Instrument Validity in Air Pollution Impact Studies*

Zhenxuan Wang † Chu A.(Alex) Yu ‡

March 16, 2026

Abstract

Instrumental variable (IV) methods dominate empirical studies of air pollution's impacts, yet their validity often rests on untested exclusion restrictions. Using comprehensive Chinese pollution and weather data, we study three recurring threats that can arise in common IV designs. First, simulations demonstrate co-pollutant bias, where instruments influence multiple pollutants. Second, correlations between pollution proxies and meteorological instruments can generate non-classical measurement error. Third, changes in local infrastructure can affect commonly used meteorological instruments and potentially confound exogeneity. We illustrate the empirical relevance of these threats in an application relating air pollution to daily crime in Chinese cities and provide practical guidance for diagnosing and addressing these risks in environmental IV designs.

Keywords: Air pollution; instrumental variables; exclusion restrictions; thermal inversions; measurement error; co-pollutant bias
JEL Codes: Q5, C3

*The authors declare no conflicts of interest and no sources of external research funding.
†Department of Agricultural and Resource Economics, North Carolina State University. Email: zhenxuan.wang@ncsu.edu.
‡Department of Economics, Wake Forest University. Email: yua@wfu.edu.

1 Introduction

Instrumental variable (IV) methods are now common in studies that estimate the effects of air pollution on health, productivity, and other socioeconomic outcomes. The shift toward IV strategies reflects growing recognition that pollution exposure is endogenous — shaped by behavioral sorting, economic activity, and monitoring placement — and that experimental or policy-based variation is often unavailable. Empirical work has relied on several sources of quasi-experimental variation, including regulatory or event shocks (e.g., Chay and Greenstone, 2003; Hanna and Oliva, 2015; Isen, Rossin-Slater and Walker, 2017), transportation flows (e.g., Moretti and Neidell, 2011; Schlenker and Walker, 2016), and meteorological conditions such as thermal inversions and wind patterns (e.g., Arceo, Hanna and Oliva, 2016; Barwick et al., 2024), which now dominate the literature.

A valid instrument must satisfy two conditions: instrument relevance and the exclusion restriction. While weak instruments can be addressed using robust inference procedures (see Stock, Wright and Yogo (2002); Andrews, Stock and Sun (2019)), there is no standard empirical procedure to detect violations of the exclusion restriction. Understanding how such violations arise in practice is therefore critical for sound causal inference in air pollution impact studies. Recent methodological work in fields such as finance and political science has begun to assess the credibility of IV designs in applied settings (e.g., Atanasov and Black, 2021; Lal et al., 2024). Yet despite the widespread use of IV methods in the air pollution literature, there has been no systematic evaluation of their validity. This paper aims to fill this gap by providing both an empirical assessment of these risks and offering practical guidance for applied work.

To characterize how IV strategies are chosen in practice, we review 65 published air pollution impact studies and use a classification tree to document systematic patterns in instrument selection. Instrument choice is closely related to observable study design features such as the pollution data source, temporal frequency, spatial scale, and study location. For

example, studies using satellite-based pollution measures rely more frequently on thermal inversion instruments, while studies using ground monitors employ a broader range of instruments. Among daily studies, wind-based instruments are more common overall, especially in the US, whereas thermal inversion instruments appear more frequently in studies of China.[1]

In this paper, we identify three recurring challenges for instrument validity in applied settings: 1). co-pollutant bias, when instruments simultaneously shift multiple pollutants; 2). non-classical measurement error, when pollution proxies, especially satellite-based measures, share atmospheric noise with meteorological instruments; and 3). confounding events, such as infrastructure development that affects local thermal dynamics and compromises the exogeneity of thermal inversion instruments.[2] China provides a natural setting for studying these issues, as it is the most common context in the recent air pollution IV literature and one in which thermal inversion instruments are widely used.

Using a unified Chinese pollution monitoring dataset, we evaluate each issue within a consistent empirical framework, focusing on day-level analysis with $PM_{2.5}$ as the primary pollutant of interest.[3] First, simulation exercises based on this dataset show how omitted co-pollutants distort IV estimates. The bias arises whether the focal pollutant is interpreted as the *stimulus* (causal exposure) or a *proxy* (for overall air quality). We find that when co-pollutant correlations are strong—such as between $PM_{2.5}$, NO_2, and SO_2—single-pollutant IV designs can produce substantial upward bias. More robust designs should jointly instrument all major pollutants or limit controls to weakly correlated species such as CO and ozone.

Second, comparison of ground-based and satellite pollution proxies reveals that measurement error in satellite data are strongly correlated with meteorological instruments,

[1] These characterizations are based on our systematic review of 65 air pollution impact studies, described in Section 2.1; among these papers, 28 use data from China and 20 use data from the United States.

[2] These examples do not exhaust all possible violations but illustrate how exclusion restrictions can fail in ways that are systematic rather than idiosyncratic.

[3] $PM_{2.5}$ serves as the focal pollutant in roughly 50% of the sampled studies, and daily-level data account for about half of all empirical designs; see Section 2.1 for details.

reflecting shared atmospheric noise that may compromise the exclusion restriction. This finding underscores the need to validate satellite-based proxies against ground observations and to test directly for correlations between instruments and measurement error, since satellite measures and meteorological IVs may capture the same atmospheric noise rather than independent signals.

Finally, a difference-in-differences analysis shows that wind-farm entry reduces the frequency and intensity of local thermal inversions, implying that infrastructure development can influence commonly used meteorological instruments. This channel may create potential confounding when infrastructure shocks alter local atmospheric dynamics, though its empirical importance appears to be context-dependent. We note that this concern should not be interpreted as a general critique of thermal inversion designs; rather, it highlights that their validity is context-dependent and turns on local conditions and potential confounding channels.

In Section 4, we illustrate these instrument-validity concerns using a standard IV application relating air pollution to daily crime in Chinese cities. Holding the outcome, instrument, and baseline controls fixed, we sequentially address co-pollutant bias, non-classical measurement error in pollution proxies, and potential confounding from wind farm entry. This structured comparison shows that different threats to the exclusion restriction are not equally consequential in practice: accounting for co-pollutant bias can fundamentally alter the interpretation of the estimated effect, and pollution proxy choice can substantially change magnitudes, whereas infrastructure-related confounding plays a more limited role in this particular setting.

We conclude by offering practical guidance for researchers using IV designs in air pollution studies. Our analysis highlights two sequential steps in applied work. First, researchers must identify instruments that are feasible and empirically relevant given their data source, temporal frequency, and geographic context. Using our review of 65 studies, we construct a data-driven classification tree that summarizes systematic relationships

between study design features and instrument choice, providing a transparent benchmark for first-stage feasibility.

Second, conditional on instrument relevance, researchers must diagnose and address threats to the exclusion restriction—especially co-pollutant bias and correlations between instruments and pollution measurement error—that can persist even when first-stage relationships are strong. To address co-pollutant bias, the preferred approach is to jointly instrument major correlated pollutants whenever feasible. When this is not possible because of under-identification, researchers should be explicit about how the estimated effect of the main pollutant should be interpreted — whether as a dose-response effect of that pollutant or as a total effect reflecting correlated co-pollutants. We also highlight potential risks for certain combinations of instruments and pollution measures that may induce non-classical measurement error, particularly when thermal inversion instruments are paired with satellite-based pollution measures that may share common atmospheric noise. In such settings, researchers may consider alternative sources of variation, such as wind-based measures or policy shocks. Section 5 provides a more detailed discussion of these considerations and their implications for applied practice.

Rather than discouraging the use of instrumental variables, our goal is to refine an influential strand of the air pollution literature. Identifying credible instruments in this setting is inherently difficult and requires both institutional knowledge and careful empirical judgment. The IV strategies widely used in this literature represent important contributions that have enabled a large body of causal research. Our analysis clarifies the conditions under which these designs are most informative, highlights settings where specific validity threats are more likely to arise, and provides practical tools for diagnosing and addressing these risks. By doing so, we aim to make commonly used IV strategies more transparent, more interpretable, and ultimately more credible for future research.

2 Background and Data

Methodological reassessments of IV practice have emerged in several applied fields. In labor and macroeconomics, studies highlight pitfalls like weak or invalid instruments in growth regressions and schooling effects, emphasizing robust diagnostics and sensitivity analyses for reliable causal inference (Altonji, Elder and Taber, 2005; Bazzi and Clemens, 2013). In political science, recent work has taken a more systematic approach by replicating published IV studies and tracing common problems in first-stage strength, inference, and exclusion restrictions (Lal et al., 2024; Mellon, 2024). In management and finance, Atanasov and Black (2021) show that many shock-based IV designs become much less persuasive once pretreatment balance and common support are enforced. Our paper extends this line of work by providing a broader and more systematic assessment of IV validity in air pollution impact studies.

2.1 Instrumental Variables in the Air Pollution Literature

We conduct a systematic search to identify air pollution impact studies published in leading general-interest and environmental economics journals as of Spring 2025. Our search yields a sample of 65 papers, with the full list reported in Appendix Table S1. Figure S1 plots the distribution of these studies by publication year, distinguishing between those using only OLS and those employing 2SLS. The air pollution literature has increasingly adopted instrumental variable methods, with 75.4% of sampled studies using some form of exogenous variation to address endogeneity concerns.

Following Benmarhnia, Bharadwaj and Romero (2018), we categorize instruments into three groups: (1) IVs based on regulatory or event shocks, (2) IVs based on variation in port and airport traffic flows, and (3) IVs based on meteorological conditions. The third category now dominates the field, accounting for 77.6% of all IV applications. Within this group, wind-based instruments are the most common (44.9% of IV studies), followed

by thermal inversions (36.7%)[4]. Given this landscape, our analysis focuses on the two dominant meteorological instruments—thermal inversions and wind patterns—although some of the issues discussed in this paper apply more broadly to other types of IVs as well.

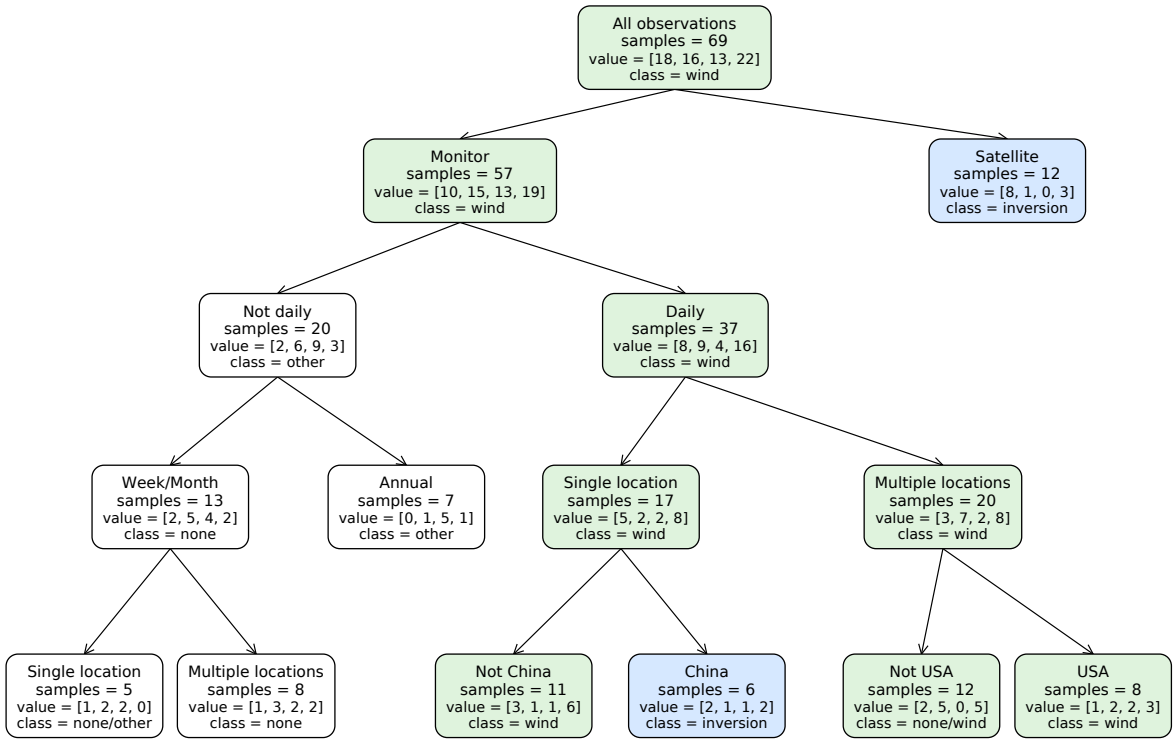


Figure 1: Instrument Decision Trees

Notes: This figure presents a classification tree describing how observable study design features—pollution data source (satellite vs. monitor), temporal frequency, spatial scale, and country—are associated with instrument choice in the sample of 65 studies. Four studies use both wind and thermal inversion instruments, so we count those papers twice in the tree. At each node, the four-number vector reports the distribution of instrument types in the order: thermal inversion, none, other (policy, traffic, event, or other meteorological instruments), and wind. The predicted class corresponds to the most frequent category within that subgroup.

To examine how observable study design features predict instrument selection, we estimate a classification tree using information extracted from each paper. The tree, shown in Figure 1, is estimated usiing a data-driven algorithm that recursively partitions studies based on pollution data source, temporal frequency, spatial scale, and location to maximize

[4]Four papers in our sample use both wind and thermal inversion instruments and are counted in both categories.

predictive accuracy. At each node, the four-number vector reports the distribution of instrument types in the order: thermal inversion, none, other (policy, traffic, event, or other meteorological instruments), and wind. The predicted class corresponds to the most frequent category within that subgroup. Each leaf (i.e., terminal node) reports the empirical distribution of instrument types among studies that share the same sequence of characteristics. The goal of the tree is not to assess instrument validity or recommend optimal choices, but to provide a compact, data-driven summary of how empirical settings are associated with IV strategies in practice.

The first split, on pollution data source, is the most prominent dimension in the tree. Studies relying primarily on satellite-based pollution measures tend to use thermal inversion instruments more often than wind-based instruments. By contrast, studies using ground monitor data exhibit substantially greater heterogeneity in instrument choice.

Temporal frequency is also informative, largely because it proxies for the type of empirical design being pursued. Studies using monthly or annual variation are more likely to be organized around longer-horizon shocks, such as regulatory changes, policy implementation, or discrete events, where identification relies on institutional timing rather than daily meteorological variation. In such settings, meteorological instruments may play a limited role.

At the daily frequency, the classification tree reveals clear spatial differentiation in instrument choice. Wind-based instruments are the most commonly used choice overall, particularly in U.S. studies. By contrast, among studies conducted in China, thermal inversion instruments are more prevalent, especially in single-location studies. These patterns likely reflect differences in instrument relevance and first-stage strength across empirical contexts (e.g., due to geographic characteristics or variation in pollution sources).

2.2 Data

Air Pollution. We obtain air pollutant concentration data from the Chinese National Urban Air Quality Real-time Publishing Platform for the years 2013-2019 across approximately 1,600 monitoring stations. For each air pollutant, we calculate the daily average by taking the mean of all available hourly readings. The air pollutants we observe include AQI, CO, $PM_{2.5}$, PM_{10}, NO_2, O_3, and SO_2. Because some monitoring stations were added in later years, we have an unbalanced panel consisting of 4.2 million station-day observations. This dataset underlies 18 sampled studies conducted in China, making it the canonical source for empirical research on air pollution in the country.

For the measurement error analysis, we also obtain satellite-based $PM_{2.5}$ data from two sources: the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) and the China High Air Pollutants (CHAP). The MERRA-2 reanalysis data product, used in four of the studies in our sample, provides hourly global estimates of aerosol mass concentrations by species at a spatial resolution of roughly 50×60 km. Following NASA's technical documentation,[5] we compute total surface-level $PM_{2.5}$ by summing modeled concentrations of five aerosol species – black carbon, organic carbon, dust, sea salt, and sulfate – each adjusted for geometric particle size and humidity-dependent growth factors. The CHAP dataset provides substantially finer spatial resolution (1 km) with daily temporal coverage. A key methodological distinction is that CHAP explicitly calibrates its estimates using ground monitoring data through machine learning approaches, whereas MERRA-2, as a physics-based reanalysis product, does not directly incorporate ground monitor calibration in the same manner.

Weather. We obtain daily weather data from the China Meteorological Administration (CMA), which reports observations from more than 2,400 weather stations across China. The dataset includes daily average, minimum, and maximum temperature, total precipita-

[5] See https://gmao.gsfc.nasa.gov/gmao-products/merra-2/faq_merra-2/.

tion, average wind speed, and the direction of the maximum wind speed. To align weather conditions with air quality observations, we interpolate station-level weather data to the locations of air pollution monitoring stations using the inverse distance weighting (IDW) method, restricting interpolation to weather stations within a 200 km radius.

Thermal Inversions. We measure thermal inversions using the M2I6NPANA product from MERRA-2. The data report air temperature at a spatial resolution of $0.5^\circ \times 0.625^\circ$ (approximately 50 km $\times$ 60 km) across 42 atmospheric layers at six-hour intervals. We extract temperature records for the three layers closest to the surface, corresponding to heights of approximately 110 m, 320 m, and 540 m above ground level. A thermal inversion between two layers occurs if the temperature in the higher layer is greater than that in the lower layer during a given six-hour period. We measure thermal inversion strength as the temperature difference between the two layers; when no inversion occurs, the strength is coded as zero. For our daily analysis, we aggregate the six-hour records by averaging inversion strength or by counting the number of inversion occurrences within a day.

Wind Farms. To examine how local wind energy projects influence thermal inversions, we use data on wind farms from the Global Wind Power Tracker (Global Energy Monitor, 2025). This dataset provides information on the geographic coordinates and commissioning year of each wind farm worldwide. We extract the subset of wind farms located in China to construct measures of local wind power development.

3 Systematic Challenges to the Exclusion Restriction

3.1 Co-Pollutants Bias

Despite the fundamental importance of multiple pollutant exposure in real-world settings, studies vary in their attention to co-pollutant issues. Only 49% of papers explicitly

recognize the potential bias caused by co-pollutants, while 51% provide unclear or no information about this critical issue. Among studies that do consider co-pollutants, approaches vary in both sophistication and rigor. One strategy involves simultaneously instrumenting other major pollutants—such as ozone, nitrogen dioxide, sulfur dioxide, and carbon monoxide—alongside the primary pollutant of interest. However, this approach may encounter under-identification problems when the number of instruments is fewer than the number of endogenous variables being instrumented. In such cases, both the choice of instruments and their functional forms become critically important.

Other papers include additional air pollutants as control variables in their main specifications, treating them as instruments for the pollutant being instrumented. This approach is problematic when these other air pollutants are themselves endogenous and correlated with the instrument – a situation that is likely in practice, as we discuss later.

A third subset of papers conducts robustness checks by examining whether the estimated coefficients remain stable when alternative air pollutants are used as the primary explanatory variable. The viability of this approach depends critically on how researchers interpret the estimated pollution effects. Consider a simple model where an outcome variable is modeled as a function of $PM_{2.5}$. Two interpretations of the estimated coefficient are possible. Under the *stimulus* interpretation, $PM_{2.5}$ is treated as the causal exposure of interest, as if the analysis measured the marginal effect of a one-unit increase in $PM_{2.5}$ in a controlled experiment. This view motivates statements such as "A one-unit increase in $PM_{2.5}$ leads to...". Under the *proxy* interpretation, $PM_{2.5}$ is viewed as an indicator of overall air quality, recognizing that other pollutants can be correlated with it. In this case, causal claims take the form "Higher levels of air pollution increase...". Under the *proxy* interpretation, robustness checks across different pollutant specifications can be informative if researchers aim to assess whether the estimated effects reflect a broader pollution mixture rather than a specific pollutant, or if dose-response relationships are similar across multiple pollutants. However, the latter assumption may not hold given the

distinct health effects of different pollutants.

In either interpretation, there is a potential violation of the exclusion restriction when co-pollutant issues are not handled carefully. The key intuition is that the chosen instruments are likely also affecting the omitted pollutants, leading to a similar omitted variable bias issue. Although there appears to be no perfect solution to this problem, we conduct simulation exercises and applications showing the trade-offs among several options and how co-pollutants should be instrumented.

3.1.1 Simulation Design

We design a simulation-based framework to examine how the presence of correlated pollutants affects the performance of IV estimation. The simulation uses real pollution and meteorological data from Chinese monitoring stations to preserve the empirical correlation structure between pollutants and instruments, while allowing us to vary the co-pollutant's impact and endogeneity in a controlled way. Pollution data include two pollutants, X_1 and X_2 (e.g., $PM_{2.5}$ and NO_2), each standardized to have mean zero and unit variance. Thermal inversion variables serve as instruments, though the co-pollutant problem generalizes to any IV design in which instruments affect multiple correlated pollutants. Because the simulation draws on real data, the first-stage relationships between instruments and pollution are entirely data-driven and capture observed empirical correlations rather than being imposed by assumption.

For monitor station i in city c on date t, the data-generating process (DGP) links pollution exposure to a synthetic outcome variable which is given by:

$$Y_{ict} = \beta_1 X_{1,ict} + \beta_2 X_{2,ict} + \rho \, \text{Endo}_{ct} + \varepsilon_{ict} \tag{1}$$

where $\varepsilon_{ict} \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 1$. The true causal parameter β_1 is set to one, while β_2 varies over the grid $\{0.1, 0.2, \ldots, 1.2\}$ to capture different degrees of co-pollutant influence.

Endogeneity enters through $\rho\,\text{Endo}_{ct}$, where $\rho \in [0,1]$ controls its magnitude. Endo_{ct} represents residual pollution variation based on the first-stage residual $X_{1,ict} - \hat{X}_{1,ict}$. Each simulation therefore corresponds to one of 132 combinations of β_2 and ρ values.

Five regression specifications are estimated in each simulation: (1) OLS–Single, which regresses Y on X_1 only; (2) IV–Single, which instruments X_1 using Z; (3) IV–with–Control, which instruments X_1 using Z and includes X_2 as a control; (4) OLS–Both, which includes both X_1 and X_2 without instruments; and (5) IV–Both, which instruments both X_1 and X_2 using Z. For each specification, the coefficient on X_1 is compared to the true value $\beta_1 = 1$, and the mean bias is computed across the 132 parameter combinations.

3.1.2 Simulation Results

Table 1 presents results from the simulation framework used to assess bias under different IV specifications. Panel A summarizes the simulation design. Panel B reports results using thermal inversion as the instrument, separately for models without and with city and year fixed effects.

We begin by considering the *stimulus* interpretation of pollution exposure, which considers the estimand as the dose-response effect of the primary pollutant. The estimated bias under the different model specifications is reported in Table 1, Panel B2. Several key insights emerge. First, the simulation confirms that co-pollutant bias remains substantial when the primary pollutant $PM_{2.5}$ is highly correlated with omitted pollutants such as NO_2 and SO_2. Using thermal inversion instruments, the IV-Single specification overestimates β_1 by 0.62 when NO_2 is the omitted co-pollutant and by 0.50 when SO_2 is omitted. In contrast, bias is close to zero for more weakly correlated pollutants such as CO and O_3. This pattern is aligned with the empirical correlations between $PM_{2.5}$ and each co-pollutant—0.59 (NO_2), 0.46 (SO_2), 0.10 (CO), and -0.14 (O_3). The inclusion of city and year fixed effects provides little improvement, suggesting that spatial and temporal controls do not mitigate omitted-variable bias arising from co-pollutant contamination.

Second, the ideal solution to address air pollution endogeneity is to instrument all major pollutants when sufficient instruments are available for identification. Under the IV-Both specification, the estimated bias is effectively zero across all pollutants, indicating that full instrumentation successfully eliminates both endogeneity and co-pollutant bias, when the underlying data-generating process is known.

Third, for pollutants that are only weakly correlated with $PM_{2.5}$, such as CO and O_3, instrumenting $PM_{2.5}$ while including these pollutants as control variables can provide a practical alternative. However, caution is warranted for highly correlated pollutants such as NO_2 and SO_2, where the IV-w/Control specification reverses the direction of bias, reflecting overcorrection when the control variable is itself endogenous.

Finally, the overall pattern is not specific to thermal inversion instruments or to the fixed effects we included. Nearly identical results emerge when using wind direction $\times$ pollution zone instruments (Appendix Table S2), following a similar approach to Deryugina et al. (2019), or when controlling for station and date fixed effects (Appendix Table S3). This suggests that the mechanisms driving co-pollutant bias are structural rather than instrument-specific. Across both meteorological instruments and model specifications, we observe a consistent hierarchy in bias magnitudes—largest for NO_2, followed by SO_2, and minimal for CO and O_3. This stability indicates that the strength of pollutant correlations, rather than the estimation strategy or instrument type, is the primary determinant of residual bias in estimated pollution effects.

3.1.3 Interpreting Pollution Effects: Stimulus versus Proxy

The simulation framework accommodates both the *stimulus* and *proxy* interpretations of pollution exposure. In both cases, we estimate the same coefficient β_1 from the same regression specification – one of the five estimators described above. What differs is the benchmark used to assess bias. Under the *stimulus* interpretation, X_1 represents the pollutant of direct interest, so the relevant ground truth is its causal effect, $\beta_1 = 1$. Under

Table 1: Simulation Framework for Evaluating Bias in IV Estimation

Panel A. Simulation Design

Variables: X_1: $PM_{2.5}$ (primary pollutant); X_2: NO_2, SO_2, CO, O_3 (co-pollutants).
All standardized (mean = 0, SD = 1).
True DGP: $Y = \beta_1 X_1 + \beta_2 X_2 + \rho(X_1 - \hat{X}_1) + \varepsilon, \varepsilon \sim \mathcal{N}(0,1)$.
Parameters: $\beta_1 = 1$; $\beta_2 \in [0.1, 1.2]$; $\rho \in [0,1]$ (increment = 0.1).

Panel B. Simulation Results using Thermal Inversion as IV

	No FE				With City/Year FE			
X_2	NO_2	SO_2	CO	O_3	NO_2	SO_2	CO	O_3
B1. Correlation structure								
Corr(X_1, X_2)	0.59	0.46	0.10	-0.14	0.59	0.46	0.10	-0.14
Corr(X_2, Endo)	0.53	0.42	0.09	-0.15	0.05	0.04	0.01	0.01
B2. Bias under stimulus interpretation $(\hat{\beta}_1 - \beta_1)$								
OLS–Single	0.84	0.76	0.52	0.37	0.81	0.73	0.52	0.37
IV–Single	0.62	0.50	0.08	-0.01	0.62	0.44	0.07	-0.04
OLS–Both	0.47	0.47	0.46	0.46	0.48	0.47	0.47	0.47
IV–w/Control	-0.50	-0.24	0.00	0.00	-0.48	-0.19	0.00	0.00
IV–Both	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
B3. Bias under proxy interpretation $(\hat{\beta}_1 - \beta_1^{proxy})$								
OLS–Single	0.46	0.46	0.46	0.46	0.47	0.47	0.47	0.47
IV–Single	0.24	0.20	0.01	0.08	0.28	0.18	0.02	0.06
B4. Key factors governing the bias								
$\pi_{x_2 \mid \hat{x}_1}$	0.96	0.78	0.12	-0.02	0.96	0.68	0.11	-0.06
B5. Instrument strength								
KP Fstat (IV-Single)	204.39	204.39	204.39	204.39	313.56	313.56	313.56	313.56
KP Fstat (IV-Both)	52.17	14.85	2.21	40.22	54.11	13.77	2.17	31.76
Observations	3,167,316	3,167,107	3,167,433	3,167,362	3,167,316	3,167,107	3,167,433	3,167,362

Notes: Bias values are calculated as the average across 132 simulation settings, corresponding to all combinations of $\beta_2 \in [0.1, 1.2]$ and $\rho \in [0,1]$. Positive values indicate overestimation relative to the true coefficient value of 1.

the *proxy* interpretation, X_1 serves as an indicator of a broader pollution bundle, reflecting its systematic correlation with other pollutants. The appropriate benchmark is then the X_1-indexed total effect:

$$X_1\text{-indexed total effect} = \beta_1 + \beta_2 \alpha_1 \tag{2}$$

where $X_2 = \alpha_0 + \alpha_1 X_1$ characterizes the dependence of X_2 on X_1. This object captures the effect of a one-unit increase in X_1 together with the predictable change in correlated

co-pollutants.[6] Thus, while the estimation procedure remains identical, the interpretation determines which "ground truth" the estimated coefficient should be compared against - direct stimulus effect versus X_1-indexed total effect. As Panel B3 of Table 1 shows, although the magnitude of the bias shrinks under the proxy interpretation, it does not disappear.

3.1.4 Quantifying Co-Pollutant Bias

Starting from the true model $Y = \beta_1 X_1 + \beta_2 X_2 + \rho \text{Endo} + \varepsilon$ and the estimated model $Y = \beta_1 X_1 + \eta$, we can decompose the estimation bias. OLS bias arises from both endogeneity and omitted co-pollutants, while IV removes the former but not the latter. Specifically,

$$\hat{\beta}_1^{IV} - \beta_1 = \beta_2 \times \pi_{X_2|\hat{X}_1} \tag{3}$$

where $\pi_{X_2|\hat{X}_1} = \frac{\text{Cov}(\hat{X}_1, X_2)}{\text{Var}(\hat{X}_1)}$ denotes the projection coefficient of the omitted co-pollutant on the instrument-induced variation in X_1. This shows that even a valid instrument may yield biased estimates when it also shifts correlated pollutants. Unlike OLS bias, which cannot be directly quantified because the degree of endogeneity is unknown, IV bias can be empirically evaluated. As shown in equation (3), the magnitude of this bias depends only on the co-pollutant effect β_2 and the projection coefficient $\pi_{X_2|\hat{X}_1}$.

To determine a plausible range for β_2, we draw on empirical and meta-analytic evidence from the environmental health literature. Studies show that both $PM_{2.5}$ and NO_2 are associated with acute health responses of comparable magnitude. WHO-supported syntheses report roughly a 0.6–0.7% increase in all-cause mortality per 10 $\mu g/m^3$ increase in $PM_{2.5}$ and a 0.4–0.5% increase per 10 $\mu g/m^3$ increase in NO_2 exposure (World Health Organization, 2021), implying a relative effect of about $\beta_2 \approx 0.8\beta_1$. Controlled human exposure (chamber) studies further demonstrate that NO_2 alone can trigger measurable respiratory effects in the absence of particulate matter (Ezratty et al., 2014), supporting this

[6] Importantly, this should not be interpreted as the effect of "overall air pollution" unless all relevant pollutants move proportionally with X_1. Pollutants that are weakly correlated or orthogonal to X_1 are not incorporated in this total effect.

range. Taken together, these findings justify using $\beta_2 \in [0.8, 1]$ as a meaningful range for the relative short-term health effect of NO_2 compared with $PM_{2.5}$.

The projection coefficient $\pi_{X_2|\hat{X}_1}$ is determined once an instrument is selected and can be directly estimated from the data, as reported in Panel B4 of Table 1. It captures how strongly the instrument-induced variation in X_1 is correlated with the omitted co-pollutant X_2, providing an empirical measure of potential IV bias. Empirically, this coefficient can be obtained by estimating a 2SLS regression of the omitted co-pollutant X_2 on X_1, instrumenting X_1 with the same instrument(s) used in the main specification. That is,

$$X_{2,it} = \pi X_{1,it} + \text{controls}_{it} + e_{it}$$

where $X_{1,it}$ is instrumented using Z_{it}. Standard errors can be computed using the same clustering structure as in the baseline IV specification.

3.2 Non-Classical Measurement Error

Another challenge in estimating the causal effect of air pollution is measurement error in the interpolated air pollution proxies. Instrumental variables provide a solution for measurement errors when these errors are classical in nature.[7] Under classical measurement error, we would expect no difference in point estimates across different proxies of air pollutants after addressing endogeneity through IV estimation — only differences in standard errors should emerge.

Consider a simple model $y = \beta x + \varepsilon$, where x is true pollution exposure and y is the outcome. In practice, researchers often observe a proxy $\tilde{x} = x + u$, introducing measurement error u. The IV estimator can then be expressed as:

$$\hat{\beta}_{IV} = \frac{\text{Cov}(y, z)}{\text{Cov}(\tilde{x}, z)} = \beta \times \frac{\text{Cov}(x, z)}{\text{Cov}(x, z) + \text{Cov}(u, z)}$$

[7] Among sampled IV studies, 43% mention measurement error as a key motivation for using instrumental variables.

assuming $\text{Cov}(\varepsilon, z) = 0$. If the measurement error u is independent of the instrument z, IV estimation remains unbiased. But when u is non-classical and correlated with the instrument, the exclusion restriction can fail. The resulting bias depends on the sign and size of $\text{Cov}(u, z)$.

We empirically examine whether measurement error in air pollution measures is correlated with commonly used meteorological instruments. We take the daily $PM_{2.5}$ reading from air pollution monitors as ground truth, and construct four different types of pollution proxies. First, we use spatial interpolation with a leave-one-out method. This approach is commonly employed in empirical studies when the location of the response/outcome variable is precisely known (e.g., the location where students take their exams), but pollution concentration at that location is not directly observed. Here, we assume the points of interest are located at the ground monitors and use readings from surrounding monitors within 50km of the target monitor with an inverse distance weighting (IDW) approach.

The second proxy is generated using an aggregation approach. Many administrative datasets or de-identified data are available at the city, county/province, or state level, requiring us to aggregate the pollution data to match the spatial resolution of the outcome variables of interest. We calculate each city's daily pollutant concentration as the arithmetic mean of all monitors located within its administrative boundaries, with monitor-to-city assignments based on official designations from the Ministry. There is an average of 4.37 monitors per city. For context, Chongqing – the largest city in China by land area – has the most monitors, with 21 active monitoring stations.

The third and fourth proxies are obtained from satellite-based $PM_{2.5}$ data. Here, we match the grids from CHAP and MERRA-2 with the ground monitor locations. The readings at these grids serve as the third and fourth proxies, respectively.

We then construct the measurement error as the difference between the ground truth and proxies. To empirically examine if the measurement errors are correlated with

instruments being used, we estimate the following model:

$$u_{it} = \sum_{k=1}^{K} \alpha_k Z_{k,it} + \mu_i + \delta_t + \nu_{it} \tag{4}$$

The outcome variable u_{it} is measurement error, i.e., the difference between the proxy estimate and the ground truth, at monitor i on date t. For the instrumental variable Z, we examine both thermal inversion and wind-related variables. All regressions control for monitor and date fixed effects.

Figure 2 reports the F-statistics from regressions of measurement errors on the instruments. When proxies are measured based on ground monitors, the measurement errors are not strongly correlated with the instruments, except for wind speed, which shows a modest but statistically significant association, likely reflecting local dispersion effects. Moving to Panels C and D, where the proxies are constructed from satellite-based measures, we find that they are highly correlated with the thermal inversion instrument, which is also derived from satellite data. This pattern suggests that satellite-based pollution estimates may share common atmospheric measurement noise with meteorological variables (Van Donkelaar et al., 2021).

3.3 Confounding Shocks to Meteorological Instruments

Infrastructure development can influence local climate conditions, raising concerns that certain meteorological instruments may be systematically correlated with infrastructure placement and expansion. Wind farm installations provide one such example. Wind farms can disrupt local thermal inversion patterns because turbine-induced turbulence mixes air layers, weakening the stable stratification that traps pollutants near the surface (Baidya Roy and Traiteur, 2010). The strategic siting of renewable energy infrastructure – often driven by geographic and climatic factors – can directly affect local pollution patterns, economic activity (Fabra et al., 2024), and health outcomes (Zou, 2020). This linkage further

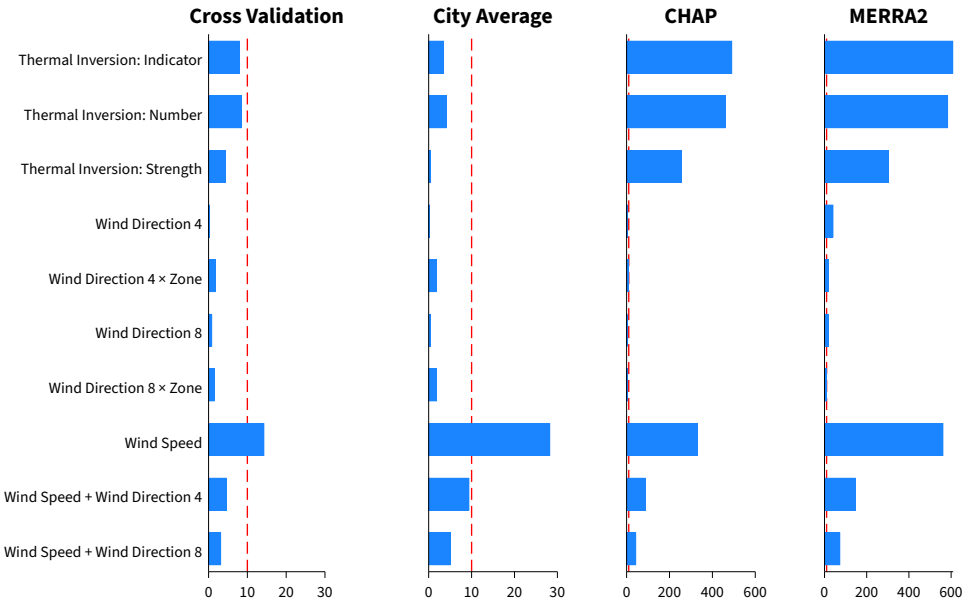


Figure 2: F-stats from regressing measurement error on IVs

Notes: This figure reports first-stage F-statistics from regressions of $PM_{2.5}$ measurement errors on meteorological instruments using equation (4). Measurement error is defined as the difference between ground-monitor $PM_{2.5}$ readings and alternative pollution proxies: spatial interpolation (leave-one-out IDW), city-level averaging, and satellite-based measures (CHAP and MERRA-2).

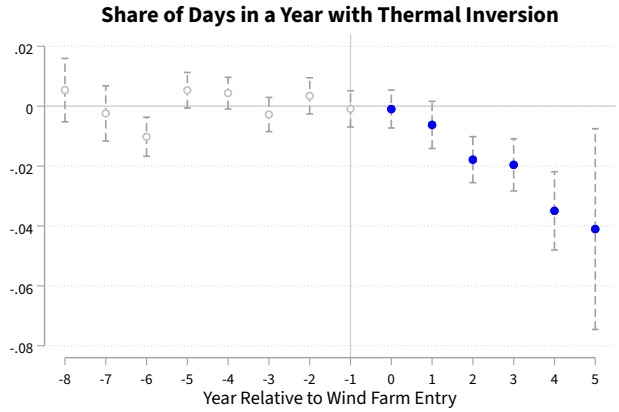
complicates the interpretation of thermal inversion as an exogenous instrument.

We estimate the effect of wind farm entry on local thermal inversion patterns using a difference-in-differences (DiD) framework. Stations located within 50 km of a wind farm are defined as treated, while those without any wind farms within 50 km serve as controls. The analysis compares changes in inversion frequency and intensity before and after the commissioning of nearby wind farms. Specifically, we estimate the following event-study specification:

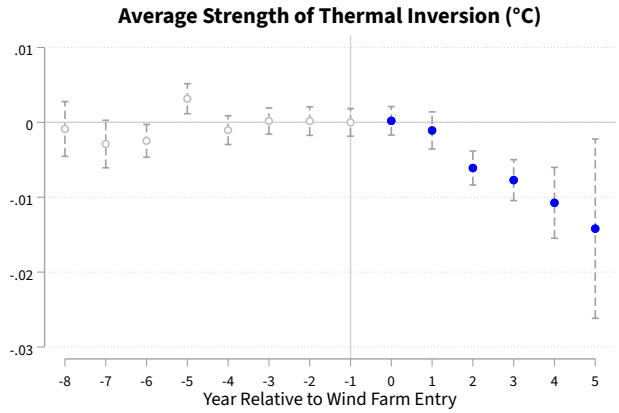
$$TI_{it} = \sum_{\tau=-8}^{5} \beta_\tau 1_{[t-\phi_i=\tau]} + \alpha_i + \delta_t + \varepsilon_{it}, \tag{5}$$

where TI_{it} denotes either the share of days in a year with thermal inversion or the mean inversion strength; ϕ_i is the first year when a wind farm within 50 km of station i was commissioned; and α_i and δ_t are station and year fixed effects. The coefficients β_τ capture differences between treated and control stations in year τ, therefore tracing the dynamic evolution of inversion frequency and intensity surrounding wind farm entry. To address

the issues with staggered treatment timing, we implement the heterogeneity robust DiD estimator of Callaway and Sant'Anna (2021). The event-study estimates presented in Figure 3 show no discernible pre-trend, while both inversion frequency and intensity decline sharply following wind farm entry.



(a) Thermal Inversion Frequency



(b) Thermal Inversion Strength

Figure 3: Wind Farm Entry and Thermal Inversion Patterns

Notes: This figure presents event-study estimates of the effect of wind farm entry on thermal inversions. Stations within 50 km of a wind farm are treated, with more distant stations as controls. The horizontal axis denotes years relative to wind farm commissioning, and estimates compare treated and control locations before and after entry with location and year fixed effects.

4 Instrument Validity in Practice: Evidence from Crime

4.1 Crime Data and Empirical Strategy

Our empirical application uses a daily panel of street and neighborhood crime in China constructed from judicial case records compiled by Zhang, Kwan and Fang (2025). The data are aggregated from incident-level criminal cases to the city–day level and cover more than 200 prefecture-level cities across mainland China during the 2013–2019 period. The sample includes common street and neighborhood offenses, such as theft and robbery, which constitute the majority of urban crimes observed in court verdicts. Because the data are based on adjudicated cases rather than police reports, they reflect judicially verified criminal activity with consistent definitions across locations. The daily frequency and broad geographic coverage of the data align naturally with our pollution and meteorological measures, making this setting well suited for assessing the practical relevance of instrument validity concerns in air pollution impact studies.

We estimate variants of the following two-stage least squares (2SLS) specification, in which $PM_{2.5}$ is instrumented using measures of thermal inversion frequency:[8]

$$PM_{2.5,it} = \alpha \mathbf{Z}_{it} + \mathbf{W}'_{it}\boldsymbol{\phi} + \gamma_i + \kappa_{dow(t)} + \xi_{my(t)} + e_{it} \tag{6}$$

$$\text{Crime}_{it} = \beta_1 \widehat{PM}_{2.5,it} + \mathbf{W}'_{it}\boldsymbol{\eta} + \mu_i + \lambda_{dow(t)} + \delta_{my(t)} + \epsilon_{it} \tag{7}$$

In the above equations, Crime_{it} denotes the number of criminal incidents in city i on day t, $PM_{2.5,it}$ is the daily average $PM_{2.5}$ concentration estimated using ground monitor data,[9] and $\mathbf{W}_{it}$ is a vector of weather controls including daily mean temperature,

[8] Thermal inversion frequency is constructed using multiple inversion indicators based on vertical temperature gradients. Specifically, T21 measures inversion occurrence using the temperature difference between the second layer and the first layer. We also use T32 (third minus second layer) and T31 (third minus first layer), where positive values indicate thermal inversions. The data are reported at six-hour intervals, yielding four observations per day. For each layer pair, we compute daily inversion frequency by counting the number of six-hour periods in which an inversion occurs.

[9] Our interpolation method follows the official guidance issued by China's Ministry of Ecology

total precipitation, and average wind speed, along with their squared terms: $\mathbf{W}_{it} = (tmean_{it}, tmean_{it}^2, prcp_{it}, prcp_{it}^2, wind_{it}, wind_{it}^2)$. City fixed effects μ_i absorb time-invariant differences across cities, day-of-week fixed effects $\lambda_{dow(t)}$ capture systematic intra-week patterns in crime, and month-by-year fixed effects $\delta_{my(t)}$ flexibly control for seasonality and national shocks. Standard errors are clustered at the city level.

Using this empirical setting, we organize the application around a single thermal-inversion 2SLS design and examine how different empirical adjustments affect the resulting estimates. The identifying variation from thermal inversions, the outcome, and the core set of controls are held fixed throughout, and each exercise modifies only one element of the specification corresponding to a specific threat to the exclusion restriction.

To examine co-pollutant bias, we move from the standard single-pollutant model to specifications that treat NO_2 and SO_2 as jointly endogenous with $PM_{2.5}$ and instrument the system using the same inversion-based measures. To assess non-classical measurement error, we keep the instrument set unchanged but replace the pollution regressor with alternative $PM_{2.5}$ proxies including two satellite-based products (CHAP and MERRA-2). To evaluate potential confounding from infrastructure-driven changes in local conditions, we augment the baseline specification with controls for wind farm entry and richer time-varying fixed effects, such as wind-farm–city-by–year or city–by-year fixed effects.[10]

4.2 Results and Implications

We then apply the conceptual threats to instrument validity discussed above — including co-pollutants bias, non-classical measurement error, and confounding shocks due to wind farm entry — to a standard empirical setting in which the outcome of interest is the number of crimes. Across all these exercises, column (1) is intentionally identical: $PM_{2.5}$

and Environment, available at: https://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/jcffbz/201309/W020250407409903720341.pdf. According to this document, Chinese air pollution monitoring stations are positioned such that the arithmetic mean of pollutant concentrations across all city grid points represents the overall average concentration of pollutants in the city's built-up area.

[10] A city is defined as a wind farm city if it has at least one wind farm.

is instrumented using thermal inversion frequency, with the same fixed effects, weather controls, and sample. This shared baseline allows a direct comparison of how different violations of the exclusion restriction affect IV estimates. Appendix Tables S4-S6 report analogous results using the crime rate (crimes per 100,000 population) as the outcome; the findings remain similar.

Table 2 examines co-pollutant bias by comparing single- and multi-pollutant IV specifications. Column (1) reproduces the standard design in which $PM_{2.5}$ alone is instrumented using thermal inversions. Column (2) instead instruments NO_2 alone, while columns (3)–(7) jointly instrument multiple pollutants. Once NO_2 is included as an endogenous regressor, the $PM_{2.5}$ coefficient becomes unstable and statistically insignificant, whereas the estimated effect loads primarily onto NO_2. To provide a consistent comparison across specifications, we compute the pollutant-indexed total effect defined in equation (2) which measures the change in the outcome associated with a one-unit increase in the indexed pollutant allowing correlated pollutants to respond according to their empirical co-movement. The NO_2-indexed total effect remains remarkably stable across specifications, ranging from 0.40 to 0.47 except in Column (5), where SO_2 is also included. In contrast, the $PM_{2.5}$-indexed total effect varies substantially in both magnitude and sign once additional pollutants are jointly instrumented.

This asymmetry indicates that the thermal-inversion instrument primarily identifies variation in a combustion-related pollution bundle for which NO_2 is the dominant component. Thermal inversions trap locally emitted combustion pollutants, and NOx emissions are closely tied to traffic and industrial activity while also contributing to particulate pollution through nitrate formation. As a result, inversion shocks increase both NO_2 and $PM_{2.5}$, but the economically relevant variation appears to operate most strongly through NO_2. In this context, single-pollutant IV designs that instrument $PM_{2.5}$ may in some settings be identifying NO_2-driven effects rather than the marginal effect of particulate mass itself.

Table 3 turns to non-classical measurement error by re-estimating the baseline specifi-

Table 2: Instrument Validity Checks - Co-Pollutant Bias

Dep. Var.:	Number of Crimes						
Instrument for:	$PM_{2.5}$	NO_2	$PM_{2.5}$, NO_2	$PM_{2.5}$, SO_2	NO_2, SO_2	$PM_{2.5}$, NO_2, SO_2	$PM_{2.5}$, NO_2, SO_2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$PM_{2.5}$	0.212***		-0.238	-0.082		-0.243	-0.189
	(0.060)		(0.181)	(0.139)		(0.194)	(0.409)
NO_2		0.403***	0.807***		0.045	0.845	0.763
		(0.102)	(0.290)		(0.495)	(0.674)	(0.711)
SO_2				0.670**	0.447	-0.036	-0.040
				(0.262)	(0.563)	(0.585)	(0.539)
O_3							0.099
							(0.081)
CO							-1.899
							(14.341)
KP Fstat	383.61	641.85	16.01	17.46	18.53	18.00	10.12
Obs	742,216	742,216	742,216	742,216	742,216	742,216	742,216
Total Effect:							
$PM_{2.5}$-indexed	0.212	-	-0.084	0.023	-	-0.088	-0.050
NO_2-indexed	-	0.403	0.457	-	0.299	0.467	0.462

Notes: The dependent variable is the daily number of crimes in a city. All specifications use the same set of daily thermal inversion frequency measures as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. All regressions include city, day-of-week and month-by-year fixed effects and weather controls. The pollutant-indexed total effect is defined as the change in the outcome associated with a one-unit increase in the corresponding pollutant, allowing correlated co-pollutants to respond endogenously. Standard errors clustered at the city level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

cation using alternative $PM_{2.5}$ proxies. Column (1) uses the city-level average of ground monitors – the same $PM_{2.5}$ measure as in Panel A – and therefore reproduces the baseline estimate by construction. In contrast, using satellite-based measures from the CHAP data leads to substantially larger estimates. Relative to ground-based measures, the CHAP proxy inflates the estimated $PM_{2.5}$ effect by approximately 52% when thermal inversion frequency is used (column 2 versus column 1) and by 50% when inversion strength is used (column 5 versus column 4). By contrast, estimates based on MERRA-2 are smaller than the ground-monitor baseline and statistically insignificant. Moreover, the differences between the ground-monitor estimates and the satellite-based estimates are statistically significant, indicating that proxy choice can materially alter both the magnitude and the interpretation of IV estimates in applied work.

Table 3: Instrument Validity Checks - Measurement Error

Dep. Var.:	Number of Crimes					
IV:	TI Frequency			TI Strength		
$PM_{2.5}$ measured by:	City Avg. (1)	CHAP (2)	MERRA-2 (3)	City Avg. (4)	CHAP (5)	MERRA-2 (6)
$PM_{2.5}$	0.212***	0.323***	0.176	0.228***	0.343***	0.188
	(0.060)	(0.097)	(0.158)	(0.055)	(0.085)	(0.177)
Obs	742,216	742,216	742,216	742,216	742,216	742,216
KP Fstat	383.61	261.93	134.16	199.32	155.92	114.24

Notes: The dependent variable is the daily number of crimes in a city. Columns (1) - (3) use the same set of daily thermal inversion frequency measures as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. Columns (4) - (6) use the same set of daily thermal inversion strength as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. All regressions include city, day-of-week and month-by-year fixed effects and weather controls. Standard errors clustered at the city level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table 4 evaluates a third potential violation: confounding events, focusing on wind farm entry. Relative to the baseline specification in column (1), adding treatment-by-time controls, dynamic treatment effects, cumulative exposure measures, or city-by-year fixed effects has little effect on either the estimated $PM_{2.5}$ coefficient or the first-stage strength. Two factors likely explain this stability. First, wind farm entry plausibly has direct effects on only a narrow set of outcomes, such as subjective well-being or mental health, while its direct impact on crime is likely limited outside of the air pollution channel. As a result, crime may not be an outcome for which infrastructure-driven confounding is particularly salient. Second, the baseline specification already includes city and month-by-year fixed effects, which absorb much of the spatial and temporal variation associated with infrastructure rollout. The additional treatment-based controls therefore provide little incremental identifying power beyond what is already accounted for in the baseline design.

Comparing the Tables 2-4 using the common baseline highlights an important asymmetry across threats to instrument validity. In Table 2, addressing co-pollutant bias fundamentally changes the interpretation of the IV estimate: the positive and precisely estimated $PM_{2.5}$ effect in the single-pollutant model largely disappears once correlated

Table 4: Instrument Validity Checks - Wind Farm Entry

Dep. Var.:	Number of Crimes				
	No control	Treat×After	Treat×Year	Cumulative wind farms	City×Year FE
	(1)	(2)	(3)	(4)	(5)
$PM_{2.5}$	0.212***	0.213***	0.218***	0.211***	0.204***
	(0.060)	(0.060)	(0.061)	(0.060)	(0.046)
Obs	742,216	742,216	742,216	742,216	742,216
KP Fstat	383.61	383.79	384.05	383.45	384.69

Notes: The dependent variable is the daily number of crimes in a city. *Treat* is an indicator equal to one if a city ever has an operating wind farm during the sample period. *After* is an indicator equal to one for years after the first wind farm begins operation in a city. Column (1) does not control for wind farm entry. Column (2) controls for an indicator equal to one if at least one wind farm has started operating in the city in a given year. Column (3) includes interactions between an indicator for cities that ever have a wind farm during the sample period and year indicators. Column (4) controls for the cumulative number of wind farms in a city. Column (5) includes city-by-year fixed effects. All specifications use the same set of daily thermal inversion frequency measures as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. All regressions include city, day-of-week and month-by-year fixed effects and weather controls. Standard errors clustered at the city level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

pollutants are jointly instrumented, with the estimated effect shifting primarily to NO_2. Non-classical measurement error in Table 3 changes both the magnitude and stability of the estimate, depending on how pollution is measured. By contrast, the confounding channel examined in Table 4 has relatively little impact in this setting. Taken together, these results suggest that co-pollutant bias and measurement error are the more immediate and empirically consequential threats to IV validity in air pollution studies, while confounding events may be context-dependent and less uniformly problematic.

5 Implications for IV Practice

Applied researchers typically face two sequential challenges when using instrumental variables. The first is feasibility: given the available data, temporal resolution, and empirical setting, which instruments are empirically viable and yield a strong first stage? The second is validity: conditional on a relevant instrument, what threats to the exclusion restriction remain? Our discussion follows this workflow. We first describe how data and context shape feasible IV strategies in practice, and then turn to exclusion-restriction concerns that

arise once relevance has been established.

5.1 Determining Feasible IVs Based on Data and Context

The first task for empirical researchers is to determine which instruments are feasible given their data and empirical setting. This is a design question. A feasible instrument must generate variation at the same temporal and spatial scale as the pollution measure and the estimating equation, and it must produce a sufficiently strong first stage in the setting under study.

Our classification tree analysis (Figure 1) provides a useful descriptive guide for this first step. It shows that papers with similar data structures and study settings tend to rely on similar families of instruments. This pattern suggests that instrument choice is not arbitrary. Instead, the available pollution measure, the frequency of observation, and the local context sharply narrow the set of instruments that can plausibly be used in practice.

At the same time, researchers should not stop at first-stage feasibility. Our results show that some instrument–data combinations are more vulnerable to exclusion-restriction problems than others. In particular, thermal inversion instruments are widely used with satellite-based pollution measures, but our analysis suggests that this pairing can be problematic because shared atmospheric conditions may induce non-classical measurement error in the pollution proxy. These results suggest that estimates from this design should be interpreted with caution.

A practical implication is that researchers should treat feasibility as an initial screening step rather than a final criterion for instrument selection. The goal is to identify instruments that not only generate relevant variation, but also fit the pollution measure and local context in a way that makes the exclusion restriction more plausible.

5.2 Co-Pollutant Bias in IV Designs

Conditional on selecting a relevant instrument, our analysis highlights exclusion-restriction threats that arise regardless of the specific IV used. Chief among these is co-pollutant bias. In environments where pollutants such as $PM_{2.5}$, NO_2, and SO_2 are strongly correlated, single-pollutant IV designs risk attributing to one pollutant variation that in fact reflects a broader pollution mixture. As shown in Table 2, when pollutants are jointly instrumented in our China application, the inversion-based variation loads primarily onto NO_2, and the NO_2-indexed total effect defined in (2) remains stable across specifications, whereas the PM2.5-indexed total effect does not. This pattern indicates that, in this context, the instrument identifies a combustion-related pollution shock rather than a stand-alone particulate mass shock. This distinction is also policy-relevant, since different pollutants correspond to different regulatory levers, and a design that implicitly identifies a combustion-driven NOx shock has different implications than one interpreted as a pure $PM_{2.5}$ mass shock.

In China during our study period, urban air pollution episodes were heavily influenced by traffic, coal-fired power generation, and industrial combustion—sources that emit large quantities of nitrogen oxides (NOx). NO_2 therefore serves as a direct marker of combustion activity and responds sharply to inversion conditions. At the same time, NOx emissions also contribute to particulate pollution through nitrate formation, implying that combustion shocks mechanically raise both NO_2 and $PM_{2.5}$. In this setting, the instrument is most naturally interpreted as shifting a combustion-driven pollution bundle, with NO_2 capturing the dominant local component of that shock.

The appropriate specification may differ across regions and time periods, since the composition of $PM_{2.5}$ varies substantially with economic structure, regulatory history, and climate. In some settings, particulate matter may reflect a larger contribution from organic aerosol, dust, or sulfate; in others, nitrate linked to NOx emissions may be more prominent. Rather than presuming a universal focal pollutant, researchers should empirically evaluate

which pollutants their instrument shifts – ideally through a diagnostic exercise similar to Table 2 that compares single- and multi-pollutant specifications and reports pollutant-indexed total effects.

A robust solution, when feasible, is to jointly instrument all major, highly correlated pollutants. When this is not feasible due to limited instruments, researchers should avoid the common practice of including highly correlated co-pollutants as simple controls (e.g., instrumenting $PM_{2.5}$ while controlling for NO_2), since this effectively treats an endogenous co-pollutant as exogenous and can reintroduce bias. A practical alternative is to restrict controls to weakly correlated species (such as CO or ozone in our setting), which are less likely to absorb the instrument-induced variation in the focal pollutant. More generally, researchers should treat the identity of the "relevant stimulus" as an empirical question tied to both the instrument and local pollution chemistry, rather than a maintained assumption of the design.

5.3 Instrument-Specific Threats to Exogeneity

Beyond co-pollutant bias, some exclusion-restriction threats depend on the specific combination of instruments and pollution measures. One such concern arises when thermal inversion instruments are paired with satellite-based pollution data. This concern is not a general indictment of thermal inversion designs. Rather, it is specific to settings in which the inversion measure and the pollution proxy may embed related atmospheric variation. In this case, shared atmospheric variation can generate non-classical measurement error, undermining the exclusion restriction even when first-stage relationships appear strong.

Accordingly, when satellite-based pollution measures are used, researchers should examine whether the instrument and pollution proxy are driven by overlapping atmospheric inputs. When that risk appears salient, researchers should exercise caution in relying on thermal inversion instruments and consider alternative sources of exogenous variation, such as wind-based measures or policy-driven shocks. By contrast, pollution

measures derived from ground monitors – through spatial interpolation or city-level averaging – exhibit weaker correlations with meteorological instruments and are therefore less susceptible to this particular concern.

Finally, the exogeneity of meteorological instruments can be affected by human-made infrastructure that alters local atmospheric conditions. Wind farm expansion provides one example: turbine-induced turbulence can reduce the frequency and strength of thermal inversions, directly affecting a commonly used instrument. In applications where wind farm entry plausibly affects outcomes through non-pollution channels, researchers should consider appropriate controls or specification checks. In our crime example, this channel appears to play a limited role, likely because wind farm entry is measured at a coarse temporal resolution and has little direct connection to daily crime outcomes. More generally, the relevance of such confounding is context- and outcome-specific and should be evaluated on a case-by-case basis rather than assumed away.

References

Altonji, Joseph G, Todd E Elder, and Christopher R Taber. 2005. "An evaluation of instrumental variable strategies for estimating the effects of catholic schooling." *The journal of human resources*, 40(4): 791–821.

Andrews, Isaiah, James H Stock, and Liyang Sun. 2019. "Weak instruments in instrumental variables regression: Theory and practice." *Annu. Rev. Econ.*, 11(1): 727–753.

Arceo, Eva, Rema Hanna, and Paulina Oliva. 2016. "Does the effect of pollution on infant mortality differ between developing and developed countries? Evidence from Mexico city." *Econ. J.*, 126(591): 257–280.

Atanasov, Vladimir, and Bernard Black. 2021. "The trouble with instruments: The need for pretreatment balance in shock-based instrumental variable designs." *Manage. Sci.*, 67(2): 1270–1302.

Baidya Roy, Somnath, and Justin J Traiteur. 2010. "Impacts of wind farms on surface air temperatures." *Proc. Natl. Acad. Sci. U.S.A.*, 107(42): 17899–17904.

Barwick, Panle Jia, Shanjun Li, Deyu Rao, and Nahim Bin Zahur. 2024. "The healthcare cost of air pollution: Evidence from the world's largest payment network." *Rev. Econ. Stat.*, 1–52.

Bazzi, Samuel, and Michael A Clemens. 2013. "Blunt instruments: Avoiding common pitfalls in identifying the causes of economic growth." *American Economic Journal Macroeconomics*, 5(2): 152–186.

Benmarhnia, Tarik, Prashant Bharadwaj, and Mauricio Romero. 2018. "Using Instrumental Variables for Assessing the Effects of Air Pollution on Health: A Cautionary Tale." *Working Paper*.

Callaway, Brantly, and Pedro H C Sant'Anna. 2021. "Difference-in-Differences with multiple time periods." *J. Econom.*, 225(2): 200–230.

Chay, K Y, and M Greenstone. 2003. "The impact of air pollution on infant mortality: Evidence from geographic variation in pollution shocks induced by a recession." *Q. J. Econ.*, 118(3): 1121–1167.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif. 2019. "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction." *Am. Econ. Rev.*, 109(12): 4178–4219.

Ezratty, Véronique, Gaëlle Guillossou, Catherine Neukirch, Monique Dehoux, Serge Koscielny, Marcel Bonay, Pierre-André Cabanes, Jonathan M Samet, Patrick Mure, Luc Ropert, et al. 2014. "Repeated nitrogen dioxide exposures and eosinophilic airway inflammation in asthmatics: a randomized crossover study." *Environ. Health Perspect.*, 122(8): 850–855.

Fabra, Natalia, Eduardo Gutiérrez, Aitor Lacuesta, and Roberto Ramos. 2024. "Do renewable energy investments create local jobs?" *Journal of Public Economics*, 239(105212): 105212.

Global Energy Monitor. 2025. "Global Wind Power Tracker."

Hanna, Rema, and Paulina Oliva. 2015. "The effect of pollution on labor supply: Evidence from a natural experiment in Mexico City." *J. Public Econ.*, 122: 68–79.

Isen, Adam, Maya Rossin-Slater, and W Reed Walker. 2017. "Every breath you take—every dollar you'll make: The long-term consequences of the clean air act of 1970." *J. Polit. Econ.*, 125(3): 848–902.

Lal, Apoorva, Mackenzie Lockhart, Yiqing Xu, and Ziwen Zu. 2024. "How much should we trust instrumental variable estimates in political science? Practical advice based on 67 replicated studies." *Polit. Anal.*, 32(4): 521–540.

Mellon, Jonathan. 2024. "Rain, rain, go away: 194 potential exclusion-restriction violations for studies using weather as an instrumental variable." *Am. J. Pol. Sci.*

Moretti, Enrico, and Matthew Neidell. 2011. "Pollution, health, and avoidance behavior: Evidence from the ports of Los Angeles." *J. Hum. Resour.*, 46(1): 154–175.

Schlenker, Wolfram, and W Reed Walker. 2016. "Airports, air pollution, and contemporaneous health." *Rev. Econ. Stud.*, 83(2): 768–809.

Stock, James H, Jonathan H Wright, and Motohiro Yogo. 2002. "A survey of weak instruments and weak identification in generalized method of moments." *J. Bus. Econ. Stat.*, 20(4): 518–529.

Van Donkelaar, Aaron, Melanie S Hammer, Liam Bindle, Michael Brauer, Jeffery R Brook, Michael J Garay, N Christina Hsu, Olga V Kalashnikova, Ralph A Kahn, Colin Lee, et al. 2021. "Monthly global estimates of fine particulate matter and their uncertainty." *Environ. Sci. Technol.*, 55(22): 15287–15300.

World Health Organization. 2021. "WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide." World Health Organization.

Zhang, Yan, Mei-Po Kwan, and Libo Fang. 2025. "An LLM driven dataset on the spatiotemporal distributions of street and neighborhood crime in China." *Sci. Data*, 12(1): 467.

Zou, Eric Yongchen. 2020. "Wind turbine syndrome: The impact of wind farms on suicide." *Working Paper*.

Online Appendix

Instrument Validity in Air Pollution Impact Studies

Zhenxuan Wang, Alex Yu

A Additional Figures and Tables **SI-1**

A Additional Figures and Tables

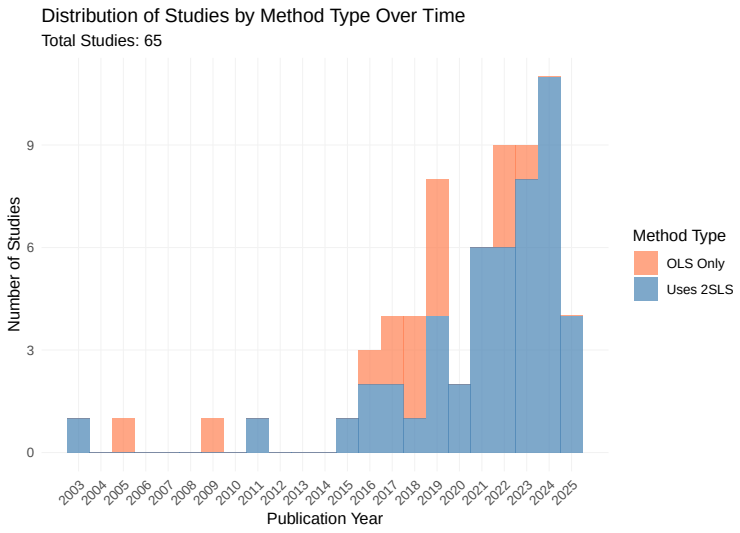


Figure S1: Summary of the Air Pollution Literature

Notes: The figure shows the distribution of studies by publication year among the 65 papers in our sample. The full list of papers is reported in Table S1.

Table S1: Instrumental Variable Strategies in the Air Pollution Impact Literature

IV Category	**Papers**
NA	Adhvaryu et al. (2022); Aragón et al. (2017); Archsmith et al. (2018); Bharadwaj et al. (2017); Chang et al. (2018); Chang et al. (2019); Chang et al. (2019); Currie and Neidell (2005); Currie et al. (2009); He et al. (2022); Heutel and Ruhm (2016); Jia and Ku (2019); Mo et al. (2023); Park and Kim (2022); Rangel and Vogl (2019); Zhang and Mu (2018)
Traffic	Huang and Gillingham (2024); Moretti and Neidell (2011); Schlenker and Walker (2016)
Event	Alexander and Schwandt (2022); Bishop et al. (2023); Chay and Greenstone (2003); Hanna and Oliva (2015); Isen et al. (2017); Salvo et al. (2024); Sun et al. (2024); Zhong et al. (2017)
Meteorological: Wind	Barwick et al. (2024); Batkeyev and DeRemer (2023); Bedi et al. (2021); Bondy et al. (2020); Carneiro et al. (2021); Chen et al. (2021); Dai et al. (2024); Deryugina et al. (2019); Fan (2024); Fu et al. (2022); Gao et al. (2023); Graff Zivin et al. (2023); Halliday et al. (2019); He et al. (2019); He and Zheng (2024); Herrnstadt et al. (2021); Hoffmann and Rud (2024); Huang et al. (2023); Kim and Gillingham (2025); Shr et al. (2023); Tanaka (2024); Wang et al. (2022)
Meteorological: Thermal Inversion	Arceo et al. (2016); Batkeyev and DeRemer (2023); Bondy et al. (2020); Bracht and Verhoeven (2025); Chen et al. (2018); Chen et al. (2022); Cook et al. (2023); Cui et al. (2023); Deschenes et al. (2020); Fu et al. (2021); Gao et al. (2024); He et al. (2019); He et al. (2025); Krebs and Luechinger (2024); Sager (2019); Xia et al. (2022); Yao et al. (2022)
Meteorological: Other	Agarwal et al. (2021); Borgschulte et al. (2024)

Table S2: Simulation Framework using Wind Instruments

Panel A. Simulation Design

Variables: X_1: $PM_{2.5}$ (primary pollutant); X_2: NO_2, SO_2, CO, O_3 (co-pollutants). All standardized (mean = 0, SD = 1).
True DGP: $Y = \beta_1 X_1 + \beta_2 X_2 + \rho(X_1 - \hat{X}_1) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0,1)$.
Parameters: $\beta_1 = 1$; $\beta_2 \in [0.1, 1.2]$; $\rho \in [0,1]$ (increment = 0.1).

Panel B. Simulation Results using wind direction × pollution zones as IV

	No FE				With City/Year FE			
X_2	NO_2	SO_2	CO	O_3	NO_2	SO_2	CO	O_3
Correlation structure								
Corr(X_1, X_2)	0.59	0.46	0.10	-0.14	0.59	0.46	0.10	-0.14
Corr(X_2, Endo)	0.54	0.43	0.09	-0.18	0.08	0.09	0.01	-0.06
$\pi_{x_2\|\hat{x}_1}$	0.91	0.67	0.15	0.26	0.75	0.25	0.13	0.63
Bias under stimulus interpretation $(\hat{\beta}_1 - \beta_1)$								
OLS–Single	0.84	0.76	0.52	0.36	0.84	0.76	0.55	0.40
IV–Single	0.59	0.43	0.09	0.17	0.49	0.16	0.09	0.41
OLS–Both	0.47	0.47	0.46	0.46	0.50	0.50	0.50	0.50
IV–w/Control	-0.24	-0.09	-0.01	0.03	-0.08	-0.03	0.00	0.07
IV–Both	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bias under proxy interpretation $(\hat{\beta}_1 - \beta_1^{proxy})$								
OLS–Single	0.46	0.46	0.46	0.46	0.50	0.50	0.50	0.49
IV–Single	0.21	0.13	0.03	0.26	0.14	-0.10	0.03	0.51
Instrument strength								
KP Fstat (IV-Single)	41.25	41.25	41.25	41.25	9.21	9.21	9.21	9.21
KP Fstat (IV-Both)	8.47	13.55	6.02	21.59	9.22	4.04	6.19	6.67

Notes: Bias values are calculated as the average across 132 simulation settings, corresponding to all combinations of $\beta_2 \in [0.1, 1.2]$ and $\rho \in [0,1]$. Positive values indicate overestimation relative to the true coefficient value of 1. Number of observations: 3,167,107.

Table S3: Robustness Checks - Simulation with Station and Date Fixed Effects

X_2	NO_2	SO_2	CO	O_3
A. Thermal inversion as IV				
Correlation structure				
Corr(X_1, X_2)	0.59	0.46	0.10	-0.14
Corr(X_2, Endo)	0.03	0.03	0.00	0.01
$\pi_{x_2\|\hat{x}_1}$	1.02	0.78	0.11	0.53
Bias under stimulus interpretation $(\hat{\beta}_1 - \beta_1)$				
OLS–Single	0.76	0.69	0.54	0.51
IV–Single	0.67	0.50	0.07	0.35
OLS–Both	0.50	0.49	0.49	0.49
IV–w/Control	-0.39	-0.14	0.00	-0.01
IV–Both	0.00	0.00	0.00	0.00
Bias under proxy interpretation $(\hat{\beta}_1 - \beta_1^{proxy})$				
OLS–Single	0.49	0.49	0.49	0.49
IV–Single	0.40	0.31	0.02	0.33
Instrument strength and sample size				
KP Fstat (IV-Single)	194.38	194.38	194.38	194.38
KP Fstat (IV-Both)	17.78	10.46	3.53	5.99
Observations	3,167,316	3,167,107	3,167,433	3,167,362
B. Wind direction × pollution zones as IV				
Correlation structure				
Corr(X_1, X_2)	0.59	0.46	0.10	-0.14
Corr(X_2, Endo)	0.03	0.04	0.01	-0.01
$\pi_{x_2\|\hat{x}_1}$	0.74	0.31	0.15	0.15
Bias under stimulus interpretation $(\hat{\beta}_1 - \beta_1)$				
OLS–Single	0.76	0.69	0.54	0.51
IV–Single	0.48	0.20	0.10	0.10
OLS–Both	0.50	0.50	0.50	0.50
IV–w/Control	-0.11	-0.03	0.00	0.00
IV–Both	-0.01	0.00	0.00	0.00
Bias under proxy interpretation $(\hat{\beta}_1 - \beta_1^{proxy})$				
OLS–Single	0.49	0.50	0.50	0.49
IV–Single	0.21	0.00	0.05	0.07
Instrument strength and sample size				
KP Fstat (IV-Single)	7.79	7.79	7.79	7.79
KP Fstat (IV-Both)	8.84	4.96	3.52	7.15
Observations	3,011,227	3,011,015	3,011,345	3,011,269

Notes: The primary pollutant is $PM_{2.5}$. Bias values are calculated as the average across 132 simulation settings, corresponding to all combinations of $\beta_2 \in [0.1, 1.2]$ and $\rho \in [0, 1]$. Positive values indicate overestimation relative to the true coefficient value of 1.

Table S4: Instrument Validity Checks Using Crime Rate - Co-Pollutant Bias

Dep. Var.:	Crime Rate						
Instrument for:	$PM_{2.5}$	NO_2	$PM_{2.5}$, NO_2	$PM_{2.5}$, SO_2	NO_2, SO_2	$PM_{2.5}$, NO_2, SO_2	$PM_{2.5}$, NO_2, SO_2
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$PM_{2.5}$	0.005***		-0.004	-0.003		-0.004	0.016
	(0.001)		(0.006)	(0.005)		(0.006)	(0.080)
NO_2		0.010***	0.017**		0.010	0.020	-0.004
		(0.002)	(0.009)		(0.017)	(0.021)	(0.098)
SO_2				0.018*	0.001	-0.003	0.001
				(0.010)	(0.022)	(0.022)	(0.030)
O_3							-0.002
							(0.017)
CO							-0.894
							(3.160)
KP Fstat	307.94	479.01	11.80	13.29	9.52	8.75	0.13
Obs	598,636	598,636	598,636	598,636	598,636	598,636	598,636
Total Effect:							
$PM_{2.5}$-indexed	0.005	-	-0.001	0.001	-	-0.001	0.016
NO_2-indexed	-	0.010	0.011	-	0.010	0.011	0.022

Notes: The dependent variable is the daily crime rate (number of crimes per 100,000 population) in a city. All specifications use the same set of daily thermal inversion frequency measures as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. All regressions include city, day-of-week and month-by-year fixed effects and weather controls. The pollutant-indexed total effect is defined as the change in the outcome associated with a one-unit increase in the corresponding pollutant, allowing correlated co-pollutants to respond endogenously. Standard errors clustered at the city level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table S5: Instrument Validity Checks Using Crime Rate - Measurement Error

Dep. Var.:	Crime Rate					
IV:	TI Frequency			TI Strength		
$PM_{2.5}$ measured by:	City Avg. (1)	CHAP (2)	MERRA-2 (3)	City Avg. (4)	CHAP (5)	MERRA-2 (6)
$PM_{2.5}$	0.005***	0.009***	0.006	0.006***	0.009***	0.005
	(0.001)	(0.002)	(0.005)	(0.001)	(0.002)	(0.005)
Obs	598,636	598,636	598,636	598,636	598,636	598,636
KP Fstat	307.94	214.87	98.92	182.28	175.64	127.51

Notes: The dependent variable is the daily crime rate (number of crimes per 100,000 population) in a city. Columns (1) - (3) use the same set of daily thermal inversion frequency measures as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. Columns (4) - (6) use the same set of daily thermal inversion strength as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. All regressions include city, day-of-week and month-by-year fixed effects and weather controls. Standard errors clustered at the city level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Table S6: Instrument Validity Checks Using Crime Rate - Wind Farm Entry

Dep. Var.:	Crime Rate				
	No control	Treat×After	Treat×Year	Cumulative wind farms	City×Year FE
	(1)	(2)	(3)	(4)	(5)
$PM_{2.5}$	0.005***	0.005***	0.006***	0.005***	0.005***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Obs	598,636	598,636	598,636	598,636	598,636
KP Fstat	307.94	308.07	308.48	308.09	305.34

Notes: The dependent variable is the daily crime rate (number of crimes per 100,000 population) in a city. *Treat* is an indicator equal to one if a city ever has an operating wind farm during the sample period. *After* is an indicator equal to one for years after the first wind farm begins operation in a city. Column (1) does not control for wind farm entry. Column (2) controls for an indicator equal to one if at least one wind farm has started operating in the city in a given year. Column (3) includes interactions between an indicator for cities that ever have a wind farm during the sample period and year indicators. Column (4) controls for the cumulative number of wind farms in a city. Column (5) includes city-by-year fixed effects. All specifications use the same set of daily thermal inversion frequency measures as instruments, including those defined between the first and second, first and third, and second and third atmospheric layers. All regressions include city, day-of-week and month-by-year fixed effects and weather controls. Standard errors clustered at the city level are reported in parentheses. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.