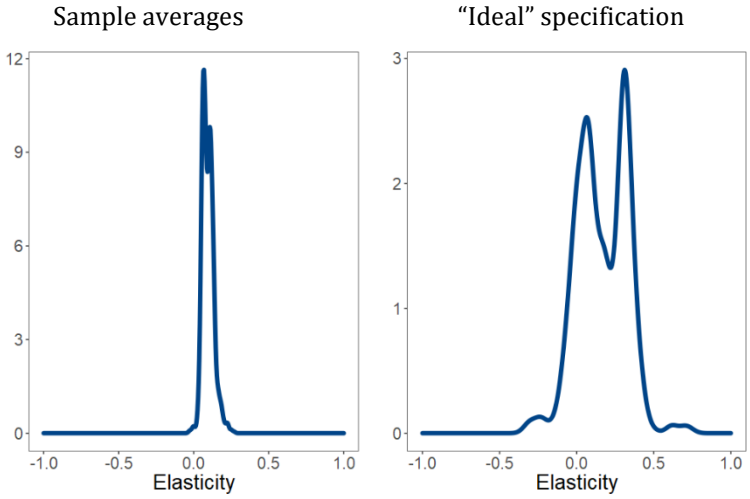
Notes. Chart shows densities for the distribution of meta-regression estimated average effect sizes. Chart on left shows estimated average effect for each specification evaluated at the sample averages. Chart on right shows estimated average effect for each specification evaluated at an "ideal" specification. X axis truncated at feasible interval of a PCC, [-1,1].

Figure 1.11 Density of Meta-Analysis Average Effect Estimates for Elasticity Subsample

Panel A – Full Elasticity Sample



Panel B –Elasticity and Addressing Endogeneity Sample



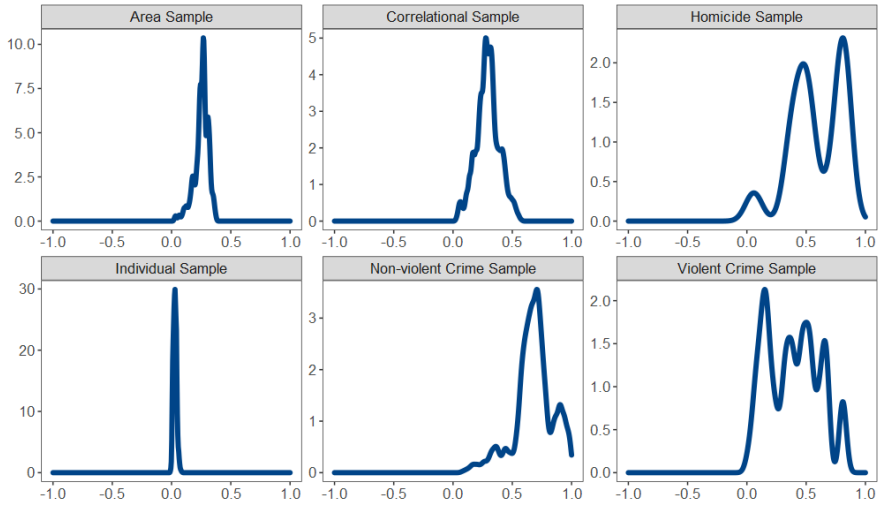
Notes. Chart shows densities for the distribution of meta-regression estimated average effect sizes for the addressing elasticity sub-sample.. Chart on left shows estimated average effect for each specification evaluated at the sample averages. Chart on right shows estimated average effect for each specification evaluated-at an "ideal" specification.

We next plot several other subsample distributions of interest in figure 1.12. The difference between the area and individual sample is striking. The area sample means and medians are much larger than the individual sample for both the sample average specification and the ideal specification. The individual sample mean and median PCCs are small and the distributions are tight around the means compared to the area sample. This suggests that covariates matter less for the individual sample effect sizes compared to the area sample. Similar to the area-individual comparison, the correlational sample has much higher means and medians than the addressing endogeneity sample.

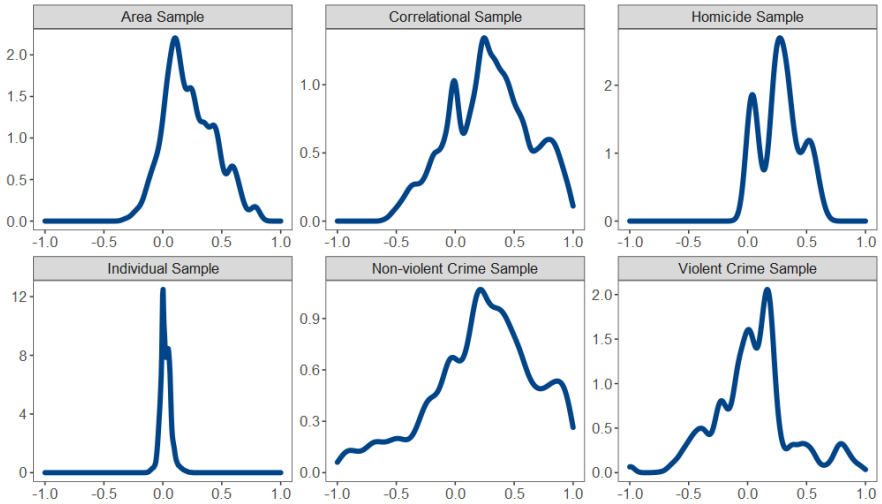
Comparing homicide, violent, and non-violent crime samples we can see they all have large mean and median PCCs, but the non-violent and violent subsamples have a portion of the distribution outside [-1,1], suggesting misspecification and that the results may not be reliable. The standard deviations for these tend to be much larger as well. Furthermore, due to the lack of homicide estimates, only 256 specifications could be run without convergence issues. Overall, the results suggest that lead affects all types of crime, but we cannot say if it has a bigger effect on some types than others. We cautiously suggest that if lead does have an effect on crime it is across all categories of crime.

Figure 1.12 Densities of Meta-Analysis Average Effect Estimates From Subsamples

Sample averages



"Ideal" specification



Notes. Chart shows densities for the meta-regression estimated average effect sizes for a number of subsamples. Top chart shows estimated average effect for each specification evaluated at the sample average for each subsample. Bottom chart shows estimated average effect for each specification evaluated at an "ideal" specification. X axes truncated at feasible interval of a PCC, [-1,1]

1.4.5 Explaining the 20th Century Crime Decline

Our calculated elasticities allow us to estimate how much of the fall in crime observed in the second half of the 20th century was caused by lead. We first use the dramatic fall in homicide in the US as an example. The median blood lead level in children in the US fell 88% from 1976-2009. Given our main elasticity estimate of 0.09 for the full elasticity sample evaluated at the sample averages(with standard deviation of 0.03), this implies a fall in homicide of 5-11% with the point estimate being 8%. The US homicide rate fell 54% from its peak in 1989 to 2014. This would mean that lead accounts for 8 of those percentage points, i.e. around 15% of the decrease in homicide. If we use the full range of elasticity mean estimates in table 1.7 we have values from 0.05-0.17. These would imply 4-15 percentage points of the 54% fall were accounted for by lead. This would mean 7-28% of the fall in homicide was due to falling lead levels.

Our estimates imply lead pollution is an important factor in reducing homicides, and lead abatement has saved lives, but it does not account for the majority of the fall. Depending on the specification, we conclude that 93%-73% of the fall in homicide in the US is unaccounted for.

We next carry out estimates of how much of the urban/rural violent crime convergence in figure 1.4 can be explained by the relatively higher blood lead levels in urban areas in the 1970s. Average blood lead levels in under 5s (geometric mean) declined by 15μg/dL in large population MSAs from 1976 to 2011, and by 12.7 in smaller MSAs and rural areas over the same period. The difference in the violent victimization rate per 1,000 people between urban and rural areas, as measured by the NCVS, was 41 in 1993, and 7 in 2015. Given the lag in time between childhood lead exposure to adult criminal acts, we believe these differing periods give enough time for the lead change to take effect. The gap in victimizations declined by 34 per 1,000 people in this period. Using an elasticity of 0.09, we estimate the relative change in blood lead levels

would account for around of these victimizations. That is, the difference in lead levels accounts for 11% of the convergence in the victimization rate. Using the 0.05-0.17 range of elasticity estimates in table 1.7, means that lead accounts for between 2 and 7 of the victimisation gap difference. This would explain 6%-20% of the convergence in violent victimisation rates. While not negligible, this leaves a large part of the convergence in urban/rural crime rates unexplained as well.

1.5 Discussion and Conclusion

Changes to the amount of lead in the environment have been put forward as one of the main causes of the decrease in crime, especially homicide, in many western countries. We performed the first meta-analysis of the effect of lead on crime. We find there is publication bias in the lead-crime literature, and that meta-analysis estimates that do not control for this will overstate the effect of lead on crime. Using meta-regression, taking into account publication bias and between-study heterogeneity, our main estimates are an average effect size of 0.16 as a partial correlation, or 0.09 as an elasticity. When using the larger PCC sample, we find that the average meta-analysis estimate for studies that address endogeneity is much smaller than for the full sample, or for the correlational sample. Similarly, the average effect size estimate for studies that have individuals as the unit of interest is much smaller than for the sample of studies that have a geographic area as the unit of interest. When we examined the differences between lead's effect on homicide, violent and non-violent crime, we could not confidently state there was any difference between them. When using the elasticity sample, the average meta-analysis estimate for studies that address endogeneity tended to be similar to the full elasticity sample, except when evaluated at the "ideal" specification, in which case it was larger.

Finally, we performed calculations to estimate the share of the decline in crime in the US that is accounted for by reductions in blood lead levels. We estimate that of the total 54% fall in homicides observed in the US in 1976-2009, reduced blood lead levels accounted for 4-15 percentage points. A substantial decrease. However, this was only a 7-28% share of the total fall, leaving 93-72% unaccounted for. Similarly, we find that the relative changes in blood lead levels account for 6-20% of the convergence in urban and rural violent crime rates observed in the US.

Overall, the results suggest that declines in lead pollution reduce crime but are not the cause of the majority of the fall in crime observed in many western countries. We are unable to provide estimates on the size of other causes here but hope our results can provide a rough benchmark for relative importance in future meta-analysis. It is possible that the large differences in our samples can be reconciled. For example, the large difference between the individual and area samples may be because crime has fallen at the extensive margin rather than the intensive margin. Tcherni-Buzzeo (2019) observe that around 5% of the population are responsible for 50% of crime, and that the fall in crime in the US is likely due to falls in this high-crime population, rather than less crimes per individual in that population. If less lead pollution only meant less probability of committing crime for this small slice in the population, it might nevertheless lead to a large fall in crime at the area level. A second possibility is that relatively small effects of lead at the individual level can be exacerbated by peer effects from other lead affected individuals[6]. Recent work has found these peer effects can even affect those without elevated blood lead levels (Gazze, *et al*., 2021). In areas with high levels of lead, the individual effects of lead may be compounded by peers also having high levels of lead, leading to a much larger impact at the area level.

[6] We thank an anonymous reviewer for this suggestion.

There are several limitations to our analysis. Most importantly, the sample size is not large. We have 24 studies and 542 estimates, this is not unusual for a meta-analysis but, particularly for our subsample estimates, this could play a part in the differences. It may explain why so much of the distribution for the different types of crime in table 1.7 were outside the feasible PCC interval of [-1,1]. We attempt to mitigate this by using various tests for publication bias, and estimating many different specifications, but we cannot rule out that the results are due to small sample effects. Secondly, the between-study heterogeneity is large in our sample. This calls into question how comparable the studies are. This is to be expected as studies use different concepts and measures of crime and lead, different units of interest, and different estimation techniques. We try to mitigate this by converting to PCCs or elasticities, using different sub-samples that have lower between-study heterogeneity, and using meta-regression with covariates. However, even with these mitigations, it may be that the literature is not comparable and therefore meta-analysis estimates will be noise. In this case it casts doubt on the external validity of the studies examining the lead-crime hypothesis. The solution would be far more studies that estimate elasticities using comparable measures of lead and crime.

For policymakers, our results are a warning against assuming the large crime levels in past decades cannot return now that lead pollution is much lower. The results are not a signal that lead abatement is fruitless. As outlined in section 1.2, the evidence of harmful biological and health changes due to lead is overwhelming. There is no known safe level of lead. Even if outcomes higher up the causal chain, such as crime, are not as affected by lead, the evidence still shows lead abatement will increase health outcomes, especially for the very young.

For future research, we have two main suggestions. The first is that there are enough low sample size, correlational studies in the lead-crime literature. What is needed now is high power, high-quality causal estimates of the effect of lead on crime. The value added of such studies would be increased by testing

the effect on different types of crime, and the possible interaction of lead with other potential causes. The second is that more high-quality causal estimates of the elasticity of other causes of crime are needed. Our results suggest lead is not responsible for the majority of the fall in crime since the 80s and therefore leaves open room for other explanations. These explanations must account for the fact homicide has fallen across many (but not all!) western countries at roughly the same time. They must also account for the fact that total crime has risen in Europe and fallen in the US, while the homicide rate has fallen in both. Further comparison of the relative shares of responsibility for the fall in crime, as well as the interaction between causes, may also be fruitful and we suggest further meta-analyses, using modern methods, would be helpful in this area.

References

Aelion, C.M., Davis, H.T., 2019. Blood lead levels in children in urban and rural areas: Using multilevel modeling to investigate impacts of gender, race, poverty, and the environment. Sci. Total Environ. 694, 133783. https://doi.org/10.1016/J.SCITOTENV.2019.133783

Andrews, I., Kasy, M., 2019. Identification of and Correction for Publication Bias. Am. Econ. Rev. 109, 2766–2794. https://doi.org/10.1257/aer.20180310

Baghurst, P.A., Tong, S., Sawyer, M.G., Burns, J., McMichael, A.J., 1999. Sociodemographic and behavioural determinants of blood lead concentrations in children aged 11-13 years. The Port Pirie Cohort Study. Med. J. Aust. 170, 63–7.

Baltagi, B.H., Song, S.H., Jung, B.C., 2002. A comparative study of alternative estimators for the unbalanced two-way error component regression model, Econometrics Journal.

Baumer, E.P., Rosenfeld, R., Wolff, K.T., 2012. Expanding the Scope of Research on Recent Crime Trends.

Becker, G.S., 1968. Crime and Punishment: an Economic Approach, in: The Economic Dimensions of Crime. Palgrave Macmillan UK, pp. 13–68. https://doi.org/10.1007/978-1-349-62853-7_2

Birkel, C., Dern, H., 2012. Homicide in Germany, in: Handbook of European Homicide Research: Patterns, Explanations, and Country Studies. Springer New York, pp. 313–328.

https://doi.org/10.1007/978-1-4614-0466-8_20

Boutwell, B.B., Nelson, E.J., Emo, B., Vaughn, M.G., Schootman, M., Rosenfeld, R., Lewis, R., 2016. The intersection of aggregate-level lead exposure and crime. Environ. Res. 148, 79–85. https://doi.org/10.1016/j.envres.2016.03.023

Boutwell, B.B., Nelson, E.J., Qian, Z., Vaughn, M.G., Wright, J.P., Beaver, K.M., Barnes, J.C., Petkovsek, M., Lewis, R., Schootman, M., Rosenfeld, R., 2017. Aggregate-level lead exposure, gun violence, homicide, and rape. PLoS One 12. https://doi.org/10.1371/journal.pone.0187953

Brody, D.J., Kramer, R.A., Flegal, K.M., Pirkle, J.L., Matte, T.D., Gunter, E.W., Paschal, D.C., 1994. Blood Lead Levels in the US Population: Phase 1 of the Third National Health and Nutrition Examination Survey (NHANES III, 1988 to 1991). JAMA 272, 277–283. https://doi.org/10.1001/JAMA.1994.03520040039038

Brubaker, C.J., Schmithorst, V.J., Haynes, E.N., Dietrich, K.N., Egelhoff, J.C., Lindquist, D.M., Lanphear, B.P., Cecil, K.M., 2009. Altered myelination and axonal integrity in adults with childhood lead exposure: A diffusion tensor imaging study. Neurotoxicology.

Buonanno, P., Drago, F., Galbiati, R., Zanella, G., 2011. Crime in Europe and the United States: dissecting the 'reversal of misfortunes.' Econ. Policy 26, 347–385. https://doi.org/10.1111/j.1468-0327.2011.00267.x

Bureau, U.C., 2009. Statistical Abstract of the United States: 2009.

Bureau of Justice Statistics, 2020. *National Crime Victimization Survey*. United States.

Carrozzi, F., Roth, S., 2020. Dirty Density: Air Quality and the Density of American Cities. IZA DP No. 13191. https://ftp.iza.org/dp13191.pdf

Cecil, K.M., Brubaker, C.J., Adler, C.M., Dietrich, K.N., Altaye, M., Egelhoff, J.C., Wessel, S., Elangovan, I., Hornung, R., Jarvis, K., Lanphear, B.P., 2008. Decreased brain volume in adults with childhood lead exposure. PLoS Med. https://doi.org/10.1371/journal.pmed.0050112

Corman, H., Mocan, H.N., 2000. A time-series analysis of crime, deterrence, and drug abuse in New York City. Am. Econ. Rev. 90, 584–604. https://doi.org/10.1257/aer.90.3.584

De Long, J.B., Lang, K., 1992. Are all Economic Hypotheses False? J. Polit. Econ. 100, 1257–1272. https://doi.org/10.1086/261860

Denno, D.W., 1990. Biology and Violence: From Birth to Adulthood. Cambridge University Press. https://doi.org/10.1017/CBO9780511752803

DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. Control. Clin. Trials. https://doi.org/10.1016/0197-2456(86)90046-2

Donohue, J.J., Levitt, S.D., 2001. The impact of legalized abortion on crime. Q. J. Econ. https://doi.org/10.1162/0033553015114405 0

Donohue, J., Levitt, S., 2019. The Impact of Legalized Abortion on Crime over the Last Two Decades. Cambridge, MA. https://doi.org/10.3386/w25863

Dore, C.J., Watterson, J.D., Murrells, T.P., Passant, N.R., Hobson, M.M., Baggott, S.L., Thistlethwaite, G., Goodwin, J.W.L., King, K.R., Adams, M., Walker, C., Downes, M.K., Coleman, J., Stewart, R.A., Wagner, A., Sturman, J., Conolly, C., Lawrence, H., Li, Y., Jackson, J., Bush, T., Grice, S., Brophy, N., Misselbrook, T., Brown, L., Dragosits, U., Sutton, M., Mile, R., Mobbs, D., 2006. UK Emissions of Air Pollutants 1970 to 2004.

Doucouliagos, C. and Stanley, T., 2013. Are all Economic Facts Greatly Exaggerated? Theory Competition and Selectivity. Journal of Economic Surveys, 27: 316-339. https://doi.org/10.1111/j.1467-6419.2011.00706.x

Doucouliagos, H., 2011. "How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics," Working Papers eco_2011_5, Deakin University, Department of Economics.

Duval, S., & Tweedie, R., 2000. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 56(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Falck, S., von Hofer, H., Storgaard, A., 2003. Nordic Criminal Statistics 1950-2000. Department of Criminology Stockholm University.

Ferraro, P., and Shukla, P., 2020. Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics? Review of Environmental Economics and Policy 2020 14:2, 339-351

Fox, J.A., 2005. Demographics and U.S. Homicide, in: The Crime Drop in America, Revised Edition. Cambridge University Press, pp. 288–317. https://doi.org/10.1017/CBO9780511616167.010

Fox, J.A., Marianne Zawitz, 2000. Homicide Trends in the United States: 2000 Update [WWW Document]. Bur. Justice Stat. URL https://www.researchgate.net/publication/246383500_Homicide_Trends_in_the_United_States_2000_Update (accessed 1.14.21).

Gazze, L., Persico, C., Spirovska, S., 2021. The Long-Run Spillover Effects of Pollution: How Exposure to Lead Affects Everyone in the Classroom. SSRN Electron. J. https://doi.org/10.2139/SSRN.3842757

GBD, Global Burden of Disease, 2019 | Institute for Health Metrics and Evaluation [WWW Document]. Glob. Burd. Dis. - IMHE. URL https://www.healthdata.org/gbd/2019 (accessed 1.26.22).

Gottfredson, M.R., Hirschi, T., 1990. A General Theory of Crime. Stanford University Press, Stanford.

Egan, K.B., Cornwell, C.R., Courtney, J.G., Ettinger, A.S., 2021. Blood Lead Levels in U.S. Children Ages 1–11 Years, 1976–2016. Environ. Health Perspect. 129. https://doi.org/10.1289/EHP7932

Havránek, T., Stanley, T.D., Doucouliagos, H., Bom, P., Geyer-Klingeberg, J., Iwasaki, I., Reed, W.R., Rost, K., Aert, R.C.M., 2020. Reporting Guidelines For Meta-Analysis In Economics. J. Econ. Surv. 34, 469–475. https://doi.org/10.1111/joes.12363

Haynes, E.N., Chen, A., Ryan, P., Succop, P., Wright, J., Dietrich, K.N., 2011. Exposure to airborne metals and particulate matter and risk for youth adjudicated for criminal activity. Environ. Res. 111, 1243–1248. https://doi.org/10.1016/j.envres.2011.08.008

Hernberg, S., 2000. Lead Poisoning in a Historical Perspective, American Journal Of Industrial Medicine.

Higgins, J.P.T., Thompson, S.G., 2002. Quantifying heterogeneity in a meta-analysis. Stat. Med. 21, 1539–1558. https://doi.org/10.1002/SIM.1186

Hollingsworth, A., Rudik, I., 2021. The Effect of Leaded Gasoline on Elderly Mortality: Evidence from Regulatory Exemptions. Am. Econ. J. Econ. Policy 13, 345–73. https://doi.org/10.1257/POL.20190654

Hong, S., Candelone, J.P., Patterson, C.C., Boutron, C.F., 1994. Greenland ice evidence of hemispheric lead pollution two millennia ago by Greek and Roman civilizations. Science (80-.). 265, 1841–1843. https://doi.org/10.1126/science.265.5180.1841

Ioannidis, J.P.A., 2005. Why most published research findings are false, in: Getting to Good: Research Integrity in the Biomedical Sciences. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J.P.A., Stanley, T.D., Doucouliagos, H., 2017. The Power of Bias in Economics Research. Econ. J. h

Keele, L., Stevenson, R., & Elwert, F., 2020. The causal interpretation of estimated associations in regression models. Political Science Research and Methods, 8(1), 1-13.

Kristensen, L.J., 2015. Quantification of atmospheric lead emissions from 70 years of leaded petrol consumption in Australia. Atmos. Environ. 111, 195–201. https://doi.org/10.1016/j.atmosenv.2015.04.012

Lanphear, B.P., 2015. The Impact of Toxins on the Developing Brain. Annu. Rev. Public Health 36, 211–230. https://doi.org/10.1146/annurev-publhealth-031912-114413

Lessler, M.A., 1988. Lead and Lead Poisoning from Antiquity to Modern Times. Ohio Journal of Science: Volume 88, Issue 3. https://kb.osu.edu/handle/1811/23252

Levin, R., Zilli Vieira, C.L., Rosenbaum, M.H., Bischoff, K., Mordarski, D.C., Brown, M.J., 2021. The Urban Lead (Pb) Burden in Humans, Animals and the Natural Environment. Environ. Res. 193, 110377. https://doi.org/10.1016/J.ENVRES.2020.110377

Levitt, S.D., 1996. The Effect of Prison Population Size on Crime Rates: Evidence from

Prison Overcrowding Litigation. Q. J. Econ. 111, 319–351. https://doi.org/10.2307/2946681

Levitt, S.D., 2004. Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. J. Econ. Perspect. https://doi.org/10.1257/089533004773563485

Levitt, S.D., 1997. Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime. Am. Econ. Rev. https://doi.org/10.2307/2951346

Lim, H.H., Ong, C.N., Domala, Z., Phoon, W.O., Khoo, H.E., 1985. Blood lead levels in Malaysian urban and rural pregnant women. Public Health 99, 23–29. https://doi.org/10.1016/S0033-3506(85)80123-2

O' Flaherty, Brendan, Sethi, Rajiv, 2015. Urban Crime, Handbook of Regional and Urban Economics, in: Gilles Duranton & J. V. Henderson & William C. Strange (ed.), Handbook of Regional and Urban Economics, edition 1, volume 5, chapter 0, pages 1519-1621, Elsevier. https://doi.org/10.1016/B978-0-444-59531-7.00023-5

Mahaffey, K.R., Annest, J.L., Roberts, J., Murphy, R.S., 2010. National Estimates of Blood Lead Levels: United States, 1976–1980. http://dx.doi.org/10.1056/NEJM198209023071001 307, 573–579. https://doi.org/10.1056/NEJM198209023071001

Marcus, D.K., Fulton, J.J., Clarke, E.J., 2010. Lead and conduct problems: A meta-analysis. J. Clin. Child Adolesc. Psychol. https://doi.org/10.1080/15374411003591455

Marvell, T.B., Moody, C.E., 1996. Specification problems, police levels, and crime rates. Criminology 34, 609–646. https://doi.org/10.1111/j.1745-9125.1996.tb01221.x

Masters, R.D., Coplan, M.J., 1999. Water treatment with silicofluorides and lead toxicity. Int. J. Environ. Stud. 56, 435–449. https://doi.org/10.1080/00207239908711215

McCulloch, C. E., & Neuhaus, J. M., 2011. Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. Statistical Science, 26(3), 388–402. http://www.jstor.org/stable/23059138

Meager, R., 2019. Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. Am. Econ. J. Appl. Econ. 11, 57–91. https://doi.org/10.1257/app.20170299

Messner, S.F., Raffalovich, L.E., Mcmillan, R., 2001. Economic deprivation and changes in homicide arrest rates for white and black youths, 1967-1998: a national time-series analysis. Criminology 39, 591–614. https://doi.org/10.1111/j.1745-9125.2001.tb00934.x

Needleman, H.L., 1991. Human Lead Exposure - Herbert L. Needleman - Google Books [WWW Document]. URL https://books.google.co.uk/books?hl=en&lr=&

id=e9fel0gM3j0C&oi=fnd&pg=PA23&dq=#v=onepage&q&f=false (accessed 12.14.19).

Needleman, H.L., 1996. Bone Lead Levels and Delinquent Behavior. JAMA J. Am. Med. Assoc. 275, 363. https://doi.org/10.1001/jama.1996.03530290033034

New Zealand Police. (2018). Historic murder offences in New Zealand 1926–2017. Retrieved from https://www.police.govt.nz/sites/default/files/publications/historic-new-zealand-murder-rates.pdf (accessed November 14, 2018).

Pope, D.G., Pope, J.C., 2012. Crime and property values: Evidence from the 1990s crime drop. Reg. Sci. Urban Econ. 42, 177–188. https://doi.org/10.1016/J.REGSCIURBECO.2011.08.008

Rainald Borck, Philipp Schrauth, Population density and urban air quality, Regional Science and Urban Economics, Volume 86, 2021, 103596, ISSN 0166-0462, https://doi.org/10.1016/j.regsciurbeco.2020.103596

Rice, K., Higgins, J.P.T., Lumley, T., 2018. A re-evaluation of fixed effect(s) meta-analysis. J. R. Stat. Soc. Ser. A (Statistics Soc. 181, 205–227. https://doi.org/10.1111/rssa.12275

Rosenfeld, R., Fornango, R., 2007. The impact of economic conditions on robbery and property crime: The role of consumer sentiment. Criminology 45, 735–769. https://doi.org/10.1111/j.1745-9125.2007.00096.x

Röver, C., 2020. Bayesian random-effects meta-analysis using the bayesmeta r package. J. Stat. Softw. 93. https://doi.org/10.18637/jss.v093.i06

Rubin, D.B., 1981. Estimation in Parallel Randomized Experiments. J. Educ. Stat. 6, 377–401. https://doi.org/10.3102/10769986006004377

Sanders, T., Liu, Y., Buchner, V., Tchounwou, P.B., 2009. Neurotoxic effects and biomarkers of lead exposure: A review. Rev. Environ. Health. https://doi.org/10.1515/REVEH.2009.24.1.15

Schafer, E.H., 1956. The Early History of Lead Pigments and Cosmetics in China. T'oung Pao 44, 413–438. https://doi.org/10.1163/156853256X00135

Schwartz, J., 1994. Low-level lead exposure and children's IQ: a meta-analysis and search for a threshold. Environ. Res. 65, 42–55. https://doi.org/10.1006/enrs.1994.1020

Schwikowski, M., Barbante, C., Doering, T., Gaeggeler, H.W., Boutron, C., Schotterer, U., Tobler, L., Van De Velde, K., Ferrari, C., Cozzi, G., Rosman, K., Cescon, P., 2004. Post-17th-Century Changes of European Lead Emissions Recorded in High-Altitude Alpine Snow and Ice. Environ. Sci. Technol. 38, 957–964. https://doi.org/10.1021/es034715o

Seeber, A., Meyer-Baron, M., Schäper, M., 2002. A summary of two meta-analyses on neurobehavioral effects due to occupational lead exposure. Arch. Toxicol. https://doi.org/10.1007/s00204-001-0315-5

Sindhu, K.K., Sutherling, W.W., 2015. Role of Lead in the Central Nervous System: Effect on Electroencephlography, Evoked Potentials, Electroretinography, and Nerve Conduction. Neurodiagn. J. https://doi.org/10.1080/21646821.2015.1043222

Stanley, T.D., 2008. Meta-Regression Methods for Detecting and Estimating Empirical Effects in the Presence of Publication Selection*. Oxf. Bull. Econ. Stat. 70, 305–9049. https://doi.org/10.1111/j.1468-0084.2007.00487.x

Stanley, T.D., Doucouliagos, H., 2014. Meta-regression approximations to reduce publication selection bias. Res. Synth. Methods 5, 60–78. https://doi.org/10.1002/jrsm.1095

Stanley TD, Doucouliagos H, Ioannidis JP, 2017. Finding the power to reduce publication bias. Stat Med. 10;36(10):1580-1598.

Statistics Canada, 2019. Number, rate and percentage changes in rates of homicide victims [WWW Document]. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510006801

Stretesky, P.B., Lynch, M.J., 2001. The Relationship Between Lead Exposure and Homicide. JAMA Pediatr. 155, 579–582. https://doi.org/10.1001/archpedi.155.5.579

Stromberg, U., Schutz, A., Skerfving, S., 1995. Substantial decrease of blood lead in Swedish children, 1978-94, associated with petrol lead. Occup. Environ. Med. 52, 764–769. https://doi.org/10.1136/OEM.52.11.764

Tanaka, S., Teshima, K., Verhoogen, E., n.d. North-South Displacement Effects of Environmental Regulation: The Case of Battery Recycling. Am. Econ. Rev. Insights. https://doi.org/10.1257/AERI.20210201

Tcherni-Buzzeo, M., 2019. The "Great American Crime Decline": Possible Explanations. Springer, Cham, pp. 309–335. https://doi.org/10.1007/978-3-030-20779-3_16

Ucrdatatool.gov, 2019. Uniform Crime Reporting Statistics. http://www.ucrdatatool.gov/abouttheucr.cfm (accessed November 17, 2019).

UK Home Office, 2012. Historical crime data - GOV.UK. https://www.gov.uk/government/statistics/historical-crime-data (accessed November 14, 2019).

Vlasak, T., Jordakieva, G., Gnambs, T., Augner, C., Crevenna, R., Winker, R., Barth, A., 2019. Blood lead levels and cognitive functioning: A meta-analysis. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2019.03.052

Westreich, D., and Greenland, S., 2013. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients, American Journal of Epidemiology, Volume 177, Issue 4, Pages 292–298, https://doi.org/10.1093/aje/kws412

WHO, 2010a. WHO | Preventing disease through healthy environments: a global assessment of the burden of disease from environmental risks. WHO.

WHO, 2010b. Childhood Lead Poisoning. WHO

Winter, A.S., Sampson, R.J., 2017. From lead exposure in early childhood to adolescent health: A Chicago birth cohort. Am. J. Public Health 107, 1496–1501. https://doi.org/10.2105/AJPH.2017.303903

Wright, J.P., Dietrich, K.N., Ris, M.D., Hornung, R.W., Wessel, S.D., Lanphear, B.P., Ho, M., Rae, M.N., 2008. Association of prenatal and childhood blood lead concentrations with criminal arrests in early adulthood. PLOS Med. 5, 732–740. https://doi.org/10.1371/journal.pmed.0050101

Zahran, S., Iverson, T., McElmurry, S.P., Weiler, S., 2017. The Effect of Leaded Aviation Gasoline on Blood Lead in Children. https://doi.org/10.1086/691686 4, 575–610. https://doi.org/10.1086/691686

Studies Included in Meta-Analysis

Aizer, A., Currie, J., 2019. Lead and Juvenile Delinquency: New Evidence from Linked Birth, School, and Juvenile Detention Records. Rev. Econ. Stat. 101, 575–587. https://doi.org/10.1162/rest_a_00814

Barrett, K.L., 2017. Exploring Community Levels of Lead (Pb) and Youth Violence. Sociol. Spectr. https://doi.org/10.1080/02732173.2017.1319307

Beckley, A.L., Caspi, A., Broadbent, J., Harrington, H., Houts, R.M., Poulton, R., Ramrakha, S., Reuben, A., Moffitt, T.E., 2018. Association of Childhood Blood Lead Levels With Criminal Offending. JAMA Pediatr. 172, 166–173. https://doi.org/10.1001/jamapediatrics.2017.4005

Billings, S.B., Schnepel, K.T., 2018. Life after lead: Effects of early interventions for children exposed to lead. Am. Econ. J. Appl. Econ. https://doi.org/10.1257/app.20160056

Curci, F., Masera, F., 2018. Flight From Urban Blight: Lead Poisoning, Crime and Suburbanization. SSRN Electron. J. https://doi.org/10.2139/ssrn.3245090

Dills, A.K., Miron, J.A., Summers, G., 2008. What do economists know about crime? NBER working paper series. https://www.nber.org/papers/w13759

Feigenbaum, J.J., Muller, C., 2016. Lead exposure and violent crime in the early twentieth century. Explor. Econ. Hist. 62, 51–86. https://doi.org/10.1016/j.eeh.2016.03.002

Fergusson, D.M., Boden, J.M., Horwood, L.J., 2008. Dentine lead levels in childhood and criminal behaviour in late adolescence and early adulthood. J. Epidemiol. Community Health 62, 1045–1050. https://doi.org/10.1136/jech.2007.072827

Grönqvist, H., Nilsson, J.P., Robling, P.O., 2020. Understanding how low levels of early lead

exposure affect children's life trajectories. J. Polit. Econ. https://doi.org/10.1086/708725

Lauritsen, J.L., Rezey, M.L., Heimer, K., 2016. When Choice of Data Matters: Analyses of U.S. Crime Trends, 1973–2012. J. Quant. Criminol. 32, 335–355. https://doi.org/10.1007/s10940-015-9277-2

Lersch, K.M., Hart, T.C., 2014. Environmental Justice, Lead, and Crime: Exploring the Spatial Distribution and Impact of Industrial Facilities in Hillsborough County, Florida. Sociol. Spectr. https://doi.org/10.1080/02732173.2014.857184

Manduca, R., Sampson, R.J., 2019. Punishing and toxic neighborhood environments independently predict the intergenerational social mobility of black and white children. Proc. Natl. Acad. Sci. U. S. A. 116, 7772–7777. https://doi.org/10.1073/pnas.1820464116

Masters, R.D., Hone, B., Doshi, A., Hone, B., Doshi, A., 1998. ENVIRONMENTAL POLLUTION, NEUROTOXICITY, AND CRIMINAL VIOLENCE 23–58. https://doi.org/10.1201/9781482287400-8

McCall, P.L., Land, K.C., 2004. Trends in environmental lead exposure and troubled youth, 1960-1995: an age-period-cohort-characteristic analysis. Soc. Sci. Res. 33, 339–359. https://doi.org/10.1016/j.ssresearch.2003.06.002

Mielke, H.W., Zahran, S., 2012. The urban rise and fall of air lead (Pb) and the latent surge and retreat of societal violence. Environ. Int. 43, 48–55. https://doi.org/10.1016/j.envint.2012.03.005

Needleman, H.L., McFarland, C., Ness, R.B., Fienberg, S.E., Tobin, M.J., 2002. Bone lead levels in adjudicated delinquents - A case control study. Neurotoxicol. Teratol. 24, 711–717. https://doi.org/10.1016/S0892-0362(02)00269-6

Nevin, R., 2000. How lead exposure relates to temporal changes in IQ, violent crime, and unwed pregnancy. Environ. Res. 83, 1–22. https://doi.org/10.1006/enrs.1999.4045

Nevin, R., 2007. Understanding international crime trends: The legacy of preschool lead exposure. Environ. Res. 104, 315–336. https://doi.org/10.1016/j.envres.2007.02.008

Nkomo, P., Mathee, A., Naicker, N., Galpin, J., Richter, L.M., Norris, S.A., 2017. The association between elevated blood lead levels and violent behavior during late adolescence: The South African Birth to Twenty Plus cohort. Environ. Int. 109, 136–145. https://doi.org/10.1016/j.envint.2017.09.004

Reyes, J., 2007. Environmental policy as social policy? The impact of childhood lead exposure on crime. B E J. Econ. Anal. POLICY 7, 51. https://doi.org/10.2202/1935-1682.1796

Reyes, J.W., 2015. Lead exposure and behavior: Effects on antisocial and risky behavior among children and adolescents. Econ. Inq. 53, 1580–1605. https://doi.org/10.1111/ecin.12202

Sampson, R.J., Winter, A.S., 2018. Poisoned development: assessing childhood lead exposure as a cause of crime in a birth cohort

followed through adolescence. Criminology 56, 269–301. https://doi.org/10.1111/1745-9125.12171

Stretesky, P.B., Lynch, M.J., 2004. The relationship between lead and crime. J. Health Soc. Behav. https://doi.org/10.1177/002214650404500207

Taylor, M.P., Forbes, M.K., Opeskin, B., Parr, N., Lanphear, B.P., 2018. Further analysis of the relationship between atmospheric lead emissions and aggressive crime: an ecological study. Environ. Heal. 17. https://doi.org/10.1186/s12940-018-0354-5

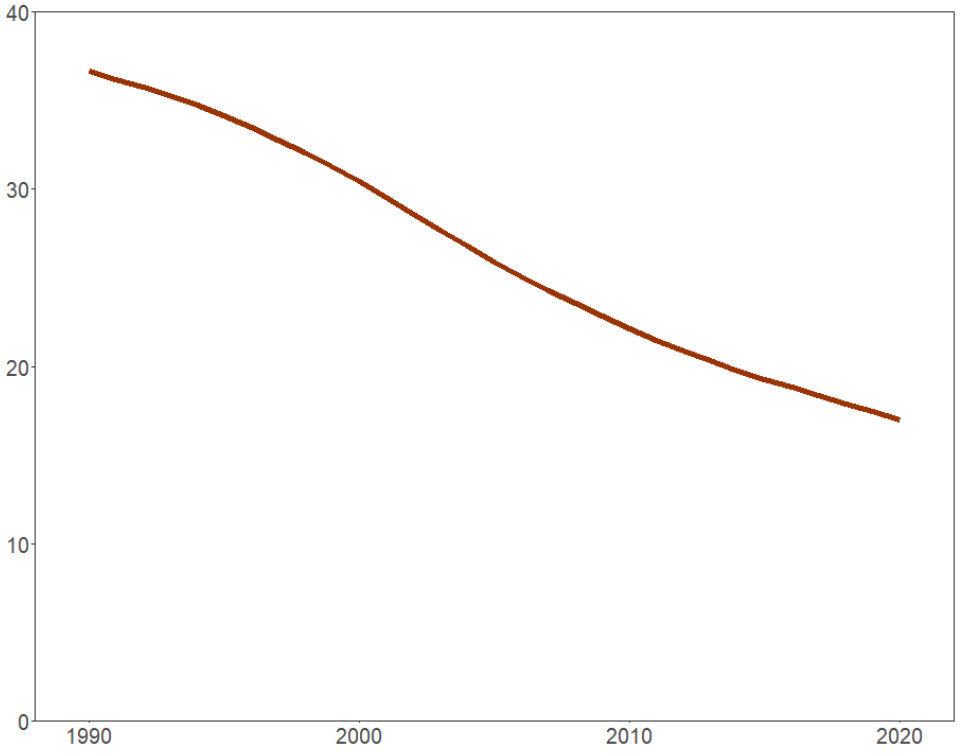
Chapter 2

The Impact of Lead Water Pollution on Birth Outcomes: A Natural Experiment in Scotland

2.1 Introduction

An estimated 2.4 million children die within their first year of life globally (UN IGME 2021). A further estimated 2 million are stillborn (UNICEF 2022a). Great strides have been made in reducing these infant deaths in recent decades (figure 2.1) thanks to improved disease and hygiene practices, as well as nutrition (WHO 2010). However, death rates remain above the level needed to meet the 2030 Sustainable Development Goals (UN IGME 2021). With an estimated 1-in-3 children having elevated levels of lead in their systems (GBD, 2019), and the global burden of lead estimated to be responsible for as many as 900,000 deaths a year (UNICEF, 2020), reducing lead pollution may be one route to prevent infant death and morbidity.

Figure 2.1 World Estimated Neonatal Death Rate, per 1000 Births



Source: UNICEF (2022b)

How does lead pollution affect birth outcomes? How big a risk factor is lead pollution for early deaths and stillbirths? What priority should reducing lead pollution be compared to other mortality reducing interventions, such as improving nutrition or medical access? Ideally, all such risk factors would be dealt with, but with political economy constraints, this is never the case. Policy makers and concerned citizens need to know how much of their time and resources lead reduction should receive compared to other mortality reducing interventions.

We follow Troesken (2006) in examining the impact of lead water pollution on health outcomes. Lead can contaminate drinking water through chemical reactions in plumbing materials containing lead. This causes the metal to dissolve or erode away from lead pipes and fixtures into the water supply. This reaction is particularly severe when the water has low mineral content or high acidity, and it is said to be highly *plumbosolvent.* Troesken (2006) provides an historical overview of how lead exposure has continually occurred through water pipes and tanks. Troesken further expanded on this in Clay et al. (2014) where they use the historical incidence of lead pipes and water pH levels and find a large effect of lead on early 20th century infant mortality. We provide a fuller description of the research on lead and birth outcomes in section 2.2.1. This research shows mixed findings on birthweights, a correlate of infant morbidity and later life outcomes, and on under-5 mortality, with most studies relying on correlational estimates of the relationship between lead pollution and birth outcomes.

Our contribution is to examine these questions using rich administrative data containing all births in Scotland's two largest cities – Glasgow and Edinburgh – and the surrounding areas from 1975-2000. We link this data with home address at time of birth, mother's characteristics, and with infant health outcomes up to 5 years later.

We combine this administrative data with plausibly exogenous variation in lead exposure from separate interventions in the Glasgow and Edinburgh water supplies. This combination allows us to credibly identify the effect of lead on birth outcomes such as birthweight, and under-5 mortality. The water intervention we examine reduced water lead levels and blood lead levels in both cities.

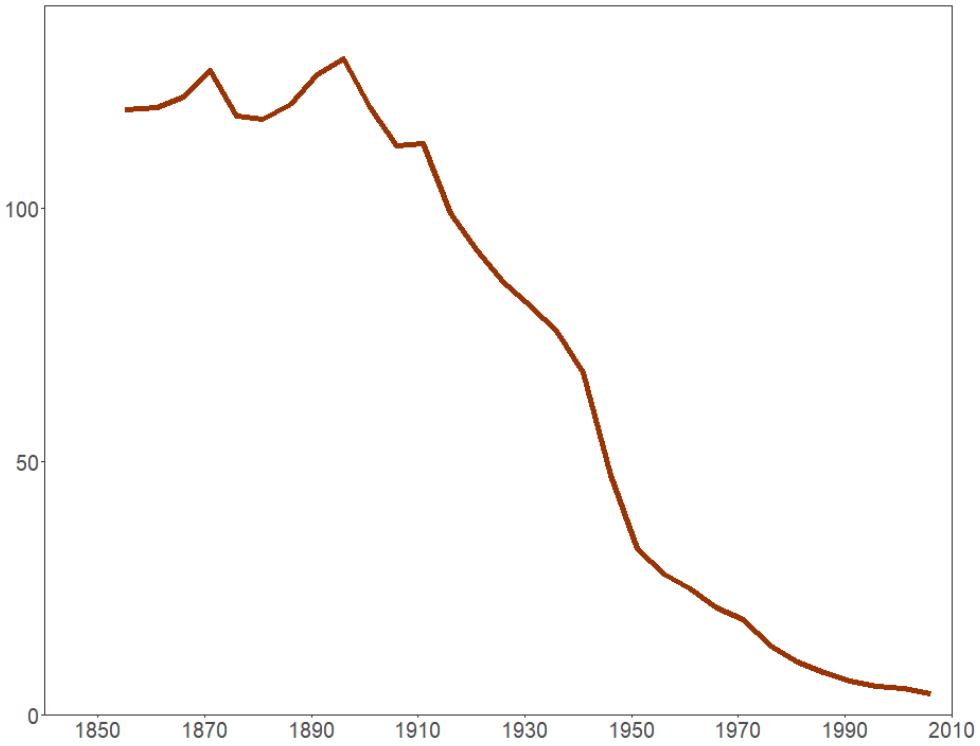
Linking the data to home address allows us to capture if a mother and child lived in an area subject to the lead reduction treatment at the time of birth. Our research design, based on a difference-in-differences approach, improves upon most of the previous literature which is based on selection on observables as an identification strategy, which can result in biased estimates, as lead pollution is correlated with socio-economic factors. Additionally, many of the studies have small sample sizes, which may be under-powered. We also employ methods that, under the assumptions in section 2.3, are unbiased in the presence of staggered treatment interventions, and allow separate treatment effects to be calculated for each treated area. In particular, we discuss and employ the two-way Mundlak regression proposed by Wooldridge (2021).

The case studies of Glasgow and Edinburgh are noteworthy as these are areas with historically high levels of lead before the reduction. Edinburgh and Glasgow were characterised by acidic soft water which made them especially plumbosolvent. In 1975, 33% of households in Scotland had water lead levels above 50μg/l, compared to only 10% in England. Glasgow was particularly affected, with 50% of households surveyed having water lead levels above 100μg/l (Quinn, 1985; Potter, 1997; Richards and Moore, 1984). We rely extensively on the long-running Glasgow (Watt et al. 1996a) and Edinburgh (Macintyre et al. 1998) lead studies, which meticulously detailed and researched the reductions in the water and blood lead levels over the 80s and 90s.

Scotland has seen greatly reduced infant deaths since 1900 (figure 2.2). Starting in the 1970s, interventions to reduce the amount of lead in the water began in Edinburgh and Glasgow, and were improved upon in the 80s and 90s. This was after infant deaths and stillbirths had already sharply reduced, thanks to improved nutrition, hygiene, and health practices.

Therefore, this is a setting where the relatively easy gains had already been exhausted, and lead might be thought to account for a larger share of the remaining deaths and pregnancy complications.

Figure 2.2 Scotland, Deaths Within First Year of Life, per 1000 Births



Source: National Records of Scotland (2022)

Contrary to much of the literature, we do not find consistent evidence for an effect of lead on birthweights. For under-5 mortality, we find weak evidence that lead reduction may have led to a maximum 0.3-0.1 percentage point decrease in deaths in Glasgow and a 0.7-0.1 percentage point decrease in

deaths in the east-side of Edinburgh, which is supplied by the Alnwickhill water plant (tap water in Edinburgh is supplied by two water plants serving the east and west of the city as we will explained more in detail in section 2.2.2) That is, given the average death rate of 1% in both areas, between 70% and 10% of the infant deaths in Alnwickhill and 30% to 10% of the infant deaths in Glasgow during our sample years would be due to lead pollution. This translates into 23-186 saved lives in Alnwickhill and 216-648 in Glasgow, over the full 25 years of the sample. This is somewhat similar to the Clay et al. (2014) estimate where they estimate an increase in pH for a city using lead pipes would reduce deaths by 7%-33%. Similarly, Edwards (2014) found that the fetal death rate increased 32−63% in the first year after lead levels spiked in Washington D.C, and Grossman and Slusky (2019) find the change in water supply in Flint, and subsequent lead water increase, responsible for a 12% decrease in the fertility rate, which they believe is due to miscarriages and fetal deaths. However, even though our main specification is in line with the literature, our results are not robust to different specifications/robustness checks.

2.2 Background

2.2.1 Lead Pollution and Birth Outcomes

A child is first exposed to lead pollution through the placenta (Dorea and Donangelo, 2006). A mother's current exposure to lead can in turn expose a foetus to lead through this route. Furthermore, due to increased bone remodelling, previous maternal lead pollution can affect the foetus, as both lead and calcium (chemically similar) are released from the bones at an increased rate during pregnancy (Yurdakök, 2012). Maternal and infant lead

levels are of similar magnitudes and highly correlated (Al-Saleh et al., 1995), but the relationship between exposure and absorption of lead is complex. For example, it is mitigated by maternal calcium intake (Dorea and Donangelo, 2006). Therefore, there are mediators between lead exposure and the damage it may cause.

A large literature has found diverse impacts of lead pollution. Biological harms include damaged nerve system and brain development when young (Cecil et al., 2008, Brubaker et al., 2009), and at higher levels abdominal pain, headaches, and seizures (WHO, 2010). Behavioural harms include aggressiveness (Marcus et al., 2010), worse memory, and lower attention span (Vlasak et al., 2019). The wider socio-economic impacts resulting from these include increased propensity to commit crime (Higney et al., 2022), lower educational attainment (Hollingsworth et al, 2022, Zheng, 2021), and possibly lower productivity due to health damage (He and Ji, 2021).

Exposure to lead pollution can have significant negative impacts on the development of children, both before and after birth. In severe cases, it can even result in stillbirth or death. Numerous studies have been conducted to determine the extent of these damages as well as other adverse effects of lead on children's health.

In this paper we focus on the effect of lead pollution on birthweight and under-5 mortality[7]. We use birthweight because it is a generally accepted proxy for baby developmental health. It is associated with a wide range of health outcomes such as higher cardiovascular and cancer deaths, diabetes, and obesity, as well as more immediate health outcomes such as infant mortality and morbidity (Law, 2002, Wilcox, 2001). Chatterji et al. (2014) find birthweight is associated with lower adult educational attainment. Behrman and Rosenzweig (2004) use a sample of identical twins and

[7] This is the standard cut-off for infant mortality, as used in the Sustainable Development Goal (SDG) target 3.2.

compare birthweight with later life outcomes, finding higher birthweight is associated with greater height and educational outcomes. Royer (2006) likewise uses a sample of twins and finds that a mother's own low birthweight is associated with them having pregnancy complications in later life, and having lower birthweight children, suggesting low birthweight has an effect for generations. However, it is generally not thought to be low birthweight itself that *causes* these harms, rather it is a proxy for underlying biological conditions, such as low nutrient ingestion in the womb or premature birth (Wilcox, 2001).

A number of studies estimate the relationship between lead exposure and birthweight. Xie et al. (2013) find a negative correlation between maternal or cord lead levels and birthweight in 252 infants. Similarly, Bornschein, R.L. et al. (1989) find a negative link between maternal blood lead levels and birthweight in 202 inner city infants. Taylor et al. (2014) find that 12% of infants whose mothers have elevated levels of lead (>5μg/dl) have low birthweight compared to 10% when lead levels are lower. In contrast, Azayo et al. (2009) find no association between maternal blood lead levels and birthweight in 150 women in Tanzania, but the average lead levels were below 5μg/dl, which is the threshold used by the WHO guidelines (WHO, 2021). Golmohammadi et al. (2007) use a sample with much higher average lead levels but also find no association in their sample of 89 infants in Iran. McMichael et al. (1986) found no association with birthweight for 749 mothers in Australia, although they do find an association with other outcomes such as spontaneous abortion.

In summary, the findings on birthweight are somewhat mixed. Both disagreeing on the presence of an effect, and on the level of lead at which an effect is found. Bellinger (2005) conducted a review concluding that there likely was an effect of both paternal and maternal blood lead levels on birthweight.

One main issue with the previous papers is that they all rely on selection on observables as an identification strategy. This likely leads to biased estimates, as lead is confounded by poverty, race, and education. Many of the studies also have low sample sizes, which may be inadequately powered. Recently, studies with better identification strategies have examined the relationship between lead and birthweights. Grossman and Slusky (2019) studied the effects of the change in Flint, Michigan's water supply on birthweight. In 2014, the city switched its water source from Lake Huron to the Flint River, causing the water distribution pipes to corrode and leach lead into the drinking water. The researchers used a difference-in-differences design and found that the change in the water supply resulted in higher lead levels, but the effect on birthweight was small and not statistically significant. This may be due to a higher number of stillbirths after the water change, which could have resulted in selection bias in the measurement of birthweight. Dave and Yang (2022) look at a similar setting, where the pH on one side of the water supply in Newark fell sharply, and therefore began leeching lead from pipes again, while it remained steady on the other side of the city. They found a small effect on birthweight, that becomes smaller and insignificant the more post treatment years are added. They rationalise this as showing the effects of mitigation strategies by mothers, such as moving to bottled water, once the increased lead levels were widely known.

For spontaneous abortion (before 28 weeks) and stillbirths (after 28 weeks) high levels of lead have long been known to have an effect. So much so, that lead oxide was described as being used as an abortifacient by working-class women in the 1800s and early 1900s (Hall and Ransom, 1906). In some cases, the amounts of lead ingested were strong enough to cause lead poisoning in the mother (Ransom, 1900). There are many papers which have examined the effects on spontaneous abortions, stillbirths and their correlates. Falcon et al (2003) find that premature births and pregnancy anomalies tended to have higher levels of lead in the placentas of 83 births

(although they find no association with birthweight). Wibberly et al (1977) found that lead levels were higher in placentas where a neonatal death occurred in Birmingham. In contrast, McMichael et al. (1986) do not find any difference in pre-pregnancy maternal blood lead levels for neonatal deaths and other births. Angell and Lavery (1982) collected cord blood lead levels in 635 cases and found no relationship with lead levels and pregnancy complications that might lead to death such as preterm delivery or premature membrane rupture, although they did not look at spontaneous abortions/stillbirths directly. Vinceti et al. (2001) examine historical birth anomalies in a heavily lead polluted area of northern Italy. They find increased oral clefts and other disabilities but no increase in neural tube defects.

Looking specifically at water lead exposure and studies which use natural experiments, Clay et al. (2014) use the differences in city water pH levels as an instrument for lead exposure, because lower pH water leeches more lead from pipes. They find, in 1900-1920, a decline in exposure equivalent to an increase in pH from 6.675 (25th percentile) to 7.3 (50th percentile) in cities with lead-only pipes would have been associated with a decrease in infant mortality of 7 to 33 percent or at least 12 fewer infant deaths per 1,000 live births. Edwards (2014) finds that a short-term spike in lead water pollution in Washington DC (due to a change in the chemical treatment) resulted in an increase of the fetal death rate. Grossman and Slusky (2019) find live births fell 12% in Flint, Michigan during the water crisis and attribute this to an increase in fetal death rates.

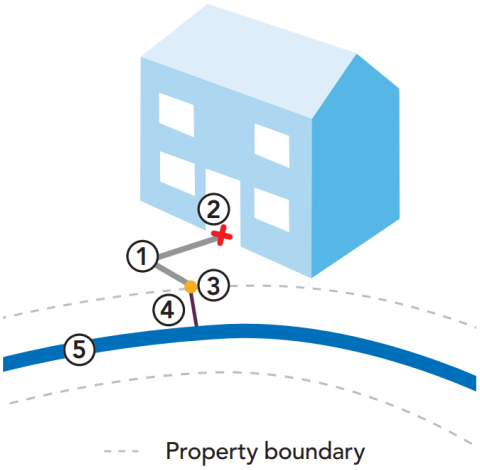
2.2.2 Lead Plumbing and Water Treatment in Glasgow and Edinburgh

Lead piping was widely used in Scotland before being banned for new work in 1968 (Richards et al., 1980). Lead is malleable, relatively cheap, and has an extremely long life as infrastructure (Feigenbaum and Muller, 2014, and Krebs, 2019). Lead piping began to be phased out from the 70s in Scotland, but still, in the 90s, as many as 589,000 homes in Scotland were estimated to contain lead pipes (Potter, 1997), around 30% of the total. This was slightly lower than the 34% in England and Wales. There were also as many as 60,000 water storage tanks made of lead, mostly in Glasgow and Edinburgh (Krebs, 2019). These were used because water service was still intermittent in the first half of the 20th century. They allowed households to store and use water during any non-flowing periods.

The reason lead water pipes have not been replaced are twofold: 1) It is expensive, and 2) Homeowners do not know they have lead pipes and that they are responsible for their replacement. Figure 2.3 shows the different parts of the water supply chain. Communication pipes are owned by the water supplier and those made of lead have now all been replaced (Akoumianaki, 2017). Internal lead piping still exists in many households but has also been gradually replaced and is estimated to only account for 20-30% of the remaining lead pollution in home water supplies in the UK (Akoumianaki, 2017). The main pollution burden is thought to come from the lead supply pipes (also called service pipes). These are the responsibility of the property owner to replace, but are underground, and therefore difficult to see. An added complication is that property owners may not know they have the responsibility to obtain replacement, even though grants are available. Today there are estimated to be 273,751 homes out of 2.6 million in Scotland with lead piping (Robertson et al., 2020). Watt et al. (2000)

estimated as many as 160,000 households out of 300,000 in Glasgow alone had a lead service pipe in 2000.

Figure 2.3 Water Infrastructure around the Home



	Responsibility
1 The water supply pipe	Homeowner
2 Stop valve	Homeowner
3 Stopcock/meter	Scottish Water
4 The communication pipe	Scottish Water
5 The water main	Scottish Water

Source: Scottish Water (2021)

The dangerous combination of certain water chemistry and lead water pipes began to be taken seriously in Europe in the 1970s. The WHO issued guideline for drinking water in 1970 with a limit of 100µg/l (WHO, 1970). Along with other European countries, the UK's Department for the Environment carried out a series of surveys of blood lead levels in the 70s and 80s. The findings of the UK survey were that "The highest blood lead concentrations were related to plumbosolvent water" (Quinn, 1985). The acidic soft water in Scotland's two largest cities made them especially plumbosolvent. In 1975 surveys found 33% of households in Scotland had water lead levels above 50µg/l, compared to 10% in England (Potter 1997). Glasgow was especially viewed as a problem, with 50% of household surveyed having water lead levels above 100µg/l (Richards and Moore, 1984).

Experiments with the Glasgow water supply in 1973 showed that lime dosing would raise the pH effectively and lower plumbosolvency (Richards et al., 1980). Following this, in both Glasgow and Edinburgh, an investment in an automatic lime-dosing system was considered worthwhile. These began operation in 1978. In Glasgow, which is supplied by Loch Katrine water, the pH was raised from 6.3 to 7.8 after this dosing. The Loch Katrine water supply area at this time is mapped in figure 2.4. This map also shows postcodes where there was a higher prevalence of lead piping in the house (Watt et al. 1996a). We exploit the difference in high and low lead prevalence areas in Glasgow as a robustness check in section 2.2.6.

In Edinburgh, the city is supplied with water from diverse sources (see figure 2.5). The north-east of the city was supplied from Alnwickhill, and the south-west supplied from Fairmilehead. The centre was supplied from both sources. Both bodies of water were fairly soft, with a pH of around 7 before dosing. In Fairmilehead, the dosing was successful, and raised the pH to above 8. In Alnwickhill the dosing was not successful due to technical difficulties and was delayed until 1985. The pH remained below 8 until after 1985, when it eventually rose to around 8.5. The mixed area in Edinburgh, supplied jointly by Alnwickhill and Fairmilehead, therefore received a partial treatment, but when measured in 1985 its pH was above 8 and closer to the Fairmilehead level. The measured pH levels before and after the lime dosing for each water supply area are given in figure 2.6.

Figure 2.4 Historical Loch Katrine Water Supply Area with High and Low Lead Piping Prevalence

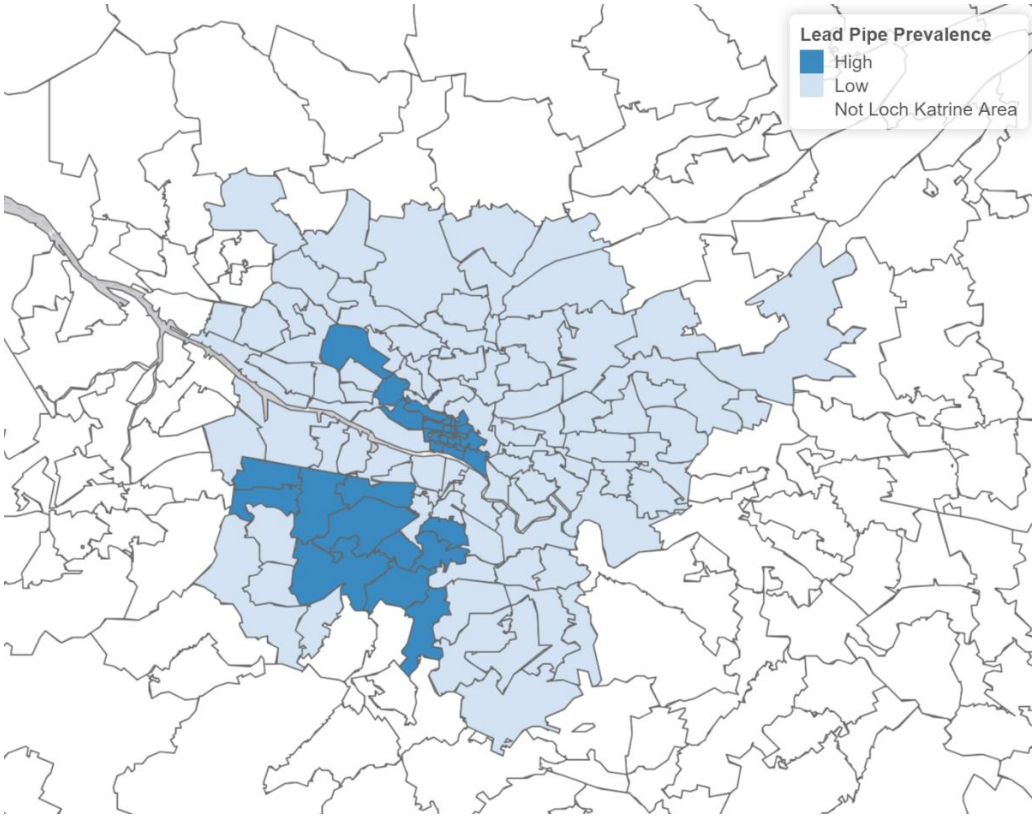


Figure 2.5 Historical Water Supply Areas in Edinburgh

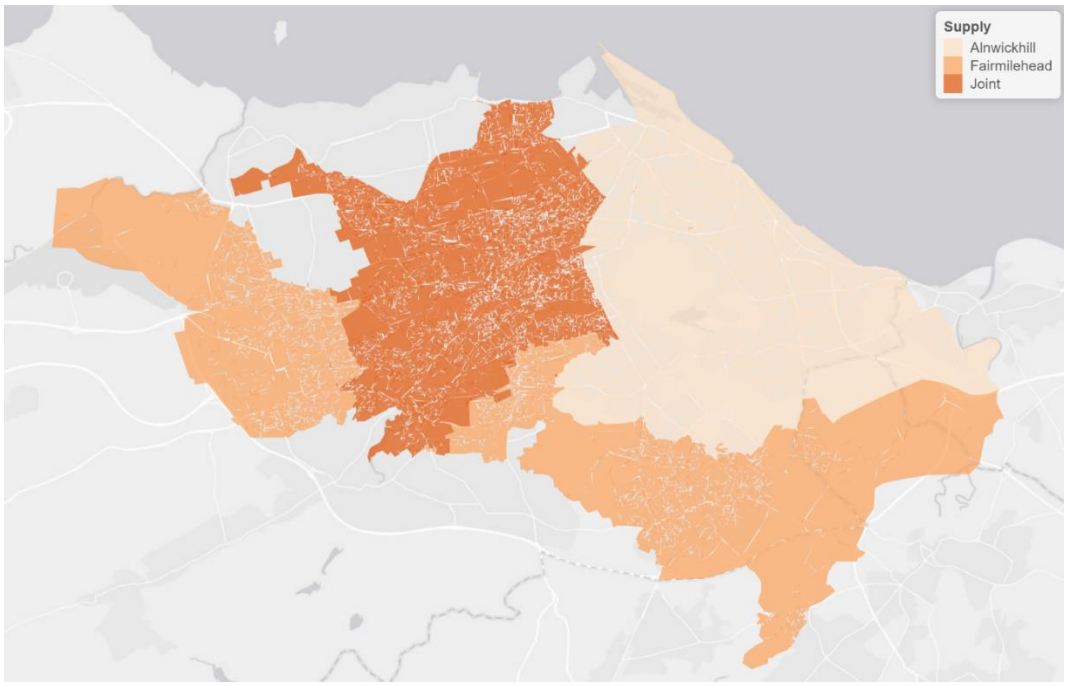
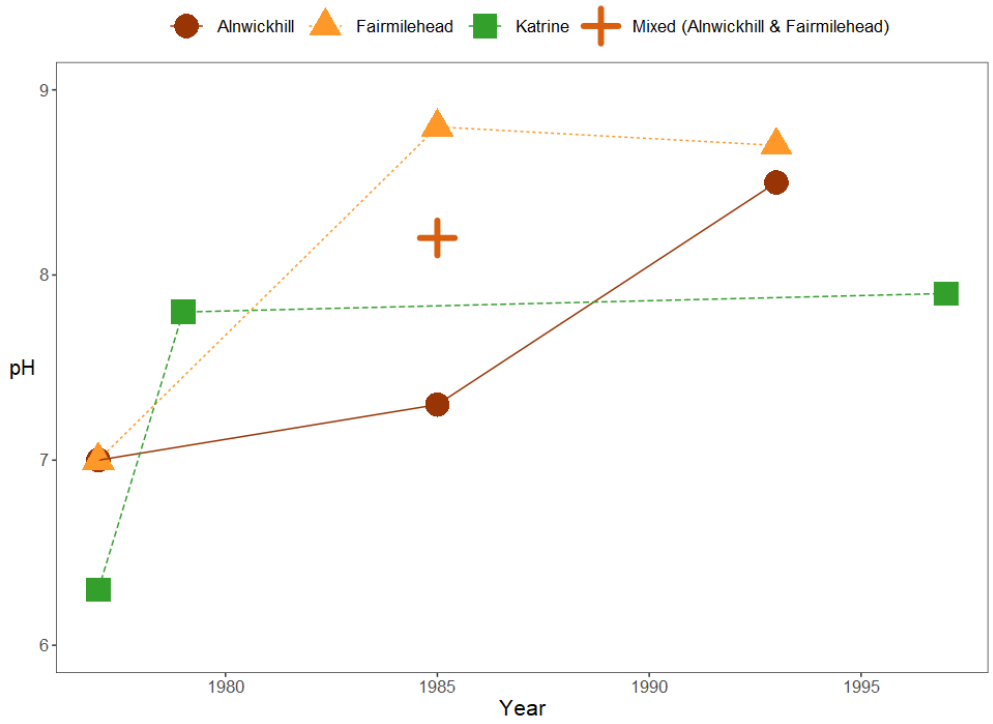


Figure 2.6 pH Levels in Each Water Supply Area



Sources: Macintyre et al., (1998), Richards et al., (1980) and Watt et al., 1996.

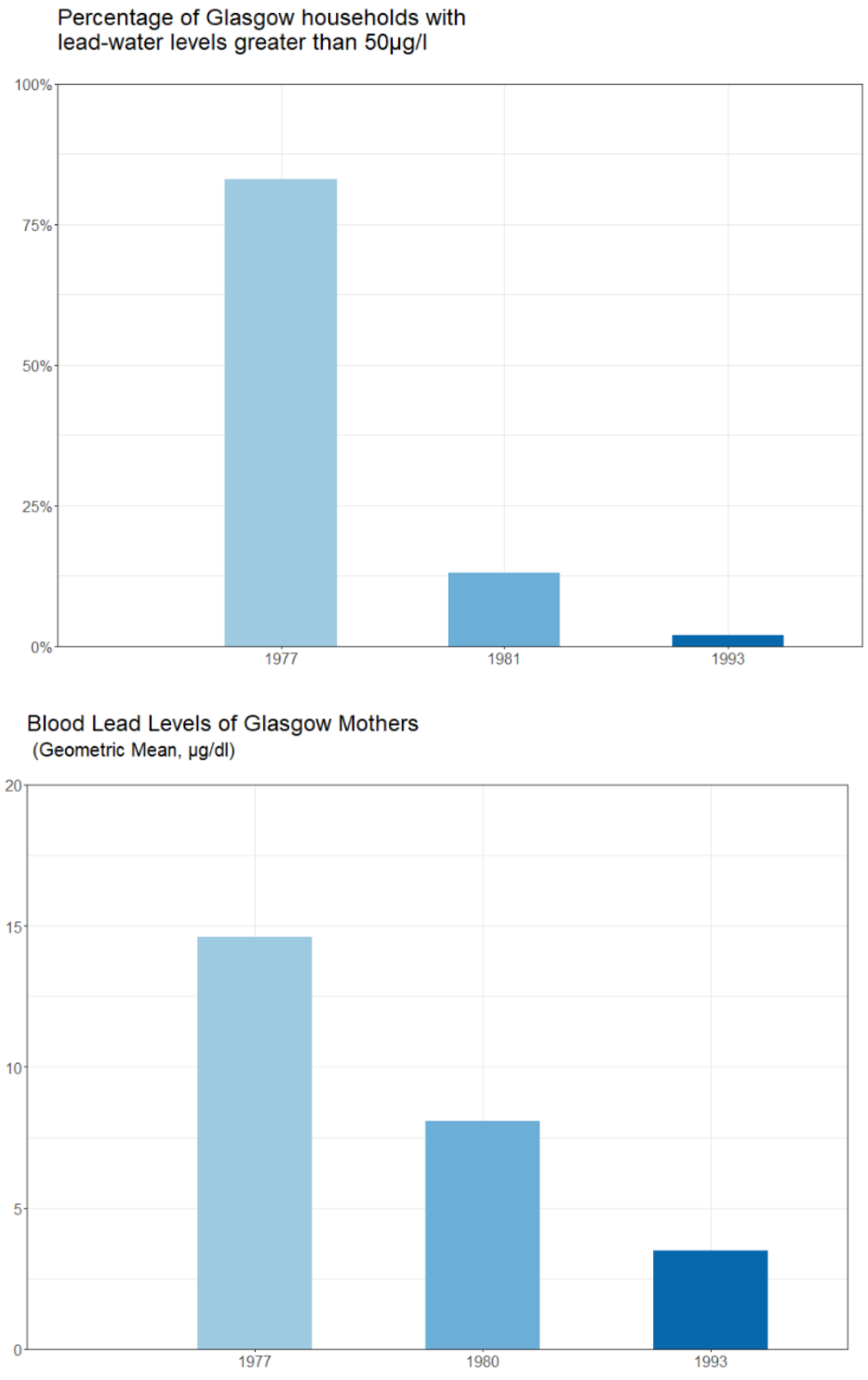
Although partially successful in reducing plumbosolvency, high lead levels continued to be present in some houses, especially in Glasgow (Watt et al, 1996a). Standards also continued to strengthen, with the EU reducing the maximum allowed lead-water levels to 50µg/l in 1980, eventually reaching 10µg/l in 2000 (Watt et al., 2000). This led to the addition of orthophosphate into the Glasgow water supply in 1989. Orthophosphate dosing reduces the solubility, and therefore bioavailability, of lead in the water supply (Comber et al, 2011). This is a different mechanism for reducing lead pollution, compared to lime treatment which raises the pH. This dosing successfully reduced blood lead levels even further (figure 2.7).

In Edinburgh, after the successful experiments in Glasgow, orthophosphate was added to both Fairmilehead and Alnwickhill supplies in 1991. This

further reduced blood lead and water lead levels in both areas (Figure 2.8). The lowering of lead-water levels in Scotland's largest cities was a success. As the UK government brought in stricter lead-water maximum levels, other areas followed. For example, water supplies across Wales began to be orthophosphate dosed in 1995. Eventually, in the 2000s, 95% of the UK's water would be treated with orthophosphate (Hayes and Hydes, 2012). However, still in the 1990s, it was debated in the UK Parliament whether water treatment was enough, with one MP saying that only full lead pipe replacement would suffice (Hansard, 1990).

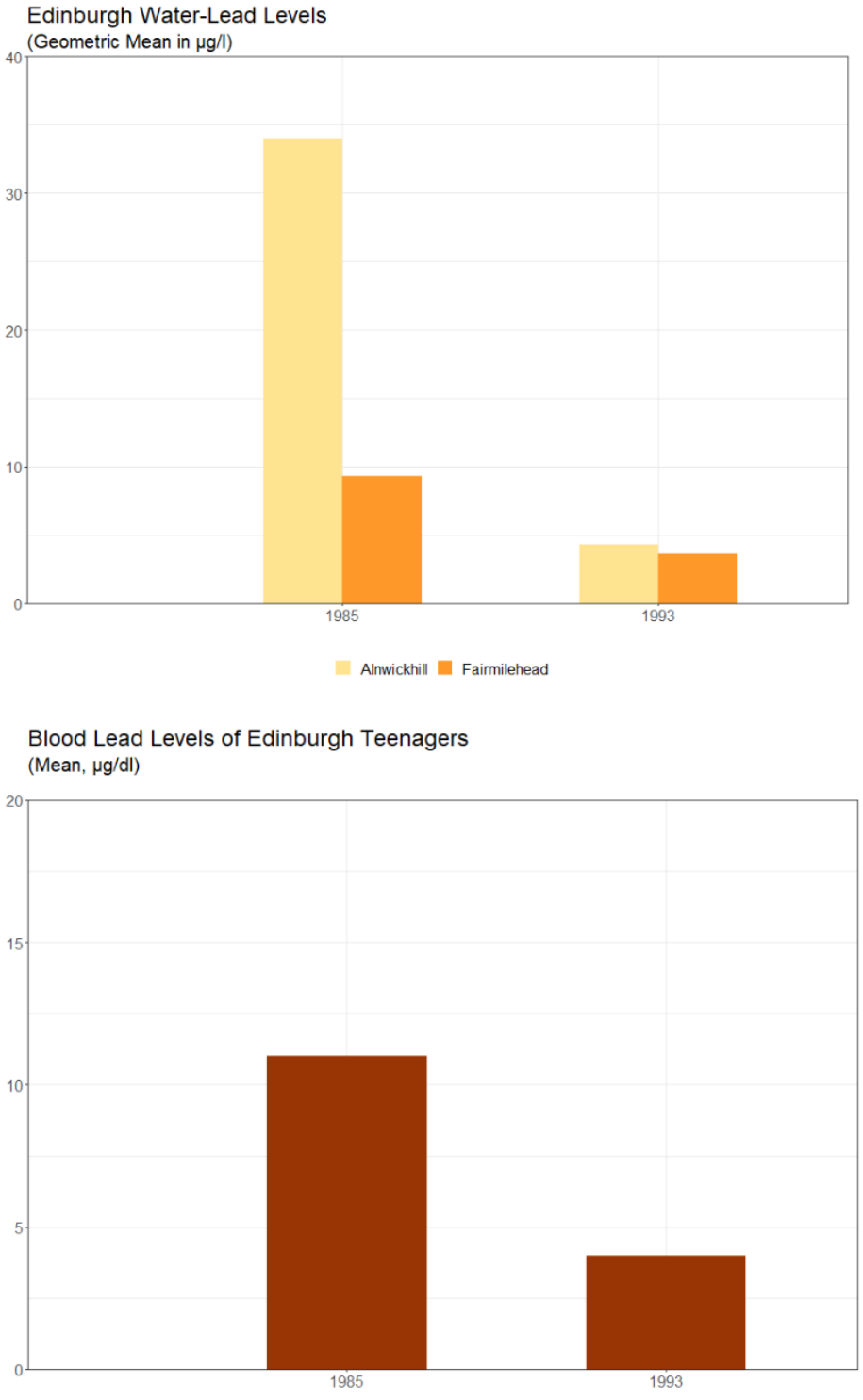
This paper exploits these interventions as natural experiments given the somewhat arbitrary assignment of water treatment to certain areas of Glasgow and Edinburgh (treatment groups) while leaving adjacent areas untouched (control group). In particular, the plausibly exogenous variation we rely on to identify treatment effects of lead on birth outcomes is first, the raising of the pH with lime dosing in both Glasgow (Katrine) and Edinburgh (Fairmilehead in the south-west of the city and Alnwickhill in the north-east) which did not occur in the surrounding areas of these cities, and secondly, the later dosing with orthophosphate in both cities which similarly was not carried out in neighbouring areas until much later. A full timeline of the treatments is given in figure 2.9.

Figure 2.7 Blood and Water Lead Levels in Glasgow



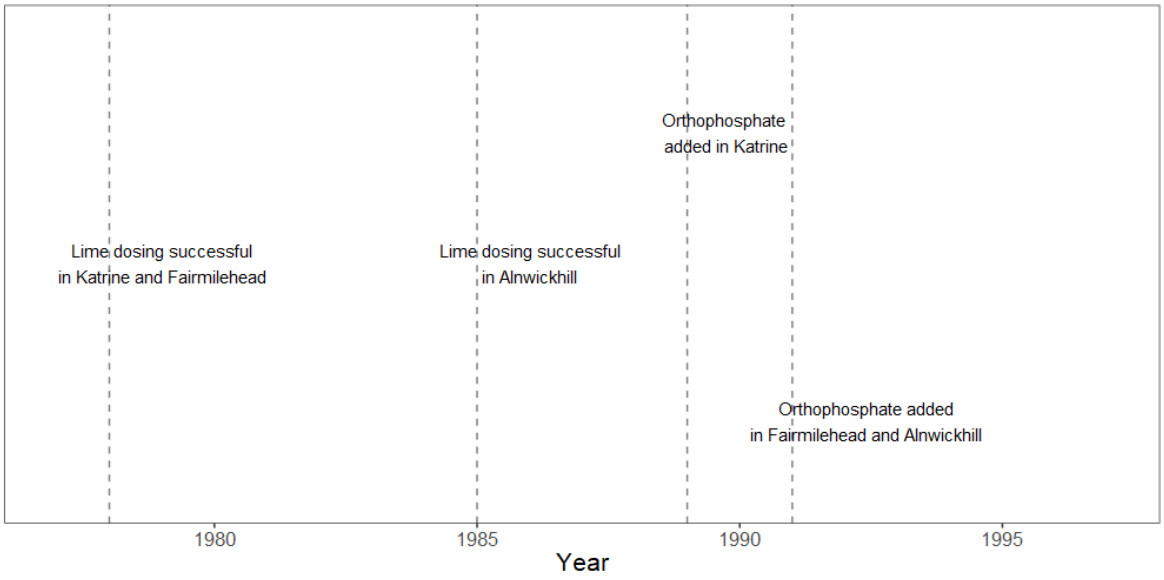
Source: Watt et al., (1996a) and Moore (1998)

Figure 2.8 Blood and Water Lead Levels in Edinburgh



Source: Macintyre et al., (1998)

Figure 2.9 Timeline of Water Treatment



Note: Katrine is the name of the water supply in the city of Glasgow. Farmilehead is the name of the water plant supplying the south-west of Edinburgh. Alnwickhill is the water plant supplying the north-east of Edinburgh. Central Edinburgh is served by both water supplies.

2.3 Data

We use health data from Public Health Scotland (PHS). The data covers all the pregnancy outcomes in Glasgow, Edinburgh, and the surrounding postcode areas of each city for the period 1975 to 2000. This data is from the Scottish Morbidity Records (SMR) and the Death, Birth and Stillbirths Registrations (NRS) Furthermore, from the NRS records, we link live births records with death registrations to identify if a child died before age five[8].

Data are matched to the Scottish Water Supply area maps for the relevant areas by using maternal postcode for the relevant period. That is, the address

[8] We would like to thank the Electronic Data Research and Innovation Service (eDRIS) of Scotland for their help with data handling and access.

of the house at the time of pregnancy is assigned to the postcode and coded as the relevant treatment/control group. The map of water supply areas, and the various treatment groups are included in figures 2.4 and 2.5. Figure 2.4 shows the Loch Katrine water supply area during the period the data cover. There is a further split in the Loch Katrine supply area between postcode sectors with relatively high levels of lead piping compared to those with relatively low levels of lead piping, as given in Watt et al. (1996a). In high lead areas, 19% reported lead piping, while in low lead areas it was 9%. Figure 2.5 shows the water supply areas in Edinburgh during the period the data cover. The Fairmilehead source supplied mostly the west of Edinburgh, while Alnwickhill the east. The "Joint" area is supplied by both water sources during this period. As explained in the methods section, the first treatment of calcium carbonate was effective in 1978 in Farimilehead areas, but not effective till 1985 in Alnwickhill areas. We therefore treat the "joint" area served by both sources as being treated at the same time as Fairmilehead in 1978 but exclude it as a robustness check in section 2.6.1.

Our two main outcomes are birthweight and under-5 mortality. We use only single births. Twins, and other multiple births are excluded as their outcomes tend to be very different, with lower birthweights in comparison to single births, as well as different probabilities of complications. However, multiple births are only around 1.5% of all births. Death episodes are used for all deaths and non-viable pregnancies, including stillbirths and spontaneous abortions. We also link the data with Scottish Morbidity Records so that it includes any deaths up to age 5. Under-5 mortality is the commonly used definition of child mortality and is the indicator used in Sustainable Development Goal (SDG) target 3.2: "Newborn and child mortality: By 2030, end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce neonatal mortality and under-5 mortality"[9].

[9] See the SDG targets and indicators here: https://sdgs.un.org/.

Additional data used as controls at the individual level include the biological sex of the baby, and a series of mother's characteristics such as age, height, and previous obstetric history, such as the number of previous spontaneous abortions, and number of previous pregnancies.

At the postcode level, the data is linked to Carstairs scores[10], which are material deprivation indices that rank different areas by using information from the 1981 census about car ownership, male unemployment, overcrowding and low social class. If a postcode is in the bottom two deciles, we code that postcode with an indicator variable as being in a deprived area. Table 2.1 includes descriptive statistics of the variables used.

Table 2.1 Descriptive Statistics

Variable	Group	Mean	Median	Std Dev	Obs
Birthweight (grams)					
	Control	**3318**	**3360**	**594**	**353,643**
	Edinburgh	**3320**	**3360**	**595**	**76,498**
	Alnwickhill	3310	3350	603	26,172
	Fairmilehead	3372	3410	567	8,315
	Joint	3317	3360	595	42,011
	Glasgow	**3257**	**3300**	**591**	**216,771**
	High	3305	3340	589	31,337
	Low	3248	3283	591	185,434
Death Indicator Variable					
	Control	**0.01**	**0**	**0.09**	**353,643**
	Edinburgh	**0.01**	**0**	**0.09**	**76,498**
	Alnwickhill	0.01	0	0.10	26,172
	Fairmilehead	0.01	0	0.09	8,315
	Joint	0.01	0	0.09	42,011
	Glasgow	**0.01**	**0**	**0.09**	**216,771**
	High	0.01	0	0.09	31,337
	Low	0.01	0	0.09	185,434

[10] These are rankings of areas by material deprivation. The variable takes into account material good ownership, such as car ownership, self-reported class, and unemployment amongst other variables to make an index. The Cairstairs scores were originally developed by Carstairs and Morris (1991) and are regularly generated and published by the MRC/CSO Social and Public Health Sciences Unit, the University of Glasgow (https://www.gla.ac.uk/schools/healthwellbeing/research/mrccsosocialandpublichealthsciencesunit/programmes/inequalities/healthinequalities/determinantsofhealthandhealthinequalitiesinscotland/carstairsscores/).

Proportion Living in Deprived Areas (Carstairs Index)					
	Control	**0.11**	**0**	**0.32**	**353,643**
	Edinburgh	**0.13**	**0**	**0.34**	**76,498**
	Alnwickhill	0.06	0	0.23	26,172
	Fairmilehead	0.04	0	0.19	8,315
	Joint	0.20	0	0.40	42,011
	Glasgow	**0.59**	**1**	**0.49**	**216,771**
	High	0.04	0	0.20	31,337
	Low	0.69	1	0.46	185,434
Total Previous Pregnancies					
	Control	**1.18**	**1**	**1.28**	**353,643**
	Edinburgh	**1.12**	**1**	**1.29**	**76,498**
	Alnwickhill	1.07	1	1.24	26,172
	Fairmilehead	1.15	1	1.27	8,315
	Joint	1.15	1	1.31	42,011
	Glasgow	**1.24**	**1**	**1.41**	**216,771**
	High	1.20	1	1.38	31,337
	Low	1.24	1	1.41	185,434
Mother's Age					
	Control	**27.72**	**28**	**5.29**	**353,643**
	Edinburgh	**28.43**	**29**	**5.46**	**76,498**
	Alnwickhill	28.25	28	5.32	26,172
	Fairmilehead	29.93	30	5.24	8,315
	Joint	28.25	28	5.54	42,011
	Glasgow	**26.90**	**27**	**5.58**	**216,771**
	High	28.56	29	5.36	31,337
	Low	26.62	26	5.57	185,434
Number of Previous Spontaneous Abortions					
	Control	**0.22**	**0**	**0.57**	**353,643**
	Edinburgh	**0.23**	**0**	**0.59**	**76,498**
	Alnwickhill	0.23	0	0.57	26,172
	Fairmilehead	0.24	0	0.61	8,315
	Joint	0.23	0	0.59	42,011
	Glasgow	**0.23**	**0**	**0.59**	**216,771**
	High	0.24	0	0.60	31,337
	Low	0.23	0	0.59	185,434
Male Infant Indicator Variable					
	Control	**0.51**		**0.50**	**353,643**
	Edinburgh	**0.51**		**0.50**	**76,498**
	Alnwickhill	0.51		0.50	26,172
	Fairmilehead	0.52		0.50	8,315
	Joint	0.51		0.50	42,011
	Glasgow	**0.51**		**0.50**	**216,771**
	High	0.51		0.50	31,337
	Low	0.51		0.50	185,434

2.4 Methods

We use the plausibly exogenous change in water treatments, at different points in time, to identify the effect of lead-water pollution on birth and early life health outcomes. Our main specifications use a difference-in-differences design. We further discuss the estimands, assumptions necessary, and specifications below. The following section is largely based on the excellent expositions in Athey and Imbens (2022), Wooldridge (2021), and Wooldridge (2010).

2.4.1 Estimands

Our main results focus on two estimands. First, the average effect of water treatment (and therefore lead reduction) at time t on the group which began treatment at time r.

We write this τ_{rt} and define it formally below.

$$(1) \qquad \tau_{rt} = E[\, y_{it}(r) - y_{it}(0) \mid d_{ir} = 1\,],\ r = q, \dots, T;\ t = r, \dots, T.$$

Where $y_{it}(r)$ is the outcome for child i at time t given their water supply began treatment at time r, and $r \leq t$, and $y_{it}(0)$ is the unobserved counterfactual outcome for child i at time t where they have not yet received treatment, q is the first period where any cohort is treated, and d_{ir} is a cohort indicator which equals 1 if individual i is in treatment group r. Simply, τ_{rt} is the average effect of treatment on the treated (ATT) for that treatment cohort in that year.

Our second main estimand is the average τ_{rt} for all the years of treatment in our data.

(2) $\bar{\tau}_r = E[\tau_{rt}], r = q, \dots, T; t = r, \dots, T.$

Which we estimate as:

(3) $\hat{\bar{\tau}}_r = \frac{\sum_{t=r}^{T} \hat{\tau}_{rt}}{(T-r+1)}$

With each τ_{rt} defined in the specifications below.

2.4.2 Difference-in-Differences Design Models

Our main results are from models relying on difference-in-differences designs. In the baseline, reduced form model, lead levels, given by the variable $Lead_{it}$, are assumed to affect the birth outcome as shown in (4). With the treatment effect of lead given by T .

(4) $y_{it} = c_j + g_t + \boldsymbol{x_i}\boldsymbol{\beta} + (\mathrm{T} \times Lead_{it}) + u_{it}$

Where y_{ijt} is the outcome for individual i, at time t. There is a time-invariant postcode-level effect, c_j, a time trend in outcome, g_t, and a vector of other variables that affect the outcome, $\boldsymbol{x_i}$, which vary by individual. The final term u_{it} is the error term.

This model cannot be estimated for a number of reasons, not least because lead exposure of each individual at each time is unknown. Even if known, other variables may covary with lead and the outcome, leading to biased estimates due to endogeneity. We could estimate a two-way fixed effects model using a $post \times treatment$ indicator but, given the staggered timing of the intervention between Glasgow, Fairmilehead, and Alnwickhill, this could lead to the effect not being identified, due to the "forbidden comparisons" problem. This is where the two-way fixed effects estimate is a weighted combination of all possible comparisons of each treatment and control group. Comparing two treated groups can lead to negative weighting of some of

these comparisons, and therefore not identify the ATT (Goodman-Bacon, 2021, and Calloway and Sant'Anna, 2021).

However, given the plausibly exogenous change in lead exposure outlined in section 2.2.2, we can identify the effect of the lead reduction for each separate treatment group, with the estimands (1) and (2), if we are willing to accept certain assumptions. Following Wooldridge (2021), these are:

Conditional No Anticipation, Staggered Treatment (CNAS)

Following Athey and Imbens (2022) and Wooldridge (2021), we define the outcome for the never-treated group as $y_{it}(\infty)$. Given this, we formally state the CNAS assumption as:

(5) $$E[\, y_{it}(r) - y_{it}(\infty) \mid d_{ir} = 1, \boldsymbol{x_i}\,] = 0,\ r = q, \dots, T;\ t < r.$$

This states that before treatment, in each treatment cohort, the average difference between the treatment cohort's outcome and the never-treated cohort's, a*fter* conditioning on covariates, is 0. That is, once we have accounted for time trends, time-invariant fixed effects (such as cohort and area fixed effects), and other covariates we have no difference on average between treated and untreated *before* treatment. This assumption is testable, and we test it in section 2.5.

Conditional Common Trends, Staggered Treatment (CCTS)

(6) $$E\big[\, y_{it}(\infty) - y_{i1}(\infty) \mid d_{ir}, \boldsymbol{x_{ij}}\,\big] = E\big[\, y_{it}(\infty) - y_{i1}(\infty) \mid \boldsymbol{x_{ij}}\,\big],$$
$$r = q, \dots, T; t = 2, \dots, T.$$

This states that for every cohort the trend in outcome if never-treated is unrelated to being in any treatment cohort, after conditioning on the covariates. This can also be tested to a degree. See section 2.5.

Linear in Parameters, Staggered Treatment (LINS)

(7) $$E\big[\, y_{i1}(\infty) \mid d_{ir}, \boldsymbol{x_{ij}}\,\big] = \eta + \boldsymbol{x_i}\boldsymbol{\kappa} + \textstyle\sum_{r=q}^{T} \lambda_r d_{ir} + \sum_{r=q}^{T} \zeta_r (d_{ir} \times \boldsymbol{x_i}),$$

$$r = q, \dots, T.$$

(8) $$E[\, g_t(\infty) \mid \boldsymbol{x_i} \,] = \sum_{s=2}^{T} \theta_s f s_t + \sum_{s=2}^{T} (f s_t \times \boldsymbol{x_i}) \boldsymbol{\pi_t}, \; t = 2 \dots, T.$$

(9) $$\tau_{rt}(\boldsymbol{x}) = \tau_{rt} + (\boldsymbol{x_i} - E[\boldsymbol{x_i} | d_{ir} = 1]) \boldsymbol{\rho_{rt}} = \tau_{rt} + \dot{\boldsymbol{x}}_{\boldsymbol{ir}} \boldsymbol{\rho_{rt}}, \; r = q, \dots, T; t = r, \dots, T.$$

Where η is the intercept and $f s_t$ are indicators for every time period that equal 1 when $s = t$. Next, p_{irt} is a post-treatment indicator. It equals 1 for every period after that group first received treatment. Formally, $p_{irt} = 1 \; \forall \; t > q_r - 1$, where q_r is the period which the group first received treatment. For example, given $t = 1,2,3$ and group 1 was first treated in period 2, then $q_1 = 2$. If group 2 first received treatment in period 3 then $q_2 = 3$. Finally, $\dot{\boldsymbol{x}}_{\boldsymbol{ir}}$ is the deviation from the cohort average for individual *i*.

Equation (7) is the average outcome if never treated in period 1. Equation (8) is the time trend, which allows for both a common trend component and heterogenous trend effects based on the predetermined covariates. Equation (9) is the treatment effect for those treated, which again allows for a common treatment effect component which is different for each cohort, and a heterogenous treatment effect for each individual based on the predetermined covariates.

Putting these together we have:

(10) $$E[\, y_{it} \mid d_{ir}, \boldsymbol{x_i} \,] = \eta + , \boldsymbol{x_i} \boldsymbol{\kappa} + \sum_{r=q}^{T} \lambda_r d_{ir} + \sum_{r=q}^{T} \zeta_r (d_{ir} \times , \boldsymbol{x_i}) + \sum_{s=2}^{T} \theta_s f s_t + \sum_{s=2}^{T} (f s_t \times \boldsymbol{x_i}) \boldsymbol{\pi_t} + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rt} (d_{ir} \times p_{irt} \times f s_t) + \sum_{r=q}^{T} \sum_{s=r}^{T} (d_{ir} \times p_{irt} \times f s_t \times \dot{\boldsymbol{x}}_{\boldsymbol{ir}}) \boldsymbol{\rho_{rt}}$$

If these assumptions hold, we can identify the effect of the lead reduction, τ_{rt}, for each group r, at time t, with the following model, using difference-in-

differences design with a two-way Mundlak regression of Wooldridge (2021):

$$(11)\quad y_{it} = \eta +, \boldsymbol{x_i}\boldsymbol{\kappa} + \sum_{r=q}^{T} \lambda_r d_{ir} + \sum_{r=q}^{T} \zeta_r (d_{ir} \times, \boldsymbol{x_i}) + \sum_{s=2}^{T} \theta_s f s_t + \sum_{s=2}^{T} (f s_t \times \boldsymbol{x_i})\boldsymbol{\pi_t} + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rt} (d_{ir} \times p_{irt} \times f s_t) + \sum_{r=q}^{T} \sum_{s=r}^{T} (d_{ir} \times p_{irt} \times f s_t \times \dot{\boldsymbol{x}}_{\boldsymbol{ir}})\boldsymbol{\rho_{rt}} + u_{it}$$

We use the two-way Mundlak over other staggered difference-in-difference estimators, such as Calloway and Sant'Anna (2021), because it is the most efficient *if* the above assumptions hold. The control group in the two-way Mundlak uses all the information for never-treated, and all the information until treatment for the later treated cohorts. Calloway and Sant'Anna (2021) estimation uses only the information from these groups in the period before treatment. Likewise, other estimators do not use the whole of the pre-treatment information and so throw away information.

If we wish to allow for more heterogeneity in time trends, we can model the time trend as simply $d_{ir} \times t$. This allows for heterogeneity in time trends and also allows us to test the common trends assumption with a Wald test jointly on all the coefficients of $d_{ir} \times t$.

2.4.3 Non-linear Difference-in-Differences Model

Equation (11) allows us to estimate the effect on birthweights, but, unless we adopt a Linear Probability Model, it is not suitable for infant deaths. These are binary outcomes. It is likely the LINS and CCTS assumptions are violated.

In this case we replace the CCTS assumption with an assumption of parallel relative trends.

Conditional Parallel Relative Trends, Staggered Treatment (CPRTS)

$$(12)\quad \frac{E\left[y_{it}(\infty)|d_{ir},\ \boldsymbol{x_i}\right]}{E\left[y_{i1}(\infty)|d_{ir},\ \boldsymbol{x_i}\right]} = \frac{E\left[y_{it}(\infty)|\boldsymbol{x_i}\right]}{E\left[y_{i1}(\infty)|\ \boldsymbol{x_i}\right]},\ t = 2,\dots,T, r = q,\dots,T$$

Equation (12) states that the ratio of average outcome if never-treated at time *t* compared to the first period average outcome only depends on the covariates. There is no selection into or out of treatment.

We replace the linear in parameters assumption with a pooled quasi-maximum likelihood logistic model.

$$(13)\quad E[\,y_{it}\mid d_{ir},\boldsymbol{x_i}\,] = \boldsymbol{\Lambda}\,[\,\eta +, \boldsymbol{x_i\kappa} + \sum_{r=q}^{T}\lambda_r d_{ir} + \sum_{r=q}^{T}\zeta_r(d_{ir}\ \times, \boldsymbol{x_i}) + \sum_{s=2}^{T}\theta_s fs_t + \sum_{s=2}^{T}(fs_t\ \times \boldsymbol{x_i})\boldsymbol{\pi_t} + \sum_{r=q}^{T}\sum_{s=r}^{T}\tau_{rt}(d_{ir}\ \times p_{irt}\times fs_t) + \sum_{r=q}^{T}\sum_{s=r}^{T}(d_{ir}\ \times p_{irt}\times fs_t\times \dot{\boldsymbol{x}}_{\boldsymbol{ir}})\boldsymbol{\rho_{rt}}\,]$$

Where $\boldsymbol{\Lambda}$ represents the logistic function. Equation (13), once again is estimated by two-way Mundlak, allows us to estimate the treatment effect on deaths. The treatment effect estimated is an average partial effect (APE) of being treated. That is, we estimate the model and then take the coefficients applicable for a particular year and cohort. We then get the expected value of the values with the treatment variable minus the values without including the treatment variable. We obtain standard errors for the APE with bootstrapping.

2.5 Results

We first plot the mean birthweight for Glasgow, Edinburgh and the control group in figure 2.10. In all groups there is a clear upward trend. The Edinburgh average tracks closely with the control group, while the Glasgow average remains below both at all times, and there is no convergence even after the 1978 and 1989 treatment. Both the Edinburgh and Glasgow

averages are more volatile than the control group. The trends appear similar for all groups. There is no clear treatment effect to be seen in the raw averages, but this may be due to differences in group characteristics that affect the treatment effect. We explore this in section 2.5 where we regress on covariates with a two-way Mundlak that controls for time-invariant fixed effects and time varying common effects.

Similarly, in figure 2.11 we plot the percentage of pregnancies that result in death for each group. As stated in section 2.3, the deaths variable includes stillbirths, neonatal deaths, spontaneous abortions, and all infant deaths up to age 5. The trends are again similar, but the percentage for Glasgow and Edinburgh is more volatile. No clear treatment effect is visible in the raw

Figure 2.10 Average Birthweight by Treatment Cohort, Grams

Notes: Chart shows the mean birthweight in grams of each birth in Edinburgh, Glasgow and the control group.

Figure 2.11 Under-5 Mortality rate by Treatment Cohort

Notes: Chart shows the total stillbirths, neonatal and under-5 death rates for all births in each birth in Edinburgh, Glasgow and the control group.

data, but once again this may be due to heterogeneity in group characteristics and therefore selection bias. We move on now to the difference-in-differences estimation.

2.5.1 Two-Way Mundlak Regressions

In table 2.2, we show the results for two-way Mundlak regressions as in equation (10) but excluding covariates $\boldsymbol{x_i}$, and all the interactions with $\boldsymbol{x_i}$. The dependent variable being birthweights and we use the full 1975-2000 sample. The columns show the estimated cohort treatment effects for each year, estimand (1), with the final row showing the average for all the years, estimand (2). The coefficients and 95% confidence intervals are also plotted in figure 2.12. The first column shows the estimated treatment effect for Glasgow, i.e. from the treatment on the Loch Katrine water supply. The majority of point-estimate coefficients are negative, which would imply the

treatment was harmful and reduced birthweights. However, almost all are not significant at the 5% confidence level. The overall average is significant and implies the average effect of treatment on the treated Glasgow group actually reduced birthweights by 13g-5g. This is, at most, 0.03 of a standard deviation and we submit that this is not socially significant. The second column shows the treatment effects for the combined Fairmilehead and the jointly supplied water areas in Edinburgh. Again the 95% intervals and point estimates are in figure 2.12. Only the effect in year 13 is statistically significant, and neither is there a clear trend in treatment effect. The overall average, however, is positive and statistically significant at the 5% level. It implies the treatment had an overall effect of increasing average birthweights by 23g-4g. Again, we would submit this as not being socially significant. Finally, the third column shows the same results for Alnwickhill. Notice there are fewer year treatment effects because Alnwickhill pH was not raised successfully until 1985. The two-way Mundlak regression method allows us to estimate the treatment effect for staggered treatments such as this. Looking at figure 2.12, we can see there is no clear effect. Although most point estimates are positive, the 95% intervals are wide and most cover zero. The average effect in table 2.2 is statistically significant at the 5% level and implies an average treatment effect of raising birthweights by 25g-3g. Similar to the effect in the Fairmilehead cohort.

We test the assumptions of common trends and no anticipation by including a cohort indicator interacted with a continuous time variable in the regression, as suggested in Wooldridge (2021). We carry out a Wald test of joint null effects on each cohort and time interaction. We reject the hypothesis of no anticipation or no common trends if the Wald test fails and finds the coefficients to be jointly statistically significant. The p-value for the test is 0.01, so we reject the hypotheses of no anticipation or common trends. It is possible these assumptions are too strong, and the data would be better described by conditional no anticipation and conditional common trends

assumptions, as described in section 2.4. Therefore, we next estimate the two-way Mundlak regression with a suit of covariates as in equation (10). The covariates are described in section 2.3.

The results of the two-way Mundlak with covariates are in table 2.3 and plotted in figure 2.13. Again, column 1 shows the Glasgow cohort treatment effects. There is no clear trend in the treatment. Most point estimates are negative but not significant. The overall average is negative and implies the treatment reduced birthweights by 25g-13g. Once again, although statistically significant we do not consider this to be socially significant. The second column show the Fairmilehead and joint treatment cohort. All the individual treatment estimates are small and not significant, as is the overall average. The third column is Alnwickhill treatment effects. The individual treatment effects show much the same pattern as the rest of the regressions, varying around zero with intervals that cover zero. The overall average is positive, but neither socially or statistically significant. We also perform a Wald test on the regressions by including the cohort and time variables. The p-value is 0.17, and therefore we do not reject the hypothesis of conditional no anticipation, or conditional common trends.

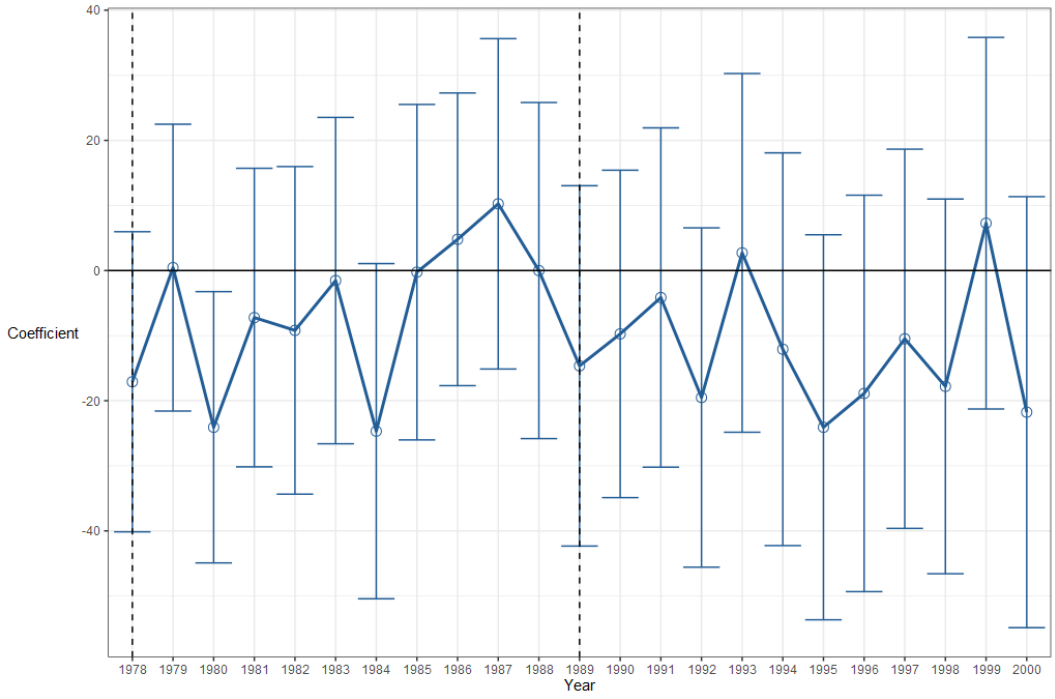
Table 2.2 Average Effect of Treatment-on-the-Treated on Birthweights in Grams, No Covariates

Year	Glasgow	Std Error	Fairmilehead	Std Error	Alnwickhill	Std Error
1978	-17.1	(11.8)	-15.2	(25.2)	-	-
1979	0.5	(11.2)	-14.8	(15.4)	-	-
1980	-24.1	(10.6)	18.0	(21.4)	-	-
1981	-7.2	(11.7)	-16.3	(21.9)	-	-
1982	-9.2	(12.8)	42.9	(25.1)	-	-
1983	-1.5	(12.8)	-12.2	(21.9)	-	-
1984	-24.7	(13.1)	10.3	(33.6)	-	-
1985	-0.2	(13.2)	25.8	(21.)	35.7	(15.9)
1986	4.8	(11.5)	-7.3	(22.9)	-5.6	(28.7)
1987	10.3	(13.)	11.7	(30.4)	31.3	(15.9)
1988	0.0	(13.2)	35.0	(24.9)	31.5	(18.4)
1989	-14.6	(14.1)	-1.4	(22.4)	-7.0	(17.5)
1990	-9.7	(12.8)	46.5	(22.8)	17.7	(22.8)
1991	-4.1	(13.3)	39.7	(20.7)	27.0	(23.6)
1992	-19.5	(13.3)	9.9	(26.2)	4.6	(25.8)
1993	2.7	(14.1)	-23.3	(21.3)	-47.2	(23.5)
1994	-12.1	(15.4)	-4.0	(27.4)	11.9	(28.3)
1995	-24.1	(15.1)	30.5	(23.9)	-3.8	(19.7)
1996	-18.9	(15.5)	46.2	(21.7)	37.3	(25.1)
1997	-10.5	(14.9)	11.1	(24.8)	20.0	(25.6)
1998	-17.8	(14.7)	4.5	(25.1)	43.8	(24.1)
1999	7.3	(14.6)	24.3	(24.5)	4.7	(23.3)
2000	-21.8	(16.9)	47.3	(30.8)	17.1	(32.4)
Average	**-9.2**	**(2.2)**	**13.5**	**(4.8)**	**13.7**	**(5.7)**

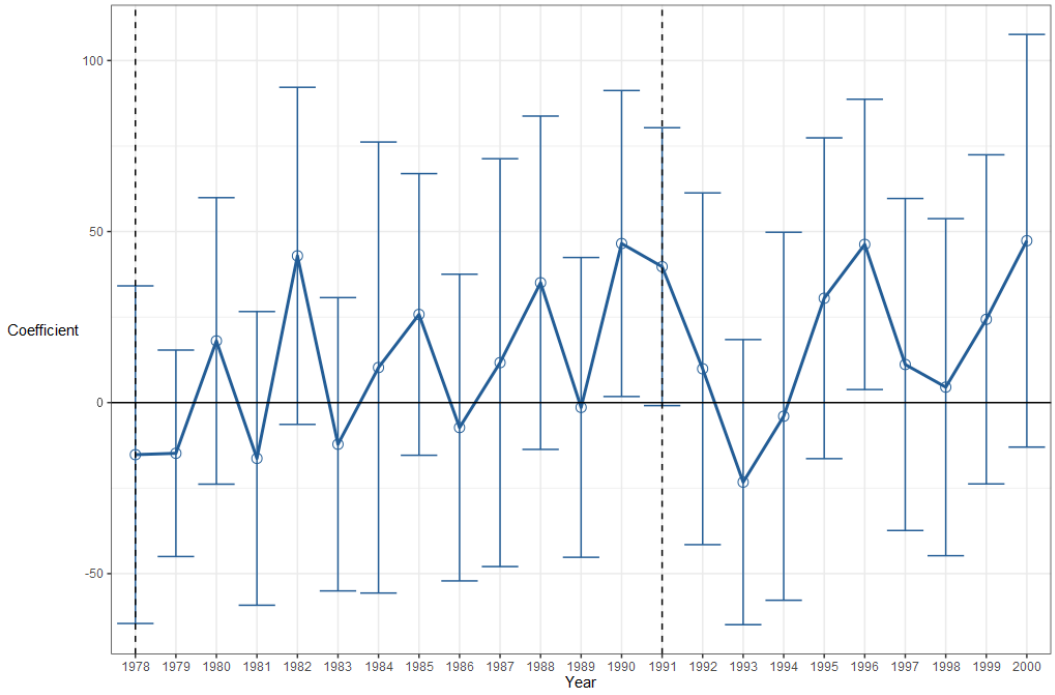
Notes: Table shows cohort specific treatment effects from two-way Mundlak regressions without covariates. Each year has an estimated treatment effect, and the bottom row is the mean of these. Robust standard errors, clustered by postcode sector, are in brackets.

Figure 2.12 Average Effect of Treatment-on-the-Treated on Birthweights, No Covariates

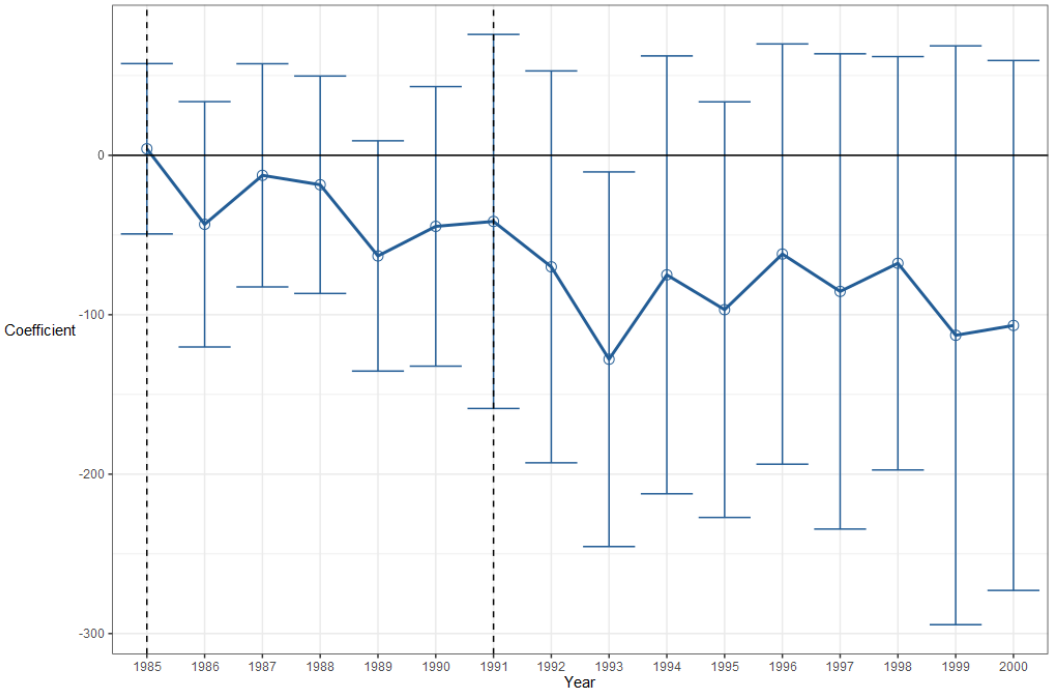
Glasgow



Fairmilehead



Alnwickhill



Notes: Figure shows each cohort's specific average treatment-on-the-treated effects for every *post-treatment* year (i.e. there are no leads in these charts) from two-way Mundlak regressions without covariates. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of treatment and has associated 95% confidence intervals calculated with robust standard errors, clustered by postcode sector.

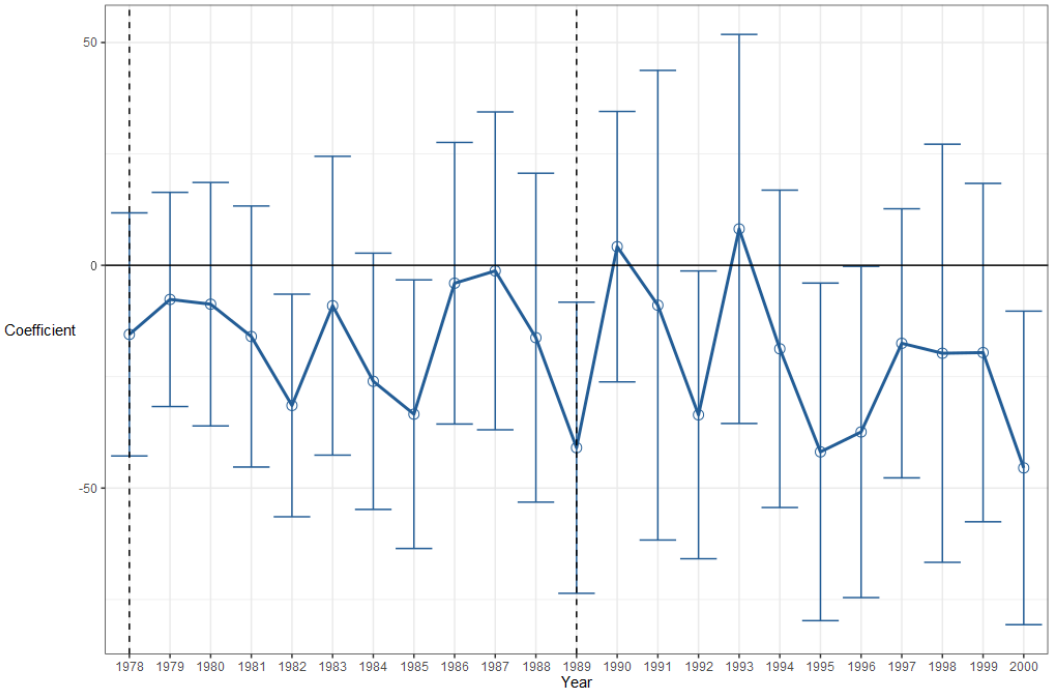
Table 2.3 Average Effect of Treatment-on-the-Treated on Birthweights, Covariates Included

Year	Glasgow	Std Error	Fairmilehead	Std Error	Alnwickhill	Std Error
1978	-15.5	(13.9)	-13.6	(22.9)	-	-
1979	-7.7	(12.3)	-5.2	(15.5)	-	-
1980	-8.7	(13.9)	12.6	(22.6)	-	-
1981	-16.0	(14.9)	-24.4	(22.)	-	-
1982	-31.5	(12.7)	33.3	(24.3)	-	-
1983	-9.1	(17.1)	-28.0	(22.7)	-	-
1984	-26.0	(14.7)	-11.5	(29.4)	-	-
1985	-33.4	(15.4)	11.3	(21.2)	44.9	(17)
1986	-4.0	(16.1)	-26.7	(19.)	-5.1	(26)
1987	-1.3	(18.2)	-2.9	(21.9)	22.7	(15.5)
1988	-16.2	(18.8)	16.8	(18.3)	33.9	(16.4)
1989	-40.9	(16.7)	-20.2	(19.)	-6.0	(18.3)
1990	4.2	(15.5)	32.8	(23.6)	21.3	(17.9)
1991	-9.0	(26.9)	27.0	(20.7)	23.0	(24.9)
1992	-33.6	(16.5)	-20.1	(24.6)	-0.8	(25.6)
1993	8.2	(22.3)	-34.8	(21.7)	-43.2	(21.5)
1994	-18.7	(18.2)	-28.3	(29.6)	7.8	(29.8)
1995	-41.9	(19.3)	-9.0	(22.9)	-17.3	(21.9)
1996	-37.4	(19.)	15.4	(22.)	25.1	(25.8)
1997	-17.5	(15.4)	-14.0	(25.9)	7.6	(27.6)
1998	-19.7	(23.9)	-28.6	(26.9)	44.5	(23.1)
1999	-19.6	(19.4)	-3.3	(19.5)	-10.5	(23.6)
2000	-45.5	(18.)	12.7	(32.1)	16.6	(31.9)
Average	**-19.2**	**(3.1)**	**-4.7**	**(4.4)**	**10.3**	**(5.9)**

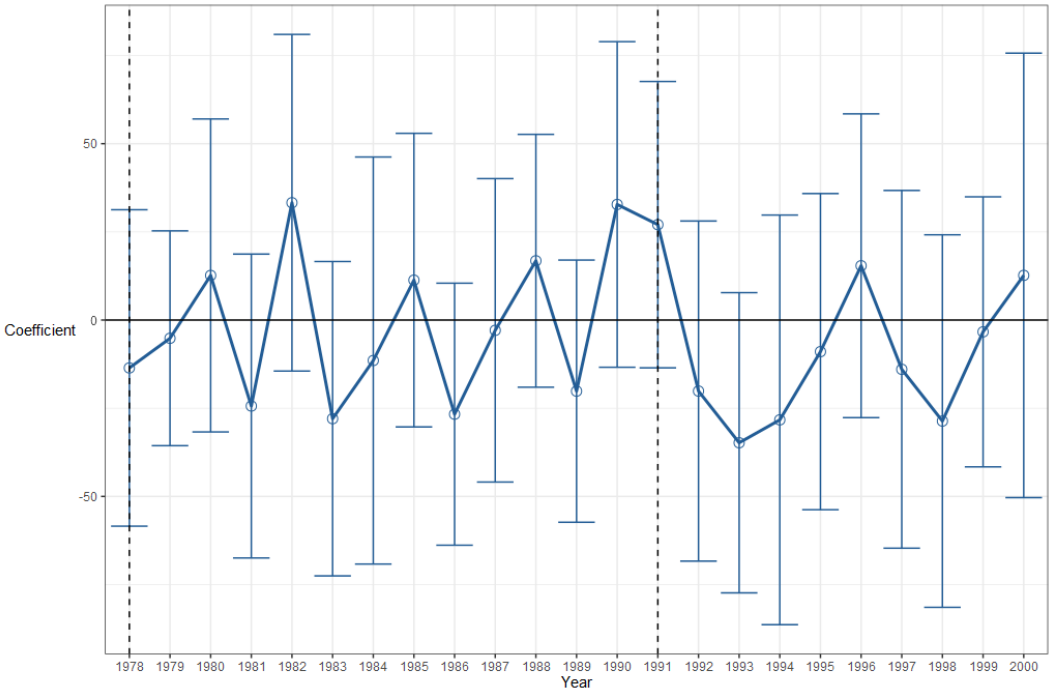
Notes: Table shows cohort specific treatment effects from two-way Mundlak regressions with covariates included in the regression. Each year has an estimated treatment effect, and the bottom row is the mean of these. Robust standard errors, clustered by postcode sector, are in brackets.

Figure 2.13 Average Effect of Treatment-on-the-Treated on Birthweights, Covariates Included

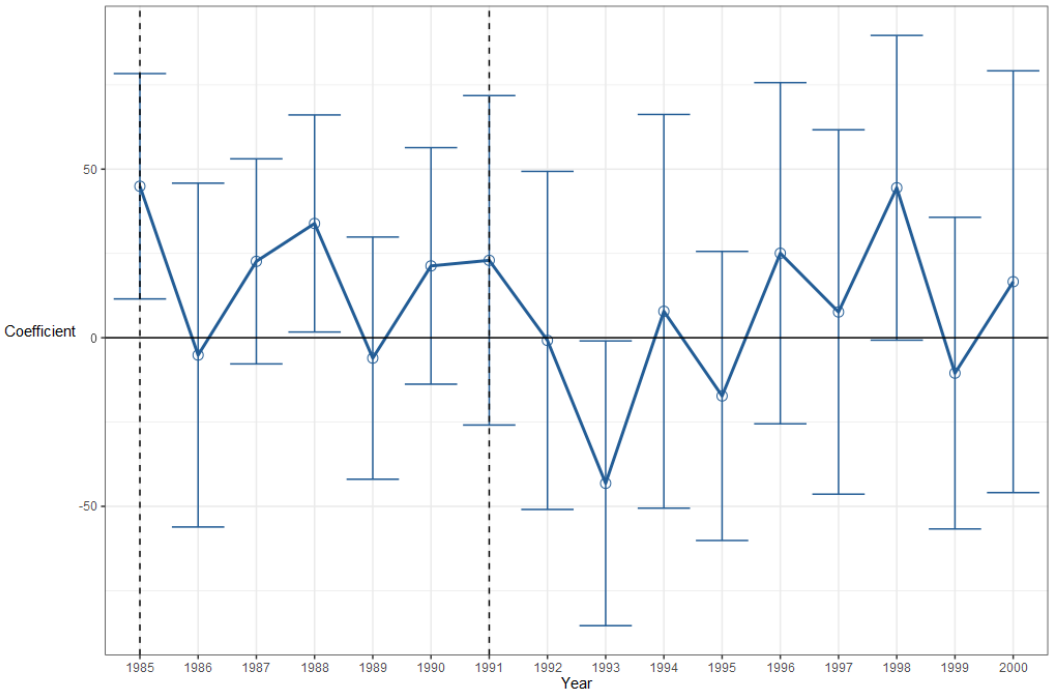
Glasgow



Fairmilehead



Alnwickhill



Notes: Figure shows each cohort's specific average treatment-on-the-treated effects for every *post-treatment* year (i.e. there are no leads in these charts) from two-way Mundlak regressions with covariates. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of treatment and has associated 95% confidence intervals calculated with robust standard errors, clustered by postcode sector.

Overall, the results imply no socially significant effect of the various water treatments on birthweights, for any cohort. We next look at the child deaths outcome.

Table 2.4 and figure 2.14 present the results from two-way Mundlak regression on mortality without any $\boldsymbol{x_i}$ covariates. All standard errors are bootstrapped to account for sampling heterogeneity in the APE. In the Glasgow results we can see a clear negative treatment average partial effect. The effect is much stronger in the earlier years before fading out. The overall APE is also negative and statistically significant. It implies the treatment lowered deaths by between 0.3 and 0.1 percentage points. Given that the

highest percent of pregnancies that result in death in Glasgow observed in our sample was around 1.25%, this is a substantial and socially significant effect. However, when we examine the Fairmilehead cohort, the majority of point estimates of the APE are positive, although not statistically significant. The overall APE is positive but not statistically significant at the 5% level (figures in table are rounded). A similar pattern is shown in the Alnwickhill cohort, where the overall APE is neither significant statistically nor socially significant. A Wald test using heterogeneous cohort trend variables as before does not reject the joint nullity of these variables, with the p-value being 0.72. We therefore do not reject the No Anticipation and Common Trends assumptions.

Next, we test the APE for the full two-way Mundlak with covariates as in equation (13). For Glasgow, the majority of year treatment effects are negative. The average estimated APE implies a smaller effect than without covariates. The estimate of the APE ranges from a large 0.4 percentage point decrease in deaths to a 0.01 increase in deaths. For the Fairmilehead cohort there is no clear pattern in the individual cohort-year treatment effects, but the overall average is negative, as we would expect, but neither not statistically significant, and much smaller than the Glasgow average treatment effect. For the Alnwickhill cohort, now almost all the cohort-year treatment point estimates are negative. The overall average is negative and implies the treatment led to a reduction in deaths of 0.7-0.1 percentage points. A substantial decrease. Once again, a Wald test does not reject the nullity of the heterogeneous cohort time trend variables, and we therefore do not reject the No Anticipation, Conditional assumption or the Common Trends, Conditional assumption (p-value 0.26).

Overall, our main results imply no evidence of a treatment effect for birthweights, but we do find evidence for a possibly strong effect on mortality reduction, at least for Glasgow and the Alnwickhill supplied area of

Edinburgh. Evidence for an effect on lower deaths in the Fairmilehead supplied area of Edinburgh is more ambiguous.

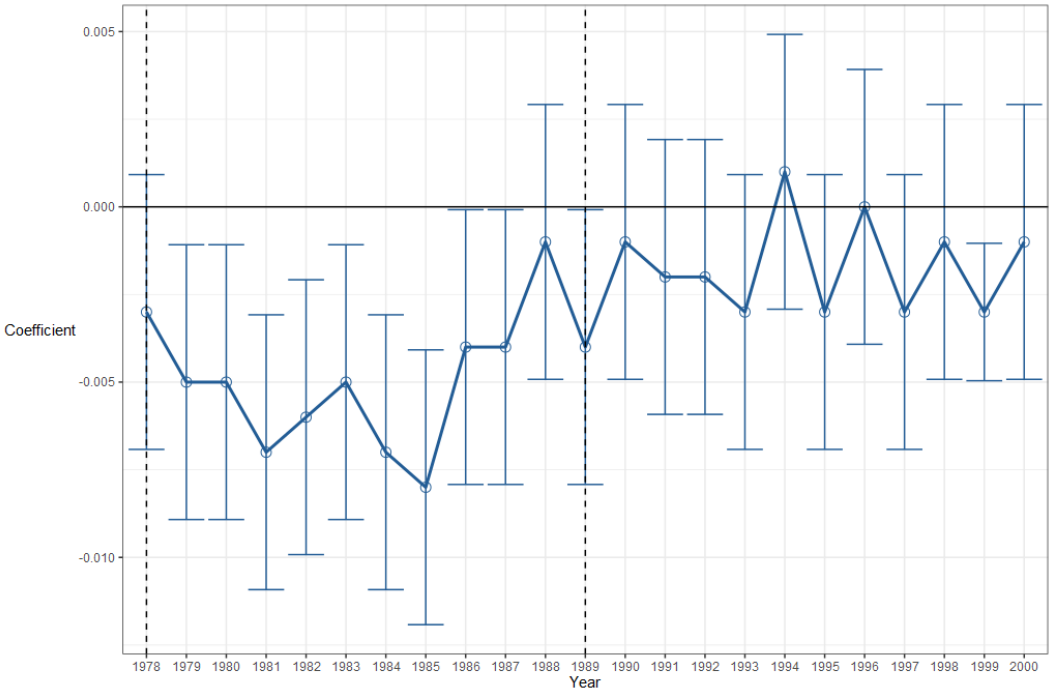
Table 2.4 Average Partial Effect of Treatment on Deaths before Age 5, No Covariates

Year	Glasgow	Std Error	Fairmilehead	Std Error	Alnwickhill	Std Error
1978	-0.003	(0.002)	0.004	(0.003)	-	-
1979	-0.005	(0.002)	0.002	(0.003)	-	-
1980	-0.005	(0.002)	-0.002	(0.003)	-	-
1981	-0.007	(0.002)	0.004	(0.003)	-	-
1982	-0.006	(0.002)	-0.003	(0.002)	-	-
1983	-0.005	(0.002)	0.002	(0.002)	-	-
1984	-0.007	(0.002)	0.007	(0.003)	-	-
1985	-0.008	(0.002)	0.000	(0.003)	-0.003	(0.003)
1986	-0.004	(0.002)	0.002	(0.003)	-0.006	(0.002)
1987	-0.004	(0.002)	0.006	(0.003)	-0.003	(0.003)
1988	-0.001	(0.002)	-0.001	(0.002)	-0.003	(0.003)
1989	-0.004	(0.002)	-0.001	(0.003)	-0.002	(0.003)
1990	-0.001	(0.002)	0.004	(0.003)	0.001	(0.003)
1991	-0.002	(0.002)	0.003	(0.003)	-0.001	(0.003)
1992	-0.002	(0.002)	0.005	(0.003)	0.006	(0.004)
1993	-0.003	(0.002)	0.001	(0.002)	-0.002	(0.003)
1994	0.001	(0.002)	0.007	(0.003)	0.003	(0.003)
1995	-0.003	(0.002)	0.002	(0.003)	0.001	(0.003)
1996	0.000	(0.002)	0.000	(0.002)	0.004	(0.003)
1997	-0.003	(0.002)	0.001	(0.002)	-0.002	(0.002)
1998	-0.001	(0.002)	0.005	(0.003)	-0.003	(0.002)
1999	-0.003	(0.001)	0.007	(0.002)	0.006	(0.003)
2000	-0.001	(0.002)	0.003	(0.002)	0.002	(0.003)
Average APE	**-0.003**	**(0.001)**	**0.003**	**(0.001)**	**0.000**	**(0.001)**

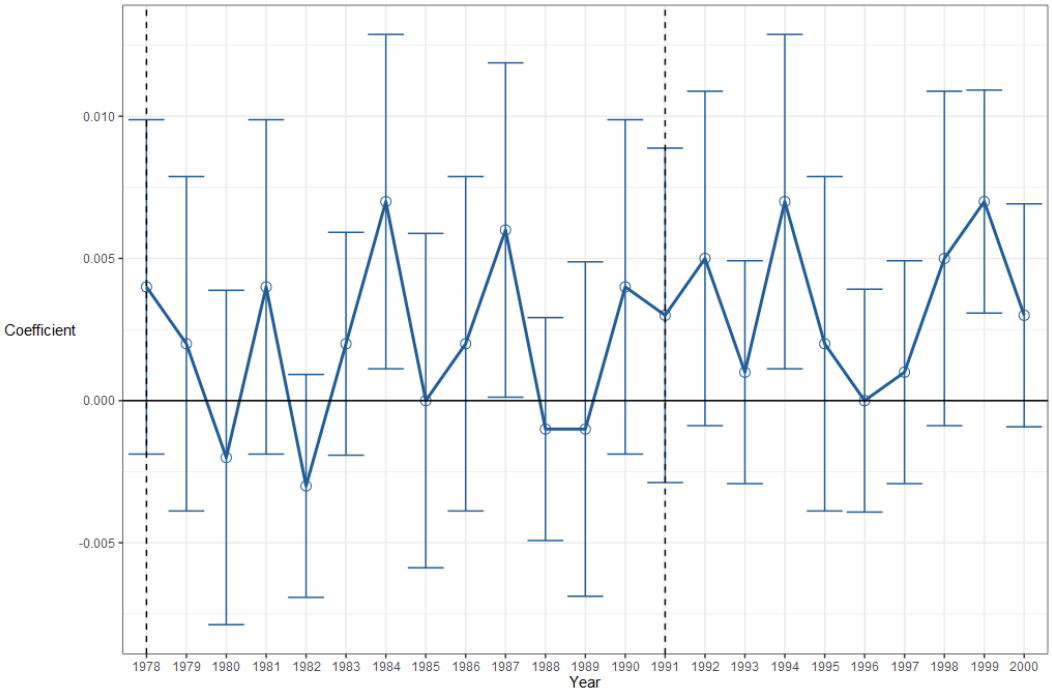
Notes: Table shows cohort specific Average Partial Effects (APE). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method without covariates. The relevant cohort and year indicators are set to 1, and the difference with and without the cohort specific treatment indicator is taken. Each year has an estimated APE, and the bottom row is the mean of these. Standard errors of the APEs are bootstrapped.

Figure 2.14 Average Partial Effect of Treatment on Deaths before Age 5, No Covariates

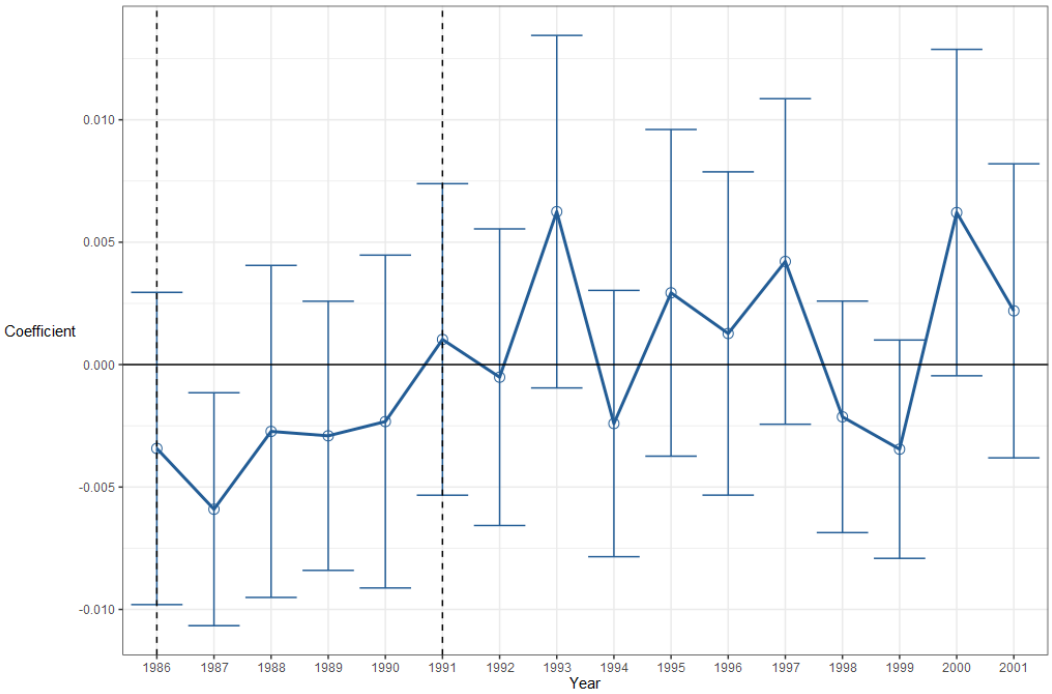
Glasgow



Fairmilehead



Alnwickhill



Notes: Figure shows each cohort's specific average partial effect (APE) for every *post-treatment* year (i.e. there are no leads in these charts). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method without covariates. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of the APE and has associated 95% confidence intervals calculated with bootstrapped standard errors.

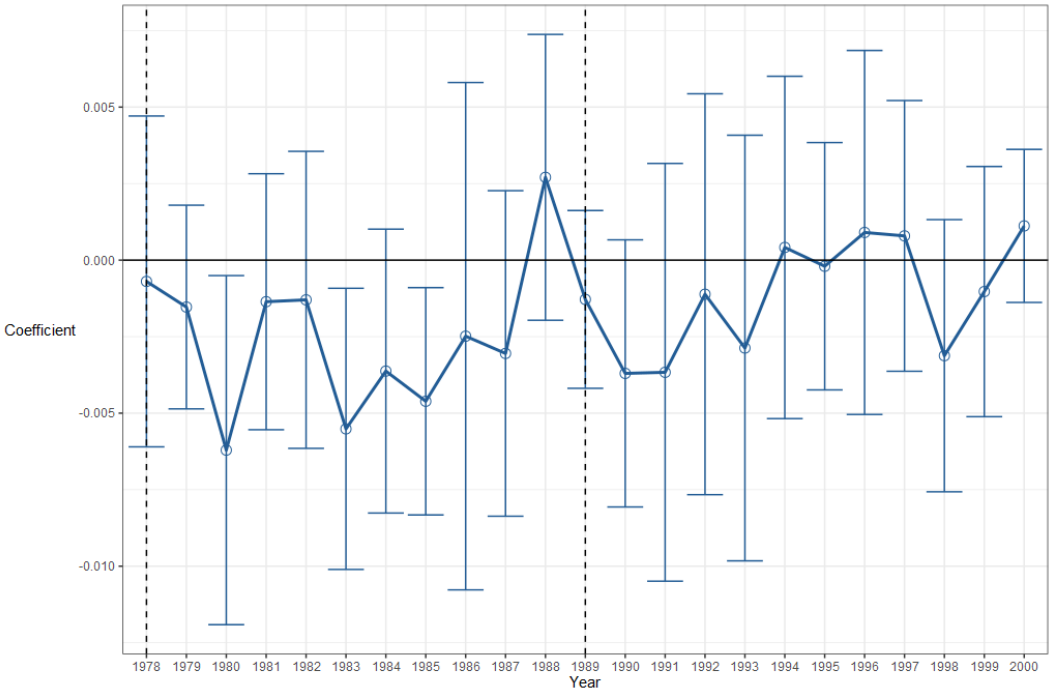
Table 2.5 Average Partial Effect of Treatment on Deaths before Age 5, Covariates Included

Year	Glasgow	Std Error	Fairmilehead	Std Error	Alnwickhill	Std Error
1978	-0.001	(0.003)	0.002	(0.003)	-	-
1979	-0.002	(0.002)	-0.001	(0.002)	-	-
1980	-0.006	(0.003)	-0.005	(0.002)	-	-
1981	-0.001	(0.002)	0.003	(0.004)	-	-
1982	-0.001	(0.002)	-0.006	(0.001)	-	-
1983	-0.006	(0.002)	0.001	(0.002)	-	-
1984	-0.004	(0.002)	-0.006	(0.001)	-	-
1985	-0.005	(0.002)	-0.006	(0.002)	-0.008	(0.003)
1986	-0.002	(0.004)	-0.005	(0.001)	-0.008	(0.002)
1987	-0.003	(0.003)	0.000	(0.003)	-0.005	(0.003)
1988	0.003	(0.002)	-0.001	(0.002)	-0.005	(0.004)
1989	-0.001	(0.001)	-0.001	(0.002)	-0.008	(0.003)
1990	-0.004	(0.002)	0.002	(0.003)	-0.008	(0.002)
1991	-0.004	(0.003)	0.002	(0.002)	-0.005	(0.003)
1992	-0.001	(0.003)	0.005	(0.003)	0.005	(0.004)
1993	-0.003	(0.004)	-0.001	(0.002)	-0.006	(0.003)
1994	0.000	(0.003)	0.005	(0.005)	-0.008	(0.002)
1995	0.000	(0.002)	-0.002	(0.002)	-0.003	(0.004)
1996	0.001	(0.003)	-0.004	(0.001)	0.003	(0.004)
1997	0.001	(0.002)	0.001	(0.002)	-0.003	(0.002)
1998	-0.003	(0.002)	0.004	(0.002)	-0.005	(0.002)
1999	-0.001	(0.002)	0.003	(0.003)	0.000	(0.002)
2000	0.001	(0.001)	0.002	(0.003)	0.000	(0.004)
Average APE	**-0.002**	**(0.002)**	**0.000**	**(0.001)**	**-0.004**	**(0.001)**

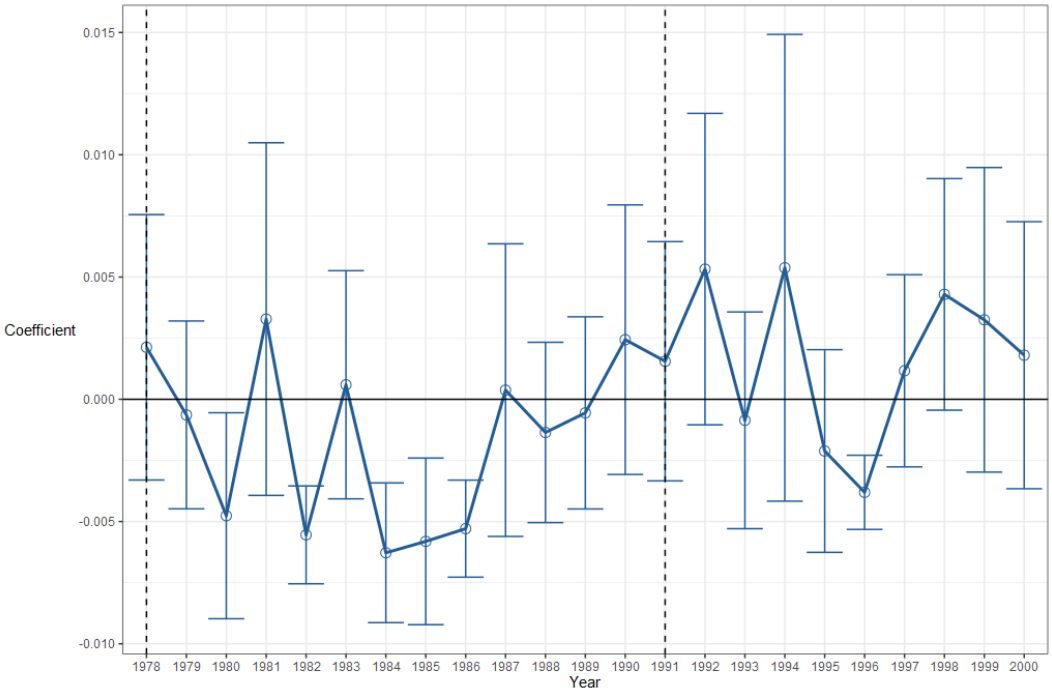
Notes: Table shows cohort specific Average Partial Effects (APE). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method and including covariates. The relevant cohort and year indicators are set to 1, the relevant covariates indicators are set to1, continuous covariate variables are set to the cohort mean value for that covariate, and the difference with and without the cohort specific treatment indicator is taken. Each year has an estimated APE, and the bottom row is the mean of these. Standard errors of the APEs are bootstrapped.

Figure 2.15 Average Partial Effect of Treatment on Deaths before Age 5, Covariates Included

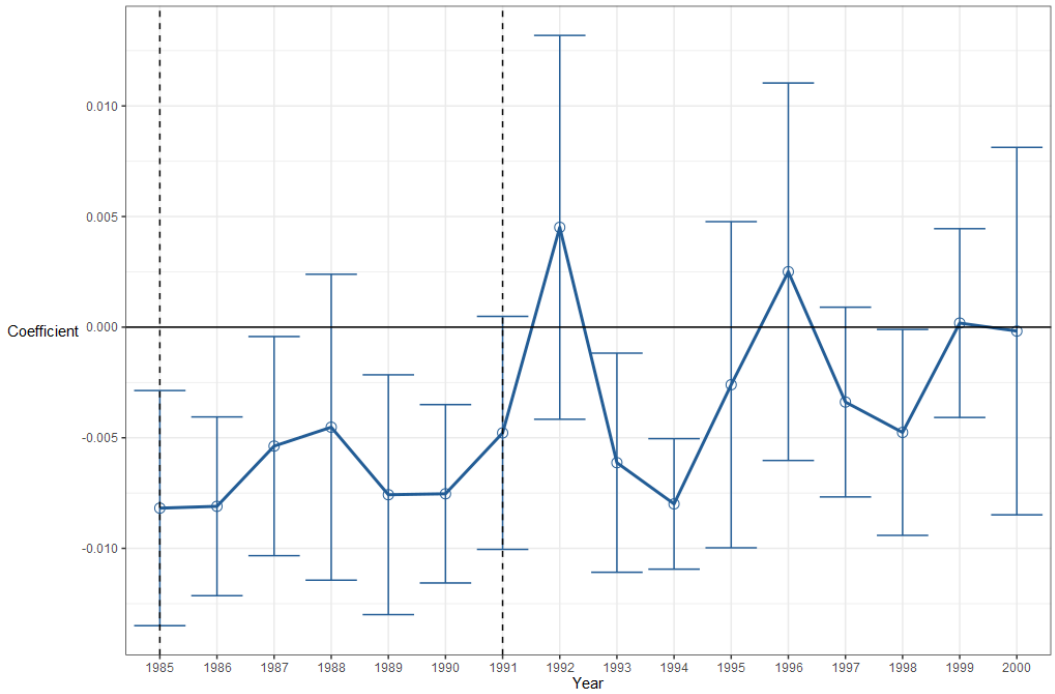
Glasgow



Fairmilehead



Alnwickhill



Notes: Figure shows each cohort's specific average partial effect (APE) for every *post-treatment* year (i.e. there are no leads in these charts). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method with covariates. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of the APE and has associated 95% confidence intervals calculated with bootstrapped standard errors.

2.6 Robustness checks

We carry out a number of robustness checks on our results.

2.6.1 Two-Way Mundlak Robustness Checks

In our main results the area in Edinburgh served with water by both the Alnwickhill and Fairmilehead plants, i.e. the "joint" area in figure 2.5, is included in the Fairmilehead cohort. This is due to households in the area having an average water pH above 8, much closer to the Fairmilehead pH, in

Table 2.6 Fairmilehead Birthweight Results Excluding Joint Water Supply Area, Average Treatment on the Treated

Year	Birthweight, No Covariates	Std Error	Birthweight, Covariates	Std Error
1978	12.4	(33)	42.8	(25.3)
1979	10.6	(25.1)	-6.1	(18.2)
1980	3.9	(36.8)	-31.3	(25.7)
1981	-26.8	(32.1)	-60.4	(24)
1982	30.5	(49.8)	-37.9	(38.2)
1983	-62.5	(19.2)	-68.8	(19.4)
1984	-25.6	(39.7)	-20.0	(32.2)
1985	-7.9	(46.2)	-75.7	(44.)
1986	-25.6	(20.6)	-79.6	(11.5)
1987	-69.3	(47.7)	-123.4	(45.1)
1988	31.8	(28.8)	-4.2	(15.4)
1989	-9.2	(30.6)	-51.2	(19.1)
1990	36.8	(50.1)	1.5	(34.6)
1991	23.1	(50.5)	-42.5	(26.6)
1992	-8.9	(50.9)	-113.5	(17.1)
1993	-66.4	(23.6)	-78.8	(22.2)
1994	-5.9	(48.5)	17.4	(45.5)
1995	26.0	(38.6)	-46.2	(21.9)
1996	84.7	(30.4)	54.1	(15.8)
1997	-21.8	(41.1)	-90.7	(18.5)
1998	-66.1	(41)	-104.0	(47.6)
1999	11.8	(36.1)	11.5	(29.2)
2000	18.8	(58.5)	-27.0	(49)
Average	**-4.6**	**(8)**	**-40.6**	**(10.1)**

Notes: Table shows cohort specific treatment effects from two-way Mundlak regressions with and without covariates included in the regression. Sample is set to exclude all observations jointly served water by the Fairmilehead and Alnwickhill plant. Each year has an estimated treatment effect, and the bottom row is the mean of these. Robust standard errors, clustered by postcode sector, are in brackets.

Table 2.7 Fairmilehead Deaths Before Age 5 Results Excluding Joint Water Supply Area, Average Partial Effect

Year	Deaths, No Covariates	Std Error	Deaths, Covariates	Std Error
1978	0.003	(0.007)	0.000	(0.000)
1979	0.002	(0.004)	0.000	(0.000)
1980	-0.006	(0.002)	0.000	(0.000)
1981	0.008	(0.007)	0.000	(0.000)
1982	-0.005	(0.002)	0.000	(0.000)
1983	0.006	(0.008)	0.000	(0.000)
1984	0.013	(0.01)	0.000	(0.000)
1985	-0.006	(0.002)	0.000	(0.000)
1986	0.013	(0.006)	0.000	(0.000)
1987	0.014	(0.011)	0.000	(0.000)
1988	0.003	(0.004)	0.000	(0.000)
1989	-0.001	(0.005)	0.001	(0.002)
1990	-0.002	(0.002)	0.000	(0.000)
1991	0.005	(0.004)	0.000	(0.000)
1992	0.003	(0.004)	0.002	(0.002)
1993	0.007	(0.006)	0.000	(0.000)
1994	-0.005	(0.002)	0.000	(0.000)
1995	0.000	(0.006)	0.000	(0.000)
1996	-0.001	(0.003)	0.000	(0.000)
1997	-0.001	(0.004)	0.000	(0.000)
1998	0.007	(0.007)	0.000	(0.000)
1999	0.009	(0.005)	0.000	(0.000)
2000	0.003	(0.004)	0.000	(0.000)
Average APE	0.003	(0.002)	0.000	(0.000)

Notes: Table shows cohort specific Average Partial Effects (APE). Sample is set to exclude all observations jointly served water by the Fairmilehead and Alnwickhill plant. These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method, with and without covariates. The relevant cohort and year indicators are set to 1, the relevant covariates indicators are set to1, continuous covariate variables are set to the cohort mean value for that covariate, and the difference with and without the cohort specific treatment indicator is taken. Each year has an estimated APE, and the bottom row is the mean of these. Standard errors of the APEs are bootstrapped.

1985 (see figure 2.6). However, including this group may mean our Fairmilehead results are not identified, as they do not receive identical treatments. We therefore exclude them from the two-way Mundlak regressions to see the effect on the results. These are shown in table 2.6 and table 2.7.

Table 2.6 shows the effect on birthweight with and without covariates. Both averages are small and negative. Implying the treatment reduced birthweights. Table 2.7 shows the results for deaths. Without covariates they show a small increase in deaths after treatment, but this is not statistically significant. With covariates we obtain a precise null effect for all years. These results are qualitatively similar to those when we include the joint water treatment area. In summary, we believe this shows it is not the inclusion of the joint treatment area in the Fairmilehead cohort that leads to no effect being found for Fairmilehead.

Next we examine the Orthophosphate intervention in isolation. That is, we only regress on outcomes that happen after all pH interventions are over. This means the treatment baseline is when all treated units have their pH raised to acceptable levels (after 1985). Therefore, the treatment of orthophosphate can be examined independently of the pH level increases. The Edinburgh group can be treated as one cohort as they all receive orthophosphate treatment at the same time, and the pH treatments have already happened.

Table 2.8 and 2.9, and figures 2.15 and 2.16 examine the regressions when we only look at the years 1986-2000, and therefore only at the orthophosphate treatment, for the birthweight outcome with and without covariates. Table 2.8, column 1 shows the treatment effects for Glasgow. The overall average is positive, but small and not significant. The biggest effects seem to be towards the end of the sample, but the 95% confidence intervals also cover zero. Table 2.9, column 1 shows the effect for Glasgow with

covariates. The results are much the same with the overall average positive but small and not significant. Column 2 shows the results for Edinburgh. Table 2.8, without covariates, shows an overall average that is negative, but small and not significant. Table 2.9 is qualitatively similar, with a small and insignificant average treatment effect.

Table 2.8 Average Effect of Treatment-on-the-Treated on Birthweights Orthophosphate Treatment Only, No Covariates

Year	**Glasgow**	**Std Error**	**Edinburgh**	**Std Error**
1989	-2.4	(12)	-	-
1990	-3.4	(12.5)	-	-
1991	8.4	(11.3)	15.7	(19.1)
1992	-8.5	(13.2)	-9.0	(19.7)
1993	14.8	(11.6)	-52.1	(13.8)
1994	-8.4	(13.9)	-18.1	(24.7)
1995	-4.7	(14.6)	2.8	(16.9)
1996	-6.7	(15.)	26.0	(18.2)
1997	6.3	(13.3)	-3.3	(20.)
1998	13.2	(13.5)	4.6	(20.4)
1999	20.5	(13.9)	-4.1	(20.1)
2000	15.8	(14.2)	23.8	(23.2)
Average	**3.7**	**(3.1)**	**-1.4**	**(7.2)**

Notes: Table shows cohort specific treatment effects from two-way Mundlak regressions without covariates included in the regression. Sample is restricted to 1985-2000. Each year has an estimated treatment effect, and the bottom row is the mean of these. Robust standard errors, clustered by postcode sector, are in brackets.

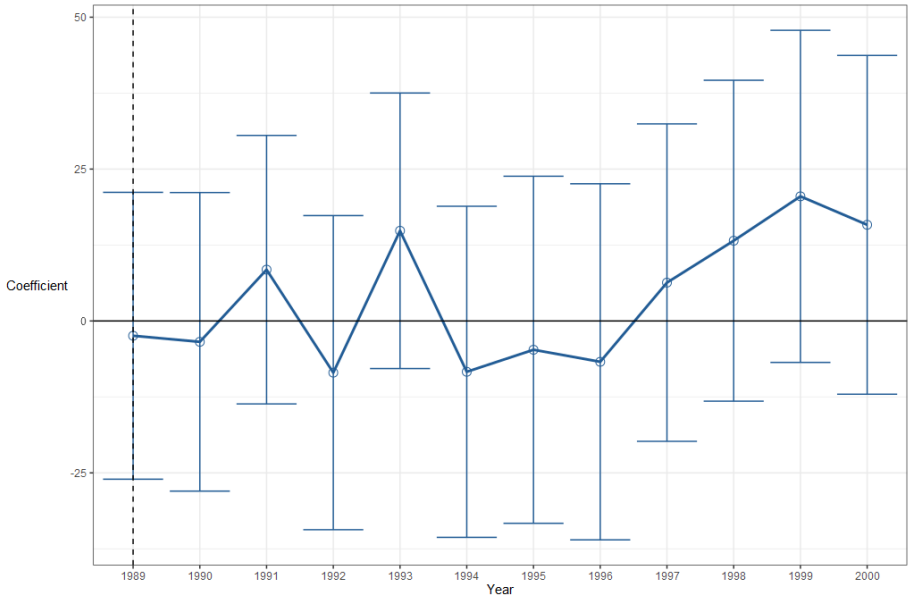
Table 2.9 Average Effect of Treatment-on-the-Treated on Birthweights Orthophosphate Treatment Only, Covariates Included

Year	Glasgow	Std Error	Edinburgh	Std Error
1989	-10.3	(12)	-	-
1990	10.1	(13.3)	-	-
1991	11.2	(12.9)	13.6	(17.1)
1992	-7.5	(12.7)	-21.4	(18.1)
1993	14.9	(12.1)	-50.8	(14.3)
1994	-10.0	(14.4)	-24.7	(26.4)
1995	-3.4	(13.8)	-17.7	(17)
1996	-9.0	(15.8)	10.6	(17.5)
1997	6.2	(13.3)	-16.8	(19.7)
1998	17.8	(14.8)	-9.5	(19.8)
1999	3.1	(15.5)	-16.4	(20.)
2000	9.8	(14.5)	10.7	(23.5)
Average	**2.7**	**(3)**	**-12.3**	**(6.2)**

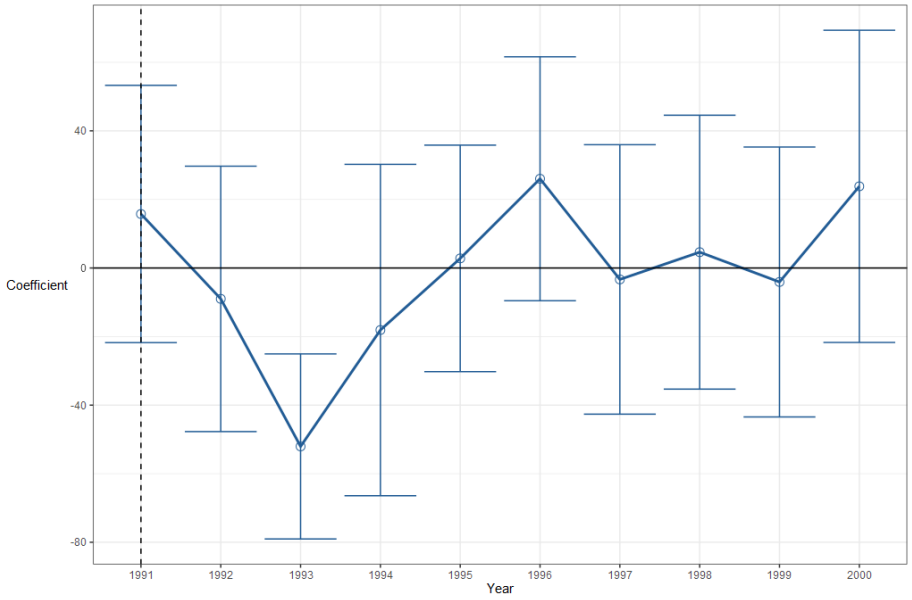
Notes: Table shows cohort specific treatment effects from two-way Mundlak regressions with covariates included in the regression. Sample is restricted to 1985-2000. Each year has an estimated treatment effect, and the bottom row is the mean of these. Robust standard errors, clustered by postcode sector, are in brackets.

Figure 2.16 Average Effect of Treatment-on-the-Treated on Birthweights Orthophosphate Treatment Only, No Covariates

Glasgow



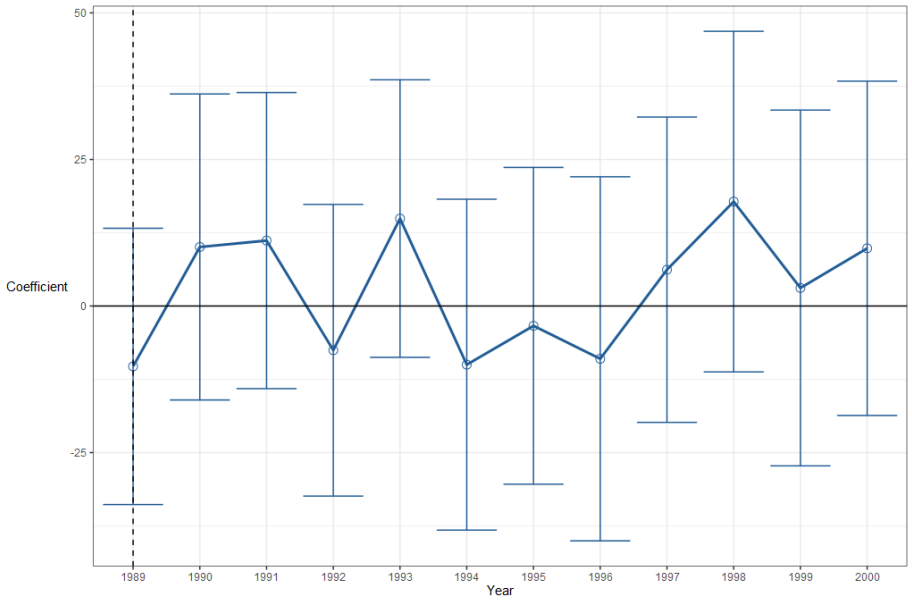
Edinburgh



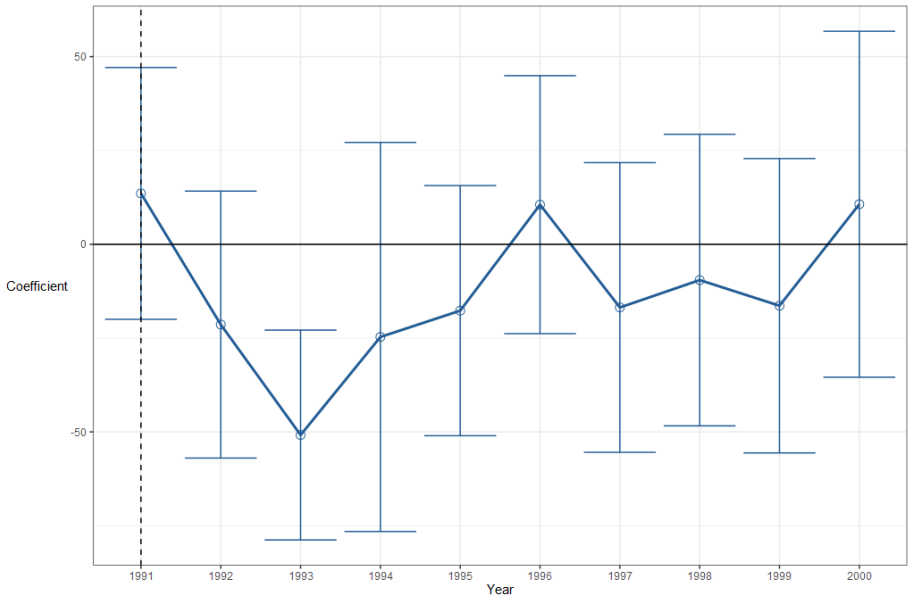
Notes: Figure shows each cohort's specific average treatment-on-the-treated effects for every *post-treatment* year (i.e. there are no leads in these charts) from two-way Mundlak regressions without covariates. Sample is restricted to 1985-2000. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of treatment and has associated 95% confidence intervals calculated with robust standard errors, clustered by postcode sector.

Figure 2.17 Average Effect of Treatment-on-the-Treated on Birthweights Orthophosphate Treatment Only, Covariates Included

Glasgow



Edinburgh



Notes: Figure shows each cohort's specific average treatment-on-the-treated effects for every *post-treatment* year (i.e. there are no leads in these charts) from two-way Mundlak regressions with covariates. Sample is restricted to 1985-2000. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of treatment and has associated 95% confidence intervals calculated with robust standard errors, clustered by postcode sector.

We repeat the logistic regression on deaths for Edinburgh and Glasgow using only the 1986-2000 sample in figure 2.17 and table 2.9, with and without covariates. Column 1 shows the APEs for the Glasgow orthophosphate treatment without covariates in the regression. Most point estimate APEs are negative as expected, the overall average point estimate is negative but close to zero. The 95% interval implies an effect from decreasing deaths by 0.1 percentage points to increasing them by 0.06 percentage points. When we include covariates, the overall APE becomes larger in magnitude and with a more precise interval, implying the orthophosphate treatment reduced deaths by 0.1-0.03 percentage points. For Edinburgh, the two point estimate overall APEs have the opposite sign from expected, implying treatment increased deaths. Without covariates it implies an increase from 0.3-0.06 percentage points. However, when covariates are includes this is no longer statistically significant, with the 95% range being from decreasing deaths by 0.02 to increasing deaths by 0.03 percentage points. In summary, when examining only the orthophosphate treatment, we find no evidence for an effect on birthweights, and any effect on deaths is confined to the Glasgow sample.

Table 2.10 Average Partial Effect of Treatment on Deaths before Age 5, Orthophosphate Treatment Only, No Covariates

Year	Glasgow	Std Error	Edinburgh	Std Error
1989	-0.001	(0.002)	-	-
1990	0.000	(0.001)	-	-
1991	-0.001	(0.001)	0.001	(0.002)
1992	0.002	(0.002)	0.005	(0.002)
1993	-0.003	(0.002)	-0.001	(0.002)
1994	0.005	(0.002)	0.005	(0.001)
1995	-0.002	(0.001)	0.001	(0.002)
1996	0.000	(0.002)	0.000	(0.002)
1997	-0.004	(0.001)	-0.001	(0.001)
1998	-0.001	(0.002)	0.001	(0.001)
1999	-0.001	(0.002)	0.007	(0.001)
2000	0.000	(0.002)	0.002	(0.001)
Average	**0.000**	**(0.000)**	**0.002**	**(0.001)**

Notes: Table shows cohort specific Average Partial Effects (APE). Sample is restricted to 1985-2000. These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method, without covariates. The relevant cohort and year indicators are set to 1, and the difference with and without the cohort specific treatment indicator is taken. Each year has an estimated APE, and the bottom row is the mean of these. Standard errors of the APEs are bootstrapped.

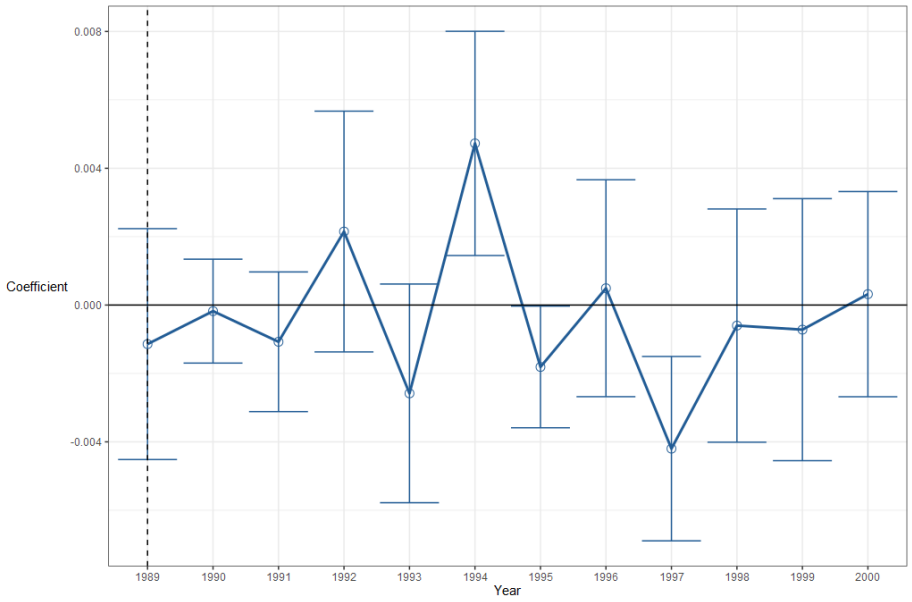
Table 2.11 Average Partial Effect of Treatment on Deaths before Age 5, Orthophosphate Treatment Only, Covariates Included

Year	Glasgow	Std Error	Edinburgh	Std Error
1989	-0.005	(0.002)	-	-
1990	-0.003	(0.002)	-	-
1991	-0.003	(0.002)	0.000	(0.002)
1992	0.003	(0.002)	0.005	(0.002)
1993	-0.003	(0.002)	-0.002	(0.003)
1994	0.003	(0.002)	0.002	(0.004)
1995	-0.002	(0.002)	-0.004	(0.001)
1996	0.000	(0.002)	0.000	(0.002)
1997	-0.004	(0.001)	-0.001	(0.001)
1998	-0.002	(0.002)	0.001	(0.002)
1999	0.000	(0.001)	0.004	(0.002)
2000	0.001	(0.001)	0.000	(0.002)
Average	**-0.001**	**(0.001)**	**0.000**	**(0.001)**

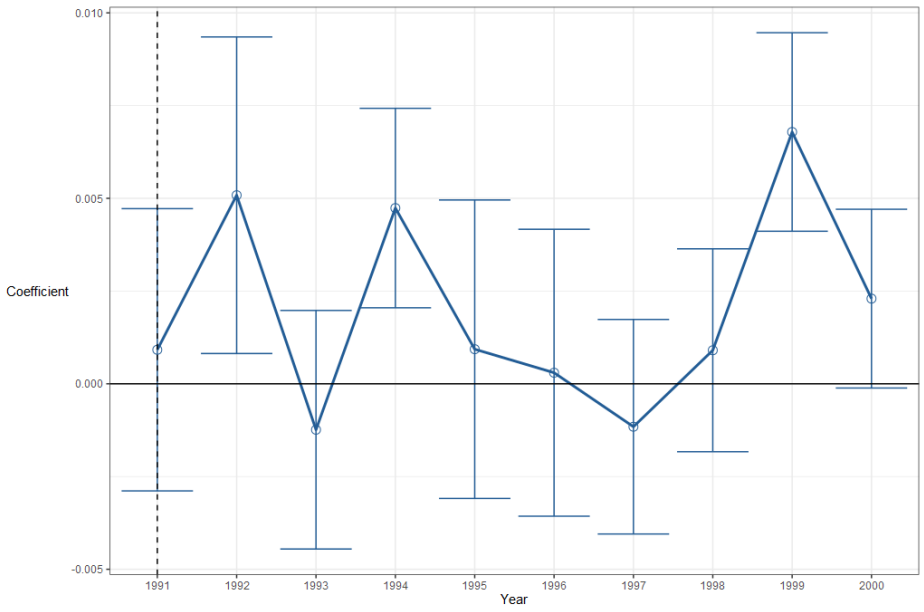
Notes: Table shows cohort specific Average Partial Effects (APE). Sample is restricted to 1985-2000. These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method and including covariates. The relevant cohort and year indicators are set to 1, the relevant covariates indicators are set to1, continuous covariate variables are set to the cohort mean value for that covariate, and the difference with and without the cohort specific treatment indicator is taken. Each year has an estimated APE, and the bottom row is the mean of these. Standard errors of the APEs are bootstrapped.

Figure 2.18 Average Partial Effect of Treatment on Deaths before Age 5, Orthophosphate Treatment Only, No Covariates

Glasgow



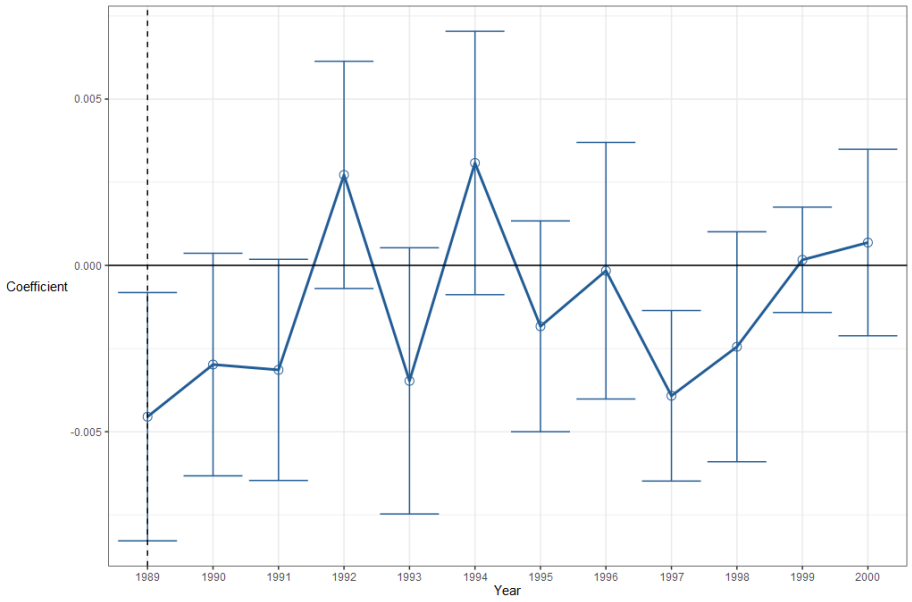
Edinburgh



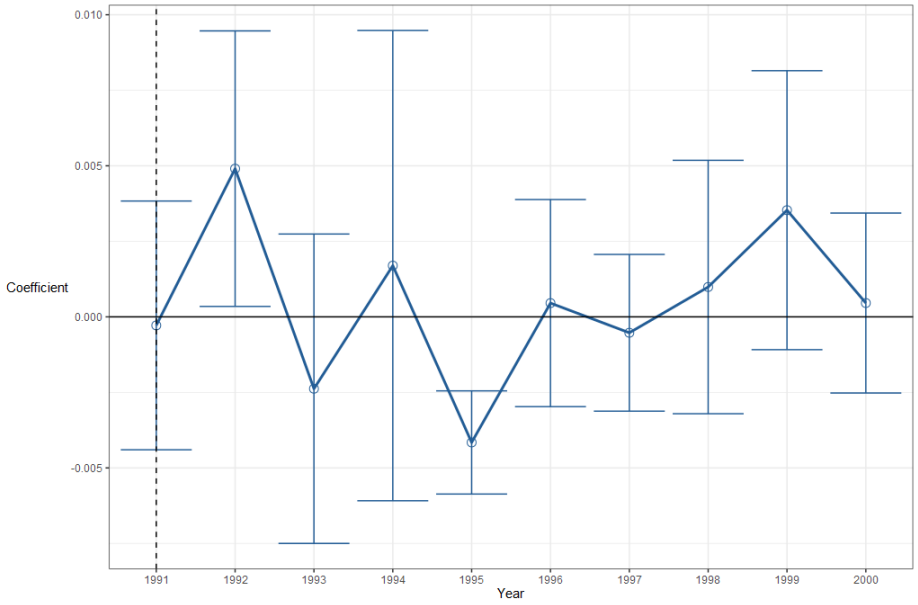
Notes: Figure shows each cohort's specific average partial effect (APE) for every *post-treatment* year (i.e. there are no leads in these charts). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method without covariates. Sample is restricted to 1985-2000. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of the APE and has associated 95% confidence intervals calculated with bootstrapped standard errors.

Figure 2.19 Average Partial Effect of Treatment on Deaths before Age 5, Orthophosphate Treatment Only, Covariates Included

Glasgow



Edinburgh



Notes: Figure shows each cohort's specific average partial effect (APE) for every *post-treatment* year (i.e. there are no leads in these charts). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method with covariates. Sample is restricted to 1985-2000. The dotted lines represent the start of either lime dosing or orthophosphate dosing in the water supply. Each circle is the point estimate of the APE and has associated 95% confidence intervals calculated with bootstrapped standard errors.

Next, we examine if the prevalence of lead piping in Glasgow affected the strength of the relationship with our outcomes and the lead reducing treatment. Different treatment dosage levels, such as we have when some areas have high lead pipe prevalence, and others have low lead pipe prevalence, can lead the treatment effect estimates being biased if there is selection into or out of the different dosage groups (see Callaway et al., 2021). To remove this threat to identification, Callaway et al., (2021) suggest regressing on each dosage group separately.

Therefore, we perform separate two-way Mundlak regressions, first removing the low lead pipe prevalence areas from the sample, then removing the high lead pipe areas (see figure 2.4). One issue is that there are far fewer births in the high lead areas, and especially few death occurrences, with only 1 or 2 in some years. Nevertheless, we include the estimation here as a robustness check.

The results for birthweights are in table 2.12. They are similar to our main results for both the high and low lead areas, with small negative effects for both areas. The under-5 mortality results are in table 2.13. For high lead areas, we see the majority of years have negative APEs, but the effects are relatively small and the overall APE is a precise null. For low lead areas, the year estimates of the APE are predominately negative. The overall APE is also negative and statistically significant.

Table 2.12 Glasgow High and Low Lead Areas Only, Average Effect of Treatment-on-the-Treated on Birthweights, Covariates Included

Year	High Lead Areas	Std Error	Low Lead Areas	Std Error
1978	-63.4	(26.5)	-10.6	12
1979	-46.5	(21.9)	7.1	11.7
1980	-90.3	(30.5)	-14.2	10.3
1981	-12.3	(24.0)	-5.9	12.3
1982	-17.9	(34.5)	-7.4	13
1983	-32.3	(32.5)	2.8	13.1
1984	-60.7	(32.6)	-19.3	13.5
1985	-50.0	(41.4)	6.9	12.5
1986	-30.5	(22.9)	9.7	11.9
1987	-10.2	(24.7)	13.2	13.5
1988	-37.5	(29.2)	5	13.7
1989	-60.5	(31.7)	-8.4	14.8
1990	-10.4	(23.9)	-11.1	13.6
1991	-25.1	(36.5)	-1.5	13.4
1992	-39.6	(32.9)	-17.9	13.6
1993	12.7	(26.6)	-0.6	14.8
1994	-55.3	(46.4)	-6.8	14.7
1995	-66.6	(31.2)	-19.2	15.8
1996	-70.9	(31.5)	-12.2	16.1
1997	-28.6	(28.1)	-10.8	15.6
1998	-87.5	(28.5)	-7.9	14.8
1999	-50.7	(19.8)	14.7	15.5
2000	-82.1	(33.3)	-14.0	17.2
Average	**-44.2**	**(5.6)**	**-4.7**	**(2.2)**

Notes: Table shows cohort specific treatment effects from two-way Mundlak regressions with covariates included in the regression. Column 1 and 3 are separate regressions, columns 2 and 4 are the standard errors. Column 1 is a regression excluding the areas with low prevalence of lead piping in Glasgow. Column 3 excludes the areas of high lead pipe prevalence in Glasgow. Each year has an estimated treatment effect, and the bottom row is the mean of these. Robust standard errors, clustered by postcode sector, are in brackets.

Table 2.13 Glasgow High and Low Lead Areas Only, Average Partial Effect of Treatment on Deaths before Age 5, Orthophosphate Treatment Only, Covariates Included

Year	High Lead Areas	Std Error	Low Lead Areas	Std Error
1978	-0.004	(0.001)	-0.003	(0.002)
1979	0.002	(0.003)	-0.006	(0.000)
1980	-0.002	(0.005)	-0.006	(0.004)
1981	0.004	(0.000)	-0.008	(0.003)
1982	0.000	(0.001)	-0.006	(0.002)
1983	-0.006	(0.001)	-0.004	(0.002)
1984	0.000	(0.004)	-0.008	(0.001)
1985	-0.005	(0.001)	-0.009	(0.002)
1986	-0.006	(0.000)	-0.003	(0.001)
1987	0.002	(0.001)	-0.005	(0.002)
1988	-0.001	(0.001)	-0.001	(0.004)
1989	-0.005	(0.000)	-0.004	(0.002)
1990	-0.001	(0.003)	-0.001	(0.003)
1991	0.001	(0.001)	-0.002	(0.003)
1992	0.002	(0.004)	-0.003	(0.002)
1993	0.001	(0.002)	-0.003	(0.004)
1994	0.005	(0.000)	0.001	(0.000)
1995	0.001	(0.001)	-0.004	(0.001)
1996	0.001	(0.002)	0.000	(0.000)
1997	-0.004	(0.002)	-0.002	(0.002)
1998	0.008	(0.002)	-0.002	(0.002)
1999	0.001	(0.002)	-0.003	(0.001)
2000	-0.001	(0.001)	-0.001	(0.002)
Average	**0.000**	**(0.001)**	**-0.004**	**(0.002)**

Notes: Table shows cohort specific Average Partial Effects (APE). These are calculated from logistic pooled quasi-maximum likelihood regressions using the two-way Mundlak method and including covariates. The relevant cohort and year indicators are set to 1, the relevant covariates indicators are set to1, continuous covariate variables are set to the cohort mean value for that covariate, and the difference with and without the cohort specific treatment indicator is taken. Each year has an estimated APE, and the bottom row is the mean of these. Standard errors of the APEs are bootstrapped. Columns 1 and 3 are separate regressions, columns 2 and 4 are the standard errors. Column 1 is a regression excluding the areas with low prevalence of lead piping in Glasgow. Column 3 excludes the areas of high lead pipe prevalence in Glasgow.

2.6.2 Regression Discontinuity Design

As an alternative identification strategy, we use a sharp regression discontinuity design (RDD). Here we separate the cohorts and regress each individual's outcome on the date of birth with the cut-off being the treatment date. We use local linear regressions with a triangular kernel. We also use the optimal non-parametric bandwidth selection method with the robust bias corrected intervals of Calonico et al. (2020). This means that it is only a direct before and after treatment comparison, within each cohort. We no longer need to rely on the assumptions in section 2.4. Instead, we assume the expected value of the outcome is continuous in the neighbourhood of the treatment cut-off for both treated and untreated units. That is, mothers cannot perfectly manipulate birth dates so as to be one side of the treatment cut-off. This would be violated if mothers knew about the upcoming water treatment and decided to delay birth until after treatment. Given there is always some randomness in birth dates (as many mothers will attest), we believe this is a reasonable assumption. See Cattaneo and Titiunik (2022) for a recent review of regression discontinuity design and its assumptions.

Given this assumption holding, we estimate the effect of treatment for individual pregnancies near the treatment cut-off. We do not use this as our main estimation strategy for two reasons, the actual difference in lead exposure near the cut-off may be miniscule. Therefore, the estimate may be too noisy to find an effect. Secondly if the assumptions in section 2.4 hold, the RDD is less efficient because we are discarding so much of the variation.

Table 2.14 and table 2.15 shows the results for the various RDD estimations. We use each cohort and we show both the pH raising treatment and the orthophosphate treatment. We also examine both outcomes, birthweights and deaths. For Glasgow, we see that none of the results are significant at the 5% level. For birthweights, both the 1978 and 1989 treatment estimates are

small in magnitude, and the 1978 is positive, while the 1989 is negative. For deaths, both the 1978 and 1989 treatment estimates imply lowered deaths due to treatment, but neither is statistically significant at the 5% level.

Table 2.14 Local Average Treatment Effect on Birthweights, Regression Discontinuity Design Results

Group	Coefficient	Std Error	Observations	Bandwidth (days)
Glasgow 1978	-3.5	(14.5)	216,556	1701
Fairmilehead 1978	-10.2	(33.2)	50,291	1488
Alnwickhill 1985	-108.0	(46.1)	26,151	1310
Glasgow 1989	8.8	(14.1)	216,556	1416
Edinburgh 1991	7.2	(28.4)	76,442	1461

Notes: This table reports the local average treatment effect from separate sharp regression discontinuity designs on birthweights. Robust, bias corrected standard errors are reported in brackets.

Table 2.15 Local Average Treatment Effect on Deaths Before Age 5, Regression Discontinuity Design Results

Group	Coefficient	Std Error	Observations	Bandwidth (days)
Glasgow 1978	0.000	(0.002)	216,771	1230
Fairmilehead 1978	0.003	(0.006)	50,326	1303
Alnwickhill 1985	-0.003	(0.005)	26,172	1781
Glasgow 1989	-0.002	(0.002)	216,771	1722
Edinburgh 1991	-0.002	(0.004)	76,498	1757

Notes: This table reports the local average treatment effect from separate sharp regression discontinuity designs on deaths. Robust, bias corrected standard errors are reported in brackets.

For Fairmilehead and the 1978 treatment, both results are the opposite sign from expected, implying treatment lowered birthweights and raised deaths, but neither is statistically significant. In Alnwickhill for the 1985 treatment, the point estimates imply it lowered birthweights and lowered deaths. The birthweights estimate is large and statistically significant at the 5% level. The deaths estimate is not significant. When we look at Edinburgh as a whole for the 1991 orthophosphate treatment, both estimates are of the expected sign. The point estimate implies birthweights increased and deaths decreased, but neither is statistically significant.

Overall, the RDD results are consistent with our main results. They show at best small effects on birthweights, but there is some possible effect on deaths, in Glasgow and in Edinburgh, but no effect in the Fairmilehead supplied area. Again, all estimates are relatively imprecise.

2.6.3 Separating Treatment Groups

As a final robustness check, we run two-way fixed effects regressions for each treatment group separately, against only the surrounding areas as a control group. This means that units treated at different times are not included in the regression, i.e. there is no staggered treatment to account for. Instead, only the surrounding never-treated areas are used as controls. The results are in table 2.16. Here the results suggest birthweights are increased by treatment in both Edinburgh groups but are not in Glasgow. For all death estimates we obtain precise nulls.

Table 2.16 Separate Two-Way Fixed Effects For Each Treatment Group

Area	Birthweights (ATT)	Std Error	Deaths (APE)	Std Error
Glasgow 1978	-0.002	(0.001)	0.000	(0.000)
Fairmilehead 1978	28.96	(5.05)	0.000	(0.000)
Alnwickhill 1985	21.37	(3.89)	0.001	(0.001)

Notes: This table shows results from three sperate regressions where each treated group is combined only with the never treated surrounding areas as the control group. Then We estimate two-way fixed effects for each group separately. Standard errors are clustered at postcode level for the Birthweights outcome, and are bootstrapped for the deaths outcome.

2.7 Discussion and conclusion

We examined the effect of lead reduction, through water treatment, on pregnancy outcomes in Glasgow and Edinburgh in the 20th century. We use a setting with plausibly exogenous staggered treatment and therefore use a difference-in-differences design that accounts for the staggered nature of the treatment. Across a variety of specifications, and with robustness checks, we do not find evidence for an effect of lead water pollution on birthweights, but we do find some evidence for a large effect on deaths, especially within the Glasgow area. Our main specification estimates that lead reduction accounts for a maximum 0.3-0.1 percentage point decrease in under-5 mortality in Glasgow and a 0.7-0.1 percentage point decrease in the Alnwickhill water plant supplied area of Edinburgh. This accounts for between 70% and 10% of the average death rate in Alnwickhill and 40% to 10% in Glasgow. This translates into 23-186 saved lives in Alnwickhill and 216-648 in Glasgow,

over the full 25 years of the sample. These findings are of a similar magnitude Clay et al. (2014) who find increasing water pH reduced deaths by 7%-33%, Edwards (2014) who finds lead in water increases the fetal death rate 32–63%, and Grossman and Slusky (2019) who find increased lead in the water responsible for a 12% decrease in the fertility rate. However, although the effect is potentially very large, it is not robust to alternative specifications.

These findings somewhat contradict the existing literature on the impact of lead and birthweights. However, while a number of studies have found a link between lead exposure and birthweights, there is still a body of literature that does not find an effect. Similarly, while a majority of studies in the literature find that lead exposure is linked to increased mortality, there are still some studies that have found no such relationship. Nevertheless, we must explain our findings in light of the plausible mechanisms laid out in section 2.2 and in light of previous findings.

The results on birthweights may be explained by a selection mechanism. It is possible that an increase in infant survival rates could also result in lower average birthweights. This is because a decrease in stillbirths and spontaneous abortions may lead to more infants born with lower birthweights surviving to birth (Goldenberg and Culhane, 2007). In this case it might be expected that the effect o treatment on average birthweights would be negative, as most of our specifications find.

As stated above, our main specification finds a positive effect of lead pollution on child mortality, but this is not robust. Hence it can be considered weak evidence at best. We propose three possible explanations for this finding.

The first is that an effect on deaths does exist, but the sample size and the low number of child deaths annually – often less than 1 percent – may not be sufficient to detect it with precision. Related to this is the possibility that there is an effect, but not at the levels of lead in our sample. We find it very

hard to believe that there is no effect of lead, no matter the level of dosage, given the quantity of scientific literature on this (see section 2.2.1, and chapter 1 and chapter 3). However, it may be the level of exposure in our sample was simply too low to have a detectable average effect. Lead pollution at high levels causes very obvious and extreme health problems, but at lower levels it is much harder to see acute lead poisoning symptoms. However, the water and blood lead levels in our sample, especially for Glasgow, are much larger than in many other studies which find an effect.

A second possibility is that the literature on lead and child mortality may be affected by publication bias. Many of the studies reviewed in section 2.2.1 have small sample sizes, or do not employ robust methods for identifying causal relationships, potentially leading to biased findings. There do remain a few studies that are better identified. In particular Clay et al. (2014), Edwards (2013) and Grossman and Slusky (2019). Nevertheless, as shown in chapter 1, the lead pollution literature suffers from publication bias. Even if every paper estimated an unbiased causal effect, in the presence of publication bias we would still be left with a bias in the literature. It is beyond the scope of this paper to estimate if there is publication bias for the lead pollution and birth outcomes literature, but it is a problem across empirical science and cannot be ruled out here.

Thirdly, it is possible that lead exposure would typically have an effect on child mortality, but a mitigating factor specific to Scotland during the period of interest may have reduced its impact. One potential factor is nutrition, specifically high milk intake. The UK and Northern Europe has some of the highest milk consumption rates in the world (FAO, 2022), more than two and a half times the global average (FAO, 2022). In the 1930s, the National Milk Scheme in Scotland promoted milk consumption and provided targeted subsidies for mothers of children under five years old. This, along with the provision of free milk in schools for much of the 20th century (Krebs, 2019), led to a significant increase in milk consumption. Studies have shown that

high milk intake is associated with lower blood lead levels (Chuang et al., 2004). Thus, it is possible that better nutrition, particularly high milk consumption, may have played a role in reducing the impact of lead on infant mortality in Scotland.

This study has several limitations. Firstly, the treatment variation is at a postcode level rather than an individual level and we also observe most of our sample only once. This means that our confidence intervals are less precise than if the treatment was exogenously applied at an individual level. However, this is common in the literature that examines the effect of lead with quasi-experiments because lead cannot be ethically given as part of a randomised control trial. We therefore must rely on coarser treatment variation. Secondly, our estimates of the average treatment on the treated mean that the effects we observe may only apply to larger urban areas. Perhaps there is something systematically different about rural areas that would mean a stronger effect. Parker and Wilby (2013) find that domestic water use per capita is much higher in rural areas compared to urban areas.

Our study has several implications. Failure to find a robust effect on infant outcomes does not mean lead remediation is pointless. Lead has been shown to affect a large variety of outcomes (see section 2.2). Although the effect on deaths we found was not always statistically significant, several point estimates imply lead reduction saved many infant lives. However, for policy makers and citizens, it may change the weighting of lead remediation compared to other actions, such as improved nutrition or neonatal healthcare. Our study also suggests there may be powerful mediating factors between lead pollution and health outcomes, and these should be investigated.

References

Akoumianaki I, 2017. Lead in drinking water: public health, mitigation, and economic perspectives. CD2016_03. Available online at: crew.ac.uk/publications.

Al-Saleh, I., Khalil, M.A., Taylor, A., 1995. Lead, erythrocyte protoporphyrin, and hematological parameters in normal maternal and umbilical cord blood from subjects of the Riyadh region, Saudi Arabia. Arch. Environ. Health 50, 66–73. https://doi.org/10.1080/00039896.1995.9955014

Angell, N.F., Lavery, J.P., 1982. The relationship of blood lead levels to obstetric outcome. Am. J. Obstet. Gynecol. 142, 40–46. https://doi.org/10.1016/S0002-9378(16)32282-7

Athey, S., Imbens, G.W., 2022. Design-based analysis in Difference-In-Differences settings with staggered adoption. J. Econom. 226, 62–79. https://doi.org/10.1016/J.JECONOM.2020.10.012

Azayo, M.M., Manji, K., Kalokola, F., 2009. Lead Levels in Women at Delivery at the Muhimbili National Hospital: A Public Health Problem. J. Trop. Pediatr. 55, 138–139. https://doi.org/10.1093/TROPEJ/FMN085

Behrman, J.R., Rosenzweig, M.R., 2004. Returns to Birthweight. Rev. Econ. Stat. 86, 586–601. https://doi.org/10.1162/003465304323031139

Bellinger, D.C., 2005. Teratogen update: Lead and pregnancy. Birth Defects Res. Part A Clin. Mol. Teratol. 73, 409–420. https://doi.org/10.1002/BDRA.20127

Bornschein, R.L., Grote, J., Mitchell, T., Succop, P.A., Dietrich, K.N., Krafft, K.M., Hammond, P.B., 1989. Effects of Prenatal Lead Exposure on Infant Size at Birth. Lead Expo. Child Dev. 307–319. https://doi.org/10.1007/978-94-009-0847-5_18

Brubaker, C.J., Schmithorst, V.J., Haynes, E.N., Dietrich, K.N., Egelhoff, J.C., Lindquist, D.M., Lanphear, B.P., Cecil, K.M., 2009. Altered myelination and axonal integrity in adults with childhood lead exposure: A diffusion tensor imaging study. Neurotoxicology. https://doi.org/10.1016/j.neuro.2009.07.007

Callaway, B., Goodman-Bacon, A., Sant'Anna, P.H.C., 2021. Difference-in-Differences with a Continuous Treatment. arXiv. https://arxiv.org/abs/2107.02637

Callaway, B., Sant'Anna, P.H.C., 2021. Difference-in-Differences with multiple time periods. J. Econom. 225, 200–230. https://doi.org/10.1016/J.JECONOM.2020.12.001

Calonico S, Cattaneo MD, Farrell MH. 2020. Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. The Econometrics Journal. 23(2):192–210. doi:10.1093/ectj/utz022.

Carstairs V, Morris R, 1991. Deprivation and health in Scotland. Aberdeen: Aberdeen University Press.

Cattaneo, M. D. and R. Titiunik, 2022. Regression Discontinuity Designs. Annual Review of Economics 14(1): 821-851.

Cecil, K.M., Brubaker, C.J., Adler, C.M., Dietrich, K.N., Altaye, M., Egelhoff, J.C., Wessel, S., Elangovan, I., Hornung, R., Jarvis, K., Lanphear, B.P., 2008. Decreased brain volume in adults with childhood lead exposure. PLoS Med. https://doi.org/10.1371/journal.pmed.0050112

Chatterji, P., Kim, D., Lahiri, K., 2014. Birth weight and academic achievement in childhood. Health Econ. 23, 1013–1035. https://doi.org/10.1002/HEC.3074

Chuang, H.Y., Tsai, S.Y., Chao, K.Y., Lian, C.Y., Yang, C.Y., Ho, C.K., Wu, T.N., 2004. The influence of milk intake on the lead toxicity to the sensory nervous system in lead workers. Neurotoxicology 25, 941–949. https://doi.org/10.1016/J.NEURO.2004.04.003

Clay, K., Troesken, W., Haines, M., 2014. Lead and Mortality. Rev. Econ. Stat. 96, 458–470. https://doi.org/10.1162/REST_A_00396

Comber, S., Cassé, F., Brown, B., Martin, J., Hillis, P., Gardner, M., 2011. Phosphate treatment to reduce plumbosolvency of drinking water also reduces discharges of copper into environmental surface waters. Water Environ. J. 25, 266–270. https://doi.org/10.1111/J.1747-6593.2010.00219.X

Dave, D.M., Yang, M., 2022. Lead in drinking water and birth outcomes: A tale of two water treatment plants. J. Health Econ. 84, 102644. https://doi.org/10.1016/J.JHEALECO.2022.102644

Dorea, J.G., Donangelo, C.M., 2006. Early (in uterus and infant) exposure to mercury and lead. Clin. Nutr. 25, 369–376. https://doi.org/10.1016/J.CLNU.2005.10.007

Doucouliagos, H., Stanley, T.D., 2009. Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis. Br. J. Ind. Relations 47, 406–428. https://doi.org/10.1111/J.1467-8543.2009.00723.X

Edwards, M., 2014. Fetal death and reduced birth rates associated with exposure to lead-contaminated drinking water. Environ. Sci. Technol. 48, 739–746. https://doi.org/10.1021/ES4034952/SUPPL_FILE/ES4034952_SI_001.PDF

Falcón, M., Viñas, P., Luna, A., 2003. Placental lead and outcome of pregnancy. Toxicology 185, 59–66. https://doi.org/10.1016/S0300-483X(02)00589-9

FAO [WWW Document], 2022 https://www.fao.org/faostat/en/#data/FBS (accessed 12.31.22).

Feigenbaum, J.J., Muller, C., 2016. Lead exposure and violent crime in the early twentieth century. Explor. Econ. Hist. 62, 51–86. https://doi.org/10.1016/j.eeh.2016.03.002

Gazze, L., Heissel, L., 2021. Infrastructure upgrades and lead exposure: Do cities face trade-offs when replacing water mains? Journal of Environmental Economics and Management, Volume 108,102472,ISSN 0095-0696, https://doi.org/10.1016/j.jeem.2021.102472.

Gazze, L., Persico, C., Spirovska, S., 2021. The Long-Run Spillover Effects of Pollution: How Exposure to Lead Affects Everyone in the

Classroom. SSRN Electron. J. https://doi.org/10.2139/SSRN.3842757

GBD, 2020. Global Burden of Disease (GBD 2019) | Institute for Health Metrics and Evaluation [WWW Document]. Glob. Burd. Dis. - IMHE. URL https://www.healthdata.org/gbd/2019 (accessed 1.26.22).

Gechter, M., Meager, 2022. Combining Experimental and Observational Studies in Meta-Analysis: A Debiasing Approach. *Working paper*.

Goldenberg RL, Culhane JF. 2007. Low birth weight in the United States. The American Journal of Clinical Nutrition. 85(2):584S-590S. doi:10.1093/ajcn/85.2.584S.

Golmohammadi, Tahere, Mohammad Reza Ansari, Abdorrahim Nikzamir, Reza S. Abhari and Shahriar Bin Elahi, 2007. The effect of maternal and fetal lead concentration on birth weight: polluted versus non-polluted areas of Iran. Medicine 74-78.

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econom. 225, 254–277. https://doi.org/10.1016/J.JECONOM.2021.03.014

Grossman, D.S., Slusky, D.J.G., 2019. The Impact of the Flint Water Crisis on Fertility. Demography 56, 2005–2031. https://doi.org/10.1007/S13524-019-00831-0

Hall, A., Ransom, W.B., 1906. PLUMBISM FROM THE INGESTION OF DIACHYLON AS AN ABORTIFACIENT. Br. Med. J. 1, 428. https://doi.org/10.1136/BMJ.1.2356.428

Hansard HC Deb 23 May 1990 Lead in Water (Scotland) [Hansard], vol 173 cc389-96. https://api.parliament.uk/historic-hansard/commons/1990/may/23/lead-in-water-scotland

Hayes, C.R., Hydes, O.D., 2012. UK experience in the monitoring and control of lead in drinking water. J. Water Health 10, 337–348. https://doi.org/10.2166/WH.2012.210

He, Q., Ji, X., 2021. The Labor Productivity Consequences of Exposure to Particulate Matters: Evidence from a Chinese National Panel Survey. Int. J. Environ. Res. Public Health 18, 12859. https://doi.org/10.3390/IJERPH182312859

Hollingsworth, A., Rudik, I., 2021. The Effect of Leaded Gasoline on Elderly Mortality: Evidence from Regulatory Exemptions. Am. Econ. J. Econ. Policy 13, 345–73. https://doi.org/10.1257/POL.20190654

Krebs, C., 2019. Lead Poisoning and Intelligence: A Search for Cause and Effect in the Scottish Mental Surveys. J. Environ. Public Health 2019. https://doi.org/10.1155/2019/8980604

Law, C.M., 2002. Significance of birth weight for the future. Arch. Dis. Child. - Fetal Neonatal Ed. 86, F7–F8. https://doi.org/10.1136/FN.86.1.F7

Macintyre, C., Fulton, M., Hepburn, W., Yang, S., Raab, G., Davis, S., Heap, M., Halls, D., Fell, G., 1998. Changes in blood lead and water lead in Edinburgh. An eight year follow-up to the Edinburgh lead study. Environ. Geochem. Health 20, 157–167.

https://doi.org/10.1023/A:1006593225113/METRICS

Marcus, D.K., Fulton, J.J., Clarke, E.J., 2010. Lead and conduct problems: A meta-analysis. J. Clin. Child Adolesc. Psychol. https://doi.org/10.1080/15374411003591455

McMichael, A.J., Vimpani, G. V., Robertson, E.F., Baghurst, P.A., Clark, P.D., 1986. The Port Pirie cohort study: maternal blood lead and pregnancy outcome. J. Epidemiol. Community Heal. 40, 18–25. https://doi.org/10.1136/JECH.40.1.18

Moore, M.R., Goldberg, A., Fyfe, W.M., Richards, W.N., 1981. Maternal Lead Levels After Alterations To Water Supply. Lancet 318, 203–204. https://doi.org/10.1016/S0140-6736(81)90384-6

Moore, M.R., Robertson, S.J., Gilmour, W.H., Murray, G.D., Britton, A., Low, R.A., Watt, G.C.M., 1998. Decline of maternal blood lead concentrations in Glasgow. J. Epidemiol. Community Heal. 52, 672–673. https://doi.org/10.1136/JECH.52.10.672

National Records of Scotland [WWW Document], 2022. https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/vital-events/general-publications/vital-events-reference-tables/2020/list-of-data-tables#section4 (accessed 12.31.22).

Parker JM, Wilby RL. 2013. Quantifying Household Water Demand: A Review of Theory and Practice in the UK. Water Resour Manage. 27(4):981–1011. doi:10.1007/s11269-012-0190-2.

Potter, S., 1997. *Lead in Drinking Water*. Science and Environment Section House of Commons Library, 1997.

Quinn, M.J., 1985. Factors Affecting Blood Lead Concentrations in the UK: Results of the EEC Blood Lead Surveys, 1979–1981. Int. J. Epidemiol. 14, 420–431. https://doi.org/10.1093/IJE/14.3.420

Ransom, W.B., 1900. Diachylon as an Abortifacient. Br. Med. J. 2, 201.

Richards WN, Britton A, Cochrane A., 1980. Reducing plumbosolvency-the effect of added lime on the Loch Katrine supply to Glasgow. Journal of the Institute of Water Engineers and Scientists 1979;34:315-333

Richards, W. N., & Moore, M. R., 1984. Lead Hazard Controlled in Scottish Water Systems. Journal (American Water Works Association), 76(8), 60–67. http://www.jstor.org/stable/41272146

Robertson, G., Wilson, A., Papastathopoulos, Y., Graham, M., Eades, L., Chatzisymeon, E., 2020. Estimating the number of homes in Scotland with internal lead piping. [WWW Document]. https://www.dwqr.scot/media/1f1n3kv4/ueh-007-19-project-management-doc-048-final-project-report-20-november-2020.pdf (accessed 12.31.22).

Royer, H., 2009. Separated at Girth: US Twin Estimates of the Effects of Birth Weight. Am. Econ. J. Appl. Econ. 1, 49–85. https://doi.org/10.1257/APP.1.1.49

Scottish Water, 2021. Your guide to water pipework. [WWW Document] https://www.scottishwater.co.uk/-/media/ScottishWater/Document-Hub/Your-

Home/Pipework-in-your-home/090821SWYourGuideToWaterPipework2021v1pageslr.pdf (accessed 12.31.22).

Taylor, C.M., Golding, J., Emond, A.M., 2015. Adverse effects of maternal lead levels on birth outcomes in the ALSPAC study: a prospective birth cohort study. BJOG An Int. J. Obstet. Gynaecol. 122, 322–328. https://doi.org/10.1111/1471-0528.12756

Troesken Werner. 2006. *The Great Lead Water Pipe Disaster*. Cambridge Mass: MIT Press.

UNICEF, 2020. The Toxic Truth: Children's exposure to lead pollution undermines a generation of potential | UNICEF [WWW Document]. URL https://www.unicef.org/reports/toxic-truth-childrens-exposure-to-lead-pollution-2020 (accessed 12.31.22).

UNICEF, 2022a. A Neglected Tragedy: The global burden of stillbirths - UNICEF DATA [WWW Document]. URL https://data.unicef.org/resources/a-neglected-tragedy-stillbirth-estimates-report/ (accessed 12.31.22).

UNICEF, 2022b. Child mortality data - UNICEF DATA [WWW Document]. https://data.unicef.org/resources/dataset/child-mortality/ (accessed 12.31.22).

UN IGME (United Nations Inter-agency Group for Child Mortality Estimation), 2021. [WWW Document]. https://www.who.int/publications/m/item/levels-and-trends-in-child-mortality-report-2021 (accessed 12.31.22).

Vinceti, M., Rovesti, S., Bergomi, M., Calzolari, E., Candela, S., Campagna, A., Milan, M., Vivoli, G., 2001. Risk of birth defects in a population exposed to environmental lead pollution. Sci. Total Environ. 278, 23–30. https://doi.org/10.1016/S0048-9697(00)00885-8

Vlasak, T., Jordakieva, G., Gnambs, T., Augner, C., Crevenna, R., Winker, R., Barth, A., 2019. Blood lead levels and cognitive functioning: A meta-analysis. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2019.03.052

Watt, G.C.M., Britton, A., Gilmour, H.G., Moore, M.R., Murray, G.D., Robertson, S.J., 2000. Public health implications of new guidelines for lead in drinking water: a case study in an area with historically high water lead levels. Food Chem. Toxicol. 38, S73–S79. https://doi.org/10.1016/S0278-6915(99)00137-4

Watt, G.C.M., Gilmour, W.H., Moore, M.R., Murray, G.D., Britton, A., Robertson, S.J., Womersley, J., 1996a. The Glasgow 93 Lead Study. Great Glasgow Health Board, Internal Report.

Watt, G.C.M., Gilmour, W.H., Moore, M.R., Murray, G.D., Britton, A., Robertson, S.J., Womersley, J., 1996b. Is lead in tap water still a public health problem? An observational study in Glasgow. BMJ 313, 979–981. https://doi.org/10.1136/bmj.313.7063.979

Wibberley, D.G., Khera, A.K., Edwards, J.H., Rushton, D.I., 1977. Lead levels in human placentae from normal and malformed births. J. Med. Genet. 14, 339. https://doi.org/10.1136/JMG.14.5.339

Wilcox, A.J., 2001. On the importance—and the unimportance— of birthweight. Int. J.

Epidemiol. 30, 1233–1241. https://doi.org/10.1093/IJE/30.6.1233

Wooldridge, J.M., 2021. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. SSRN Electron. J. https://doi.org/10.2139/SSRN.3906345

Wooldridge, J. M., 2010. Econometric analysis of cross section and panel data (2nd ed.). MIT Press.

WHO, 1970. European standards for drinking-water, 2nd ed. World Health Organization. https://apps.who.int/iris/handle/10665/40025

WHO, 2010. Preventing Disease through Healthy Environments: a Global Assessment of the Burden of Disease from Environmental Risks. [WWW Document]. https://www.who.int/publications/i/item/9789241565196 accessed 12.31.22).

WHO, 2016. The public health impact of chemicals: knowns and unknowns. [WWW Document]. https://www.who.int/publications/i/item/WHO-FWC-PHE-EPE-16-01 (accessed 12.31.22).

WHO, 2021. Exposure to lead: a major public health concern, 2nd edition [WWW Document]. https://www.who.int/publications/i/item/9789240037656 (accessed 12.31.22).

Xie, X., Ding, G., Cui, C., Chen, L., Gao, Y., Zhou, Y., Shi, R., Tian, Y., 2013. The effects of low-level prenatal lead exposure on birth outcomes. Environ. Pollut. 175, 30–34. https://doi.org/10.1016/J.ENVPOL.2012.12.013

Yurdakök, K., 2012. Environmental pollution and the fetus. J. Pediatr. Neonatal Individ. Med. 1, 33–42. https://doi.org/10.7363/010116

Zheng, J., 2021. Lead exposure, human capital formation, and inequality: the impacts of drinking water lead on short-run and long-run educational outcomes. Working paper. https://www.jiamengzheng.com/assets/files/lead_JMP_zheng.pdf

Chapter 3

The Impact of Lead Pollution on Human Capital Formation: Size of Dose Matters[11]

[11] We would like to thank Tanya Wilson, Hector Rufrancos, participants at the European Association of Environmental and Resource Economists (EAERE) conference 2022, at the Scottish Economic Society (SES) conference 2022, and at the Economics of Housing (EcHo) workshop at Royal Holloway 2022 for their comments.

Note on Chapter 3

Unfortunately, data access to the Scottish Qualification Authority data was lost during the pandemic. Their organisation brought in new rules subsequently that has meant we have not been able to get access to matched sibling data as we had when the analysis in this paper was done. Therefore, this paper does not have the same level of robustness checks as the chapters 1 and 2. Nor will it be possible to get access to the data again to perform more analysis for the foreseeable future.

3.1 Introduction

Recent estimates indicate 1 in 3 children suffer from high blood lead levels (GBD, 2019). The short-term health consequences of lead pollution on children are well known, but in recent decades studies have shown it may also have a variety of long-term, higher order outcomes, from anti-social behaviour and crime to lowering educational attainment (see, e.g., Aizer and Currie, 2019; Aizer et al 2018; Reyes, 2007). Given there are potentially 800 million children worldwide with high lead levels, this implies huge future costs we are imposing now, and consequently large returns on investment in abatement.

But lead is not the only pollutant, nor the only long-term cause of harm to children. Given production and political-economy constraints, lead abatement must compete against other policy needs, such as poverty reducing transfers, or greenhouse gas abatement. Balancing the long-term gains from lead abatement with other concerns requires knowledge of what long-term harm it does and, crucially, the relationship between that harm and the level of lead. This is called the *dose-response* relationship.

Understanding the dose-response relationship is crucial for cost-benefit analysis and for understanding the trade-offs society faces when choosing lead abatement over other initiatives. If lead causes higher marginal damage at lower levels, such as those observed in high income countries today (typically below 5μg/dl), then reducing lead levels may not be as beneficial for low- and middle-income countries, as even high-income countries can struggle to reduce levels below this threshold. High income countries, on the other hand, will still see large gains from further reducing remaining lead pollution, even though lead has already fallen greatly from levels seen in the 20th century. The reverse is true if the does-response relationship goes the other way. When marginal harms are larger at high levels of lead, then even small lead reduction measures may be attractive to lower income countries,

whereas high income countries may see more value in targeting the remaining pockets of high exposure, rather than in broad-based attempts to eliminate exposure.

In this paper, we focus on the dose-response impact of lead pollution when born on children's academic achievement 15-16 years later, exploiting a natural experiment in 1989 in Scotland. Previous studies have found infant lead exposure to be associated with long-term harmful outcomes, but there is disagreement in the literature on the dose-response relationship. Grönqvist, Nilsson and Robling (2019) find a threshold effect when investigating the impact of pollution from leaded gasoline on children school's achievements using local moss lead levels as an instrument for blood lead in Sweden. In particular, they find that the effects of early childhood exposure on the chances of graduating, or on the grade point average, are minimal or non-existent until blood lead levels reach around 5μg/dl. Somewhat similarly, Sampson and Winter (2018) show no effect of lead on anti-social behaviour when blood lead is below 5μg/dl and increasing marginal effects at higher levels. Reyes (2007) finds that for some crime outcomes the effects of lead are 20 times as large for the 4th quartile of lead exposure than they are for the 1st quartile. Gazze, Persico, and Spirovska (2021) find pre-school blood lead is linearly associated with worse education outcomes at the individual level, but also find large spill over effects from having peers with high lead levels. This potentially means spiralling adverse consequences once individual and networks effects are combined. In contrast, Evens et al., (2015) find higher marginal effects of lead on reading ability at lower levels of lead than at higher levels. Mielke and Zahran (2012) find the relationship between lagged air lead concentration levels and assault rates to be linear. Hollingsworth et al., (2022) similarly find a linear dose-response relationship, with lead lowering test scores. Miranda et al., (2007) find the relationship between blood lead and reading and mathematics ability to be linear, while Canfield et al., (2003) find that marginal effects on lead on IQ are

greater below 5μg/dl blood lead levels. Reyes (2015), using differences in state allowances of lead levels in gasoline as an instrument, finds higher elasticities of blood lead on behavioural problems at 5μg/dl blood lead than at >10μg/dl.

We use rich administrative data showing exam results for every state school pupil in Scotland from 2000-2009. Uniquely, we link the data by name and address to enable us to identify siblings within the same household. Our identification strategy uses the plausibly exogenous variation resulting from a treatment of the water supply in Glasgow, Scotland in November 1989 to estimate the effects of infant lead water ingestion at different doses on long-term education outcomes. Glasgow had elevated levels of lead in its water supply since the mid-19th century, due to a change in the water source used by the city (Troesken, 2006). Before treatment, Glasgow had water lead levels far in excess of those of Flint, Michigan in 2015, and the highest average blood lead levels of any city surveyed in the UK (Quinn, 1985). After treatment, the percentage of households with lead-water levels greater than 50μg/l fell from 13% to 2% (Watt et al., 1996a) and the blood lead level of mothers from Glasgow decreased from 11.9μg/dl to 3.7μg/dl (Watt et al., 1996a).

To investigate the dose-response relationship, we split our sample into high, low, or control group doses based on the prevalence of lead piping in the local area and the water lead levels observed. We used a detailed map that distinguished between high and low lead prevalence that was constructed by researchers working on the Glasgow lead study between the end of the 1980s and beginning of the 1990s (Watt et al., 1996a). Residential sorting across locations is not random, and we deal with potential biases arising from correlated unobservable characteristics in a variety of ways. Our main identification strategy uses the difference between siblings within the same household born either side of treatment (i.e., before and after 1989), with difference-in-differences estimated between the dosage groups. This enables

us to control for family and housing fixed effects, such as the amount of lead piping in the household. Treated areas (high and low) are compared with the rest of Scotland, and with each other separately. This follows the Callaway, Goodman-Bacon, and Sant'Anna (2021) approach on comparing different doses with difference-in-differences.

As a robustness check, our second strategy also uses a difference-in-differences between dosage groups but with outcomes averaged at the school level, and using all pupil outcomes, not just those for siblings.

Across econometric strategies, specifications and datasets, we find little evidence of an adverse lead effect on educational achievement for the low dosage group, but we do find evidence of socially significant effects for the high dosage group.

Our findings are in contrast to the literature stating lead pollution has the highest marginal effects on human capital formation when it is at low levels (see above discussion).

Our results contribute to the growing literature on the non-health effects of pollution (see e.g., Aguilar Gomez et al., 2022). Our paper uses a clean natural experiment to identify treatment groups in a place, Glasgow, which in the 1980s suffered from health problems and inequalities that are comparable to some of today's cities in middle-income countries. Much of the existing literature focuses on contemporaneous effects using cross-sections or, when looking at the long-run effects, by focusing on individual's outcomes over time. Our paper improves upon this, because our rich administrative data enables us to control for family fixed effects by computing siblings differences across areas before and after the water treatment. Family background and resulting unobserved investments may be an important and often overlooked confounder in previous studies.

Our main contribution is to show that the marginal effects on education outcomes are not greater at low levels of lead, instead it is at the high levels where the greatest marginal effects can be found. The implications are that countries with low average lead levels cannot expect large gains in educational attainment from lead abatement, except in targeted programmes aimed at the highest lead polluted areas. However, countries and areas with high infant lead ingestion such as India, where as many as 60% of its 470m children may have lead levels greater than 5μg/dL, can expect huge future educational gains from lead abatement policies.

3.2 Background

3.2.1 Lead Pollution and Human Capital

Lead has been recognised as harmful for thousands of years (see Needleman, 1992), but the long-term effects of infant lead ingestion on educational outcomes have only been investigated in recent decades. Lead water pollution when young is thought to be especially harmful for three reasons: firstly, children absorb up to 50% of ingested lead compared to 10% in adults (WHO, 2010); secondly, the blood-brain barrier is the main defence against large, water-soluble molecules, and this is not fully developed until after the first year of life, with in utero absorption being the most dangerous period (Goldstein, 1990); thirdly, a much higher share of infant diet tends to come from water, either through breast milk and their mother's water lead ingestion (Ettinger et al., 2004), or, more directly, from bottles of milk formula mixed with water (Baum and Shannon, 1997). In a 1993 survey, 84% of infants in Glasgow were bottle-fed (Watt et al., 1996b).

Lead impairs nerve conduction (Sindhu and Sutherling, 2015), damages myelination in the nerve system (Brubaker et al., 2009), and can impede brain development (Lanphear, 2015). This may affect educational outcomes

directly, through nerve and brain injury. Lead has been associated with impaired cognitive functioning (Vlasak et al., 2019) and lower IQ scores (Schwartz, 1994). A second possible mechanism is through behavioural changes. Blood lead levels are associated with aggressiveness, anti-social behaviour, and delinquency (Thomson et al., 1989, Needleman, 1996, and Reyes, 2015). These behaviours may have spill over effects on peers, so that even children with low-lead levels may experience worse educational outcomes due to peer behaviour (Gazze, Persico, and Spirovska, 2021). Given the likely strong relationship between infant lead levels and water lead, due to bottle feeding, water lead may be particularly harmful during early development, yet few studies look primarily at water lead levels and human capital. Zheng (2021) uses an instrumental variable estimation and finds increases in water lead levels reduce both mathematics and reading scores. Ferrie, Rolf, and Troesken (2012) find childhood water lead exposure lowers intelligence scores in US army enlistees.

3.2.2 The Glasgow Water Treatment

Glasgow's population grew from around 90,000 citizens in 1801, to 300,000 in 1841 (University of Portsmouth, 2022). Even in 1801, the water supply of 30 wells was inadequate being "impregnated with sewage and other deleterious matter" (Burnet, 1869). The Council and several private companies in turn attempted to improve matters, by taking water from the Clyde River and water to the south.

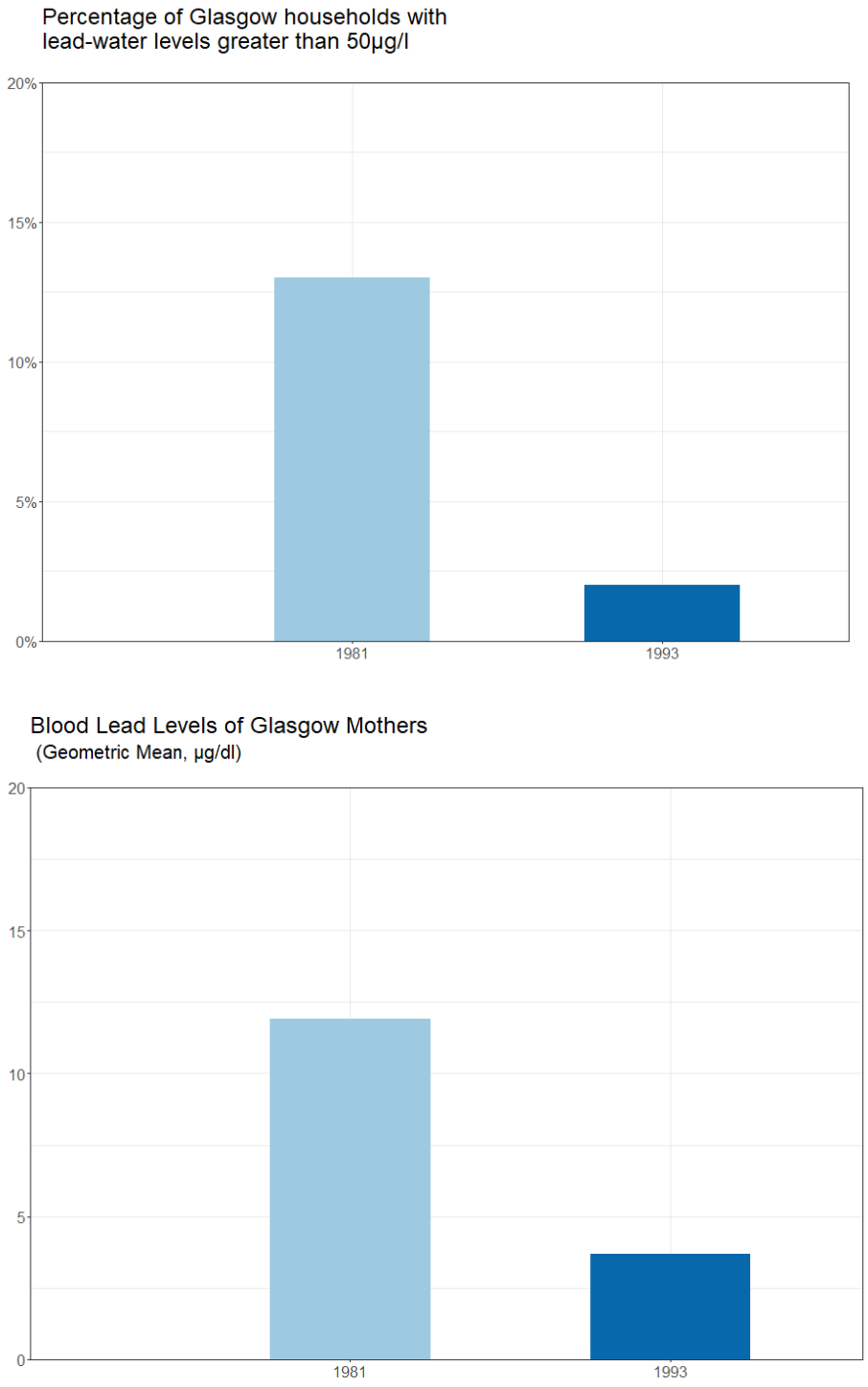
The Clyde water was pumped without being filtered and, due to the industrial use of the water, it was considered of poor quality, while the water supplied to the south of the city by the Gorbals Gravitation Water Company was of better quality. Even combined, however, these waterworks were not sufficient to keep the city supplied, especially not with water of good quality, so the Council eventually decided upon a new water supply: Loch Katrine. Katrine, a large and picturesque mountain lake, was considered more than

adequate to supply Glasgow's growing population with good quality water. Indeed, the quality of the water was much remarked upon. A report to the council from a chemist, one Dr Smith, reported that the water "was almost absolutely pure, clear to the utmost and without colour…[and] needs no purification". He then recommended this water to the council over any other option saying, "no town will have an equal abundance of such remarkably pure water" (Burnet, 1869).

The water is very soft and pure. Soft water lacks the mineral content found in harder waters and this means it has a low pH. Low pH water reacts with lead pipes (high plumbosolvency), dissolving the metal into the water supply (Kim *et al*., 2011). This was known at the time, due to the experiments of Robert Christison (1844). In 1854 one chemist, a Professor Penny, found that Loch Katrine water, after travelling through lead pipes, was "highly charged with lead", and believed it would be hazardous to supply such water to Glasgow (Burnet, 1869). The city council collected statement from various professors, engineers, and inhabitants of cities with soft water, before deciding there was no health risk. Glasgow has been supplied with Loch Katrine water ever since.

Professor Penny's worries over the Glasgow water lead levels would not be returned to until the mid-20th century. UK blood lead monitoring surveys in the 1970s found that Glasgow had the highest geometric mean blood lead level in any city surveyed (Quinn, 1985) at 18μg/dl. Six separate lead working sites were monitored in the same survey, and the Glasgow mean blood-lead level was higher than the mean of people living near those sites in all but one case. It was higher than the mean level of the lead workers themselves in 4 out of 6 sites (Quinn, 1985). Using these monitoring results, Quinn (1985) found that local plumbosolvency was much more closely related to local blood lead levels than distance to a road, and that lead-water intake was likely the biggest factor in the UK.

Figure 3.1 Water Lead and Blood Lead Levels in Glasgow



Source: Watt et al., 1996a.

By this time, the health impacts of "moderate" levels of lead were being taken seriously, and an EU directive in 1980 set the maximum water supply concentration of lead to 50μg/l from the previous 100μg/l (Watt *et al*., 2000). The Glasgow water supply was treated in 1978 with lime (Calcium hydroxide) to raise the pH and reduce plumbosolvency. This raised the pH from 6.3 to 7.8, and reduced the water lead levels (Moore *et al*., 1981).

However, in the late 1980s, the remaining levels of lead were deemed to still be too high. Surveys of water lead levels in residences found that in 1981, after the initial treatments, 13% of Glasgow households had water lead levels greater than 50μg/l (Moore et al, 1998), and 5% of homes had lead-water levels greater than 100µg/l (Moore et al., 1982). For comparison, the 90th percentile of lead-water samples in Flint, Michigan in 2015 was 31μg/l (Pieper et al., 2018).

Therefore, a second treatment of adding orthophosphate to the water was begun in November 1989 (Watt *et al*., 1996a). Correspondence with the engineering team involved with the project indicate the treatment was successful within a few weeks (author correspondence, 2020[12]), and lead-water levels fell. The percentage of households with lead-water levels greater than 50μg/l fell from 13% to 2% (Watt *et al*., 1996a). A long-term survey of mothers giving birth from Glasgow shows a decline in geometric mean blood lead levels from 11.9μg/dl in 1981 to 3.7μg/dl in 1993 (Watt *et al*., 1996a). Figure 3.1 shows these declines. Although the water reduction can be mostly attributed to the water treatments, the blood lead level reduction was coterminous with reductions in leaded gasoline in the UK. Therefore, additional methods are needed to identify the effect of the water treatment, and reduction in lead.

We use the plausibly exogenous reduction in lead intake resulting from the 1989 water treatment to identify the effect of lead on education outcomes.

[12] We would like to thank Stuart Robertson for his time and efforts.

However, we also distinguish between areas with a high lead pipe prevalence and areas with a low lead pipe prevalence withing Glasgow.

In Glasgow at this time there were estimated to be 160,000 housing units with some lead piping out of the 300,000 in the city (Watt *et al*., 1996a), but this was not equally concentrated. Far more of the older housing units had lead piping, either as service pipes under the ground, internal piping, or lead water tanks. Surveys of the population in Glasgow showed that 19% in the high lead areas said they had lead piping compared to 9% in the other areas of Glasgow (Watt *et al*., 1996a). They also had far higher concentrations of lead in their water supply even after the 1989 treatment (table 3.1). Crucially, table 3.1 shows that "low lead" does not mean "no lead". Even after treatment, around 13% of households in low lead areas had water lead levels above the modern UK limit of 10μg/l (Drinking Water Inspectorate, 2021). Around 46% had lead levels greater than 2μg/l, which would indicate even higher lead levels were present before treatment. McDonell, Campbell, and Stone (2000) found that the reduction in neural tube defects in the years after treatment was much greater in the high lead areas than in the low lead areas (table 3.2). These facts indicate a dose response relationship, where the effect of treatment will be higher in areas with higher lead-water levels. We therefore divide our sample into "High Lead" and "Low Lead" for our main estimates.

Table 3.1 Water Lead Concentrations in High and Low Lead Pipe Prevalence Areas, Glasgow, 1993

	Percent of Households	
µg/l	High Lead Areas	Low Lead Areas
<2	37.4	53.8
2-9	35.5	31.8
10-24	17.7	8.3
25-49	5.6	3.9
≥50	3.7	1.5
Observations	*785*	*941*

Notes: Data from table 5 in Watt *et al.* (1996a).

Table 3.2 Pregnancy prevalence of neural tube defects for each 1000 live births, Glasgow

	1983-95	1990-95
High Lead Area	2.1	0.69
Low Lead Area	2.4	1.8

Notes: Data from table 2 in McDonell, Campbell, and Stone (2000). Neural tube defects are early stage in utero damage to the brain, spine, or spinal cord.

3.3 Data

Our education data, sourced from the Scottish Qualifications Authority (SQA), includes crucial information such as the date of birth and postcodes of each schoolchild who sat exams during the period of interest. This data is essential for determining which pupils were affected by the water treatment and distinguishing between treatment and control areas. The SQA data also includes the education outcomes for every pupil in Scotland (97% are 15-16 years old on year of test), the year of examination, and the centre they attended (usually a school or college). The datasets include pupils born

between 1984 and 1993. We only have data for state schools, not for private schools or academies[13]. However, 96% of pupils in Scotland use state schools[14]. We exclude schools in Edinburgh from our sample, as they underwent a similar treatment in 1991, partly due to the findings of the Glasgow lead monitoring studies. SQA also provide matching indicators for siblings, where children are matched to the same family by surname, postcode and first line of address, but with different ages and first names. Finally, they provide the Scottish Index of Multiple Deprivation (SIMD) 2009 quintile for each child's postcode. The SIMD is a ranked index of deprivation on multiple dimensions (Income, Health, Education, Housing etc). The index is recalculated every 3 years. Although the ranking of each postcode moves around somewhat, the quintiles are relatively stable. We also use youth unemployment data for each year at the local authority level.

All children in Scotland during this period sit exams in their fourth year of secondary school at Standard Grade. 97% are 15 or 16 years old when they sit their exams. They sit exams in several subjects. We observe exam results from the years 2000-2009. The passing grades for these exams go from 1, the highest, to 7, a fail. A grade of 1 or 2 is called a "Credit" grade and allows one to go on to study the next level in the following year (called a "Higher"). Points are also awarded for each grade in each subject, and these are used as a marker for progressing to tertiary level education. The better the marks received, the higher the number of points. We only include the first examination year where a child sits Standard Grades in our sample (i.e., we do not include resits or repeated years).

We consider three outcome variables. The first is the total Standard Grade points achieved in that examination year. More points are better, but some

[13] Academies are state funded schools, but they are not controlled by the local council. Instead, they are usually controlled by an independent non-profit organisation.
[14] Scottish Council of Independent Schools census: https://www.scis.org.uk/facts-and-figures/#:~:text=SCIS%20uses%20the%20information%20collected,4%25%20of%20pupils%20in%20Scotland

subjects, such as Physics or Chemistry, are considered harder, but nevertheless taken as they are a prerequisite for some university courses (e.g., medicine often requires at least two science subjects). Some other subjects may be chosen instead if they are believed to be easier to get a Credit grade in, and the child does not wish to study medicine or engineering for example. Therefore, we consider two other outcomes: whether a child achieves a Credit grade in Mathematics, or in English. These are two subjects every child must sit, and therefore may give a better indication of change in ability rather than tastes in subjects. We use a Credit grade because this is the level needed to progress to "Highers" (a more difficult level of study, and a prerequisite for university) in the following year.

We sort each child into the "High Lead", "Low Lead" or "Control" based on their 1993 school postcode, using the plan of Loch Katrine supplied households, and high and low leaded pipe prevalence used in the map of Watt et al. (1996a). See figures 3.2 and 3.3 for a map of the schools and the high and low lead areas. Postcodes are UK government administrative boundaries used for a variety of purposes including the sorting of mail. Almost all postcodes have no more than 100 addresses, with the average being 15[15]. In urban areas they tend to be smaller than in rural areas. Although the matching is done based on household postcodes and then siblings were marked in the data for us, we were not given the household postcodes themselves. Therefore, children are sorted into treatment areas based on the school postcode. This means that children are assumed to be in the treatment area if their school is in the treatment area. In some cases, this will not be correct. For example, a "High Lead" school catchment area may include some areas that are "Low Lead". However, state school catchment areas are based on the postcodes surrounding the school so we expect most child postcodes to be nearby, especially in urban areas. There may also be placing requests.

[15] See https://www.ons.gov.uk/methodology/geography/ukgeographies/postalgeography.

These are requests by parents to have a child attend a certain school even though they are not in the school catchment area. Placing requests are generally a very small percentage of total pupil numbers but the combination of placing requests and school treatment not matching household treatment area means there will be some measurement error. The sign of the bias is difficult to know *a priori*. Schools perceived as "good schools" will likely have more placing requests, and therefore more pupils from outside their treatment area. Population changes in postcodes surrounding the schools can affect how many in a school are from the same treatment area as the school postcode.

If the treatment increases schooling outcomes, then pupils from "Low Lead" or "Control" areas going to "High Lead" schools, or "Control" to "Low Lead" would bias treatment estimates downwards, as would "High Lead" moving to the other areas. However, this may not be the only selection bias effect. For other kinds of bias to affect our estimated, it would require a violation of the parallel trends assumption in section 3.4. For example, if, over time, the pupils from "Control" areas tending to move to treatment areas were higher quality than before, this could lead to positive bias in the treatment effect. This is somewhat testable with event studies, which we carry out in section 3.5.2.

Summary statistics for our data are included in table 3.3.

Table 3.3 Descriptive Statistics

Panel A - Full Sample			
Variable	N	Mean	Std Dev
Outcomes			
Standard Grade Points	522661	163.106	80.269
Mathematics Credit Pass (=1 if passed with credit score)	468490	0.307	0.461
English Credit Pass (=1 if passed with credit score)	490212	0.425	0.494
Covariates			
Child SIMD quintile 1	544041	0.213	0.41
Child SIMD quintile 2	544041	0.204	0.403
Child SIMD quintile 3	544041	0.201	0.401
Child SIMD quintile 4	544041	0.198	0.399
Sex (1 = Male)	558379	0.505	0.5
Year of Birth	558379	1988.526	2.839
Area Youth Unemployment (%)	558379	62.625	6.33

Panel B - High Lead Areas Sample			
Variable	N	Mean	Std Dev
Outcomes			
Standard Grade Points	18248	155.499	74.628
Mathematics Credit Pass (=1 if passed with credit score)	17004	0.259	0.438
English Credit Pass (=1 if passed with credit score)	17650	0.337	0.437
Covariates			
Child SIMD quintile 1	17763	0.41	0.492
Child SIMD quintile 2	17763	0.234	0.423
Child SIMD quintile 3	17763	0.155	0.362
Child SIMD quintile 4	17763	0.117	0.322
Sex (1 = Male)	18616	0.518	0.5
Year of Birth	18616	1988.436	2.852
Area Youth Unemployment (%)	18616	52.079	3.034

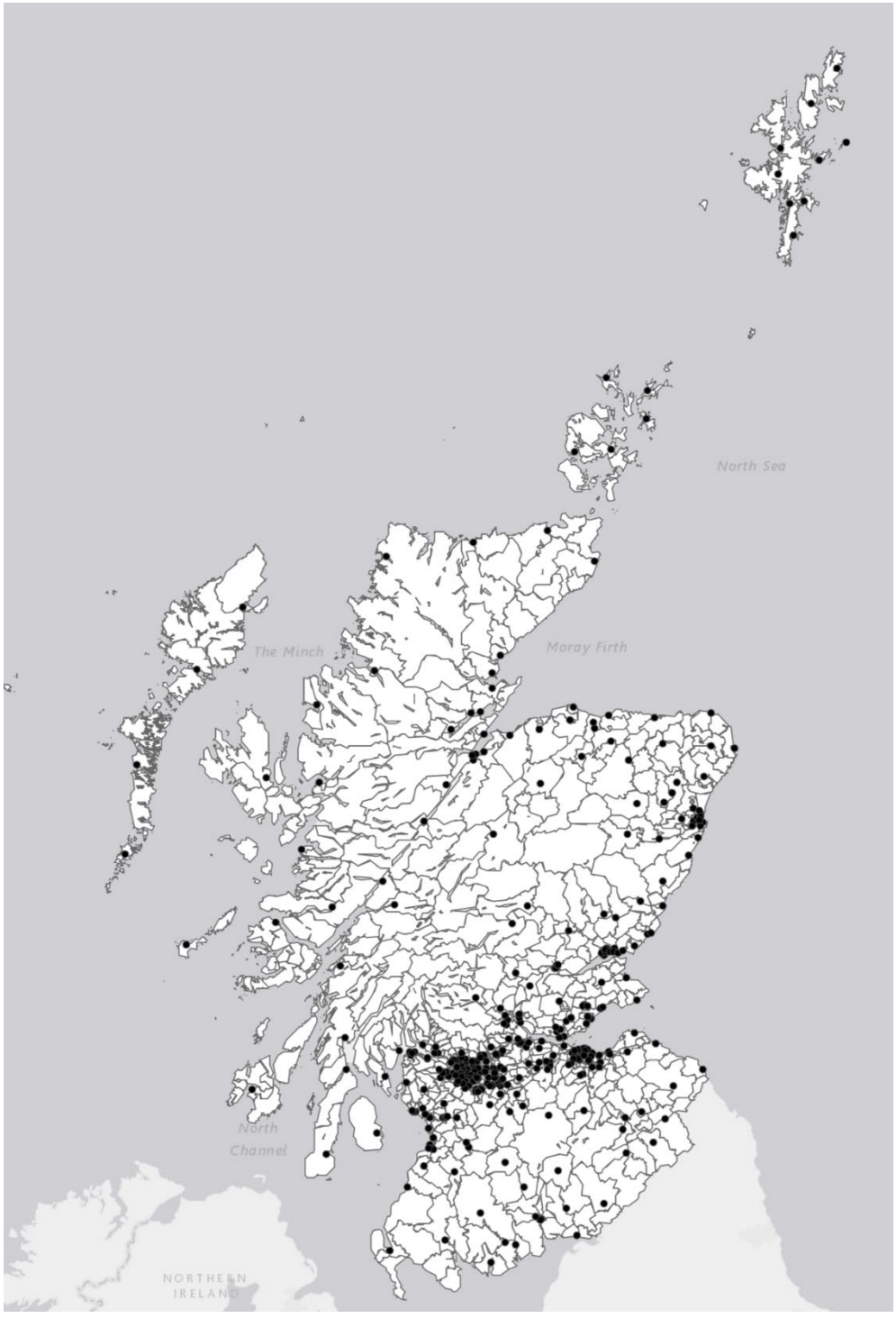
Panel C - Low Lead Areas Sample			
Variable	N	Mean	Std Dev
Outcomes			
Standard Grade Points	35012	144.936	75.292
Mathematics Credit Pass (=1 if passed with credit score)	31039	0.233	0.423
English Credit Pass (=1 if passed with credit score)	33628	0.335	0.472

Covariates			
Child SIMD quintile 1	35153	0.564	0.496
Child SIMD quintile 2	35153	0.145	0.352
Child SIMD quintile 3	35153	0.1	0.301
Child SIMD quintile 4	35153	0.093	0.29
Sex (1 = Male)	37144	0.48	0.5
Year of Birth	37144	1988.535	2.841
Area Youth Unemployment (%)	37144	54.563	6.016

Panel D - Control Sample

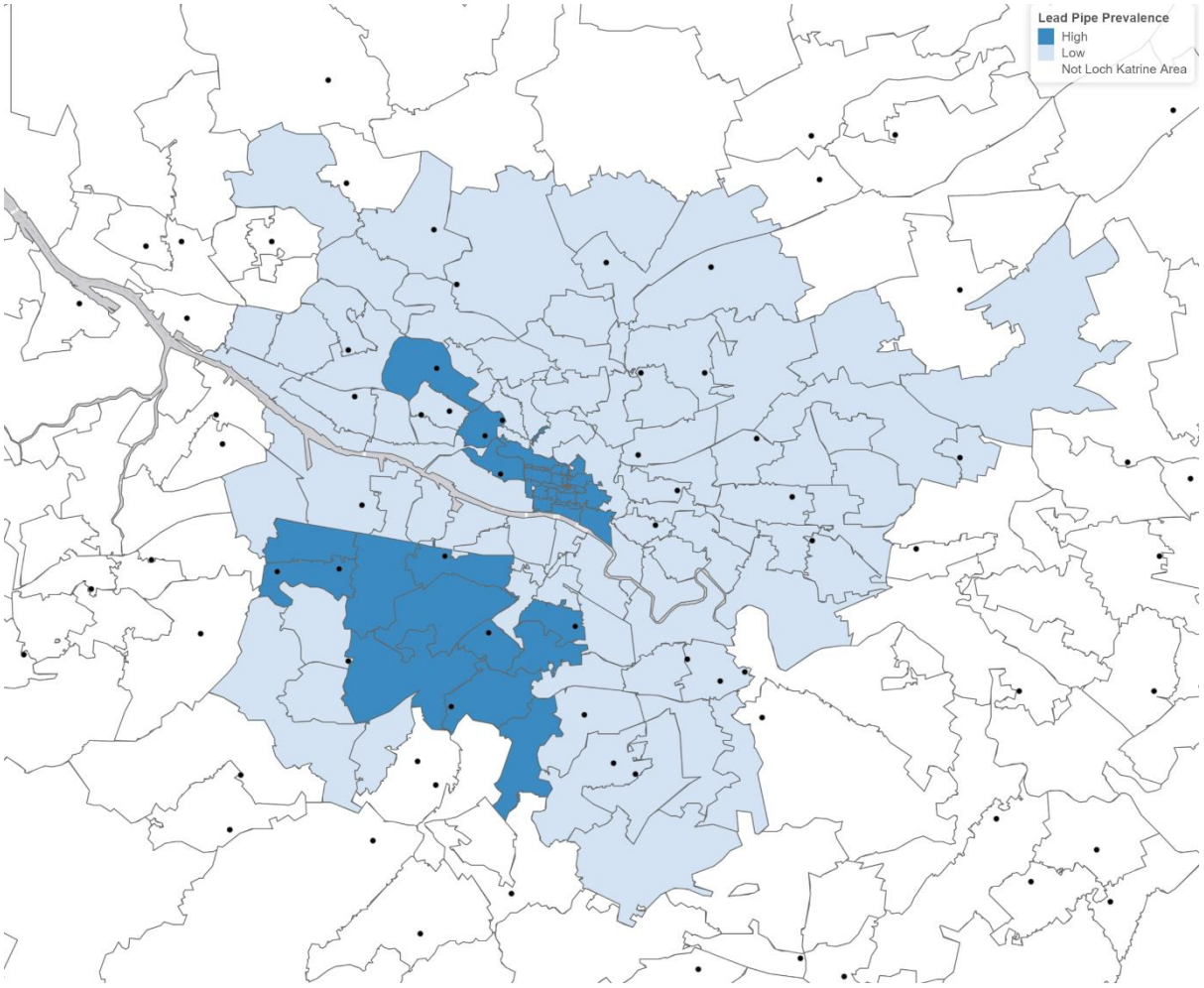
Variable	N	Mean	Std Dev
Outcomes			
Standard Grade Points	469401	164.757	80.655
Mathematics Credit Pass (=1 if passed with credit score)	420447	0.315	0.464
English Credit Pass (=1 if passed with credit score)	438934	0.435	0.496
Covariates			
Child SIMD quintile 1	491125	0.181	0.385
Child SIMD quintile 2	491125	0.207	0.405
Child SIMD quintile 3	491125	0.21	0.407
Child SIMD quintile 4	491125	0.209	0.407
Sex (1 = Male)	502619	0.506	0.5
Year of Birth	502619	1988.529	2.838
Area Youth Unemployment (%)	502619	63.646	5.616

Figure 3.2 Distribution of Schools in Scotland



Notes: Dots represent school locations. Due to travel distances and sparse population several markings for "Schools" on the islands may represent the same nominal school but are given different IDs in the data.

Figure 3.3 Distribution of Schools, High and Low Lead Areas in Loch Katrine Water Supply Area



Notes: shaded area is the Loch Katrine water supply area. Darker shading indicates a high prevalence of lead piping. Lines represent different postcode sectors. Dots represent school locations.

3.4 Empirical Strategy

Our identification strategy relies on the plausibly exogenous variation in lead ingestion, in the womb and in childhood, resulting from the orthophosphate treatment of the Loch Katrine water supply to Glasgow in November 1989. As lead ingestion has been shown to be particularly harmful in the womb, our treatment start date is for children conceived after treatment. Of course, we do not have data on when our sample was conceived, only the date of birth.

We take as our start date 1st of September 1990. That is, given the treatment would be effective by end of November (author correspondence, 2020), we take as the treatment group children born 9 months after this. The distribution of birth time from conception for term births is unimodal and symmetrical within the 10th-90th percentiles (Jukic et al., 2013). We believe there will only be minor classical measurement error resulting from this, but it may attenuate our estimates. Therefore, our estimates, if unbiased, may be a lower bound of the effect.

Our main identifying assumption is that the water treatment in November 1989 is exogenous variation in the lead intake of children conceived within the Loch Katrine supply area in Glasgow. Therefore, we assume that this treatment has an effect on education outcomes and is not associated with any confounding variables. We estimate our main results with a variety of difference-in-differences specifications. Our estimand is the Average Effect of Treatment on the Treated (ATT). This requires an assumption of parallel trends, the change in outcomes would be the same for treated and untreated without treatment. Given Glasgow is an urban area, with much higher concentrations of poverty than the Scottish average, we therefore also condition on a variety of covariates so that we assume parallel trends conditional on these covariates in some specifications.

As we have only one treatment period, that is common to all treated units, and a control group that is always untreated, we do not have to consider potential negative weighting arising from comparing earlier treated to later treated groups. Therefore, certain elements of the modern difference-in-difference literature, such as Goodman-Bacon decomposition (Goodman-Bacon, 2021) or reweighting of estimates (Calloway and Sant'Anna, 2021), do not apply. However, we do need to factor in another facet of the recent difference-in-difference literature: continuous treatment.

Different treatment doses combined into one treatment group can mean biased estimates of the ATT. Callaway, Goodman-Bacon, and Sant'Anna (2021) show that when you combine different doses you need stronger assumptions than with standard two-way fixed effects. The common identifying assumption with standard two-way fixed effects difference-in-differences is parallel trends (or conditional parallel trends). With continuous treatment (i.e. different dosages) combined, this assumption will be violated if there is selection into different dosage units. For example, if people within the treatment area begin to move to housing that has more lead piping, due to the water now being safer. With combined dosage difference-in-differences, we require stronger assumptions of either no selection into different dosage areas on average, or homogenous treatment effects.

These stronger assumptions can be relaxed back to standard parallel trends if, following the advice in Callaway, Goodman-Bacon, and Sant'Anna (2021), we separate the dosage units and compare them individually with the never-treated groups. Now all that is required is parallel trends between each dosage level separately with the control group. This is analogous to the traditional parallel trends assumption and can be made conditional on covariates. Callaway, Goodman-Bacon, and Sant'Anna (2021) show that this approach recovers an unbiased estimate of the ATT for that group and dose, but we sacrifice some efficiency by excluding some of the sample.

Our best proxies for treatment dosage are the high and low lead areas of Glasgow as described in section 3.2. Following the advice in Callaway, Goodman-Bacon, and Sant'Anna (2021), we compare each separate dose group to the never treated group in separate regressions.

In our first approach, we use the matched sibling-household data and carry out a simple difference-in-differences estimation. Given matched siblings live in the same household, with the same lead piping exposure before and after

treatment, household and area characteristics will be the same between siblings, and we should be able to recover the ATT with this approach, given our assumptions.

First we exclude households without siblings either side of the treatment divide. That is, we only consider households which have at least one older sibling born before 1st of September 1990, and at least one younger sibling born after this date. We take the difference between the siblings' outcomes within the household. If there are more than one sibling on one side of the treatment divide we average their outcomes as shown in below:

$$(1) \qquad Household\_Difference_h = \frac{\sum_{j=1}^{J} Y_{hj}}{n_{h1}} - \frac{\sum_{i=1}^{I} Y_{hi}}{n_{h0}}$$

Where the *Y* is one of three outcomes outlined in the data section, *h* is the household identifier, and *j* is the individual identifier of a sibling born before treatment, and *i* for an individual born after treatment, n_{h0} is the number of siblings in household *h* born before treatment and n_{h1} the number born after. We expect this to be negative on average, as older siblings tend to outperform younger ones (see Keller, Troesch, and Grob, 2015; Lehmann, Nuevo-Chiquero, and Vidal-Fernandez, 2016; or Havari and Savegnago, 2022).

We then average these household differences for the Control sample, the Low Lead sample, and the High Lead sample for all three outcomes. Finally, we take the difference-in-differences using these means.

$$(2) \qquad \hat{\theta} = \frac{\sum_{h \in G_1} Household\_Difference_h}{N_{G_1}} - \frac{\sum_{h \in G_2} Household\_Difference_h}{N_{G_2}}$$

Where $\hat{\theta}$ is our difference-in-differences estimate of the ATT, *G* is the sample group (Control, Low Lead, or High Lead), and N_G the number of households in that group. One potential source of confounding is family size. If family size

changes differentially both over time and in different treatment groups, this may lead to biased estimates. For example, if control areas tended to have larger families after the treatment, given younger sibling tend to perform worse than older ones, this could lead to positively biased estimates. We therefore use a second approach, that does not rely on sibling differences.

Our second approach uses the whole sample, but as we only observe each child once we must average outcomes at the school level. The baseline two-way fixed effects difference-in-difference specification is in (3).

$$(3) \qquad Y_{st} = \alpha + \theta\, Treat_s \times Post + \boldsymbol{X}_{st}\boldsymbol{\beta} + \gamma_s + \lambda Post + \epsilon_{st}$$

Where α is an intercept term, $Treat$ is an indicator variable for if a school lies within the Loch Katrine water supply area, $Post$ is an indicator for the time periods after 1st of September 1990, $\boldsymbol{X}_{st}$ is a vector of school level characteristics, γ_s are school fixed effects, and ϵ_{st} is the error term. We cluster our estimated errors by school. The variable of interest is θ , the coefficient on the interaction $Treat_s \times Post$. This is also an estimate of the ATT, but at the school level.

Given the different dosage groups, we also split the school treated areas into High Lead and Low Lead. For example, the estimate of the causal effect on the High Lead group can be recovered from (4).

$$(4) \qquad Y_{st} = \alpha + \theta\, High_s \times Post_{st} + \boldsymbol{X}_{st}\boldsymbol{\beta} + \gamma_s + \lambda Post + \epsilon_{st}$$

Where the $Treat$ variable has been replaced with an indicator for if a school is in the High Lead area. (4) is estimated by excluding the Low Lead sample. Similarly, we can estimate the casual effect of Low Lead dosage by excluding the High Lead sample and estimating (4) but using an indicator for if a school is in a Low Lead area.

We also use an event study specification, to see the placebo effects of $Treat$ interacted with years before the treatment, and to see if the effect is monotonic after treatment. This specification is outlined in below.

$$(5) \qquad Y_{st} = \alpha + \lambda_t + \sum_{\tau=-q}^{-1} \delta_t \, Treat_{st} \; + \sum_{\tau=0}^{m} \theta_t \; Treat_{st} + \boldsymbol{X}_{st}\boldsymbol{\beta} \; + \; \gamma_s + \epsilon_{st}$$

Where *m* and *q* are the leads and lags. To check the effects of different dosages we exclude either High or Low Lead groups as before and check the event studies individually compared to the control group.

3.5 Results

3.5.1 Matched Sibling Difference-in-Differences

Table 3.4 panel A shows the average difference between siblings within a household, pre and post treatment for each dose group, as calculated in (1). Figure 3.4 also shows the same figures. As expected, older siblings tend to perform better than their younger siblings, within the same household across all outcomes, as can be seen by the negative signs. However, younger siblings in the High Lead area, which has a higher prevalence of lead piping and lead-water levels, appear to perform better than their peers in the other dosage groups across all three outcomes after treatment. In contrast, the Low lead dosage group does not appear to perform better than the control group in Standard Grade points, and performs worse for the English credit outcome, but is better in the Mathematics Credit outcome.

In panel B of table 3.4 we calculate the difference-in-differences as in (2). The difference-in-differences point estimates, and their associated 95%, Bonferroni corrected, intervals are in figure 3.5. In the first column we compare the Low Lead sibling differences to the control group sibling differences. Wide standard errors mean that Bonferroni corrected 95% confidence intervals cover zero for all outcomes except the Mathematics Credit outcome, where the effect of treatment on the Low Lead group is

estimated to have increased the probability of achieving a Mathematics credit pass by 3.6 percentage points.

Table 3.4 Differences Between Siblings in Same Housing, Pre and Post-Treatment

Panel A – Mean Sibling Differences

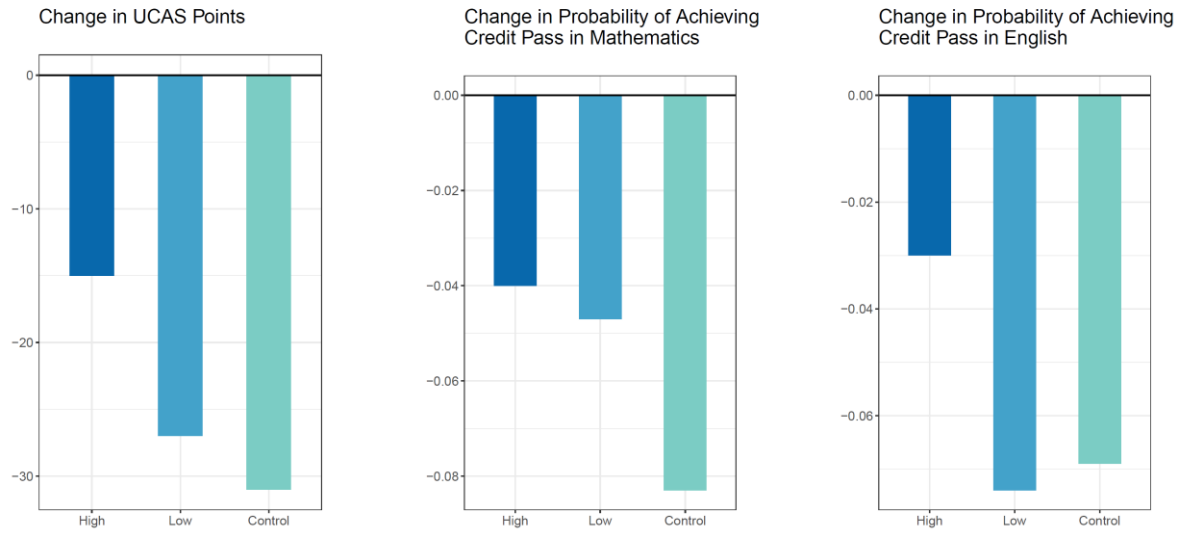
	Control	Low Lead	High Lead
Standard Grade Points			
Mean Sibling Difference	-31	-27	-15
Standard Deviation	(83)	(76)	(70)
Observations	37302	2228	1232
Mathematics Credit			
Mean Sibling Difference	-0.083	-0.047	-0.040
Standard Deviation	(0.518)	(0.516)	(0.510)
Observations	35361	2123	1222
English Credit			
Mean Sibling Difference	-0.069	-0.074	-0.030
Standard Deviation	(0.546)	(0.560)	(0.550)
Observations	36314	2202	1229

Panel B – Difference in Differences

	Low - Control	High - Control	High - Low
Standard Grade Points			
Difference-in-Differences	4	17	12
Standard Error	(2)	(2)	(3)
Mathematics Credit			
Difference-in-Differences	0.036	0.043	0.007
Standard Error	(0.012)	(0.015)	(0.018)
English Credit			
Difference-in-Differences	-0.005	0.038	0.043
Standard Error	(0.012)	(0.016)	(0.02)

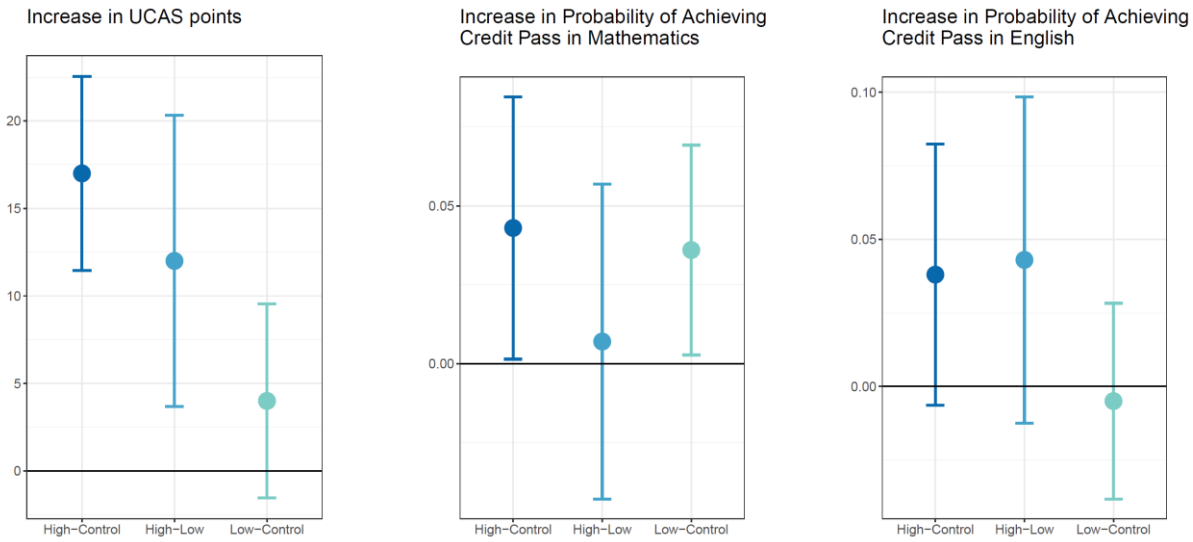
Notes: Panel A shows difference between siblings born before treatment and siblings born after, averaged by dosage group. Panel B shows the difference-in-differences estimate between the averaged differences in panel A.

Figure 3.4 Average Sibling Differences, Before and After Treatment Date



Notes: Figure shows the average within-household difference between siblings born before the treatment date and siblings born after the treatment date. High refers to households living within treated areas that had a high prevalence of lead piping. Low refers to households living within treated areas that had a low prevalence of lead piping. Control refers to non-treated areas in Scotland.

Figure 3.5 Average Treatment on the Treated, Difference-in-Differences Estimates



Notes: Figure shows point estimates as circles with associated 95% Bonferroni corrected confidence intervals. High refers to households living within treated areas that had a high prevalence of lead piping. Low refers to households living within treated areas that had a low prevalence of lead piping. Control refers to non-treated areas in Scotland.

In the second column we take estimate difference-in-differences for the High Lead minus the control group sibling difference. Here the results are clearer. The treatment is estimated to increase the Standard Grade achieved by 17 points, and the probability of achieve a mathematics credit pass by 4.3 percentage points. The point estimate for the increase in probability of an English credit pass is 3.8 percentage points, but Bonferroni corrected 95% confidence intervals cover zero for this outcome.

In the third column we compare High and Low lead treatment areas. Here, the point estimates are all positive, suggesting dosage does make a difference and younger siblings in High Lead areas perform relatively better after the treatment than those in Low Lead areas. However, Bonferroni corrected 95% confidence intervals cover zero for all but the Standard Grade points outcome.

The sibling difference results in table 3.4 suggest it is only in High Lead areas that there is a socially significant difference in education outcomes after treatment. We next move on to the school level difference-in-differences.

3.5.2 School Level Difference-in-Differences

In table 3.5 we present the two-way fixed effect estimates for the full treatment and control sample. We present estimates for all three outcome variables, with and without school level covariates. All the point estimates are positive, suggesting the lower lead resulting from the water treatment may have had an effect, but Bonferroni corrected 95% confidence intervals cover zero for all estimates.

In table 3.6 we present the two-way fixed effect estimates for only the High Lead and Control schools, excluding those in the Low Lead zone of the Loch Katrine water supply area. This decreases the potential bias and lowers the identification risk resulting from differential treatment dosage as shown in

Callaway, Goodman-Bacon, and Sant'Anna (2021). In all cases the point estimates are higher than in table 3.5, where we use the whole of the Loch Katrine water supply area. The estimates in Panel A suggest that the treatment increased by 18 points the average Standard Grade points achieved in High Lead area schools. This is approximately the difference between getting an A instead of a B in one of the, typically eight, subjects taken at standard grade. Bonferroni corrected 95% confidence intervals do not cover zero in either case. The estimates in panel B suggest the treatment increase the proportion of pupils achieving a credit pass in mathematics by around 5 percentage points. Bonferroni corrected 95% confidence intervals do not cover zero for the estimate without school level covariates but do when these covariates are added. Panel C suggests the treatment increases the proportion of students achieving a credit pass in English by around 1-2 percentage points, but the Bonferroni corrected 95% confidence intervals cover zero in both cases.

In table 3.7, we show the same results but comparing the Low Lead areas to control areas, excluding the High Lead area schools. Here again all points estimates are positive, but Bonferroni corrected 95% confidence intervals cover zero in all cases. The point estimates are lower than for the High Lead sample for Standard Grade points and for Mathematics Credit passes, but higher for English Credit passes.

Table 3.5 Total Loch Katrine Water Supply Area, School Level Difference-in-Differences

	(1)	(2)
Panel A – Standard Grade Points		
Treatment × Post	11.029	10.505
	(5.72)	(5.818)
Observations	727	663
Unit Level Covariates	No	Yes
Panel B – Mathematics Credit Pass Share		
Treatment × Post	0.049	0.048
	(0.022)	(0.023)
Observations	718	654
Unit Level Covariates	No	Yes
Panel C – English Credit Pass Share		
Treatment × Post	0.007	0.008
	(0.017)	(0.019)
Observations	722	658
Unit Level Covariates	No	Yes

Notes: Table shows difference-in-differences estimation of school level average outcomes between treated schools and control schools. Standard errors are clustered by school and presented in brackets. Column (1) is estimate without school level covariates, and column (2) with. Covariates include index of multiple deprivation quintile, share of boys in school, and the local youth unemployment rate.

Table 3.6 High Lead Areas, School Level Difference-in-Differences

	(1)	(2)
Panel A – Standard Grade Points		
Treatment × Post	18.662	18.204
	(4.931)	(4.936)
Observations	680	616
Unit Level Covariates	No	Yes
Panel B – Mathematics Credit Pass Share		
Treatment × Post	0.056	0.054
	(0.019)	(0.022)
Observations	671	607
Unit Level Covariates	No	Yes
Panel C – English Credit Pass Share		
Treatment × Post	0.016	0.017
	(0.014)	(0.016)
Observations	675	611
Unit Level Covariates	No	Yes

Notes: Table shows difference-in-differences estimation of school level average outcomes between High Lead area treated schools and control schools. Standard errors are clustered by school and presented in brackets. Column (1) is estimate without school level covariates, and column (2) with. Covariates include index of multiple deprivation quintile, share of boys in school, and the local youth unemployment rate.

Table 3.7 Low Lead Areas, School Level Difference-in-Differences

	(1)	(2)
Panel A – Standard Grade Points		
Treatment × Post	8.373	7.854
	(7.007)	(7.085)
Observations	711	647
Unit Level Covariates	No	Yes
Panel B – Mathematics Credit Pass Share		
Treatment × Post	0.047	0.045
	(0.027)	(0.028)
Observations	702	638
Unit Level Covariates	No	Yes
Panel C – English Credit Pass Share		
Treatment × Post	0.004	0.005
	(0.022)	(0.023)
Observations	706	642
Unit Level Covariates	No	Yes

Notes: Table shows difference-in-differences estimation of school level average outcomes between Low Lead area treated schools and control schools. Standard errors are clustered by school and presented in brackets. Column (1) is estimate without school level covariates, and column (2) with. Covariates include index of multiple deprivation quintile, share of boys in school, and the local youth unemployment rate.

As explained in section 3.4, the identifying assumption with difference-in-difference and continuous treatment when comparing each dose level individually with the control is parallel trends between the control group and the group that has a given dose level. This is analogous to the traditional parallel trends assumption and can be made conditional on covariates. When regressing with all dosages, stronger assumptions are required of either no selection into different does areas on average, or homogenous treatment effects.

In this section we show the school-level average means for each dose group, and event studies for each outcome and group, estimates using (5). Figure 3.6 shows the school-level mean outcomes over time. The Treatment group means are more volatile as they have far lower sample sizes. For the High Lead group, starting from a lower base, there is convergence in outcome post-treatment with the control group. The High Lead group has a higher mean standard grade points achieved, and mean Mathematics Credit passes achieved by the end of the period but remains lower for English. There is less convergence for the Low Lead group, and this group appears to match the patterns on the control group before and after treatment.

To test the conditional parallel trends and no-anticipation assumptions in our two-way fixed effects, we perform event studies. These show the coefficient on treatment interacted with each year of our sample. Figure 3.7 shows the results for the full sample. There does appear to be some increase in coefficient size after treatment, but the confidence intervals are wide.

Figure 3.8 shows the event study for High Lead compared to control. Here we see larger increases in the coefficient after treatment than in figure 3.7 for Standard grade points and Mathematics credit passes, but not for English Credit passes. The confidence intervals are also tighter, widening at the end as we get further from the reference year (1989), as is standard with event studies. Figure 3.9 presents the same results for the Low Lead sample. Here there is some upturn in the coefficients after treatment year, but the effects are more muted than for the high lead sample and the confidence intervals are extremely wide throughout.

Figure 3.6 Mean School Outcomes per Dose Group

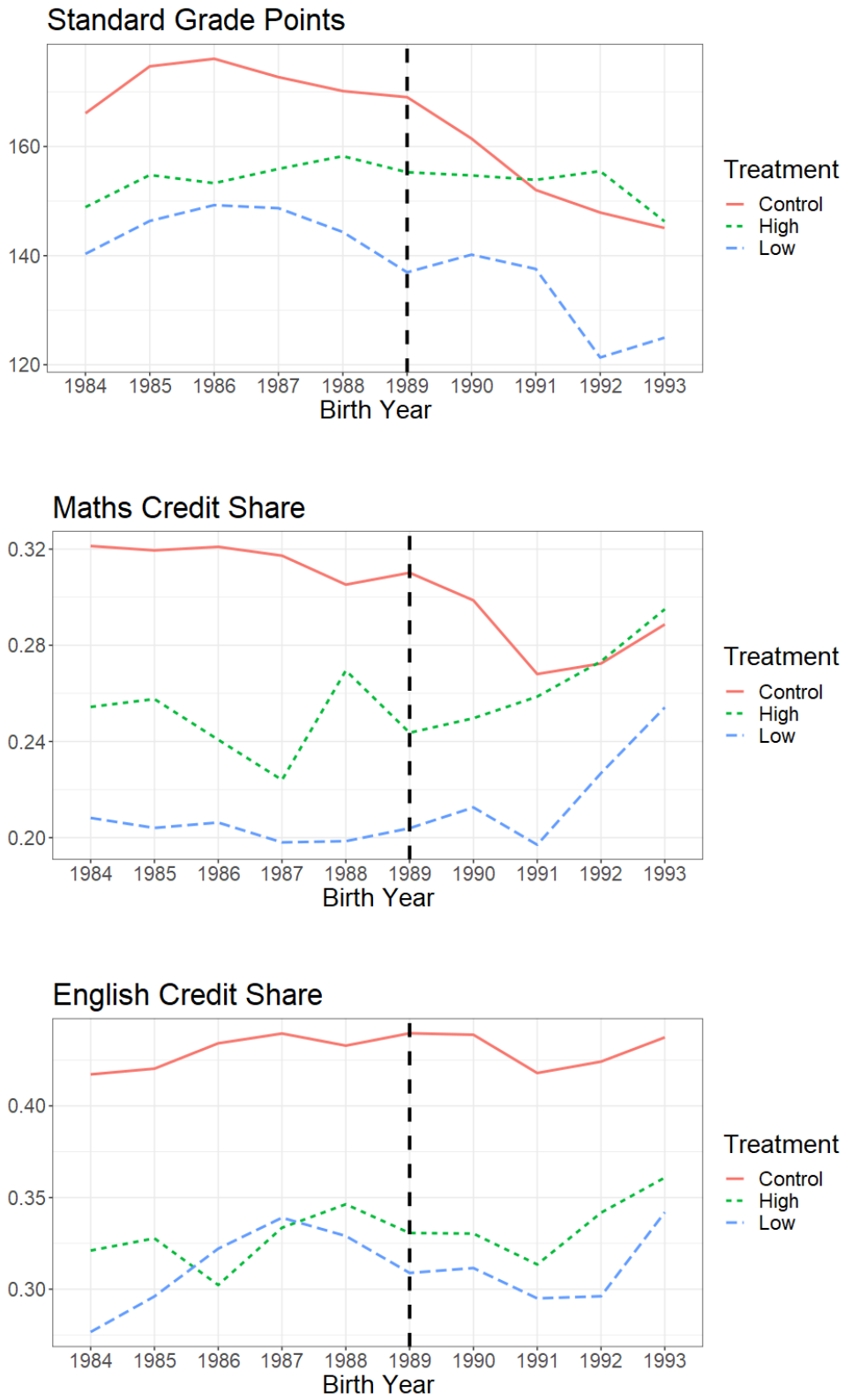
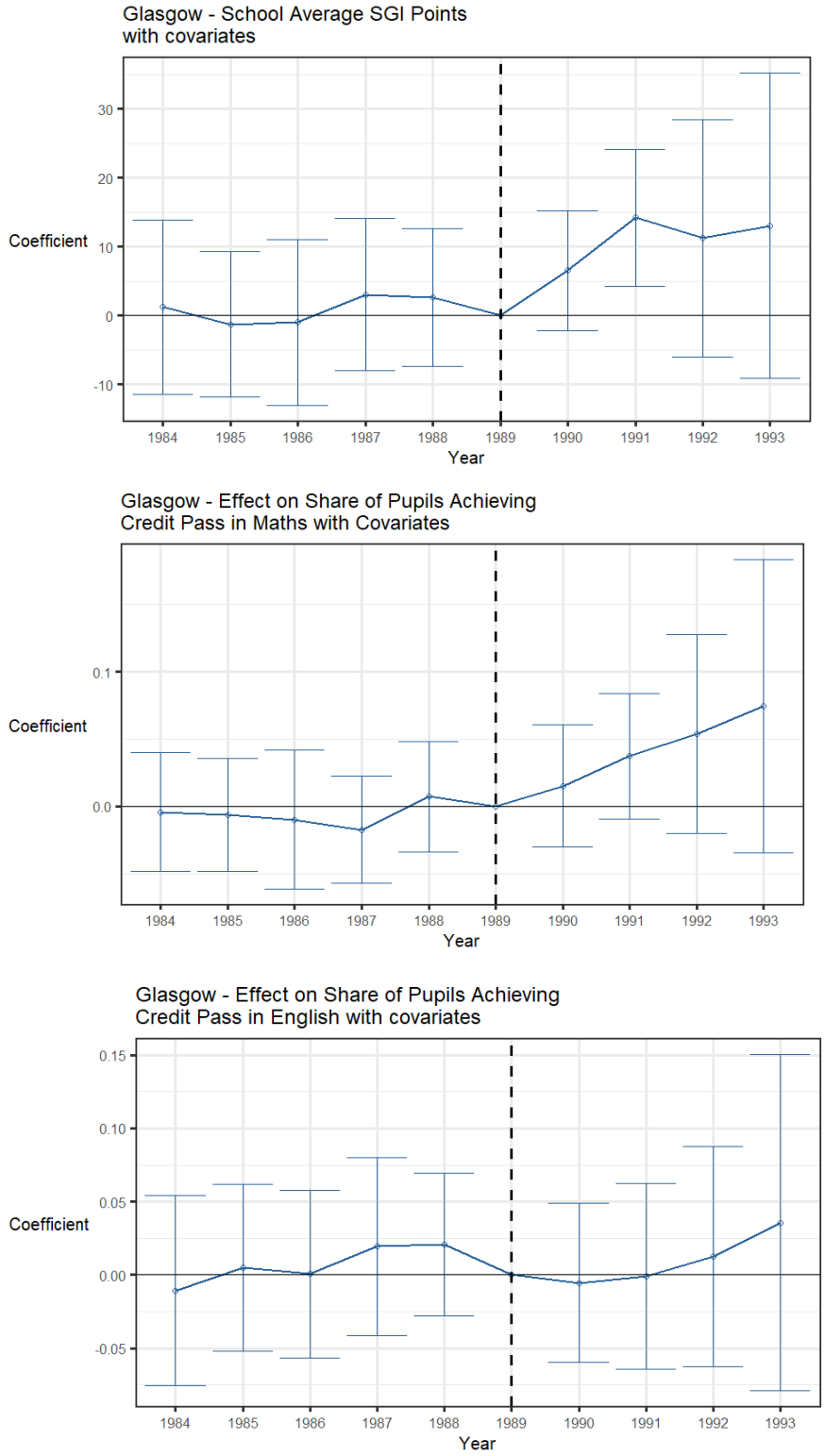
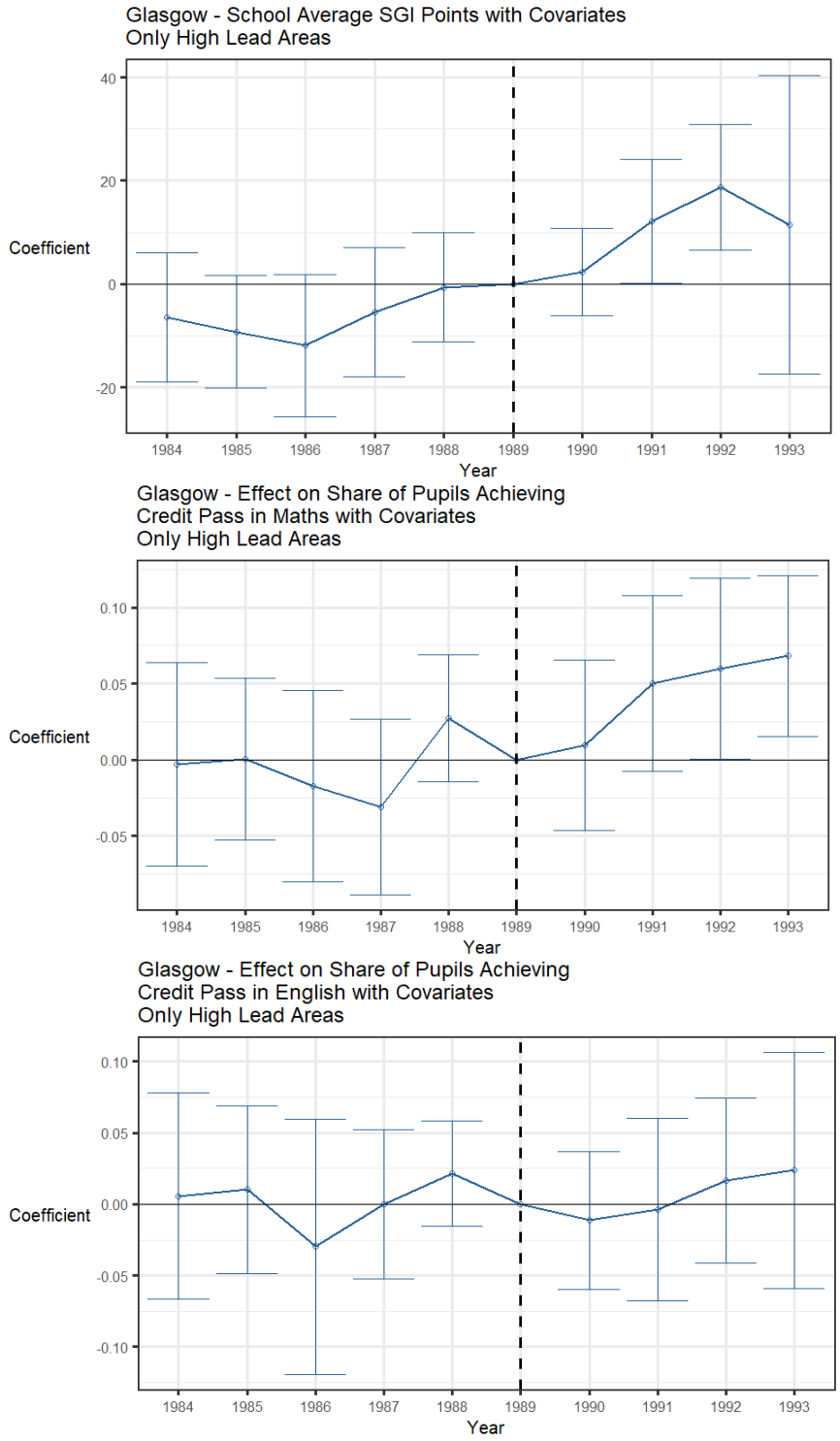


Figure 3.7 Event Studies, All Treatment Areas



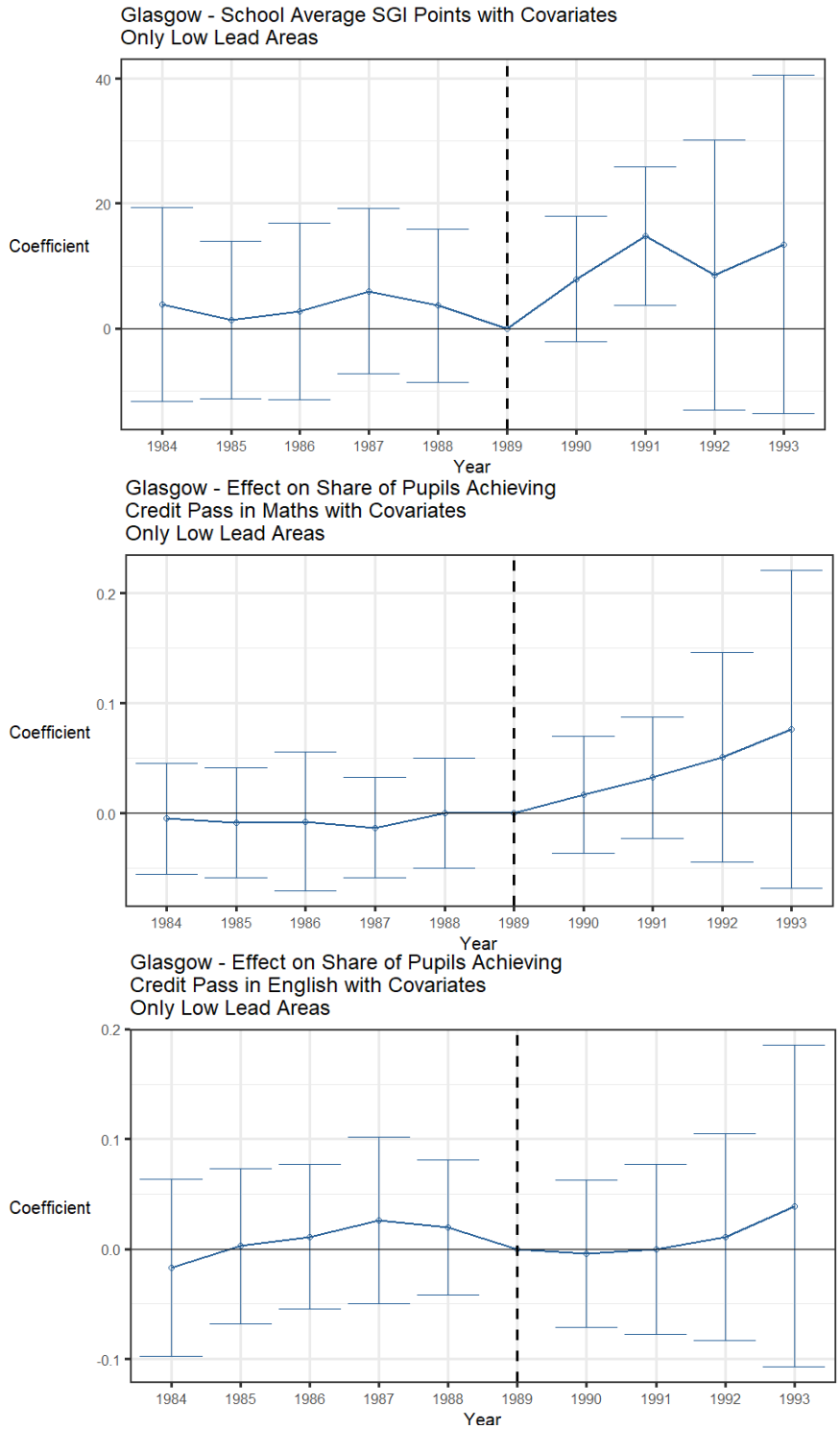
Notes: Charts show event study estimations of an indicator variable for treatment interacted with the year of birth. The outcome variable is the school level-average indicated in each chart heading. Outcomes are averaged by school, and then dosage group. Errors are clustered at school level.

Figure 3.8 Event Studies, High Lead Areas



Notes: Charts show event study estimations of an indicator variable for treatment interacted with the year of birth, but Low Lead areas of the treatment group are excluded from the estimation. The outcome variable is the school level-average indicated in each chart heading. Outcomes are averaged by school, and then dosage group. Errors are clustered at school level.

Figure 3.9 Event Studies, Low Lead Areas



Notes: Charts show event study estimations of an indicator variable for treatment interacted with the year of birth, but High Lead areas of the treatment group are excluded from the estimation. The outcome variable is the school level-average indicated in each chart heading. Outcomes are averaged by school, and then dosage group. Errors are clustered at school level.

3.6 Discussion and Conclusion

We estimated the impact of reduced lead consumption in drinking water among children on their later educational outcomes by using the exogenous variation from a 1989 water treatment program in Glasgow, Scotland as a natural experiment. Our results suggest that lower lead ingestion from water when an infant, and lower maternal lead ingestion when a child is in the womb, leads to better grades at ages 15-16. However, our results show that the positive effects are concentrated within the areas of high lead pipe prevalence, and therefore higher lead-water levels. The levels of lead in the water in Glasgow before the 1989 treatment were generally higher than those seen in Flint, Michigan in 2015. Even after treatment, the distribution of lead-water levels in the High Lead area was similar to that of Flint in 2015 (Table 3.1). This implies that socially significant improvements in education outcomes will only be seen when the reduction in lead pollution is large.

This is in line with the literature on lead outcomes that shows the does-response effects are non-linear. Grönqvist, Nilsson and Robling (2019), show that the effects of lead are low until a threshold of around 5 μg/dl blood lead levels. Reyes (2007) shows that effects are far stronger for the 4th quartile of lead exposure, in some cases 20 times as large as for the 1st quartile. Sampson and Winter (2018) show a clear non-linear increasing relationship between infant blood lead levels and anti-social behaviour in teenagers, with no effect below 5 μg/dl. Our results are in contrast to those arguing the marginal effects are higher at low levels of lead, or that the effects of lead are linear.

There are a number of limitations to this study. Firstly, the treatment group is concentrated in one urban centre, with treatment at one point in time. This potentially limits the external validity of our results. Glasgow notoriously has a number of unexplained poor health outcomes (known as the "Glasgow Effect") and although we estimate the ATT, this tells us little about the effects of treatment on the control group. Secondly, although there is a reasonably

large sample at the individual-sibling level differences in table 3.4, at the school level averages the sample is very limited. It may be that our failure to find larger effects for the low leads sample is related to the small number of schools within Glasgow, and the measurement error in assigning High or Low lead to children within a school's catchment area. A third limitation is that education is a high-order outcome, with many contributing factors. Our results say nothing about other, more direct effects, such as those on health, which may have a different dose-response relationship.

The implications of our findings are that the gains from lead abatement on education are non-linear. Therefore, lead abatement programmes and infrastructure spending should first be targeted at those with the highest levels of lead ingestion when this is possible. This is especially important in low and middle-income countries, where the average blood lead levels are far higher (GBD, 2019). By some estimates 1 in 3 children have blood lead levels above 5 μg/dL, and 280 million in India alone (GBD, 2019). Therefore, while lead abatement in low lead areas may have some benefit, when discretion is possible, resources should be targeted at areas and countries with much higher blood lead levels. We recommend future research on the effect of lead on higher order outcomes, like education, not only test whether an effect is socially significant, but also attempt to map the shape of the dose-response relationship.

References

Aizer, Anna, Janet Currie. 2019. Lead and Juvenile Delinquency: New Evidence from Linked Birth, School, and Juvenile Detention Records. The Review of Economics and Statistics 101 (4): 575–587. https://doi.org/10.1162/rest_a_00814

Aizer, Anna, Janet Currie, Peter Simon, and Patrick Vivier. 2018. Do Low Levels of Blood Lead Reduce Children's Future Test Scores? American Economic Journal: Applied Economics, 10 (1): 307-41. https://doi.org/10.1257/app.20160404

Aguilar-Gomez S., Holt Dwyer, Joshua Graff Zivin, Matthew Neidell, 2022.This Is Air: The Non-health Effects of Air Pollution. Annual Review of Resource Economics, Annual Reviews, vol. 14(1), pages 403-425, October.

Baum, C.R., Shannon, M.W., 1997. The lead concentration of reconstituted infant formula. J. Toxicol. Clin. Toxicol. 35, 371–375. https://doi.org/10.3109/15563659709043369

Brubaker, C.J., Schmithorst, V.J., Haynes, E.N., Dietrich, K.N., Egelhoff, J.C., Lindquist, D.M., Lanphear, B.P., Cecil, K.M., 2009. Altered myelination and axonal integrity in adults with childhood lead exposure: A diffusion tensor imaging study. Neurotoxicology. https://doi.org/10.1016/j.neuro.2009.07.007

Burnet, J., 1869. History of the Water Supply to Glasgow, from the Commencement of the Present Century. With Descriptions of the Water Works Projected, Executed, and from Time to Time in Operation. And an Appendix Containing Tables of Capital, Revenue, and Expenditure of t. Bell \& Bain.

Callaway, B., Goodman-Bacon, A., Sant'Anna, P.H.C., 2021. Difference-in-Differences with a Continuous Treatment. arXiv. https://arxiv.org/abs/2107.02637

Callaway, B., Sant'Anna, P.H.C., 2021. Difference-in-Differences with multiple time periods. J. Econom. 225, 200–230. https://doi.org/10.1016/J.JECONOM.2020.12.001

Canfield, R.L., Henderson, C.R., Cory-Slechta, D.A., Cox, C., Jusko, T.A., Lanphear, B.P., 2003. Intellectual impairment in children with blood lead concentrations below 10 microg per deciliter. N. Engl. J. Med. 348, 1517–1526. https://doi.org/10.1056/NEJMOA022848

Christison, R., 1844. XVI. On the Action of Water upon Lead. Earth Environ. Sci. Trans. R. Soc. Edinburgh 15, 265–276. https://doi.org/10.1017/S0080456800029951

Drinking Water Inspectorate, 2021. Long-term Strategies to Reduce Lead Exposure from Drinking Water. [accessed 2023 Jan 16]. https://www.dwi.gov.uk/research/completed-research/premise-risk/long-term-strategies-to-reduce-lead-exposure-from-drinking-water/.

Ettinger, A.S., Téllez-Rojo, M.M., Amarasiriwardena, C., Bellinger, D., Peterson, K., Schwartz, J., Hu, H., Hernández-Avila, M., 2004. Effect of Breast Milk Lead on Infant Blood Lead Levels at 1 Month of Age. Environ. Health Perspect. 112, 1381. https://doi.org/10.1289/EHP.6616

Evens, A., Hryhorczuk, D., Lanphear, B.P., Rankin, K.M., Lewis, D.A., Forst, L., Rosenberg, D., 2015. The impact of low-level lead toxicity on school performance among children in the Chicago Public Schools: a population-based retrospective cohort study. Environ. Heal. 14. https://doi.org/10.1186/S12940-015-0008-9

Ferrie, J.P., Rolf, K., Troesken, W., 2012. Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence test scores among World War Two U.S. Army enlistees. Econ. Hum. Biol. 10, 98–111. https://doi.org/10.1016/J.EHB.2011.09.003

Gazze, L., Persico, C., Spirovska, S., 2021. The Long-Run Spillover Effects of Pollution: How Exposure to Lead Affects Everyone in the Classroom. SSRN Electron. J. https://doi.org/10.2139/SSRN.3842757

GBD, 2020. Global Burden of Disease (GBD 2019) | Institute for Health Metrics and Evaluation [WWW Document]. Glob. Burd. Dis. - IMHE. URL https://www.healthdata.org/gbd/2019 (accessed 1.26.22).

Goldstein, G.W., 1990. Lead poisoning and brain cell function. Environ. Health Perspect. 89, 91–94. https://doi.org/10.1289/EHP.908991

Goodman-Bacon, A., 2021. Difference-in-differences with variation in treatment timing. J. Econom. 225, 254–277. https://doi.org/10.1016/J.JECONOM.2021.03.014

Grönqvist, H., Nilsson, J.P., Robling, P., 2019. Understanding How Low Levels of Early Lead Exposure Affect Children's Life-Trajectories, Journal of Political Economy. University of Chicago Press. https://doi.org/10.1086/708725

Havari, E., Savegnago, M., 2022. The intergenerational effects of birth order on education. J. Popul. Econ. 35, 349–377. https://doi.org/10.1007/S00148-020-00810-5/FIGURES/3

Hollingsworth A, Huang JM, Rudik I, Sanders NJ. 2022 Oct 7. A Thousand Cuts: Cumulative Lead Exposure Reduces Academic Achievement. J Human Resources.:0222. doi:10.3368/jhr.0222-12169R2.

Jukic, A.M., Baird, D.D., Weinberg, C.R., Mcconnaughey, D.R., Wilcox, A.J., 2013. Length of human pregnancy and contributors to its natural variation. Hum. Reprod. 28, 2848. https://doi.org/10.1093/HUMREP/DET297

Keller, K., Troesch, L.M., Grob, A., 2015. First-born siblings show better second language skills than later born siblings. Front. Psychol. 6, 705. https://doi.org/10.3389/FPSYG.2015.00705/BIBTEX

Kim, E.J., Herrera, J.E., Huggins, D., Braam, J., Koshowski, S., 2011. Effect of pH on the concentrations of lead and trace contaminants in drinking water: A combined batch, pipe loop and sentinel home study. Water Res. 45, 2763–2774. https://doi.org/https://doi.org/10.1016/j.watres.2011.02.023

Lanphear, B.P., 2015. The Impact of Toxins on the Developing Brain. Annu. Rev. Public Health 36, 211–230. https://doi.org/10.1146/annurev-publhealth-031912-114413

Lehmann, J.Y.K., Nuevo-Chiquero, A., Vidal-Fernandez, M., 2016. The Early Origins of Birth Order Differences in Children's Outcomes and Parental Behavior. J. Hum. Resour. 53, 0816–8177. https://doi.org/10.3368/JHR.53.1.0816-8177

Macdonell, J.E., Campbell, H., Stone, D.H., 2000. Lead levels in domestic water supplies and neural tube defects in Glasgow. Arch. Dis. Child. 82, 50–53. https://doi.org/10.1136/adc.82.1.50

Macdonell, J.E., Campbell, H., Stone, D.H., 2000. Lead levels in domestic water supplies and neural tube defects in Glasgow. Arch. Dis. Child. 82, 50–53. https://doi.org/10.1136/ADC.82.1.50

Macintyre, C., Fulton, M., Hepburn, W., Yang, S., Raab, G., Davis, S., Heap, M., Halls, D., Fell, G., 1998. Changes in blood lead and water lead in Edinburgh. An eight year follow-up to the Edinburgh lead study. Environ. Geochem. Health 20, 157–167. https://doi.org/10.1023/A:1006593225113

Masten, S.J., Davies, S.H., McElmurry, S.P., 2016. Flint Water Crisis: What Happened and Why? J. Am. Water Works Assoc. 108, 22. https://doi.org/10.5942/JAWWA.2016.108.0195

Mielke, H.W., Zahran, S., 2012. The urban rise and fall of air lead (Pb) and the latent surge and retreat of societal violence. Environ. Int. 43, 48–55. https://doi.org/10.1016/j.envint.2012.03.005

Miranda, M.L., Kim, D., Galeano, M.A.O., Paul, C.J., Hull, A.P., Morgan, S.P., 2007. The Relationship between Early Childhood Blood Lead Levels and Performance on End-of-Grade Tests. Environ. Health Perspect. 115, 1242. https://doi.org/10.1289/EHP.9994

Moore, M.R., Goldberg, A., Fyfe, W.M., Richards, W.N., 1981. Maternal Lead Levels After Alterations To Water Supply. Lancet 318, 203–204. https://doi.org/10.1016/S0140-6736(81)90384-6

Moore, M.R., Robertson, S.J., Gilmour, W.H., Murray, G.D., Britton, A., Low, R.A., Watt, G.C.M., 1998. Decline of maternal blood lead concentrations in Glasgow. J. Epidemiol. Community Heal. 52, 672–673. https://doi.org/10.1136/JECH.52.10.672

Needleman, H.L., 1992. Human lead exposure. CRC Press.

Needleman, H.L., 1996. Bone Lead Levels and Delinquent Behavior. JAMA J. Am. Med. Assoc. 275, 363. https://doi.org/10.1001/jama.1996.03530290033034

Pieper, K.J., Martin, R., Tang, M., Walters, L., Parks, J., Roy, S., Devine, C., Edwards, M.A., 2018. Evaluating Water Lead Levels During the Flint Water Crisis. https://doi.org/10.1021/acs.est.8b00791

Quinn, M.J., 1985. Factors Affecting Blood Lead Concentrations in the UK: Results of the EEC Blood Lead Surveys, 1979–1981. Int. J. Epidemiol. 14, 420–431. https://doi.org/10.1093/IJE/14.3.420

Rau, T., Urzúa, S., Reyes, L., Ashenfelter, O., Beaman, L., Card, D., Jacob, B., Kline, P., Lange, F., Palloni, G., Smith, J., Thomas, D., 2015. Early Exposure to Hazardous Waste and Academic Achievement: Evidence from a Case of Environmental Negligence. JAERE 2. https://doi.org/10.1086/683112

Reyes, J., 2007. Environmental policy as social policy? The impact of childhood lead exposure on crime. B E J. Econ. Anal. POLICY 7, 51. https://doi.org/10.2202/1935-1682.1796

Reyes, J.W., 2015. Lead exposure and behavior: Effects on antisocial and risky behavior among children and adolescents. Econ. Inq. 53, 1580–1605. https://doi.org/10.1111/ecin.12202

Sampson, R.J., Winter, A.S., 2018. Poisoned development: assessing childhood lead exposure as a cause of crime in a birth cohort followed through adolescence. Criminology 56, 269–301. https://doi.org/10.1111/1745-9125.12171

Schwartz, J., 1994. Low-level lead exposure and children's IQ: a meta-analysis and search for a threshold. Environ. Res. 65, 42–55. https://doi.org/10.1006/enrs.1994.1020

Sherlock, J.C., Ashby, D., Delves, H.T., Forbes, G.I., Moore, M.R., Patterson, W.J., Pocock, S.J., Quinn, M.J., Richards, W.N., Wilson, T.S., 1984. Reduction in exposure to lead from drinking water and its effect on blood lead concentrations. Hum. Toxicol. 3, 383–392. https://doi.org/10.1177/096032718400300503

Sindhu, K.K., Sutherling, W.W., 2015. Role of Lead in the Central Nervous System: Effect on Electroencephlography, Evoked Potentials, Electroretinography, and Nerve Conduction. Neurodiagn. J. https://doi.org/10.1080/21646821.2015.1043222

Thomson, G.O.B., Raab, G.M., Hepburn, W.S., Hunter, R., Fulton, M., Laxen, D.P.H., 1989. Blood-lead levels and children's behaviour--results from the Edinburgh Lead Study. J. Child Psychol. Psychiatry. 30, 515–528. https://doi.org/10.1111/J.1469-7610.1989.TB00265.X

Troesken, W., 2006. The great lead water pipe disaster. MIT Press.

Vlasak, T., Jordakieva, G., Gnambs, T., Augner, C., Crevenna, R., Winker, R., Barth, A., 2019. Blood lead levels and cognitive functioning: A meta-analysis. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2019.03.052

Wani, A.L., Ara, A., Usmani, J.A., 2015. Lead toxicity: a review. Interdiscip. Toxicol. 8, 55. https://doi.org/10.1515/INTOX-2015-0009

Watt, G.C.M., Britton, A., Gilmour, H.G., Moore, M.R., Murray, G.D., Robertson, S.J., 2000. Public health implications of new guidelines for lead in drinking water: a case study in an area with historically high water lead levels. Food Chem. Toxicol. 38, S73–S79. https://doi.org/10.1016/S0278-6915(99)00137-4

Watt, G.C.M., Gilmour, W.H., Moore, M.R., Murray, G.D., Britton, A., Robertson, S.J., Womersley, J., 1996. The Glasgow 93 Lead Study. Great Glasgow Health Board, Internal Report.

Watt, G.C.M., Gilmour, W.H., Moore, M.R., Murray, G.D., Britton, A., Robertson, S.J., Womersley, J., 1996. Is lead in tap water still a public health problem? An observational study in Glasgow. BMJ 313, 979–981. https://doi.org/10.1136/bmj.313.7063.979

Zheng, J., 2021. Lead exposure, human capital formation, and inequality: the impacts of drinking water lead on short-run and long-run educational outcomes. Working paper. https://www.jiamengzheng.com/assets/files/lead_JMP_zheng.pdf

Chapter 1 Appendices

The Lead-Crime Hypothesis: A Meta-Analysis[16]

[16] Data and code available at https://anthonychigney.github.io/home/Research/

A. Review of literature used in meta-analysis

There are 24 total studies included in this meta-analysis. The studies use different methods to examine the lead-crime relationship. Longitudinal studies, which track the same people over time, are common. Fergusson, Boden and Horwood (2008) use a longitudinal sample and find a positive association between dentine lead levels at 6-9 years of age and later offending while including race and family socioeconomic status covariates. However, the effect was smaller once variation in education grades was added. They reasoned that the effect of lead was in reducing education outcomes, leading to more crime. Overall, they find that lead only explains 1% of the variation in crime. Nkomo *et al*. (2017) used a longitudinal sample in South Africa and found a positive association between blood lead levels at age 13 and violent crime in later life. Beckley *et al*. (2018) find only a small positive effect of childhood lead levels and both violent and non-violent crime in their longitudinal sample of New Zealand residents. They conclude other factors are much more important for determining crime rates. Finally, Sampson and Winter (2018) follow a longitudinal sample in Chicago and find school age lead levels are not associated with an increase in arrests in later life. Overall, longitudinal studies show a mixed picture, both on whether there is an effect and whether it is a strong one.

A different strand of research looks at the correlation of lead levels and crime across time and areas, rather than at an individual level. Three studies look at time series of lagged lead levels and crime for the US. Nevin (2000) finds a positive effect, but McCall and Land (2004) find no effect on the age cohorts most affected in youth by the increase in leaded gasoline. They reason that increased lead levels at one time should only affect the crime rates of that cohort, not earlier cohorts, and so only look at crime rates for those certain age ranges. Lauritsen, Rezey, and Heimer (2016) look at two different data series

of crime: the National Crime Victimization Survey (NCVS) and the Uniform Crime Reports (UCR). They find that lead is positively correlated with violent crime in the UCR but not the NCVS, which they consider a better measure of violent crime. However, they consider both data sources equally valid for property crime. Stretesky and Lynch (2004) find a strong effect when looking across US countries for both property and violent crime using the UCR. Mielke and Zahran (2012) find a strong effect across six US cities, Lersch and Hart (2014) find the same looking at Florida census tracts. Both Barrett (2017) and Manduca and Sampson (2019) find a strong positive relationship in census tracts in Chicago using different methods. Looking outside the US, Taylor *et al*. (2018) find positive results for violent crime in Australia, and across six suburbs in New South Wales. Nevin (2007) estimates the relationship for many OECD countries and finds pre-school blood levels are strongly associated with a whole range of violent and non-violent crime. On the whole, studies which look at geographic areas as the unit of interest tend to find the strongest positive associations between lead and crime.

The final strand of the literature are those studies that attempt to identify a casual effect while accounting for endogeneity from unobserved variables correlated with both crime and lead. These could bias the estimate of the effect of lead on crime. Lead exposure is correlated with poverty (Baghurst *e al*. 1999) and race (Sampson and Winter, 2016) and likely with other, unobservable, variables. We cannot rule out that these variables may cause individuals to commit more crime and be more exposed to lead, rather than lead being the cause. Even panel data designs with controls may not account for this endogeneity. The endogeneity threat has led to some, more recent, studies using quasi-experimental methods. Needleman (2002) carried out a "case control" study where young offenders were matched to a "control" group chosen for similar observable characteristics. The offender group was found to have higher bone lead levels. Although this this is an improvement beyond

looking at correlation alone, the likelihood of unobservable group differences means that the problem of endogeneity was not adequately resolved.

Reyes (2007) is the first study to use quasi-experimental methods to derive a causal estimate. She uses the different grades and concentration of lead in gasoline in US states as an instrumental variable for lead levels. She finds an effect of lead on violent crime but not property crime. In a later paper (2015) she uses a similar identification strategy with individual-level data. Here she finds a positive effect on both property and violent crime. Feigenbaum and Muller (2016) also use an instrumental variable strategy. They instrument for the presence of lead water pipes in US cities using the distance to the nearest lead refinery in 1899, a period in which thousands of US cities built their water supplies. They find a positive causal effect on homicides in 1921-1936. Aizer and Currie (2018) use nearby traffic volume interacted with year of birth as an instrument for lead and include sibling fixed effects. They find a positive relationship between lead and incarceration. Curci and Masera (2018) also find a positive association when they look across 300 US cities. Most of the estimates from this paper do not fall under the "addressing endogeneity" category, but in one chart of estimates they use soil quality as an instrument for lead. Grönqvist *et al*. (2019) use a sample of 800,000 Swedish children grouped by neighbourhoods and cohorts. They instrument for blood lead levels by the lead measured in moss in the areas. The estimates are mixed but tend to show a small positive effect on crime. Finally, Billings and Schnepel (2018) match a treatment group of children who had blood lead levels above a 10μg/dL threshold in two tests, with a control group of children who were above the threshold in the first test and just below in the second test, thus failing to qualify for treatment. This, close to randomised control trial, study finds a positive effect of lead on crime, with a stronger effect on property crime than violent crime. Overall, the few studies that use quasi-experimental methods all find a positive effect on crime, but they tend to find a smaller effect than the studies that look at correlations across geographic areas.

B. Converting to common estimates

To conduct a meta-analysis all estimates must be converted to a common metric. We use both elasticities and partial correlation coefficients (PCCs). We calculate the PCC as shown in equation (I):

$$\text{(I)} \qquad PCC_{ij} = \frac{t_{ij}}{\sqrt{t_{ij+df_{ij}}^2}}$$

Where t_{ij} is the t-ratio for estimate *i* of study *j*, and df_{ij} is the degrees of freedom. The standard error of each PCC is calculated according to equation (II):

$$\text{(II)} \qquad SE_{ij} = \frac{PCC_{ij}}{t_{ij}}$$

Some papers reported odds ratios rather than correlation coefficients. Following Polanin and Snilstveit (2016), we converted these to PCCs.

$$\text{(III)} \qquad PCC_{ij} = \frac{ln(OR_{ij}) \times \left(\frac{\sqrt{3}}{\pi}\right)}{\sqrt{\left(\ln(OR_{ij}) \times \left(\frac{\sqrt{3}}{\pi}\right)\right)^2 + a_{ij}}}$$

Where OR_{ij} is the odds ratio *i* for study *j* and $a_{ij} = \frac{\left(n_{ij1}+n_{ij2}\right)^2}{n_{ij1}n_{ij2}}$. Here a_{ij} is a correction factor which depends on the sample size in the control and treatment groups (n_{ij1} and n_{ij2}). If the sample sizes are unknown, or there are no treatment and control groups, we follow Borenstein *et al*. (2009) and set them to be equal, which gives $a = 4$.

In a similar way we calculate standard error equivalents for odds ratio estimates. Following the Cochrane Handbook (Higgins and Green, 2011), first we convert the 95% confidence intervals to odds ratio standard errors (ORSE).

$$\text{(IV)} \quad ORSE_{ij} = \frac{(\ln(\overline{CI}) - \ln(\underline{CI}))}{3.92}$$

Where $\overline{CI}$ is the upper confidence interval limit and $\underline{CI}$ is the lower confidence interval limit. I then convert this into partial correlation coefficient standard errors.

$$\text{(V)}\ SE_{ij} = \sqrt{\frac{(a^2 \times ORSE_{ij}^2 \times \left(\frac{3}{\pi^2}\right)}{\left(\left(\log(OR_{ij}) \times \left(\frac{\sqrt{3}}{\pi}\right)\right)^2 + a\right)^3}}$$

Only one study (Billings and Schnepel, 2018) has estimates which are similar to randomised control trial estimates, with a mean difference shown between control and treatment groups. These can also be converted to PCCs. For these we follow Borenstein *et al*. (2009) and first compute the within-groups standard deviation $SD_{ij}$for estimate *i* of study *j,* as shown in (VI).

$$\text{(VI)}\ SD_{ij} = \sqrt{\frac{(n_{ij1}-1) \times S_{ij1}^2 + (n_{ij2}-1) \times S_{ij2}^2}{n_{ij1}+n_{ij2}-2}}$$

Here, n_{ij1} is the sample size for the control group for *i* of study *j*, S_{ij1} is the standard deviation for the control group, while n_{ij2} and S_{ij2} are the same from the treatment group.

We use this to calculate Cohen's D:

$$\text{(VII)}\ D_{ij} = \frac{\bar{X}_{ij1} - \bar{X}_{ij2}}{SD_{ij}}$$

Where $\bar{X}_{ij1}$ is the sample mean for the control group and $\bar{X}_{ij2}$ for the treatment group. Finally, we convert Cohen's D to a PCC by equation (VIII).

(VIII) $$PCC_{ij} = \frac{D_{ij}}{\sqrt{D_{ij}^2 + a_{ij}}}$$

Here a_{ij} is the same as that for equation (III) except we have the sample sizes for each group so we do not set it to equal 4. The variance for Cohen's D is calculated as in (IX).

(IX) $$DVar_{ij} = \frac{n_{ij1} + n_{ij2}}{n_{ij1} \times n_{ij2}} + \frac{D_{ij}^2}{2(n_{ij1} + n_{ij2})}$$

This is then used to calculate the standard error of the PCC.

(X) $$SE_{ij} = \sqrt{\frac{a_{ij}^2 \times DVar_{ij}}{\left(D_{ij}^2 + a_{ij}\right)^3}}$$

One further study only uses simple correlations (Lauritsen *et al*., 2016). The standard errors for these must be approximated. We use the approximation of one divided by n-3 for the correlation standard errors, as n is the same for all estimates, the standard errors are the same for all these estimates.

C. Common effects and random effects meta-analysis

C.1 Common and Random Effects Weighted Averages

This section explains how common and random effects meta-analysis estimates are calculated.

Before calculating commom or random effects estimates, first we convert all PCCs to normalised estimates with equation (XI), so that correct confidence intervals can be calculated.

(XI) $Z_{ij} = 0.5\ ln\left(\frac{1+PCC_{ij}}{1-PCC_{ij}}\right)$

Where Z_{ij} is the normalised effect size of a PCC. The process is that first PCCs are converted to normalised estimates, we estimate using either common effects or random effects, then the estimates are converted back to a PCC with equation (XII).

(XII) $PCC = \frac{e^{2z}-1}{e^{2z}+1}$

Where in this case the PCC is the meta-analysis estimate as a correlation coefficient, and Z is the estimate obtained from the normalised PCCs.

To calculate the common effects averages we weight each estimate by the inverse of the variance, and then divide the sum of these weighted estimates by the sum of the weights as shown in following two equations:

(XIII) $W_{ij} = \frac{1}{V_{ij}}$

(XIV) $FE = \frac{\sum_{i=1}^{N} W_{ij} Z_{ij}}{\sum_{i=1}^{N} W_{ij}}$

Where V_{ij} is the variance of estimate i of study j , FE is the fixed effects average, and Z_{ij} is normalised PCC. This average is converted back into a PCC by equation (XII). Along with the averages I calculate 95% confidence intervals, first by obtaining the standard errors of FE.

(XV) $SE_{FE} = \sqrt{\frac{1}{\sum_{i=1}^{k} W_{ij}}}$

Then obtaining lower and upper limits in the normal fashion. The fixed effect averages and standard error can be used to calculate Z-scores for hypothesis testing as normal.

Random effects estimates are estimated in the same way as fixed effects, except we replace V_{ij} in equation (XIII) with V_{ij}^{*}. Where $V_{ij}^{*} = V_{ij} + T^2$, and T^2 is an estimate of the between-study variation. There are different methods of estimating T^2, we use the DerSimonian-Laird (1986) method.

C.2 Estimating Heterogeneity

We use three measures of heterogeneity in our meta-analysis H^2, I^2, and τ^2. Each attempts to quantify the heterogeneity in study effect sizes. Estimating these is inference on the dispersion of θ_j, as outlined in the main text.

These methods all use Cochran's Q statistic in their calculations. The Q statistic is a estimate of the variation in the true effect sizes θ_j, compared to the sampling variation. It is calculated as below:

(XV) $\quad Q = \sum_{i=1}^{N} W_{ij} Z_{ij}^2 - \frac{\left(\sum_{i=1}^{N} W_{ij} Z_{ij}\right)^2}{\sum_{i=1}^{N} W_{ij}}$

If Q is large, it means that a relatively larger share of the variation in observed effect sizes is due to differences in each study's true effect size θ_j, rather than due to sampling variation. Under the null hypothesis of no

difference in θ_j the Q statistic will be Chi-square distributed with N-1 degrees of freedom.

Simply testing for completely homogeneous effects is extreme, given we assume effect size heterogeneity throughout the analysis (see section 3.4). Therefore we move on to testing how heterogeneous the effects are with the three statistics we use.

τ^2 is an estimate of variance of θ_j, the "true" effect size distribution. It is calculated as:

(XVI) $\tau^2 = \frac{Q-df}{C}$

Where df is the degrees of freedom and C, a variable that transforms the Q statistic back into the original units of analysis (either PCCs or elasticities in our case). It is calculated as:

(XVII) $C = \sum_{i=1}^{N} W_{ij} - \frac{\left(\sum_{i=1}^{N} W_{ij}\right)^2}{\sum_{i=1}^{N} W_{ij}}$

The larger τ^2 is, the larger the estimated variance in "true" effect sizes between studies.

I^2 attempts to quantify what proportion of the observed variance is due to sampling errors, against the proportion due to study effect size heterogeneity. It is a figure between 0% and 100%. Very high I^2 means that most of the observed variation is due to effect size variation between studies. I^2 is calculated as:

(XVIII) $I^2 = \left(\frac{Q-df}{Q}\right) \times 100\%$

Finally, H^2 is:

(XIX) $H^2 = \frac{Q}{df}$

If H^2 = 1 then there is no variation in study effect sizes. It has no upper bound, and the greater it is the larger the between-study heterogeneity.

D. Publication bias adjustment

We use seven methods to obtain an estimate of the average effect after adjusting for publication bias. This section describes those methods in more detail.

All publication bias methods either test or assume that the observed sample distribution is a truncated version of the underling population distribution.

We have no details about the missing values (i.e. this is not a censored distribution). Therefore, selection models using observations (such as in Heckman, 1976) are not possible.

The publication bias methods rely on assumptions about the truncation process that generates the selection bias which causes the observed distribution to differ from the population distribution. The observed sample and the selection bias assumptions are combined in some estimation procedure, and this produces an estimate which is adjusted for the publication bias, if it is found to be present. In some cases when tests reject publication bias there is no adjustment, and the estimate collapses into either the common or random effects estimate.

Linear Methods

The first four methods are all linear regressions based on the PET-PEESE method. The PET-PEESE is itself an extension of the Egger (1997) test. The Egger test is a simple regression of the effect size on the standard error. A t-test on the standard error coefficient is a test of publication bias where ***H0*** = no publication bias, and ***H1*** = there is publication bias.

Stanley and Doucouliagos (2014) note the heteroskedasticity in the Egger test, as more precise effect sizes (assuming a shared effect size distribution and that estimates also have sampling error) will tend to be closer together. Therefore, they extend the Egger test by using weighted least squares, with the weights being the inverse of the standard errors themselves, which are an estimate of this heteroskedasticity. The coefficient on the precision (1/SE) is the Funnel Asymmetry Test (FAT). The intercept in this model becomes the Precision Effect Test (PET). The FAT is an estimate of the bias, the sign of which indicates the direction of the bias. The PET is an estimate of the

average effect size when publication bias is zero, i.e., the effect size population mean.

The coefficient on the FAT approximates the inverse Mills' ratio. However, this is not a constant, it varies with the standard error. Therefore, Stanley and Doucouliagos (2014) propose using a Taylor expansion around the standard error to better approximate the inverse Mills' ratio. In theory, any number of additional polynomials could be included in the regression, but sample size restrictions in meta-analysis, and the decreasing returns on including more polynomials, mean that few meta-analyses go beyond a cubic term. Stanley and Doucouliagos (2014) propose constraining the linear term on the standard error to be zero and using a squared term. This is the Precision Effect Estimate with Standard Error (PEESE) test. They find in simulations that this performs better than the FAT-PET when the "true" mean of the population of estimates is not equal to zero. This is the second method we use.

The third method is simply the FAT-PET but including study fixed effects. This is more efficient than the standard the FAT-PET, assuming the common effects model is not true for the population. This is estimated with restricted maximum likelihood, which adjusts the degrees of freedom downward for each study fixed effect, without which the variance of the error is biased downwards.

The fourth method we use is the FAT-PET with an instrumental variable. There are other reasons beyond publication bias why the effect size might be correlated with the standard error. For example, regression discontinuity designs (although there are none in our sample) converge at a rate at least as slow as the cubed root of the sample size. Whereas OLS converges at a rate of the root of the sample size. A regression discontinuity with the same sample size will tend to have larger standard errors than the simple OLS regression.

The effect size will also be different, perhaps because they estimate different estimands, or perhaps because the bias is larger in the OLS sample. Similarly, two stage least squares will tend to have larger errors even if it is estimating the same estimand as OLS. Therefore, the coefficient on the standard error may not be a good approximation of the inverse Mills' ratio.

An alternative strategy is to use the inverse of the square root of the sample size as an instrumental variable for the standard error. The sample size is correlated with the standard error. Assuming no relationship between sample size and the effect size beyond its relationship to the standard error (the exclusion restriction), then it will give a better estimate of publication bias and therefore a better PET estimate.

Non-linear methods

The weighted average of adequately powered estimates (WAAP) developed by Stanley, Doucouliagos, and Ioannidis (2016) estimates a common effects weighted average using only high-powered studies. Studies are discarded if they do not meet some power threshold given by:

(D.1) $$\frac{\widehat{\mu}_w}{2.8}$$

Where $\hat{\mu}_w$ is some estimate of the average effect, and the 2.8 denominator comes from the sum of two t-distributed test standard deviations, $t_{1-\frac{\alpha}{2}} + t_{(1-\beta)}$. Following convention, the critical value of the test of the null is set as $\alpha = 0.05$, and the power of the test is set as 80%, so that $\beta = 20\%$. This gives a sum of 1.96 + 0.84 = 2.8. Stanley, Doucouliagos, and Ioannidis (2016) suggest using the common effects estimate as the value $\hat{\mu}_w$. Given the very small common effects estimate in our sample this would only leave only one study, that of Grönqvist, Nilsson and Robling (2019). This would mean the WAAP collapses into the weighted average estimate in table 1. To be more

generous to the Lead-Crime hypothesis, we instead use the larger random effects estimate as $\hat{\mu}_w$. The studies and number of estimates from each considered to be adequately powered under this method is given in table D.1.

Table D.1 – Studies and estimates used in WAAP

Study	**Estimates**
Aizer & Currie (2019)	6
Beckley et al. (2018)	10
Billings & Schnepel (2018)	3
Curci & Masera (2018)	97
Feigenbaum & Muller (2016)	43
Fergusson et al. (2008)	6
Grönqvist, Nilsson and Robling (2019)	54
Lersch & Hart (2014)	2
Manduca & Sampson (2019)	2
Masters et al. (1998)	3
Mielke & Zahran (2012)	1
Nevin (2000)	1
Nevin (2007)	26
Nkomo et al. (2017)	10
Reyes (2007)	65
Reyes (2015)	13
Stretesky & Lynch (2004)	20

Trim and Fill first ranks studies by the absolute value of their effect sizes, then estimates how many effect sizes are missing from either the positive or negative side of the distribution (the negative side in our case). Importantly, these studies are assumed to be not observed with probability one. This contrasts with other methods which estimate the publication probabilities over certain intervals (such as Andrews-Kasy). The trim-and-fill method then uses an iterative algorithm to obtain an average effect estimate.

1. First obtain the random effects estimate from the full sample, use this to estimate the number of missing studies (they propose three different estimators for this).
2. Using the estimate for number of missing studies on the negative side, an equal number of studies are "trimmed" from the sample on the positive side, starting with the largest and moving down.
3. Now obtain another random effects estimate from the trimmed sample and use this to again estimate a number of missing studies.
4. Continue until the random effects estimate of iteration j is equal to the estimate of iteration $j - 1$.
5. Now add the "fill", where imputed values are added to the negative side of the distribution, using the estimates obtained in the last iteration and the most positive values in the sample left after "trimming" (see section 5 in their paper).
6. Finally, obtain a new random effects estimate using the full initial sample, plus the imputed "filled" values.

This method adds 226 estimates to the full sample trim and fill, 82 to the elasticity sample, and 11 to the representative estimates sample.

In the Andrews and Kasy (2019) method, they use a step function to estimate the probability of observing an effect over various intervals of the distribution. This contrasts with the trim and fill, where some observations are assumed missing with probability one, and the FAT-PET, which uses an approximation of the inverse Mills' ratio to deal with the truncation.

They observe, however, that the publication probabilities can only be identified up to scale. That is, we cannot know that absolute probability of publication over any one interval. Therefore, we must estimate relative publication probabilities. We do this by setting one publication probability as the reference probability, and then identifying the others up to scale, i.e.,

relative to this one. In our case the reference probability is the probability of observing a positive effect size that is significant at the 5% level. This probability is set at some arbitrary value (one in our case) and the other probabilities estimated relative to this. If the estimated probabilities are less than one, then they are less likely to be observed than positive values significant at the 5% level, and vice versa.

With relative probabilities estimated, the distribution is reweighted using the relative probabilities to reconstruct the true untruncated distribution. We can use this to get an estimate of the population mean, adjusting for the publication bias. We use the maximum likelihood approach and algorithm in Hedges (1992) as recommended by Andrews and Kasy (2019) to do this. In the case of only using representative estimates, we did not achieve convergence.

The estimates publication probabilities over different z-score intervals are shown below for the full sample and the elasticity sample.

Figure D.1 – Estimated relative publication probabilities, partial correlations

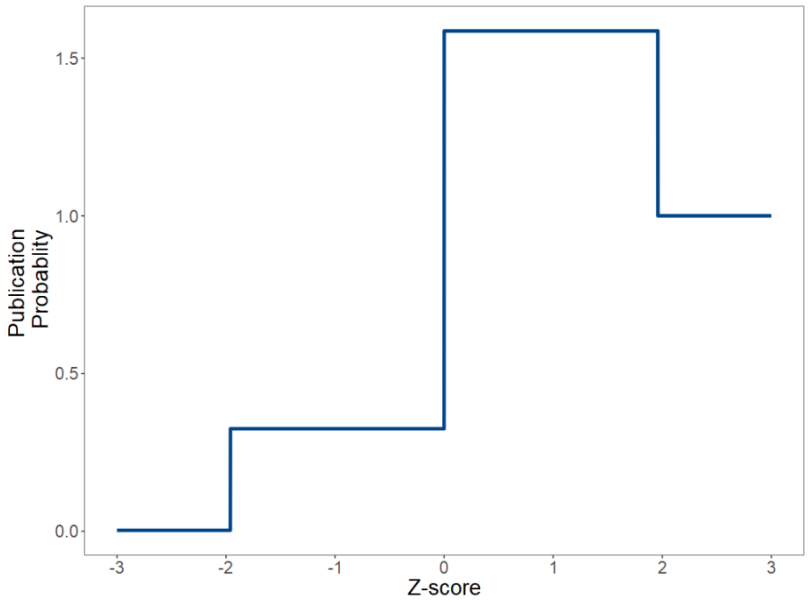
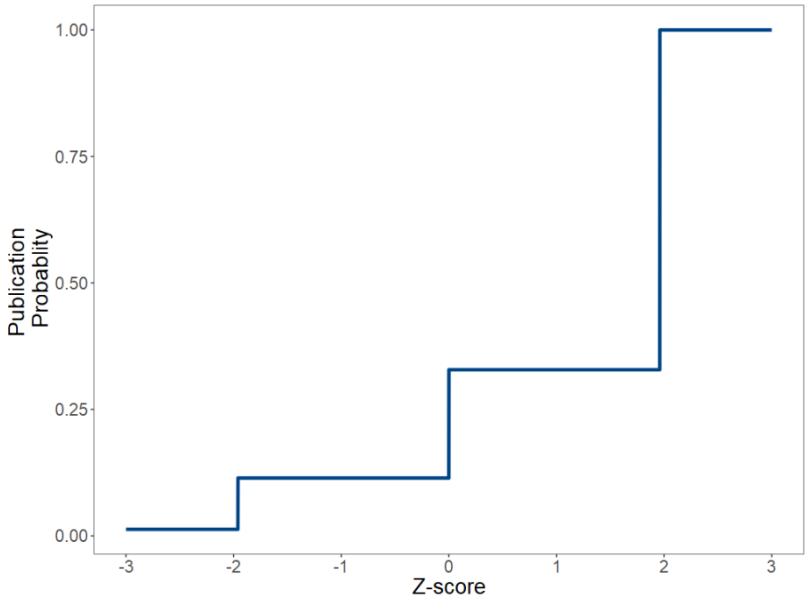


Figure D.2 – Estimated relative publication probabilities, elasticities



E. Analysis using only representative estimates

In most of our analysis we use all estimates. As a robustness check, here we use only one representative estimate from each paper. There was not always a clear representative estimate from each study. Therefore, choosing the estimates involves some subjective judgement. We tried to choose results mentioned in the abstract or as the main result. In general, we chose representative estimates which were less specific (i.e., totals preferred to subsample male/female, white/black results etc.), and estimates obtained using more covariates for correlational results.

In section 1.4.3 we test for publication bias using all estimates. In table E.1 we repeat the exercise using only the representative estimates. However, we cannot estimate the hierarchical model, or cluster errors as we only have one estimate per study. Furthermore the Andrews-Kasy method, using maximum likelihood, did not converge.

Table E.1 – Effect beyond bias and publication bias estimates using representative estimates, partial correlations

	FAT-PET	FAT-PEESE	IV	WAAP	TF
Full Sample, PCCs					
Effect Beyond Bias	-0.001	0.007	-0.001	0.007	0.015
	(0.002)	(0.004)	(0.002)	(0.004)	(0.059)
Publication bias	3.717	12.152	3.733	.	.
	(0.894)	(6.998)	(0.880)		
Groups	24	24	24	.	.

Notes. Estimates are PCCs presented with their standard errors in brackets. FAT-PET is Funnel Asymmetry test and Precision Effect Test (Stanley and Doucouliagos, 2014). FAT-PEESE is Funnel Asymmetry Test and Precision Effect Estimate with Standard Error. The multi-level FAT-PET is a mixed effects-multi-level model with a different slope coefficient for each study. IV is a FAT-PET regression with square root of sample size used as an instrumental variable for the precision using two stage least squares. WAAP (Stanley, Doucouliagos, & Ioannidis, 2017) is the Weighted Average of Adequately

Powered Estimates, where studies below a certain estimated power are removed before calculating the effect. Trim and fill (Duval & Tweedie, 2000), removes outlier studies and then adds imputed studies before calculation an average effect. The Andrews-Kasy (Andrews & Kasy,2019) method is a step function selection model which reweights the observed sample with estimated publication probabilities. See Online appendix D for full explanation of each method.

F. Bayesian Model Averaging (BMA)

We carry out two forms of Bayesian model averaging: 1) we obtain an ensemble estimate of the effect beyond bias, using both linear and non-linear publication bias correction models, 2) we take model averages over all covariates used in the meta-regressions.

Table F.1 presents Bayesian model averages of publication bias correction models. We use the RoBMA R package of Bartoš *et al*. (2021).

Table F.1 – Effect beyond bias, Bayesian model averages

	Full Sample, PCCS	**Endogeneity Sample, PCCS**	**Representative Estimates, PCCs**	**Elasticities**	**Elasticities, Endogeneity sample only**
Effect Beyond Bias	-0.17 (-0.288,0.000)	0.001 (0.000,0.006)	-0.091 (-0.667, 0.000)	0.087 (0.059, 0.114)	0.003 (0.000,0.037)
Observations	542	220	24	312	211

Notes. Table shows Bayesian model average estimates from various publication bias models. Upper and lower bounds of 95% credibility estimates presented in brackets.

We also carry out Bayesian model averaging with all variables used in our meta-regression analysis. We estimate a normal-gamma conjugate model

with a uniform model prior and unit information g-prior. These are the same as in Bajzik *et al*. (2019), see there for more information. The results are given below in table F.2.

Table F.2 – Posterior results from Bayesian model averaging, PCC

Variable	**Posterior Mean**	**Posterior Standard Deviation**	**Posterior Inclusion Probability**
Precision	0.35	0.03	1.00
Control gender	0.04	0.03	0.71
Control race	0.00	0.01	0.09
Control income	-0.04	0.03	0.70
Control education	0.00	0.00	0.05
Homicide	-0.03	0.03	0.54
Violent	0.00	0.02	0.14
Non_Violent	-0.01	0.03	0.34
Area	0.24	0.03	1.00
OLS	0.03	0.03	0.55
ML	0.04	0.03	0.81
Odds_Ratio	-0.04	0.07	0.35
Panel dummy	-0.17	0.02	1.00
Addressing Endogeneity	0.00	0.00	0.05
North_America	-0.41	0.03	1.00
Europe	0.00	0.02	0.07
Direct Lead Measure	-0.39	0.04	1.00
Publication Year	0.00	0.01	0.08
Covariates	-0.07	0.01	1.00
Sample size	0.00	0.00	0.09
FAT	3.40	NA	1.00
Observations	542		

We evaluate the posterior means at the sample averages for each variable (excluding the FAT as normal). This gives a point estimate PCC of 0.09.

We do the same for the elasticity sample in table F.3.

Table F.3 – Posterior results from Bayesian model averaging, elasticity

Variable	**Posterior Mean**	**Posterior Standard Deviation**	**Posterior Inclusion Probability**
Precision	0.24	0.07	1.00
Control gender	-0.14	0.09	0.83
Control race	0.00	0.00	0.05
Control income	0.00	0.00	0.05
Control education	-0.01	0.09	0.24
Homicide	0.00	0.01	0.05
Violent	0.06	0.01	1.00
Non_Violent	0.00	0.00	0.05
Area	0.00	0.03	0.07
OLS	0.00	0.02	0.10
ML	0.00	0.02	0.08
Panel dummy	-0.22	0.09	0.98
Addressing Endogeneity	0.00	0.01	0.07
North_America	-0.01	0.04	0.18
Direct Lead Measure	-0.01	0.08	0.08
Publication Year	0.06	0.04	0.77
Covariates	0.00	0.01	0.06
Sample size	0.03	0.01	0.94
FAT	1.66	NA	NA
Observations	312		

Again, we evaluate the posterior means at the sample averages for each variable (excluding the FAT as normal). This gives a point estimate elasticity of 0.07.

G. Alternative elasticity estimates

Our full sample includes studies that we could not obtain elasticity estimates from. However, it is a larger and possibly more representative sample of the literature. In this section we therefore convert the PCC estimates from the full sample into plausible elasticities. The PCC and the elasticity are related, but not in a straightforward manner. This forces us to make some strong assumptions in the interests of welfare analysis.

Given a PCC and the change in a given measure of crime for a given measure of lead, $\frac{\delta Crime}{\delta Lead}$, then the relationship between the two is given in (7).

$$(8) \qquad PCC = \frac{\delta Crime}{\delta Lead}\frac{sd(Lead)}{sd(Crime)}\frac{sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)}{sd(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2)}$$

Where $sd(.)$ means the standard deviation. $\widetilde{Lead} - \tilde{\boldsymbol{z}}\gamma_1$ are the residuals from a regression of *Lead* on $\boldsymbol{z}$, a vector of variables related to lead and crime, where both lead and $\boldsymbol{z}$ have been standardised. Similarly, $\widetilde{Crime} - \tilde{\boldsymbol{z}}\gamma_2$ are the residuals from a regression of *Crime* on $\boldsymbol{z}$**,** where both have been standardised. If we wish to attach a causal interpretation to the elasticity, we can think of $\boldsymbol{z}$, following Peters, Bühlmann, and Meinshausen (2016), as the minimum set of variables under which the distribution of *Crime* is invariant when conditioned on both $\boldsymbol{z}$ and *Lead*.

It can be seen that a PCC will always share the same sign as $\frac{\delta Crime}{\delta Lead}$ but will be inflated or deflated according to the relative size of the standard deviations in (7). $\frac{\delta Crime}{\delta Lead}\frac{sd(Lead)}{sd(Crime)}$ is equivalent to a standardised coefficient. The intuition for the last ratio is as follows: the greater the variation in *Lead* that is not explained by $\boldsymbol{z}$, the larger the PCC, because the overlapping variation between

the independent effect of *Lead* and *Crime* is relatively greater. The PCC is also greater the larger the amount of variation in *Crime* explained by $\boldsymbol{z}$**.** This is because the share of unexplained variation in *Crime* becomes smaller, so the share of variation jointly explained by *Lead* and $\boldsymbol{z}$ increases. As more of the variation in *Crime* is explained by both *Lead* and $\boldsymbol{z}$**,** their PCCs will tend to 1 or -1.

To evaluate an elasticity at the sample means we multiply both sides by $\frac{\overline{Lead}}{\overline{Crime}}$, where the bar indicates the mean. We can then rearrange (7) to put it in terms of the elasticity η.

$$(9) \qquad \eta = \frac{\overline{Lead}}{\overline{Crime}} \frac{sd(Crime)}{sd(Lead)} \frac{sd(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2)}{sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)} PCC$$

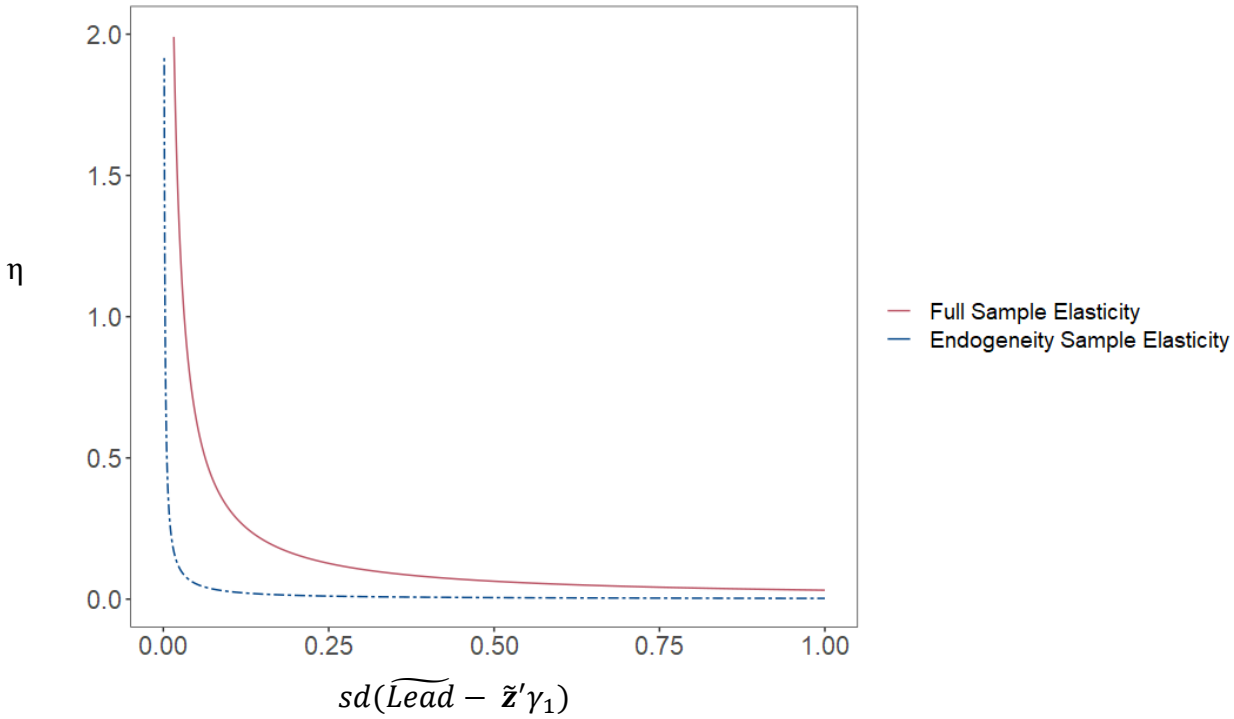
We can see that the size of the PCC relative to the elasticity depends on three ratios. The first two, the relative means and standard deviations, depend on the measures of crime and lead. We use homicide and blood lead data from the US as an illustrative example to examine plausible elasticities, given the fall in both violent and non-violent crime was particularly pronounced there. The means, standard deviations, and sources are given in table IX. Given these, the relative size of the PCC to the elasticity depends upon the third ratio of residual standard deviations. This ratio could theoretically take any value between zero and infinity, and therefore so could the elasticity (assuming the PCC is positive). We therefore look at what are plausible values for this ratio and what is the range of the elasticity given these values.

The maximum value the numerator $sd\left(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2\right)$ can take is one, representing no common variation between $\boldsymbol{z}$ and *Crime*. We hold it at one, to inflate the PCC as much as possible. The final element of the equation is $sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)$. This is the residual variation in *Lead* not explained by $\boldsymbol{z}$**.** The lower this is, the more the PCC will be inflated, and therefore the greater the elasticity. The elasticity is convex in $sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)$, decreasing at a decreasing rate.

Figure F.1 plots the relationship between the elasticity and $sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)$, given the estimated mean PCCs, the values in table IX, and holding $sd\left(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2\right)$ constant at the maximum value of one. The elasticities drop sharply with an increase in the denominator $sd\left(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2\right)$, with the elasticity for the addressing endogeneity sample approaching close to zero almost immediately. The elasticity for the full sample slopes down more gently but even so does not suggest a large elasticity except at extremely small values of $sd\left(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2\right)$.

We can now propose a range of plausible values for the elasticity. Given the uncertainties around the ratio of unexplained variations in (9), this is somewhat arbitrary, but we hope, given the discussion above, not unreasonably so. There is no compelling reason to suppose $\boldsymbol{z}$ would explain more of the variation in *Lead* than in *Crime*. Nevertheless, if we take as a lower bound that $sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)$ is ten times as large as $sd\left(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2\right)$, and as a conservative upper bound that they are equal, then we can give a range of values based on our estimated PCCs. For the full sample PCC, this gives an elasticity of 0.32-0.03. For the addressing endogeneity sample PCC, the range is 0.03-0.00, to two decimal places. The median blood lead level in children fell 88% from 1976-2009. The full sample elasticity estimates therefore would suggest the fall in lead has decreased homicide in the US by between 28% and 3%. The equivalent decrease for the addressing endogeneity sample is between 3% and 0%. The US homicide rate fell 54% from its peak in 1989 to 2014. This would mean that lead accounts for between 52% and 6% of the decrease in homicide using the full sample elasticity, and 5%-0% using the addressing endogeneity elasticity. Our generous assumptions of the lower bound on the ratio of residual variation in (8) imply that lead may be the most important factor in the fall in homicide. Our upper bound on that same ratio implies lead accounts for very little of the fall in crime.

Figure G.1 – Estimated Elasticity of on lead on crime



Notes. Chart shows how η, the calculated elasticity of lead on crime, varies with changes in $sd(\widetilde{Lead} - \tilde{\mathbf{z}}'\gamma_1)$, the standard deviation of the residual in a regression of a set of standardised variables $\tilde{\mathbf{z}}$, and the standardised measure of lead $\widetilde{Lead}$.

Table G.1 – Descriptive statistics of data used for elasticity estimation

Variable	**Mean**	**Standard Deviation**
Median blood lead level for children ages 1-5 in US	3.39	4.42
US Homicide rate	6.98	1.81

Sources. NHANES data for blood lead and FBI uniform crime reports for the homicide data.

H. Chapter 1 Appendix References

Polanin and Snilstveit, 2016. Converting between effect sizes: Campbell policy note 3. https://doi.org/10.4073/cmpn.2016.3

Bajzik, J., Havranek, T., Irsova, Z., Schwarz, J., 2019. Estimating the Armington Elasticity: The Importance of Data Choice and Publication Bias.

Borenstein, M., Hedges, L. V., Higgins, J.P.T., Rothstein, H.R., 2009. Introduction to Meta-Analysis. John Wiley & Sons, Ltd,

Chichester, UK. https://doi.org/10.1002/9780470743386

DerSimonian, R., Laird, N., 1986. Meta-analysis in clinical trials. Control. Clin. Trials. https://doi.org/10.1016/0197-2456(86)90046-2

Higgins JPT, G.S. (editors), 2011. Cochrane Handbook for Systematic Reviews of Interventions, Version 5. ed. The Cochrane Collaboration.

Peters, J., Bühlmann, P., Meinshausen, N., 2016. Causal inference by using invariant prediction: identification and confidence intervals. J. R. Stat. Soc. Ser. B (Statistical Methodol. 78, 947–1012. https://doi.org/10.1111/rssb.12167