Appendix D: Implications of Bias due to Measurement Error

This section describes the reasoning for our claims about the prospective biases in our coefficients due to measurement error. The formalization is based on Bound et al. (1994). Let the measured speech complexity index and pollution levels, respectively, be written as:

$$y = y^* + \nu \quad (6)$$

$$p = p^* + u \quad (7)$$

where y is the speech complexity index (i.e., the Flesch-Kincaid index) and p is a measure of particulate matter (i.e., $PM_{2.5}$). An * indicates the true value of the data – i.e., the spoken level of complexity and individual level of pollution exposure. ν and u are error terms that are (potentially) correlated with these true values. Our preferred regressions are in levels with a full suite of time and individual fixed effects included. These parameters remove any time invariant individual measurement error (e.g., if an MP repeatedly uses a word incorrectly) or variation that is common to all MPs (e.g., if pollution is systematically higher on, say, Mondays). For current purposes, we ignore these effects and focus on the two aforementioned sources measurement error arising from the data-interpretation mismatch. For the dependent variable, this is Hansard editing and, for the independent variable, inexact assignment of pollution exposure.

We start with the error in the Flesch-Kincaid index and make several observations. First, editors do not transcribe texts on the same day on which the words are spoken. While we do not have information on specific transcription dates, we believe it is reasonable to treat ν as independent of pollution, p and u. Next, our research hypothesis is that MPs will be affected by pollution. This phenomenon may manifest itself in several ways. MPs may stumble or have a greater propensity to use verbal ticks. These "ums" and "ahs" are then systematically edited out of the recorded text in non-classical fashion. This means that the level of editing applied to a specific

speech is correlated with the true level of speech complexity. Hence we rewrite v from (6) as:

$$\nu = \delta y^* + \nu^* \quad (8)$$

where v^* is uncorrelated with the dependent (and independent) variable and δ is the coefficient from a hypothetical regression of v on the true speech quality index, y^*. Based on what we know about the editing process, we expect that $\delta \leq 0$. This is because when MPs increase the frequency of "ums" and "ahs" editors will be active (i.e., short, single syllable words are deleted from the official text) and y^* will be low. Recording a $y > y^*$ implies that $v > 0$ for low values of y^*, so regressing v on y^* gives a coefficient, δ, which is less than zero.

We next turn to pollution assignment. Given our design, we maintain that assignment of pollution exposure is conditionally independent of potential outcomes – politicians are making speeches for citizens who live across the country and their statements are formally documented within the official record (the database that we exploit). Still, there may be error in the measurement of the pollution assigned to specific MPs. Prior to making a speech in the House of Commons, an MP may have travelled to a heavily polluted location or may have time-varying health issues that make her more susceptible to ambient concentrations on a particular day. Pollution levels vary throughout the day, so averages may over- or under-state true exposure. Moreover, we focus on contemporaneous pollution and lagged exposure may matter.

Overall however we treat the error in pollution assignment, the independent variable of interest, as classical errors-in-variables – i.e., u is uncorrelated with p^*. This errors-in-variables specification implies attenuation bias that is proportional to the ratio of the variance of u to the variance of the measured p. The magnitude of this bias is captured by the coefficient a hypothetical regression of u on p. Define λ as the estimated coefficient from this (hypothetical) regression. And as we are dealing with attenuation bias, we expect that $\lambda \leq 1$.

We now combine these two biases. Let the true parameter from a linear least squares regression of the Flesch-Kincaid index on pollution concentration equal β (i.e., this is the parameter that we would estimate without measurement error) and what we actually estimate be $\hat{\beta}$.

Using (8) and (7), we can write the bias in the estimated parameter as:[21]

$$\begin{aligned} \text{plim}\hat{\beta} - \beta &= -\lambda\beta + \delta\beta \\ &= -\beta(|\delta| + \lambda) \end{aligned} \quad (9)$$

where the second line follows from $\delta \leq 0$. The bias in (9) equals $-\beta(|\delta| + \lambda)$ and shows the attenuation arising from the biases in the dependent variable and pollution assignment. Error in the dependent variable leads to an downward (toward zero) bias of δ whereas λ reflects the conventional attenuation bias (also towards zero) of the standard errors-in-variables model. Both biases are then scaled by the true effect size, β.