Online Appendices

The Lead-Crime Hypothesis: A Meta-Analysis

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):

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

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):

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

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

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

Where OR_{ij} is the odds ratio *i* for study *j* and $a_{ij} = \frac{(n_{ij1}+n_{ij2})^2}{n_{ij1}n_{ij2}}$. 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).

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

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.

(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).

(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:

(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 fixed or random effects meta-averages, 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 meta-averages 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) $Q = \sum_{i=1}^{N} W_{ij} Z_{ij}^2 - \frac{\left(\sum_{i=1}^{N} W_{ij} Z_{ij}\right)^2}{\sum_{i=1}^{N} W_{ij}}$

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 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:

$$\text{(XVII)}\quad C = \sum_{i=1}^{N} W_{ij} - \frac{\left(\sum_{i=1}^{N} W_{ij}\right)^2}{\sum_{i=1}^{N} W_{ij}}$$

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:

$$\text{(XVIII)}\qquad I^2 = \left(\frac{Q-df}{Q}\right) \times 100\%$$

Finally, H^2 is:

$$\text{(XIX)}\quad H^2 = \frac{Q}{df}$$

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

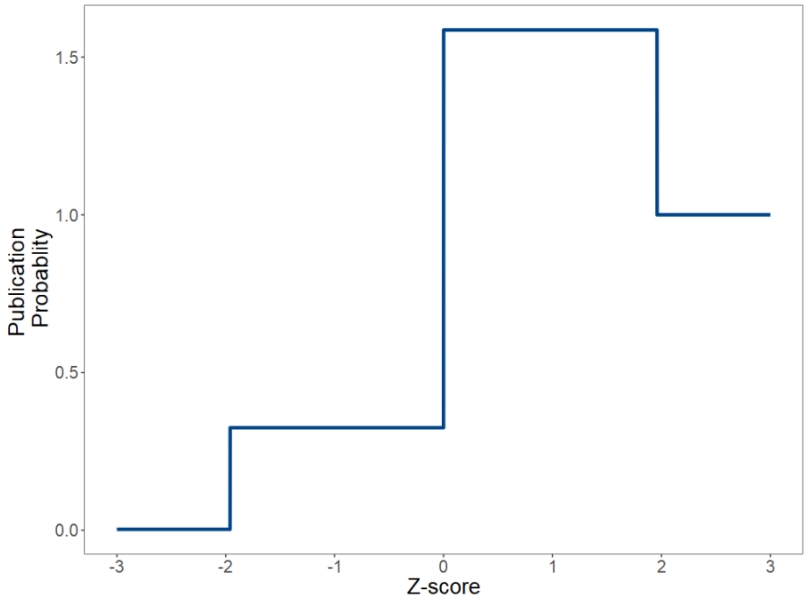
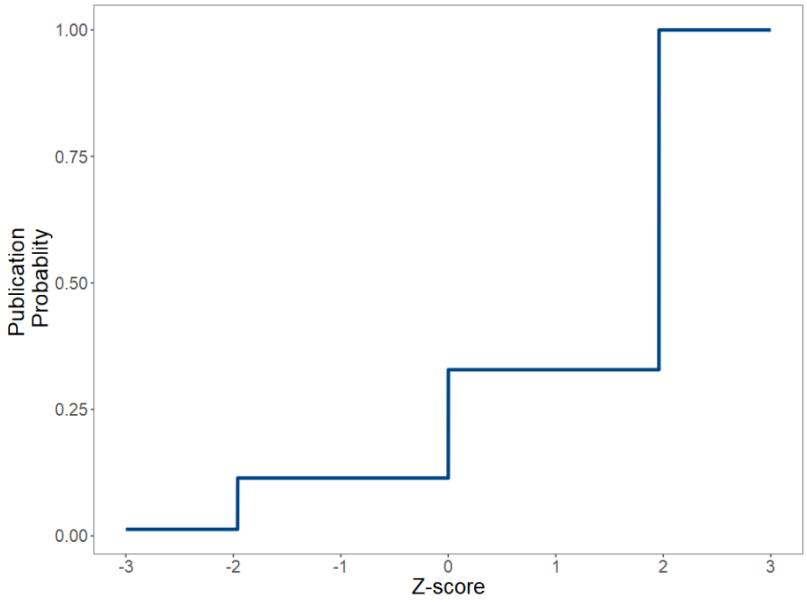


Figure D.2 – Estimated relative publication probabilities, elasticities



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 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		-0.09	0.09	0.03
Observations	542	220	24	312	211

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.

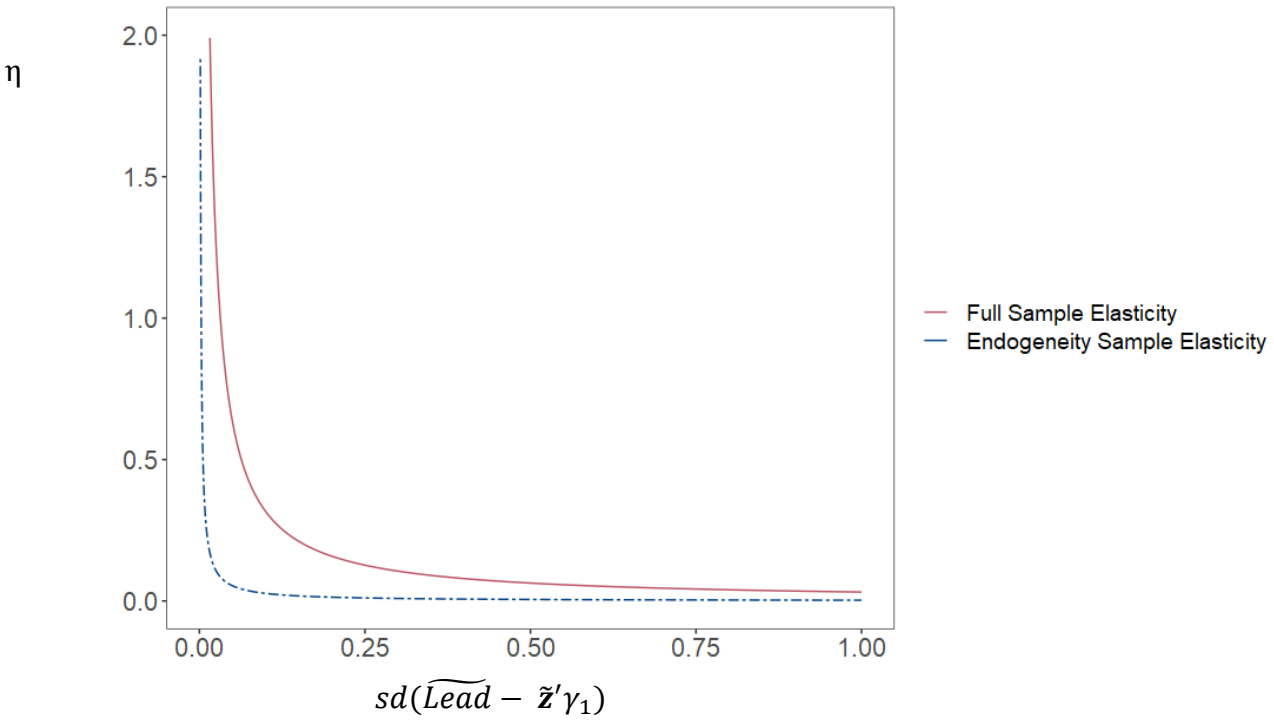
The maximum value the numerator $sd\left(\widetilde{Crime} - \tilde{\boldsymbol{z}}'\gamma_2\right)$ can take is one, representing no common variation between $\boldsymbol{z}$ and *Crime*. We hold it at one, to inflate the PCC as much as possible. The final element of the equation is $sd(\widetilde{Lead} - \tilde{\boldsymbol{z}}'\gamma_1)$. 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. Online 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