Millimet, Daniel L.

Working Paper

(Don't) Walk This Way: The Econometrics of Crosswalks

IZA Discussion Papers, No. 17154

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Millimet, Daniel L. (2024) : (Don't) Walk This Way: The Econometrics of Crosswalks, IZA Discussion Papers, No. 17154, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/302671

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17154

(Don't) Walk This Way: The Econometrics of Crosswalks

Daniel L. Millimet

JULY 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17154

(Don't) Walk This Way: The Econometrics of Crosswalks

Daniel L. Millimet
Southern Methodist University and IZA

JULY 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

(Don't) Walk This Way: The Econometrics of Crosswalks*

It is increasingly common in empirical research to merge data sets containing different units of observation. When the units are not nested, a *crosswalk* specifying how the units from one data source are allocated to the units of the other is needed. Unfortunately, most crosswalks are *ad hoc*, a fact that is often ignored by researchers and has not caught the attention of econometricians. Here, I show that use of an incorrect crosswalk induces measurement error that is necessarily nonclassical and can be consequential. I discuss and illustrate the ramifications of using a flawed crosswalk, present two specification tests, offer potential solutions, and provide an application to the effects of social media on political polarization.

JEL Classification: C18, C81, D72, J16

Keywords: crosswalk, nonclassical measurement error, social media, polarization

Corresponding author:
Daniel L. Millimet
Department of Economics
Box 0496
Southern Methodist University
Dallas, TX 75275-0496
USA

E-mail: millimet@smu.edu.

* The author is grateful to Hao Dong, Parker Fleming, James MacKinnon, John Mullahy, Justin Pierce, Daniel Reynolds, and Denni Tommasi for helpful discussions, and Francine Blau and Peter Brummond for sharing their crosswalks. The author is grateful to conference participants at Texas Econometrics Camp and the New York Econometrics Camp, as well as seminar participants at Syracuse University and University of Minnesota. Nafisa Esha provided valuable research assistance.

Any road followed precisely to its end leads precisely nowhere.

– Bene Gesserit, *Dune* (1965)

1 Introduction

Researchers have become more creative and ambitious in constructing data sets for empirical analysis. This often includes combining data from multiple sources or from multiple time periods. But, to take liberties with an old adage, with great ambition comes great responsibility. Specifically, the responsibility to ensure that the assembled data are accurate. Accuracy often comes in to question when data are linked across sources or over time and the *unit of observation changes*. In such cases, it is necessary to convert the data to a *common* unit of observation to enable statistical analysis. Conversion to a common unit of observation is accomplished with the use of a *crosswalk* (or *concordance*).

Crosswalks detail how the researcher maps one unit of observation into a different unit of observation. If units are nested across the different data sources, then the mapping is straightforward. However, when they are not nested, complications arise. In such cases, crosswalks specify how each original unit is apportioned to units in the alternative data source.

To fix ideas, Table I highlights some commonly used crosswalks. It is by no means exhaustive. Geography in the United States (and other countries) can be classified in many unique ways across data sources. Moreover, even when using a common geographic unit, the borders of such units often change (in a non-nested way) over time. For example, congressional districts and school districts do not need to follow zip code or county borders, and the boundaries of both are frequently redrawn. Industry and occupation classifications are also revised every few years to keep up with changes in technology and production. This also applies to sector classifications available in international trade and tariff data. Medical data use codes to denote diseases, medical procedures, or drugs that evolve over time and are inconsistent across institutions.

Consider the example of obtaining attributes of US congressional districts from county-level data. If counties are nested within districts, then the attributes of a district are the aggregation of the attributes of the counties that fall within its borders.[1] However, counties are often divided among multiple districts. Texas has 254 counties, 30 of which are split into between two and nine districts for the $118^{\texttt{th}}$ Congress (Figure I). Viewed from the other side, none of Texas' 38 congressional districts are comprised solely of entire counties; each contains a portion of one to four counties. Currently, the boundaries of only 12 of 435 congressional districts completely follow county borders.

[1] Of course, the 'correct' aggregation scheme may not always be obvious.

The solution to date, often attributed to Hornbeck (2010), is to specify a crosswalk that apportions counties to districts assuming that a county's contribution is proportional to the land area or population contained in the district.[2] For instance, to derive a count of foreign-born in each district given data on the foreign-born population in each county, a researcher might assume that 10% of the county's foreign-born reside in a given district if 10% of the land area of a county lies in the district. This assumes that the foreign-born are uniformly distributed within a county, which is not realistic given the presence of ethnic enclaves (e.g., Abramitzky and Boustan, 2017). The final *estimate* is a weighted sum of the foreign-born populations in the counties that it overlaps.

Alternative crosswalks rely on different weighting schemes. For example, population-based weights may be used to map income or education from counties to districts. However, even in this case, there are choices to be made as population may be measured at the individual level or the household level or only including certain demographic groups. In the case of industry-level data, value-added-, employment-, establishment-, shipment-, or payroll-based weights may be used to map industries from one classification to another.

This brief description should make it clear that every crosswalk contains a degree of arbitrariness. For example, Goldschmidt and Schmieder (2017, p. 1174) state: "We used crosswalks together with our best judgment to consistently classify business service firms and outsourcing over time." Bisbee and Zilinsky (2023, p. 289) describe the "choices and challenges of aggregating the same measures to different geographic units." Thus, crosswalks *necessarily* introduce some measurement error into the data. Moreover, such errors are often compounded as more than a single crosswalk is necessary to arrive at the final data set. For instance, Carlino and Drautzburg (2020, p. 773) write:

> "After imputing employment according to the prevailing classification scheme in each year, we use cross-walks from the 1977 SIC classification to the 1987 SIC classification and from future NAICS classifications to the 1997 six-digit NAICS classification, which we then, in turn, transform to the 1987 four-digit SIC classification and aggregate up to the three-digit level."

Nonetheless, the use of crosswalks is exploding in top economics journals and NBER working papers (Figure II), but has yet to catch the attention of econometricians.[3]

With this backdrop in mind, the goal of this paper is to shine a light on the imperfections in the process

[2]Such issues have a long history in demography research. Wright (1936) considers the issue of disaggregating population estimates to finer geographic levels. Spoer *et al.* (2023) construct county and census tract crosswalks using population shares created from data at the census block level as census blocks do not span multiple tracts, counties, or districts.

[3]The count of articles is based on a manual search for the word "crosswalk" in December 2023. NBER working papers found at `https://www.nber.org/papers`.

that are often glossed over or relegated to supplemental appendices.[4] The lack of attention paid to the use of imperfect crosswalks likely stems from several factors. First, perhaps the common perception is that the errors introduced are 'small' in some sense. Second, without crosswalks, our ability to ask and hopefully answer interesting and important research questions is severely hampered. For instance, the US Department of Housing and Urban Development (HUD) states on its website[5]:

> "One of the many challenges that social science researchers and practitioners face is the difficulty of relating United States Postal Service (USPS) ZIP codes to Census Bureau geographies. There are valuable data available only at the ZIP code level that, when combined with demographic data tabulated at various Census geography levels, could open up new avenues of exploration."

Finally, researchers often devote extraordinary amounts of time producing crosswalks that then become freely available (see Table I). Researchers contributing to this public good surely strive for accuracy.

In this paper, I delve into the first and second factors. I explore how 'small' the errors must be to be inconsequential, whether errors in existing crosswalks are likely to meet this level of accuracy, and alternatives to current research practice. Such alternatives are necessary not only to improve the validity of research relying on crosswalks, but also research that forsakes the use of imperfect crosswalks. For example, Agarwal *et al.* (2018) map the zip codes of mortgages to congressional districts, *excluding* mortgages in zip codes that span multiple districts of the $111^{\texttt{th}}$ Congress. This amounts to discarding 5,632 of 32,846 (17.1%) zip codes, introducing possible sample selection bias.

The remainder of the paper is organized as follows. Section 2 provides several motivating examples. Section 3 provides a literature review. Sections 4 and 5 assess the consequences of using imperfect crosswalks as well as potential solutions. Section 6 presents a simulation exercise. Section 7 provides an application to the impact of social media on political polarization. Section 8 concludes.

2 Motivating Examples

Multiple Crosswalks. To explore the practical consequences of using an error-laden crosswalk, I begin by examining a few cases where multiple crosswalks are available for the same data exercise. First, I examine different crosswalks from counties in 1990 to districts in the $103^{\texttt{rd}}$ US Congress. The crosswalks come from Ferrara *et al.* (2022) and the Missouri Census Data Center (MCDC).[6] Ferrara *et al.* (2022)

[4] Certainly some researchers give crosswalks the required attention. For example, Levinson (2015) carefully considers robustness across crosswalks, but finds only 'small' changes. Section 2 discusses a few others as well.

[5] See `https://www.huduser.gov/portal/datasets/usps_crosswalk.html`.

[6] See `https://mcdc.missouri.edu/applications/geocorr.html`.

provide six crosswalks based on different weighting schemes. One is based on land area and the remainder are based on population with various adjustments for urban or rural status, uninhabitable land, etc. The MCDC provides three different weighting schemes, based on land area, individual population, and number of household units. The standard deviations of the pairwise differences across the six weighting schemes in Ferrara *et al.* (2022) range from 0.043 to 0.212. The standard deviations of the pairwise differences across the three MCDC weighting schemes range 0.014 to 0.103. These values are used to benchmark the errors introduced in the Monte Carlo study in Section 6 and the empirical Monte Carlo study in Appendix B.

Next, I examine multiple industry crosswalks from the Standard Industrial Classification (SIC) to the North American Industrial Classification System (NAICS) 1997 provided in a single data source by Schaller and DeCelles (2022). Three different weighting schemes are available based on total establishments, employment, and payroll. Here, the standard deviations of the pairwise differences in the weights range 0.037 to 0.109.

Lastly, I examine the crosswalks used in Blau *et al.* (2013) to study trends in occupational segregation by gender. The authors construct three crosswalks from 1990 Census occupation codes to 2000 Census occupation codes by using (i) aggregate population counts, (ii) female population counts, and (iii) male occupation counts. The study illustrates the importance of using gender-specific crosswalks when examining occupational segregation as "trends in occupational segregation ... are masked" when using a gender-neutral crosswalk (Blau *et al.*, 2013, p. 471). The standard deviation of the differences between the female- (male-) specific weights and the pooled crosswalk is roughly 0.04 (0.02). The standard deviation of the differences between the female- and male-specific weights is about 0.05.

Validation Data. I next compare data obtained via crosswalk to validation (i.e., ground truth) data. To do so, I rely on data from the American Community Survey (ACS) provided by the US Census Bureau.[7] The ACS provides aggregate data for both counties and congressional districts. As the basic geographical unit in the ACS is a census tract, and tracts do not cross either county or congressional district borders, ACS data aggregated to both counties and congressional districts are subject only to sampling error, but otherwise represent the truth. Moreover, since both aggregations are derived from the same underlying data across census tracts, the sampling error similarly affects both aggregations. Thus, we can consider the data at the congressional district level to be correct.

With this in mind, I start with county-level counts of the number of households living in poverty and the number of households participating in the Supplemental Nutrition Assistance Program (SNAP) for

[7]Data obtained from `https://data.census.gov/`.

2020 obtained from the 5-year ACS. I then map this into counts for congressional districts using the three MCDC weighting schemes from 2020. Finally, I compare the crosswalked counts to the counts reported by the US Census Bureau for congressional districts directly.

Figure A.I in Appendix A displays kernel densities of the four versions (three crosswalked counts and the true counts). It is evident that the crosswalked versions have a smaller variance; they do not fit the tails of the true distribution. Figure A.II plots the measurement errors (crosswalked counts minus the true counts) against the true counts. The errors are negatively correlated with the truth, and display a larger variance under the land area weighting scheme. Overall, the correlations between the three crosswalked counts of households in poverty (participating in SNAP) and the true counts range from 0.34 to 0.86 (0.45 to 0.86). The correlations between the errors from three crosswalked counts of households in poverty (participating in SNAP) and the true counts range from -0.31 to -0.59 (-0.37 to -0.61). The errors do have sample means very close to zero; ranging from -0.19 to 0.08 (-0.17 to 0.05) for poverty (SNAP).[8]

To see the bias that these weighting errors can introduce into a linear regression estimated via Ordinary Least Squares (OLS), I conduct a simple illustration. As detailed in Appendix A, I generate a district-level outcome as a function of the *true* number of households in poverty. I then regress the outcome on each of the three *crosswalked* counts. Next, I simulate ten correctly measured covariates that are correlated with the true count and generate a new outcome that depends on the true count and these ten covariates. The true coefficient on all covariates is one. In the simple regression where the crosswalked count is the only covariate, the bias is modest except when using the land area weighting scheme (Table A.I). In the multiple regression, however, the bias is extremely large and the estimated coefficient on each crosswalked count is less than 0.1. The bias is exacerbated in the multiple regression since what matters are the *partial* variances and covariances.

Empirical Monte Carlo. As a final exercise, I perform an empirical Monte Carlo study based on Che *et al.* (2022). The analysis and results are relegated to Appendix B, but illustrate the impact that choosing a different crosswalk can have on a study's findings, as well as the impact of perturbations of crosswalks within the range of the differences documented here.

In sum, errors introduced due to reliance on an imperfect crosswalk cannot be universally dismissed as 'small' in practice.

[8]The average number of households in poverty (participating in SNAP) across congressional districts is roughly 35,000 (32,000).

3 Literature

To my knowledge, there is *no* literature on the econometric implications of using an inexact crosswalk, which is shocking given the dramatic rise in their use (Figure II). Nonetheless, reliance on erroneous crosswalks has similarities to other data issues. First, there are parallels to the choice of weight matrix in spatial econometric models (Kelejian, 2008; Herrera *et al.*, 2019, among others). Second, it is similar to the problem of spatial misalignment (e.g., Pouliot, 2023) and mixed data sampling (MIDAS) (e.g., Yang *et al.*, 2023). With spatial misalignment, the dependent and independent variables do not align geographically (such as the location of farms and weather stations). In MIDAS models variables are measured at different temporal frequencies, with high frequency variables often being aggregated to a lower frequency. Third, the problem relates to the literature on multiple proxies (e.g., Lubotsky and Wittenberg, 2006). A special case of this is when matching data sets using strings and multiple matches are possible (Poirier and Ziebarth, 2019). This is part of a general problem known as probabilistic record linkage (e.g., Ridder and Moffitt, 2007). Finally, there are similarities with the problem of choosing weights in synthetic control studies (e.g., Abadie and L'Hour, 2021). A formal analysis of crosswalk errors is much needed.

4 Model

4.1 Setup

To assess the impact of using an incorrect crosswalk, let the data-generating process (DGP) be given by the following simple regression model

$$y_i = \alpha + \beta x_i^* + \varepsilon_i, \quad i = 1, ..., N \tag{1}$$

where y_i is the outcome, x_i^* is the covariate, and ε_i is a mean zero error term for observation i. We are interested in estimating β. This model can be derived from a multiple regression where the other covariates – including fixed effects – have been partialled out; the Frisch-Waugh-Lovell theorem still applies despite the particulars of the issue being analyzed (see Appendix C). However, as illustrated in Section 2, it is important to not lose sight of the fact that partialling out covariates can make the bias from even small weighting errors quite consequential.

In contrast to the standard regression setup, the covariate, x_i^*, is unobserved. Instead, the same covariate

is observed for a *different* unit of observation.[9] Let z_j^*, $j = 1, ..., M$, denote the value of the same covariate captured by x^* except for observation units indexed by j. For example, i may index congressional districts and j may index counties, or i may index NAICS industries and j may index SIC industries. The covariates are related by

$$x_i^* \coloneqq \sum\nolimits_j \omega_{ij}^* z_j^*, \tag{2}$$

where ω_{ij}^* is the true weights that map the data from z^* for unit j to x^* for observation i. The collection of weights is known as a *crosswalk*.

Even if ω_{ij}^* is known, assessing the asymptotic properties of the OLS estimates of Equation (1) is nonstandard as x^* and z^* cannot both be independent and identically distributed (iid). From Equation (2) it follows that if z^* is iid, then x^* will be heteroskedastic and cross-sectionally dependent if some units j span multiple units i. Alternatively, due to cross-sectional dependence (and other reasons), it is not realistic to assume x^* is iid. Following Poirier and Ziebarth (2019) I proceed under the assumption that x^* and z^* are iid. Appendix D considers this assumption in more detail.

I make the following assumption.

Assumption 1 (*Data-Generating Process*)**.**

(i) *The population model is* $y_i = \alpha + \beta x_i^* + \varepsilon_i$ *for all* i.

(ii) $x_i^* \coloneqq x_i(\omega^*) = \sum_j \omega_{ij}^* z_j^*$, *where* $\omega^* = (\omega_{11}^*, ..., \omega_{1M}^*, \omega_{21}^*, ..., \omega_{2M}^*, ..., \omega_{1M}^*, ..., \omega_{NM}^*)$ *is the true weighting scheme.*

(iii) $\{x_i^*\}, \left\{z_j^*\right\}$ *are iid across* $i = 1, ..., N$ *and* $j = 1, ..., M$, *respectively, each with finite first and second moments.*

(iv) $X^* \coloneqq [\iota \; x_i^*]$ *is an* $N \times 2$ *matrix of full rank where* ι *is an* $N \times 1$ *vector of ones and* $x(\omega^*)$ *is an* $N \times 1$ *vector with representative element* $x_i(\omega^*)$.

(v) $\texttt{plim} \, \frac{1}{N} [X^{*\prime} X^*] = Q$, *where* Q *is a positive definite matrix.*

(vi) $\texttt{plim} \, \frac{1}{N} [X^{*\prime} \varepsilon] = 0$.

4.2 Weights

I assume x^* is a convex combination of z^*.

[9] The case where the true dependent variable, y^*, is an unknown function of z^* is left for future research.

Assumption 2 (*Weights*)**.** *The weights, ω^*, satisfy*

(i) $\omega^*_{ij} \in [0,1]\ \forall i,j$

(ii) $\sum_i \omega^*_{ij} = 1\ \forall j$.

Consider the example of mapping counties (j) to congressional districts (i). Assumption 2 requires that the allocation of parts of county j to districts must be mutually exclusive and exhaustive. If a county is entirely contained within one district, ω^*_{ij} equals one for that district, zero otherwise. If a county is divided amongst more than one district, $\omega^*_{ij} \in (0,1)$ for districts containing at least some of county j, zero otherwise. This is illustrated in Figure III for the case of three counties and three districts.

4.3 Unknown Weighting Scheme

In practice, the true weighting scheme, ω^*, is typically unknown. A crosswalk weighting scheme, denoted by ω, is used. The relationship between the crosswalk and true weights is given by

$$\omega_{ij} \coloneqq \omega^*_{ij} + \delta_{ij}, \tag{3}$$

where δ_{ij} is the measurement error in the weights. The observed covariate is

$$x_i(\omega) \coloneqq \sum\nolimits_j \omega_{ij} z^*_j. \tag{4}$$

I make the following assumption.

Assumption 3 (*Crosswalk*)**.** *The crosswalk weights, ω, satisfy*

(i) $\omega_{ij} \in [0,1]\ \forall i,j$

(ii) $\sum_i \omega_{ij} = 1\ \forall j$

(iii) $\omega_{ij} = \omega^*_{ij}$ *if* $\omega^*_{ij} \in \{0,1\}$.

Assumptions 3(i) and 3(ii) restrict attention to crosswalks where x is also a convex combination of z^*. Assumption 3(iii) is not guaranteed to hold, but is often reasonable. It implies that if unit j is entirely contained in observation i or entirely absent from observation i, then this is known. This leads to the following remark.

Remark 1. *Under Assumptions 2 and 3, the weighting errors, δ_{ij}, have the following properties*

(D1) $\delta_{ij} \in [-\omega^*_{ij}, 1 - \omega^*_{ij}] \; \forall i, j$

(D2) $\sum_i \delta_{ij} = 0 \; \forall j$.

D1 follows from the fact that the crosswalk and true weights lie in the unit interval. The bounded nature of ω^* and ω ensure that δ is not normally distributed and that the weighting errors are negatively correlated with the true weights (Black *et al.*, 2000). **D2** follows from Equation (3) and the fact that $\sum_i \omega^*_{ij}$ and $\sum_i \omega_{ij}$ sum to one for all j. Moreover, not only is δ mean zero for each j, but **D2** also implies that the weighting errors exhibit negative spatial correlation across j. If a crosswalk allocates too much of z^*_j to unit i, then the crosswalk must necessarily allocate too little to some other unit(s) i', $i \neq i'$.

Figure III provides a hypothetical illustration of the weighting errors. Within each county the crosswalk and true weights must sum to one, the weighting errors must sum to zero, and the weighting errors are negatively correlated across districts within a given county (negative spatial correlation). Nothing precludes the weighting errors from being independent across counties, even within the same district. In the figure, the weights for county j' are assumed to be correct since it is nested in a single district.

Using Equations (2), (3), and (4), the weighting errors map into errors in the desired covariate, $x_i(\omega^*)$. This yields

$$\begin{aligned} x_i(\omega) &= x_i(\omega^*) + \sum\nolimits_j \delta_{ij} z^*_j \\ &= x_i(\omega^*) + \mu_i, \end{aligned} \tag{5}$$

which resembles the typical errors-in-variables setup. However, the measurement error, μ, is nonclassical. I add the following assumption.

Assumption 4 (*Crosswalk Errors*)**.**

(i) $\mathtt{E}[\delta] = 0$.

(ii) $\mathtt{plim}\, \frac{1}{N}(\Delta'\Delta) = \frac{1}{N}\mathtt{diag}\left(\sum_i \delta^2_{i1}, ..., \sum_i \delta^2_{iM}\right)$, *where* Δ *is an* $N \times M$ *matrix with representative element* δ_{ij}.

(iii) $\mathtt{plim}\, \frac{1}{N}(\Delta z^*) = 0$, *where* z^* *is an* $M \times 1$ *vector with representative element* z^*_j.

Assumption 4(i) formalizes **D2** and assumes that the weighting errors are mean zero not only in the sample for each unit j, but also in the population. This is *not* a strong assumption given the bounded nature of

the weights as shown in Remark 1.[10] Assumption 4(ii) states that δ_{ij} and $\delta_{i'j'}$ are independent for all i, i' when $j \neq j'$. The weighting errors cannot be independent when $j = j'$ given **D2**. Assumption 4(iii) implies that the weighting errors are independent of the covariate, z^*. This is not as restrictive as it may seem. For example, say z^* measures median income and the weighting scheme, ω is based on population shares. Even if income is correlated *population*, this does not imply that income is correlated with the *weighting errors*, δ.[11] A crucial implication of this assumption is that $\texttt{plim}\ \frac{1}{N}(\mu'\varepsilon) = 0$ even if $\texttt{Cov}(z^*, \varepsilon) \neq 0$.[12]

This leads to the following remark.

Remark 2. *Under Assumptions 1 – 4, the following properties hold.*

(M1) $\texttt{E}(\mu_i) = 0$

(M2) $\texttt{Var}(\mu_i) = \left[\texttt{Var}(z^*) + \texttt{E}(z^*)^2\right] \sum_j \texttt{Var}(\delta_{ij}) \leq \left[\texttt{Var}(z^*) + \texttt{E}(z^*)^2\right] \sum_j \left(\omega_{ij}^* - \omega_{ij}^{*2}\right)\ \forall i$

(M3) $\texttt{Cov}[x_i(\omega^*), \mu_i] = \texttt{Cov}\left(\sum_j \omega_{ij}^* z_j^*, \sum_j \delta_{ij} z_j^*\right) < 0\ \forall i$

(M4) $\texttt{Cov}(\mu_i, \mu_{i'}) \leq 0\ \forall i \neq i'$.

M1 states that the measurement error is mean zero. **M2** states that the measurement error is likely heteroskedastic since the upper bound of the variance of δ is a function of the true weights using the Bhatia–Davis inequality (Bhatia and Davis, 2000). **M3** states that the measurement error is negatively correlated with the true covariate, $x(\omega^*)$, since the weighting errors are negatively correlated with the true weights. **M4** states that the measurement errors exhibit negative spatial correlation as the covariance will be strictly less than zero when i and i' share a common unit j.

4.4 Measurement Error in z^*

If z^* is mismeasured, then there is a second source of measurement error. It is critical to allow for measurement error in z^* in the main analysis given the discussion in Sections 1 and 2 and Appendix B. It appears more common than not that z^* is derived using a prior crosswalk or imputation or is an imperfect proxy (as in the application in Section 7 where a sample of Twitter users in a locale is used as a proxy for social media usage). Let the observed covariate be given by

$$z_j \coloneqq z_j^* + \psi_j, \tag{6}$$

[10] To be clear, since the true weights and the proposed weights must sum to one for each unit j, the weighting errors *cannot* be one-sided as they *must* sum to zero. The validation exercise in Section 2 confirms this in practice.

[11] Nonetheless, this assumption may not hold in all applications. Relaxing this assumption is left to future work.

[12] $\texttt{Cov}(z^*, \delta) = 0$ implies that $\texttt{Cov}(\mu, \varepsilon) = 0$ even if $\texttt{Cov}(z^*, \varepsilon) \neq 0$ since μ is a function of the *products* between δ and z^*.

where ψ_j denotes the measurement error. Combining Equation (6) with (2), (3), and (4) yields

$$\begin{aligned} x_i(\omega) &= \sum\nolimits_j \left(\omega_{ij}^* + \delta_{ij}\right)\left(z_j^* + \psi_j\right) \\ &= x_i(\omega^*) + \sum\nolimits_j \left[\omega_{ij}^*\psi_j + \delta_{ij}\left(z_j^* + \psi_j\right)\right] \\ &= x_i(\omega^*) + \check{\mu}_i. \end{aligned} \tag{7}$$

To derive the properties of $\check{\mu}$, add the following assumption.

Assumption 5 (*Covariate Errors*)**.**

(i) $\mathtt{E}(\psi|z^*) = 0$ *where* ψ *and* z^* *are each* $M \times 1$ *vectors.*

(ii) $\mathtt{E}\left(\psi'\psi\right) = \frac{1}{N}\mathtt{diag}(\sigma_{\psi_1}^2, ..., \sigma_{\psi_M}^2)$.

(iii) $\mathtt{plim}\ \frac{1}{N}\left(\Delta\psi\right) = 0$.

(iv) $\mathtt{plim}\ \frac{1}{N}\left(W^*\psi\right) = 0$, *where* W^* *is a* $N \times M$ *matrix of the true weights.*

Under Assumption 5 ψ is classical measurement error in z^* and also uncorrelated with the weighting errors, δ, and the true weights, ω^*.

This leads to the following remark.

Remark 3. *Under Assumptions 1 – 5 the following properties hold.*

(T1) $\mathtt{E}(\check{\mu}_i) = 0$

(T2) $\mathtt{Var}(\check{\mu}_i) = \mathtt{Var}(\mu_i) + \sum_j \left(\omega_{ij}^{*2}\sigma_{\psi_j}^2 + \sigma_{\psi_j}^2\mathtt{Var}(\delta_{ij})\right) \leq \mathtt{Var}(\mu_i) + \sum_j \left[\omega_{ij}^{*2}\sigma_{\psi_j}^2 + \sigma_{\psi_j}^2\left(\omega_{ij}^* - \omega_{ij}^{*2}\right)\right]\ \forall i$

(T3) $\mathtt{Cov}[x_i(\omega^*), \check{\mu}_i] = \mathtt{Cov}\left(\sum_j \omega_{ij}^* z_j^*, \sum_j \delta_{ij} z_j^*\right) < 0\ \forall i$

(T4) $\mathtt{Cov}(\check{\mu}_i, \check{\mu}_{i'}) \leq \mathtt{Cov}(\mu_i, \mu_{i'}) \leq 0\ \forall i \neq i'$.

T4 follows from **M4** and that $\mathtt{Cov}\left(\sum_j \omega_{ij}^*\psi_j, \sum_j \omega_{i'j}^*\psi_j\right) < 0$. Compared to the case of no measurement error in z^* considered in Remark 2, classical measurement error in z^* amplifies the variance of the composite error and the negative spatial correlation in the regression error.

4.5 Properties of OLS

To assess the impact of crosswalk errors, I restate the full DGP for clarity.

$$\begin{aligned} y_i &= \alpha + \beta x_i(\omega^*) + \varepsilon_i \\ x_i(\omega^*) &= \textstyle\sum_j \omega^*_{ij} z^*_j \\ \omega_{ij} &= \omega^*_{ij} + \delta_{ij} \\ z_j &= z^*_j + \psi_j \\ x_i(\omega) &= x_i(\omega^*) + \textstyle\sum_j \left[\delta_{ij} z^*_j + (\omega^*_{ij} + \delta_{ij})\psi_j\right] = x_i(\omega^*) + \breve{\mu}_i. \end{aligned}$$

The estimating equation is

$$y_i = \alpha + \beta x_i(\omega) + (-\beta \breve{\mu}_i + \varepsilon_i) \tag{8}$$

This leads to the following result.

Proposition 1. *Let Assumptions 1-5 hold and let $\widehat{\beta}_{ols}$ denote the OLS estimate of β in Equation (8). Then*

$$\begin{aligned} \texttt{plim}\, \widehat{\beta}_{ols} &= \frac{\texttt{Cov}\left[x(\omega), y\right]}{\texttt{Var}\left[x(\omega)\right]} \\ &= \beta \left\{ \frac{\texttt{Var}\left[x(\omega^*)\right] + \overbrace{\texttt{Cov}\left(\sum_j \omega^*_j z^*_j, \sum_j \delta_j z^*_j\right)}^{<0}}{\texttt{Var}\left[x(\omega^*)\right] + \texttt{Var}\left(\breve{\mu}\right) + 2\underbrace{\texttt{Cov}\left(\sum_j \omega^*_j z^*_j, \sum_j \delta_j z^*_j\right)}_{<0}} \right\} := \beta\Pi \end{aligned}$$

If z^* is observed, then ψ equals zero and $\texttt{Var}\left(\breve{\mu}\right)$ is replaced with $\texttt{Var}\left(\mu\right)$ above.

Proof: See Appendix E.

Figure IV characterizes the bias for different combinations of the variances of the true covariate, $x(\omega^*)$, and the composite measurement error, $\breve{\mu}$. Importantly, there are regions where OLS suffers from attenuation bias ($\Pi \in (0,1)$), expansion bias ($\Pi > 1$), and even sign reversal ($\Pi < 0$). Outcomes other than attenuation bias occur when there is a large difference between the variances $x(\omega^*)$ and $\breve{\mu}$ such that $\texttt{Cov}\left[x(\omega^*), \breve{\mu}\right] < \texttt{min}\{\texttt{Var}[x(\omega^*)], \texttt{Var}(\breve{\mu})\}$. When $\texttt{Var}\left[x(\omega^*)\right]$ ($\texttt{Var}(\breve{\mu})$) is smaller, then sign reversal (expansion bias) occurs. The usual attenuation bias arises when the variance of the measurement error and the variance of the true covariate are both larger than their covariance. While there is no great intuition behind this result, to

state it less formally: (i) expansion bias arises when the variance of the measurement error is 'small', yet still 'strongly' covaries with the true covariate, and (ii) sign reversal arises when the variance of the true covariate is 'small', yet still 'strongly' covaries with the measurement error.

Note, the analysis here is similar to Black *et al.* (2000) except with a particular structure on the nonclassical measurement error in the regression covariate. It is also similar to the analysis of OLS in the presence of mixed data sampling (e.g., Yang *et al.*, 2023).

5 Pathways Forward

5.1 Single Candidate Weighting Scheme

Prior to discussing potential solutions for researchers given access to a single error-laden crosswalk, I briefly mention two possibilities that are *not* satisfactory solutions. First, one might aggregate the data to a level where a crosswalk is no longer needed. For example, if the changes over time in an industrial classification only affects, say, classifications at the 4-digit level but not the 3-digit level, then estimating a regression model at the 3-digit level precludes the need to use a crosswalk. Positing this as a solution is unsatisfactory for three reasons. First, researchers typically do not do this. Thus, guidance for researchers using crosswalks is needed. Second, it is inefficient to aggregate if one does not have to; there are fewer industries at the 3-digit level than the 4-digit level. The solutions discussed below offer alternatives to researchers. Finally, aggregation is not always possible. Consider again the case of counties and congressional districts. Here, aggregation to avoid the use of crosswalks entails changing the geographic unit of observation to regions that are comprised only of entire counties and congressional districts. A close look at Figure I indicates that this entails aggregating to the entire state. While researchers could abandon county- and district-level analyses in favor of state-level analyses, these face their own complications. For example, in the application in Section 7, the outcome of interest is the political ideology of US Congressional representatives. Aggregating this to the state-level would introduce a new weighting issue.

Second, in the presence of a single covariate suffering from classical measurement error, OLS estimates from the forward and reverse regressions can be used to bound the true coefficient asymptotically (Black *et al.*, 2000). This is the so-called Frisch bounds. Because the measurement error due to reliance on an error-laden crosswalk is nonclassical, this is no longer necessarily the case (see Appendix E). In Figure IV, the OLS estimate of β from the reverse regression suffers from expansion bias in the orange region and attenuation bias in the magenta region. Thus, while there are regions where Frisch bounds do contain the

truth asymptotically, this is not guaranteed.

I now assess possible solutions.

Decomposition. If some of the true weights, ω_{ij}^*, equal one and these are measured correctly in the crosswalk (Assumption 3(iii)), the covariates $x(\omega^*)$ and $x(\omega)$ can each be decomposed into two components. Defining $\mathcal{J}_i = \left\{j \ : \ \omega_{ij}^* > 0\right\}$ and $\mathcal{J}_i^1 = \left\{j \ : \ \omega_{ij}^* = 1\right\}$, it follows

$$\begin{aligned} x_i(\omega^*) &= \sum\nolimits_j \omega_{ij}^* z_j^* = \sum\nolimits_{j \in \mathcal{J}_i \setminus \mathcal{J}_i^1} \omega_{ij}^* z_j^* + \sum\nolimits_{j \in \mathcal{J}_1^1} z_j^* \\ &:= \widetilde{x}_i(\omega^*) + \breve{x}_i^* \end{aligned} \tag{9}$$

$$\begin{aligned} x_i(\omega) &= \sum\nolimits_j \omega_{ij} z_j = \sum\nolimits_{j \in \mathcal{J}_i \setminus \mathcal{J}_i^1} \omega_{ij} z_j + \sum\nolimits_{j \in \mathcal{J}_i^1} z_j \\ &:= \widetilde{x}_i(\omega) + \breve{x}_i \end{aligned} \tag{10}$$

where $\widetilde{x}_i(\omega^*)$ and $\widetilde{x}_i(\omega)$ are the parts of each covariate derived from units j that are only partially included in unit i and $\breve{x}_i^*$ and $\breve{x}_i$ are the parts derived from units j that are entirely included in unit i.

From Equation (7), it follows that

$$\widetilde{x}_i(\omega) = \widetilde{x}_i(\omega^*) + \sum\nolimits_{j \in \mathcal{J}_i \setminus \mathcal{J}_i^1} \left[\omega_{ij}^* \psi_j + \delta_{ij} \left(z_j^* + \psi_j\right)\right] = \widetilde{x}_i(\omega^*) + \widetilde{\mu}_i \tag{11}$$

$$\breve{x}_i = \breve{x}_i^* + \sum\nolimits_{j \in \mathcal{J}^1} \psi_j = \breve{x}_i^* + \breve{\mu}_i, \tag{12}$$

where $\breve{\mu}_i = \widetilde{\mu}_i + \breve{\mu}_i$. Importantly, $\breve{x}_i$ is error-free unless $z_j \neq z_j^*$ for some $j \in \mathcal{J}_i^1$. Substituting Equations (9) - (12) into (1), the estimating equation is

$$y_i = \alpha + \widetilde{\beta}\widetilde{x}_i(\omega) + \beta\breve{x}_i + \left(-\widetilde{\beta}\widetilde{\mu}_i - \beta\breve{\mu}_i + \varepsilon_i\right), \tag{13}$$

where $\widetilde{\beta} = \beta$. However, one should not impose the restriction that $\widetilde{\beta} = \beta$ during estimation as this reduces the model to Equation (8) and the result in Proposition 1. This leads to the following result.

Proposition 2. *In addition to Assumptions 1-5, assume $\mathcal{J}_i^1$ is non-empty for some i. Let $\widehat{\beta}_{ols}$ denote the*

OLS estimate of β in the bivariate regression model in Equation (13). Then

$$\begin{aligned}
\texttt{plim}\,\widehat{\beta}_{ols} &= \frac{\texttt{Var}(\widetilde{x})\texttt{Cov}(\breve{x},y)-\texttt{Cov}(\breve{x},\widetilde{x})\texttt{Cov}(y,\widetilde{x})}{\texttt{Var}(\breve{x})\texttt{Var}(\widetilde{x})-\left[\texttt{Cov}(\breve{x},\widetilde{x})\right]^2} \\
&= \frac{\beta}{D}\left\{\begin{array}{l} D-\underbrace{\texttt{Var}(\breve{\mu})\texttt{Var}(\widetilde{x})}_{\neq 0} \\ \quad +\underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j^*,\sum_{j\in\mathcal{J}\setminus\mathcal{J}^1}\omega_j^* z_j^*\right)}_{?}\times \\ \qquad\left\{\texttt{Var}\left(\breve{\mu}\right)+\underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}\setminus\mathcal{J}^1}\omega_j^* z_j^*,\sum_{j\in\mathcal{J}\setminus\mathcal{J}^1}\delta_j z_j^*\right)}_{<0}\right\}\end{array}\right\}
\end{aligned}$$

where $D \coloneqq \texttt{Var}(\breve{x})\texttt{Var}(\widetilde{x})-\left[\texttt{Cov}(\breve{x},\widetilde{x})\right]^2$. If z^* is observed, then $\psi=0$ for all j, $\texttt{Var}(\breve{\mu})=0$, and $\texttt{Var}\left(\breve{\mu}\right)$ is altered.

Proof: See Appendix E.

Proposition 2 shows that $\widehat{\beta}_{ols}$ is consistent if z^* is observed and $\texttt{Cov}(z_j^*,z_{j'}^*)=0$ for all $j\neq j'$. Consistent estimation is possible because β is identified from variation in the part of $x(\omega)$ that is measured without weighting error.

As mentioned, one should not impose the restriction $\widetilde{\beta}=\beta$ during estimation. However, this suggests a *specification test* for the presence of crosswalk weighting errors. If z^* is observed, then $\widehat{\widetilde{\beta}}$ and $\widehat{\beta}$ are both consistent estimates of β under the null that $\omega_{ij}=\omega_{ij}^*$ for all ij. Thus, testing the null hypothesis $\texttt{H}_\texttt{o}:\widetilde{\beta}=\beta$ using a two-sided alternative constitutes a conservative specification test in that it may reject even when the crosswalk is correct if z^* is unobserved.

A final comment is warranted. I am assuming a homogeneous effect of $x(\omega^*)$, β. If this is not the case, then estimation of Equation (13) may suffer from the type of contamination bias discussed in Goldsmith-Pinkham *et al.* (forthcoming). A parametric solution of the type proposed by the authors may be applicable here as well. For brevity, I leave this for future exploration.

Instrumental Variables. A common solution to measurement error is Instrumental Variables (IV). I obtain the following result.

Proposition 3. *Let Assumptions 1-5 hold and let $\widehat{\beta}_{iv}$ denote the IV estimate of β in Equation (8). Then*

for a generic instrumental variable q

$$\begin{aligned}\texttt{plim}\,\widehat{\beta}_{iv} &= \frac{\texttt{Cov}\left[q,y\right]}{\texttt{Cov}\left[q,x(\omega)\right]} \\ &= \frac{\beta\texttt{Cov}\left[q,x(\omega^*)\right]+\texttt{Cov}\left(q,\varepsilon\right)}{\texttt{Cov}\left[q,x(\omega^*)\right]+\texttt{Cov}\left(q,\sum_j\omega_j^*\psi_j\right)+\texttt{Cov}\left(q,\sum_j\delta_j z_j^*\right)+\texttt{Cov}\left(q,\sum_j\delta_j\psi_j\right)}.\end{aligned}$$

Proof: See Appendix E.

This leads to the following remark.

Remark 4. *Under the conditions in Proposition 3, q must satisfy the following conditions*

(IV1) $\texttt{Cov}\left(q,\sum_j\omega_j^*z_j^*\right)\neq 0$

(IV2) $\texttt{Cov}\left(q,\sum_j\omega_j^*\psi_j\right)=0$

(IV3) $\texttt{Cov}\left(q,\sum_j\delta_j z_j^*\right)=0$

(IV4) $\texttt{Cov}\left(q,\sum_j\delta_j\psi_j\right)=0$

(IV5) $\texttt{Cov}\left(q,\varepsilon\right)=0.$

for $\widehat{\beta}_{iv}$ to be consistent.

With access to a single crosswalk, possible instruments include

(Q1) $q_i \coloneqq c_i$, where c is a $N\times 1$ vector and $\texttt{plim}\,\frac{1}{N}\left[c'x(\omega^*)\right]\neq 0$

(Q2) $q_i \coloneqq \sum_j\omega_{ij}b_j$, where b is a $M\times 1$ vector and $\texttt{plim}\,\frac{1}{M}\left(b'z^*\right)\neq 0$

(Q3) $q_i \coloneqq \sum_{j\in\mathcal{J}^1} z_j = \breve{x}_i$

(Q4) $q_i \coloneqq \sum_{j\in\mathcal{J}_i^1} b_j$

Instrument **Q1** is derived from an excluded covariate, c, that is at the same unit of observation as $x(\omega^*)$. Instrument **Q2** is derived from an excluded covariate, b, that is available at the same unit of observation as z. The same crosswalk is used to map b into a usable instrument. Instruments **Q3** and **Q4** are the parts of $x(\omega^*)$ and b, respectively, that are measured without weighting errors. The instrument is only available if $\mathcal{J}_i^1$ is non-empty for some i.

Add the following assumption.

Assumption 6 (Instruments)**.**

(i) $\texttt{Cov}\left(q,\sum_j\omega_j^*z_j^*\right)>0.$

(ii) $\texttt{Cov}(b, z^*) > 0$.

(iii) $\texttt{Cov}(b, \psi) = 0$.

Assumptions 6(i) and 6(ii) are without loss in generality. Assumption 6(iii) requires b to be independent of the measurement error in z^*. This leads to the following result.

Corollary 1. *Let Assumptions 1-6 hold. Instruments* ***Q1*** *–* ***Q3*** *are unlikely to satisfy the requirements in (IV1) – (IV5). Instrument* ***Q3*** *is a valid instrument if z^* is observed. Instrument* ***Q4*** *is a valid instrument.*

Proof: See Appendix E.

For **Q1** it is unlikely $\texttt{Cov}\left(c, \sum_j \delta_j z_j^*\right) = 0$ if **IV1** holds – since $\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^*\right) < 0$ – leading to violation of **IV3**. For **Q2** $\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right)$ will generally be positive, violating **IV3**. **Q3** and **Q4** are valid since the weights are known, however **Q3** also requires z^* to be observed.

5.2 Multiple Candidate Weighting Schemes

I also consider potential options when researchers possess a set $\mathcal{G}$ containing G candidate crosswalks. Denote the weighting scheme under crosswalk $g \in \mathcal{G}$ by ω_g and the weighting errors by δ_g.

Model Selection. Access to multiple crosswalks permits a second specification test. Following the logic in Kelejian (2008), the non-nested J-test of Davidson and MacKinnon (1981) can be used to select the 'true' weighting scheme among the set $\mathcal{G}$. The proposed test is given in Algorithm 1. Ideally, the test identifies a single weighting scheme, say g^*, among the candidates by failing to reject the null when $g = g^*$ and rejecting the null when $g \neq g^*$. If the candidate weighting schemes are highly correlated, then such a clean result is unlikely due to the resulting high multicollinearity between $x_i(\omega_g)$ and $x_i(\omega_{g'})$.

Algorithm 1 Non-Nested J-test

1: Choose a candidate weighting scheme $g \in \mathcal{G}$.

2: Estimate

$$y_i = \alpha_{g'} + \beta_{g'} x_i(\omega_{g'}) + \varepsilon_{i,g'}, \quad \forall g' \in \mathcal{G} \setminus g$$

3: Estimate the augmented model

$$y_i = \alpha_g + \beta_g x_i(\omega_g) + \sum\nolimits_{g' \in \mathcal{G} \setminus g} \zeta_{g'} \left[\widehat{\alpha}_{g'} + \widehat{\beta}_{g'} x_i(\omega_{g'}) \right] + \varepsilon_{i,g}$$

4: Test $\mathtt{H_o} : \zeta = 0$ against $\mathtt{H_a} : \zeta \neq 0$, where $\zeta = (\zeta_1, ..., \zeta_{g-1}, \zeta_{g+1}, ..., \zeta_G)$, using a standard Wald test.

5: Repeat Steps 1-4 allowing each of the G alternatives to serve as the null model.

Model Averaging. With multiple crosswalks available, several model averaging approaches are possible. These include

(MA1) Average weights: $\bar{\omega}_{ij}(\kappa) = \sum_g \kappa_g \omega_{ij,g}$, where $\kappa_g \in [0,1]\ \forall g$ and $\sum_g \kappa_g = 1$.

(MA2) Average covariates: $\bar{x}_i(\kappa) = \sum_g \kappa_g x_i(\omega_g)$, where $\kappa_g \in [0,1]\ \forall g$ and $\sum_g \kappa_g = 1$.

(MA3) Average coefficients: $\bar{\beta}(\kappa) = \sum_g \kappa_g \beta(\omega_g)$, where
$\beta(\omega_g) \coloneqq \mathtt{Cov}\left[x(\omega_g), y\right] / \mathtt{Var}\left[x(\omega_g)\right]$, $\kappa_g \in [0,1]\ \forall g$, and $\sum_g \kappa_g = 1$.

MA1 entails using a weighted average of the crosswalk weights, with weights given by κ_g, to map units j into units i, and then estimating a single regression model. The average crosswalk weights, $\bar{\omega}_{ij}(\kappa)$, will satisfy the properties in Remark 1. **MA2** entails using a weighted average of the covariates after using each crosswalk to map units j into units i and then estimating a single regression model. In this setup, linearity in Equation (2) implies that **MA1** and **MA2** are equivalent. **MA3** entails using each crosswalk to estimate a separate regression model and then taking a weighted average of the coefficient estimates. **MA3** differs from **MA1** and **MA2** due to the nonlinearity of $\beta(\omega_g)$.

Implementation requires the κ-weights to be chosen. For simplicity I use equal weights ($\kappa_g = 1/G$). For **MA2** I also follow Lubotsky and Wittenberg (2006) who shows that the sum of the coefficients on multiple proxies is identical to the optimal weighted average. Regardless, the averaging estimators are inconsistent but may reduce the bias as in Black *et al.* (2000).[13]

[13] Poirier and Ziebarth (2019) consider the estimator **MA2** in a different context with multiple proxies and show that OLS is consistent. The proof relies on three assumptions that do not hold here: (i) The first two moments of $x_i(\omega_g)$ and $x_i(\omega'_g)$ are identical $\forall\ g \neq g'$, (ii) $\mathtt{Cov}\left[x_i(\omega_g), x_i(\omega'_g)\right] = 0\ \forall\ g \neq g'$, and (iii) the true covariate is among the proxies being averaged. Requirements (ii) and (iii) are particularly problematic here.

Decomposition. If some of the true weights, ω_{ij}^*, equal one and and these are measured correctly in all crosswalks (Assumption 3(iii)), then the decomposition approach in Section 5.1 can be amended. Specifically, I augment (13) and estimate

$$y_i = \alpha + \sum\nolimits_g \widetilde{\beta}_g \widetilde{x}_i(\omega_g) + \beta \breve{x}_i + \upsilon_i + \varepsilon_i, \tag{14}$$

where υ_i is the composite error term. $\widehat{\beta}_{ols}$ will be consistent under the same conditions in Proposition 2. Moreover, testing the null hypothesis $\texttt{H}_\texttt{o} : \sum_g \widetilde{\beta} = \beta$ provides a specification test.[14]

Instrumental Variables. With multiple candidate weighting schemes, researchers might consider additional instrumental variables

(Q5) $q_i = \sum_j \omega_{g',ij} z_j$

(Q6) $q_i = \sum_j \omega_{g',ij} b_j$

where $\omega_{g'}$ is an alternative weighting scheme to the one used to construct the covariate, $x(\omega_g)$. I add the following assumption.

Assumption 6 (cont.) (Instruments)**.**

(iv) $\delta_{g,ij}$ *and* $\delta_{g',ij}$ *are independent conditional on* ω_{ij}^* *for all* $g, g' \in \mathcal{G}$*.*

Assumption 6(iv) requires the weighting errors to be uncorrelated across crosswalks conditional on ω^* and is identical to Assumption A1 in Black *et al.* (2000). This leads to the following result.

Corollary 1 (cont.). *Instruments* ***Q5*** *and* ***Q6*** *are unlikely to satisfy the requirements in (IV1) – (IV5). In fact, no potential instrument is likely to satisfy the requirements if* $\omega_{ij}^* < 1$ *for all* ij*.*

Proof: See Appendix E.

For **Q5** and **Q6**, $\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right) \neq 0$ violating **IV3**. The fact that any instrument must covary with the true covariate which depends on products of the true weights and z^*, but be independent of the product of the weighting errors and z^* despite the weighting errors being negatively correlated with the true weights, implies that finding a valid instrument is exceptionally difficult. This result has been been pointed out previously (e.g., Loewenstein and Spletzer, 1996; Black *et al.*, 2000). However, **Q3** and **Q4** are plausibly valid since they do not depend on mismeasured weights.

[14] Note, the test involves $\sum_g \widetilde{\beta}_g$ since one can derive (14) by substituting $\bar{x}_i(\kappa)$ into (8) and using the decomposition in (9) and (10), implying that $\widetilde{\beta}_g = \kappa_g \beta_g$ and $\sum_g \widetilde{\beta}_g = \beta$.

6 Simulations

6.1 Setup

To assess the performance of the various estimators and specification tests, as well as provide guidance on whether crosswalk errors are likely to be 'small', I consider the following DGP.

$$
\begin{aligned}
y_i &= \beta x_i\left(\omega^*\right) + \varepsilon_i & \omega_{ij}^* &= \frac{\texttt{exp}\left(c_{ij}^*\right)}{\sum_{i'} \texttt{exp}\left(c_{i'j}^*\right)} \\
x_i\left(\omega^*\right) &= \sum\nolimits_j \omega_{ij}^* z_j^* & \omega_{1,ij} &= \frac{\texttt{exp}\left(c_{ij}^* + \delta_{1,ij}\right)}{\sum_{i'} \texttt{exp}\left(c_{i'j}^* + \delta_{1,i'j}\right)} \\
x_i\left(\omega_1\right) &= \sum\nolimits_j \omega_{1,ij} z_j & \omega_{2,ij} &= \frac{\texttt{exp}\left(c_{ij}^* + \delta_{2,ij}\right)}{\sum_{i'} \texttt{exp}\left(c_{i'j}^* + \delta_{2,i'j}\right)} \\
q_{1i} &= \texttt{N}\left(x_i\left(\omega^*\right), 1\right) & c_{ij}^* &\sim \texttt{N}(0,1), \quad \forall j \in \mathcal{J}_i \\
q_{2i}\left(\omega_1\right) &= \sum\nolimits_j \omega_{1,ij} b_j^* & \delta_{1,ij}, \delta_{2,ij} &\sim \texttt{N}\left(0, \sigma^2\right) \\
q_{3i} &= \sum\nolimits_{j \in \mathcal{J}_i^1} z_j & z_j^*, b_j^* &\sim \texttt{N}_2\left(0, \begin{bmatrix} 10^2 & 0.75\left(10^2\right) \\ & 10^2 \end{bmatrix}\right) \\
q_{4i} &= \sum\nolimits_{j \in \mathcal{J}_i^1} b_j^* & z_j &= \texttt{N}\left(z_j^*, \sigma_\psi^2\right) \\
q_{5i}\left(\omega_2\right) &= \sum\nolimits_j \omega_{2,ij} z_j & \varepsilon &\sim \texttt{N}(0, 0.25) \\
q_{6i}\left(\omega_2\right) &= \sum\nolimits_j \omega_{2,ij} b_j
\end{aligned}
$$

where $i = 1, ..., N$, $j = 1, ..., M$, and the true value of β is one. $\mathcal{J}_i$ denotes the set of units j that are included in i. If $j \notin \mathcal{J}_i$, then $\omega_{1,ij} = \omega_{2,ij} = \omega_{ij}^* = 0$. If $j \in \mathcal{J}_i$, then $\omega_{1,ij},\ \omega_{2,ij},\ \omega_{ij}^* > 0$. If $j \in \mathcal{J}_i^1 \subset \mathcal{J}_i$, then j is entirely contained in i and $\omega_{1,ij} = \omega_{2,ij} = \omega_{ij}^* = 1$.

The true covariate is $x_i\left(\omega^*\right)$. The observed covariate, $x_i\left(\omega_1\right)$, is derived from first crosswalk, ω_1, and the observed underlying covariate, z. As ω_1 and ω_2 are generated identically (but based on different random draws), there is no gain to considering $x_i\left(\omega_2\right)$ as the observed covariate. Instead, ω_2 is used in the generation of the instrumental variables. The instruments, q_k, $k = 1, ..., 6$, correspond to **Q1**–**Q6** in Section 5. Note, all instruments are designed to have very strong first-stages; there is no weak instrument issue. This permits assessment of the performance of these instruments under ideal circumstances.

The sample size, N, is 100, while M is 398. The size of M is determined by the definition of $\mathcal{J}_i$. Here, $\mathcal{J}_i$ includes four to six units (i.e., $\#\mathcal{J}_i \in \{4, 5, 6\}$ for all i). Specifically, some units j span three units i, implying $\omega_{ij}^* \in (0, 1)$ for three values of i for each j. In addition, each unit i contains three units j that do

not span other units (i.e., $\#\mathcal{J}_i^1 = 3$ for all i). This is illustrated in Figure V. In this case, $\#\mathcal{J}_i = 4$ for $i = 1$ and N, $\#\mathcal{J}_i = 5$ for $i = 2$ and $N-1$, and $\#\mathcal{J}_i = 6$ for the remainder. This setup resembles situations such as mapping counties to congressional districts or one industrial classification to another.

The parameters that vary are σ^2 and σ_ψ^2. The parameter σ varies from 1 to 3 in increments of 0.25 and determines the extent of the weighting errors. When σ goes to zero, the researcher-provided weights, ω_1 and ω_2, converge to the true weights, ω^*. The parameter σ_ψ corresponds to dispersion of the classical measurement error in z_j. Values are chosen to fix the reliability ratio at values between 0.7 and one in increments of 0.1.

The attributes of the simulated data are shown in Table F.II for the case where z^* is observed. Importantly, the standard deviation of the weighting errors ranges from 0.018 to 0.033. This is at the *low end* of the differences across weighting schemes discussed in Section 2 and thus capture plausible deviations. The reliability ratio of $x(\omega)$ varies from 0.93 to 0.98. Appendix Table F.II also reports the median first-stage F-statistics in the IV regressions; the median first-stage F-statistic ranges from 83 to more than 2500.

In each simulated data set, I perform the following

1. OLS estimation of the true model in Equation (1) (`OLS` : $x(\omega^*)$)
2. OLS estimation of Equation (8) (`OLS` : $x(\omega)$)
3. OLS estimation of Equation (14) (`OLS` : `decomp`)
4. IV estimation of Equation (8) using a single instrument from **Q1** - **Q6** (`IV` : `instrument`)
5. OLS estimation of Equation (8) except including the covariate obtained from each weighting scheme and then summing the coefficients to obtain the optimal model averaging estimate (`OLS` : `MA2`)
6. Average the OLS estimates from Equation (8) and the same model except replacing the covariate with one derived from the alternative weighting scheme using equal weights as in **MA3** (`OLS` : `MA3`)

I compute the bias, absolute bias, and root mean squared error (RMSE) for the estimates of β, as well as the coverage rate and width of the Frisch bounds, and perform the two specification tests described in Sections 5.1 and 5.2.

6.2 Results

Select results are shown graphically in Figure VI; additional results are provided in Appendix F. Panels (A) and (B) plot the bias on identical scales with and without measurement error in z^*, respectively. Panels (C) and (D) are identical but plot the RMSE. For comparison, each panel provides the results obtained using the correct crosswalk, $x(\omega^*)$ (solid black line).

There are several striking findings. First, the J-test performs exceptionally well. The proportion of samples where the test *fails to reject* the null that ω^* is the correct weighting at the $p < 0.05$ level, but *rejects* the nulls that ω_1 or ω_2 is the correct weighting scheme at the $p < 0.05$ level, ranges from 0.946 to 0.951. Thus, the test is correctly sized. This also holds if z^* is measured with error; the probabilities range from 0.948 to 0.958 when the reliability ratio of z is 0.7. For the alternative specification test, the proportion of samples where $\texttt{H}_\texttt{o} : \sum_g \widetilde{\beta}_g = \beta$ is rejected at the $p < 0.05$ level using a two-sided alternative, which corresponds to the power of the test as the crosswalk is incorrect, varies from 0.542 to 0.989 depending on the severity of the weighting errors. However, the power declines as the measurement error in z^* worsens.

Second, the Frisch bounds include the true value in *at most* 75% of the simulations. The coverage rate of the Frisch bounds worsens as z^* is measured with error. Thus, the bounds are of limited practical use with even small errors in the crosswalk weights.

Third, the consequences of ignoring even small crosswalk errors are severe. When z^* is correctly observed, the bias of $\texttt{OLS} : x(\omega)$ (solid red line) ranges from roughly -0.03 to -0.1 as σ varies from 1 to 3 (Figure VI). A bias of -0.1 is meaningful in that the true value of β is 1 and $\sigma = 3$ is at the low end of variation across weighting schemes discussed in Section 2. Moreover, measurement error in z compounds the bias. The bias increases to nearly -0.4 when the reliability ratio is 0.7. $\texttt{OLS} : x(\omega)$ also fairs poorly in terms of RMSE (Figure VI). The ratio of the RMSEs of $\texttt{OLS} : x(\omega)$ to $\texttt{OLS} : x(\omega^*)$ varies from about 13 (when $\sigma = 1$) to 37 (when $\sigma = 3$) when z^* is observed. The ratios increase to roughly 114 (when $\sigma = 1$) and 131 (when $\sigma = 3$) when the reliability ratio of z is 0.7. This makes it clear that crosswalk errors should not be dismissed as 'small', particularly when multiple crosswalks are used.

Fourth, three estimators are essentially unbiased when z^* is observed: $\texttt{OLS} : \texttt{decomp}$, $\texttt{IV} : \texttt{Q3}$, and $\texttt{IV} : \texttt{Q4}$ (Figure VI). Each of these estimators identifies β solely from variation in $x(\omega^*)$ arising from units j where the weights are known to be one. However, the RMSE of these estimators is large relative to the case where the true crosswalk weights are known due to discarding some of the variation in $x(\omega^*)$ (Figure VI). The efficiency loss depends on the frequency of weights that are equal to one. The ratio of the RMSEs of each to $\texttt{OLS} : x(\omega^*)$ is lowest for $\texttt{OLS} : \texttt{decomp}$, ranging from six (when $\sigma = 1$) to ten (when $\sigma = 3$). The ratios

for `IV : Q3` (`IV : Q4`) vary from 11 to 14 (15 to 19). When z^* is mismeasured, `IV : Q4` continues to have a bias close to zero whereas `OLS : decomp` and `IV : Q3` do not since these are functions of z. The ratio of the RMSEs of `OLS : Q4` to `OLS` : $x(\omega^*)$ now varies from 39 (when $\sigma = 1$) to 45 (when $\sigma = 3$). Thus, the loss in efficiency is substantial.

Fifth, `IV : Q1`, `IV : Q5`, and `IV : Q6` have very small biases if z^* is observed. When z^* is not observed, the bias of `IV : Q5` increases dramatically; the others are largely unaffected as the instruments are not functions of z. In terms of RMSE, `IV : Q1` performs the best (Figure VI). In fact, across all estimators considered, it has the lowest RMSE when z^* is not correctly observed. Thus, there is merit in using a (strong) instrument that does not need to be crosswalked.[15]

Finally, the remaining estimators do not offer much upside. `OLS : MA3` performs identically to `OLS` : $x(\omega)$. `OLS : MA2` offers some improvement over `OLS` : $x(\omega)$ in terms of both bias and RMSE, analogous to the result in Black *et al.* (2000). Thus, averaging the covariate across multiple crosswalks is a simple improvement that researchers can make over current practice. `IV : Q2` performs nearly identically to `OLS` : $x(\omega)$ when z^* is observed since both use the same mismeasured crosswalk weights. However, when z^* is unobserved, the performance of `OLS` : $x(\omega)$ worsens, but the bias of `IV : Q2` is unchanged since the instrument does not depend on z. The RMSE does worsen due to the lower first-stage F-statistic, but much less than `OLS` : $x(\omega)$.

7 Application

Political polarization – a term that encompasses both affective (ill-will towards the political opposition) and ideological (divergence in political positions) polarization – is rising. Autor *et al.* (2020, p. 3140) states that "the ideological divide in American politics is at an historic high," while Callander and Carbajal (2022, p. 826) write that "political polarization is an important and enduring puzzle." Here, I assess the impact of social media on political polarization while addressing potential issues with mapping counties into congressional districts. Specifically, I build on Fujiwara *et al.* (2024) and Müller and Schwarz (2023) who use county-level data to assess the effect of Twitter (now 𝕏) on county-level electoral outcomes and hate crimes, respectively. Instead, I map county-level data on Twitter usage to congressional districts to explore its impact on the political ideology of US Congressional representatives. Thus, I am explicitly focusing on the *ideological* polarization of *elected* officials (Graham and Svolik, 2020; Kubin and von Sikorski, 2021).

The relationship between social media and political polarization is unsettled.[16] Kubin and von Sikorski

[15] Note, `IV : Q1` has the largest first-stage F-statistics (Table F.II)

[16] See Allcott *et al.* (2020) review the impact of social media on welfare.

(2021, p. 194) provide an extensive review, concluding that "the true effect of social media exposure on political polarization remains unclear." However, the fear is that social media creates echo chambers leading to greater polarization. Conover *et al.* (2021, p. 89) writes: "The concern is that when politically active individuals can avoid people and information they would not have chosen in advance, their opinions are likely to become increasingly extreme as a result of being exposed to more homogeneous viewpoints and fewer credible opposing opinions." Similarly, Levy (2021, p. 832) states: "As social media becomes a major news source, there are growing concerns that individuals are exposed to more pro-attitudinal news, defined as news matching their ideology, and as a result, polarization increases." Social media also provides direct contact between voters and politicians and Callander and Carbajal (2022, p. 861) shows theoretically that interactions beteen voters and politicians are a "necessary ingredient" for polarization.

To examine the connection between social media and political polarization, I estimate the following specification

$$y_{it} = \alpha + \beta \ln \mathbb{X}_i + W_{it}\gamma + \varepsilon_{it}, \tag{15}$$

where y_{it} is the ideology of the (elected) representative in district i during Congress t, $\mathbb{X}_i$ is a time invariant measure of Twitter usage, W_{it} is a vector of controls, and ε_{it} is a mean zero error term. The sample includes the 110^{th} (2007-8) through 116^{th} (2019-20) Congresses, and excludes Alaska, Hawaii, and Washington, D.C.

To measure ideology I use the DW-Nominate (Dynamic Weighted NOMINAl Three-step Estimation) scores available from UCLA Social Science Division's Voteview.[17] The scores vary from -1 to 1, with higher values indicating greater conservatism. Each politician receives two scores per two-year Congressional session, referred to as the first and second dimensions. Dimension 1 (2) represents economic (social) ideology. From this, I use three measures of ideology derived from each dimension: (i) raw score, (ii) absolute value of the score, and (iii) squared score. Using the raw score, the model assesses whether Twitter usage leads to representatives that are more liberal or conservative on average. Using the other measures, the model assesses whether Twitter usage leads representatives to hold more extreme views, regardless of whether those extreme views are liberal or conservative. As the scores are derived from roll call votes, the second and third outcomes are measures of *issue* polarization.

The controls in W come from Fulton and Dhima (2021) and include district-level measures of the share voting Democrat in the last presidential election, land area, median income, female population share, school-age population share, minority population share, share with at least a four-year college degree, unemployment rate, share in blue collar occupations, whether the district was redistricted since the last election,

[17] See https://www.voteview.com.

relative expenditures of the Democrat and Republican candidates, relative experience of the Democrat and Republican candidates, and whether the Democrat and Republican candidates are female. District fixed effects are excluded since districts change over time.

The variable of interest is Twitter usage within the district. While ideally one would have ground truth data originating at the congressional district level, such data does not exist to my knowledge. Instead, it must be constructed from data available at the county-level, corresponding to z in Section 4. The county-level variable is constructed in Fujiwara *et al.* (2024) and Müller and Schwarz (2023) after assigning 475 million geo-coded tweets to users and then counties. The original data were collected by Kinder-Kurlanda *et al.* (2017). The result is a measure of the number of individuals using Twitter in each county in 2015 based on 3.7 million users (or roughly 7% of Twitter users). Basing users on this limited sample implies that it is measured with some error; actual Twitter usage, z^*, is unobserved. The fact that it is time invariant is an additional source of error. Fujiwara *et al.* (2024) and Müller and Schwarz (2023) propose an instrument for Twitter usage based on the location of followers of the 2007 South by Southwest (SXSW) Festival. As articulated in these studies, SXSW is a plausibly exogenous shock to Twitter usage that affected some locations more than others. Specifically, the instrument is the *county-level* number of new followers of the SXSW account in March 2007, controlling for the number of followers of the SXSW account prior to March 2007. This corresponds to b in Section 5.

I map the Twitter variables to congressional districts using four crosswalks assigning weights based on land area, housing units in 2010, population in 2010, and population in 2000. Using each weighting scheme, three models are estimated. The estimates should be seen as the naïve results a researcher would obtain by selecting one crosswalk to use during data construction. The results are shown in Table II. The columns labelled "Actual" are estimated by OLS with $\ln(\mathbb{X})$ as the covariate of interest. The columns labelled "RF" are the reduced forms estimated by OLS with log new SXSW followers in March 2007 as the covariate of interest. Finally, the columns labelled "IV" are estimated by IV using new SXSW followers in March 2007 to instrument for $\ln(\mathbb{X})$.

Table III present the results from alternatives to the naïve approach. Columns (1) and (2) include the actual and RF covariates, respectively, from each of the four crosswalks in a single model and report the sum of the coefficients.[18] This is equivalent to the optimal `OLS : MA2` estimator. Columns (3) to (5) report the coefficient on the equally weighted average covariate, which corresponds to the equally weighted version of `OLS : MA2`. Finally, Columns (6) to (8) report the coefficients on the relevant covariate aggregated

[18]The IV results are missing because there are not instruments for all four covariates.

over counties that lie entirely within a single congressional district. This corresponds to the estimator `OLS : decomp` in Section 5.2. The "IV" specification uses log new SXSW followers in March 2007 aggregated over counties that lie entirely within a single congressional district to instrument for $\ln(\mathbb{X})$ aggregated over all counties using the 2010 population crosswalk. This corresponds to the estimator `IV : Q4`. Based on the simulation results, `OLS : decomp` and `IV : Q4` are the preferred estimators.

A few important results stand out. First, while not shown, I perform the non-nested J-test. The test is performed multiple times where each weighting scheme has a turn as the true weighting scheme under the null. With one of the weighting schemes as the null, twelve tests are performed corresponding to the "Actual" and "RF" models in Table II. When the weighting scheme under the null is land area, I reject the null at the $p < 0.05$ level in 10 of 12 cases. When the weighting scheme under the null uses the number of housing units, I reject the null in only two cases. When the weighting scheme under the null uses the population in 2000 (2010), I reject the null in six (four) cases. Thus, the population crosswalks are preferable to land area, with population measured in housing units having the strongest empirical support.

Second, the results in Table II indicate a positive, statistically significant effect on conservatism along Dimension 1 (Panel A), but not Dimension 2 (Panel B), across all crosswalks. However, the "Actual" and "IV" estimates in Panel A are about 40% smaller when using the land area crosswalk; the "IV" estimates using the population crosswalks yield elasticities at the sample mean of about 0.015. In contrast, the results in Table III suggest little effect of social media on the raw ideology score, particularly when using the decomposition approach. To the extent there is evidence of a positive effect on conservatism, it is stronger along Dimension 2 (Panel B). Thus, the naïve results are not robust once measurement error in the crosswalk is addressed.

Third, the results in Table II indicate a positive, statistically significant effect on both measures of polarization along Dimension 1 (Panel A) across all crosswalks and specifications except Column (1). Only the "IV" estimates produce a positive, statistically significant effect on polarization along Dimension 2 (Panel B). Again, the point estimates are considerably smaller when using the land crosswalk; the "IV" estimates using the population crosswalks yield elasticities at the sample mean below 0.02. In contrast, the results in Table III provide less robust evidence that social media increases polarization along Dimension 1 (Panel A), but stronger evidence that it increases polarization along Dimension 2 (Panel B). Specifically, the decomposition approach points to a small, positive effect on polarization along Dimension 2.[19] . Again, the conclusions are altered once measurement error in the crosswalk is addressed.

[19]The "IV" results are quite imprecise. This arises because there is relatively little variation in the instrument once it is restricted to counties entirely contained in a single congressional district.

In sum, using the decomposition estimators found to perform well in the simulations in Section 6, I find that social media has a small amplifying effect on polarization among Congressional representatives related to social issues. However, results obtained when ignoring the possibility of measurement error in the crosswalks point to effects that are not only larger in magnitude, but also apply to polarization related to economic issues.

8 Conclusion

The use of crosswalks to map data from one unit of observation to another is rapidly proliferating. The main contribution of this paper is to show the empirical relevance of the bias that results from using imperfect crosswalks; 'small' weighting errors matter. Moreover, traditional approaches to measurement error such as Frisch bounds, model averaging, and instrumental variables are unlikely to resolve the issue.

The second contribution of the paper is to provide two specification tests to help researchers evaluate the accuracy of crosswalks, as well as an econometric solution in cases where some units are correctly known to map into a single unit. This situation is very common, as some counties lie exclusively within congressional or school districts and some industries or occupations map to a single category under an alternative classification system. The specification tests and econometric solution perform very well in simulations.

The final contribution of the paper is to explore the impact of social media on political polarization in the US. Data on Twitter usage is only available at the county-level, whereas the political ideology of Congressional representatives is defined at the district level. The analysis points to a small amplifying effect of social media on polarization related to social, but not economic, issues. In addition, ignoring weighting errors in commonly used crosswalks is consequential; naïve results point to effects that are not only larger in magnitude, but also apply to polarization related to economic issues.

As the use of crosswalks is likely to only increase moving forward, there are many extensions that must be addressed in the future. Some of these have been mentioned here, such as allowing for heterogeneous coefficients, correlation between z^* and the weighting errors, and measurement error due to the use of a crosswalk to obtain the outcome, y. Additional topics include allowing for endogenous weights, ω^*, allowing for multiple covariates to be crosswalked where each has its own 'true' weighting scheme, and whether the weights can be estimated. In the interim, researchers need to be much more cognizant of the issues that arise when using crosswalks.

References

Abadie, A. and L'Hour, J. (2021). A penalized synthetic control estimator for disaggregated data. *Journal of the American Statistical Association*, **116** (536), 1817–1834.

Abramitzky, R. and Boustan, L. (2017). Immigration in american economic history. *Journal of Economic Literature*, **55** (4), 1311–1345.

Agarwal, S., Amromin, G., Ben-David, I. and Dinc, S. (2018). The politics of foreclosures. *The Journal of Finance*, **73** (6), 2677–2717.

Allcott, H., Braghieri, L., Eichmeyer, S. and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, **110** (3), 629–676.

Autor, D., Dorn, D., Hanson, G. and Majlesi, K. (2020). Importing political polarization? The electoral consequences of rising trade exposure. *American Economic Review*, **110** (10), 3139–3183.

Bhatia, R. and Davis, C. (2000). A better bound on the variance. *The American Mathematical Monthly*, **107** (4), 353–357.

Bisbee, J. and Zilinsky, J. (2023). Geographic boundaries and local economic conditions matter for views of the economy. *Political Analysis*, **31** (2), 288–294.

Black, D. A., Berger, M. C. and Scott, F. A. (2000). Bounding parameter estimates with nonclassical measurement error. *Journal of the American Statistical Association*, **95** (451), 739–748.

Blau, F. D., Brummund, P. and Liu, A. Y.-H. (2013). Trends in occupational segregation by gender 1970-2009: Adjusting for the impact of changes in the occupational coding system. *Demography*, **50** (2), 471–494.

Callander, S. and Carbajal, J. C. (2022). Cause and effect in political polarization: A dynamic analysis. *Journal of Political Economy*, **130** (4), 825–880.

Carlino, G. and Drautzburg, T. (2020). The role of startups for local labor markets. *Journal of Applied Econometrics*, **35** (6), 751–775.

Che, Y., Lu, Y., Pierce, J. R., Schott, P. K. and Tao, Z. (2022). Did trade liberalization with china influence US elections? *Journal of International Economics*, **139**, 103652.

Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F. and Flammini, A. (2021). Political polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, **5** (1), 89–96.

Davidson, R. and MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, **49** (3), 781–793.

— and — (2004). *Econometric Theory and Methods*. Oxford University Press.

Ferrara, A., Testa, P. and Zhou, L. (2022). *New Area- and Population-based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790-2020*. Tech. Rep. V4, Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], available at https://doi.org/10.3886/E150101V4.

Fujiwara, T., Müller, K. and Schwarz, C. (2024). The effect of social media on elections: Evidence from the united states. *Journal of the European Economic Association*, **22** (3), 1495–1539.

Fulton, S. A. and Dhima, K. (2021). The gendered politics of congressional elections. *Political Behavior*, **43**, 1611–1637.

Goldschmidt, D. and Schmieder, J. F. (2017). The rise of domestic outsourcing and the evolution of the german wage structure. *Quarterly Journal of Economics*, **132** (3), 1165–1217.

Goldsmith-Pinkham, P., Hull, P. and Kolesár, M. (forthcoming). Contamination bias in linear regressions. *American Economic Review*.

Graham, M. H. and Svolik, M. W. (2020). Democracy in America? partisanship, polarization, and the robustness of support for democracy in the united states. *American Political Science Review*, **114** (2), 392—409.

Herrera, M., Mur, J. and Ruiz, M. (2019). A comparison study on criteria to select the most adequate weighting matrix. *Entropy*, **21** (2).

Hornbeck, R. (2010). Barbed wire: Property rights and agricultural development. *Quarterly Journal of Economics*, **125** (2), 767–810.

Kelejian, H. H. (2008). A spatial J-test for model specification against a single or a set of non-nested alternatives. *Letters in Spatial and Resource Sciences*, **1** (1), 3–11.

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J. and Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, **4** (2), 2053951717736336.

Kubin, E. and von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, **45** (3), 188–206.

Levinson, A. (2015). A direct estimate of the technique effect: Changes in the pollution intensity of us manufacturing, 1990–2008. *Journal of the Association of Environmental and Resource Economists*, **2** (1), 43–56.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, **111** (3), 831–870.

Loewenstein, M. A. and Spletzer, J. R. (1996). *Belated Training: The Relationship Between Training, Tenure and Wages*. Tech. rep., Bureau of Labor Statistics Working Paper 296.

Lubotsky, D. and Wittenberg, M. (2006). Interpretation of regressions with multiple proxies. *Review of Economics and Statistics*, **88** (3), 549–562.

Müller, K. and Schwarz, C. (2023). From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, **15** (3), 270–312.

Pierce, J. R. and Schott, P. K. (2012). A concordance between ten-digit u.s. harmonized system codes and sic/naics product classes and industries. *Journal of Economic and Social Measurement*, **37** (1-2), 61–96.

Poirier, A. and Ziebarth, N. L. (2019). Estimation of models with multiple-valued explanatory variables. *Journal of Business & Economic Statistics*, **37** (4), 586–597.

Pouliot, G. A. (2023). Spatial econometrics for misaligned data. *Journal of Econometrics*, **232** (1), 168–190.

Ridder, G. and Moffitt, R. (2007). The econometrics of data combination. In J. J. Heckman and E. E. Leamer (eds.), *Handbook of Econometrics*, vol. 6, Elsevier, pp. 5469–5547.

Schaller, Z. and DeCelles, P. (2022). *Weighted Crosswalks for NAICS and SIC Industry Codes*. Tech. Rep. V1, Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], available at https://doi.org/10.3886/E145101V2.

Spoer, B. R., Chen, A. S., Lampe, T. M., Nelson, I. S., Vierse, A., Zazanis, N. V., Kim, B., Thorpe, L. E., Subramanian, S. V. and Gourevitch, M. N. (2023). Validation of a geospatial aggregation method for congressional districts and other us administrative geographies. *SSM - Population Health*, **24**, 101511.

Wright, J. K. (1936). A method of mapping densities of population: With cape cod as an example. *Geographical Review*, **26** (1), 103–110.

Yang, Y., Jia, F. and Li, H. (2023). Estimation of panel data models with mixed sampling frequencies. *Oxford Bulletin of Economics and Statistics*, **85** (3), 514–544.

TABLE I: Common Crosswalks

Category	Classification Systems	Example Crosswalks
Geography (US)	Census Tract, Census Block, Zip Code, County, Public Use Micro Areas, Commuting Zones, School Districts, Congressional Districts Health Service Areas	https://www.ddorn.net/data.htm https://www.huduser.gov/portal/datasets/usps_crosswalk.html https://www.census.gov/geographies/reference-files/2000/geo/relationship-files.html https://nces.ed.gov/programs/edge/geographic/relationshipfiles https://mcdc.missouri.edu/applications/geocorr.html https://cdmaps.polisci.ucla.edu/ https://data.dartmouthatlas.org/supplemental/#crosswalks Ferrara *et al.* (2022)
Industry	NAICS (2022), NAICS (2017), NAICS (2012), NAICS (2007), NAICS (2002), NAICS (1997), Standard Industrial Classification (SIC) Census Industrial Classification	https://www.bls.gov/ces/naics/ https://www.ddorn.net/data.htm https://www.nber.org/research/data/nber-ces-manufacturing-industry-database Schaller and DeCelles (2022)
Occupation	Standard Occupational Classification (SOC) 1950, SOC (1990), SOC (2000), SOC (2010), SOC (2018) Classification of Instructional Programs (CIP) American Community Survey (ACS) O*NET-SOC (2000), O*NET-SOC (2006), O*NET-SOC (2009), O*NET-SOC (2010), O*NET-SOC (2019) International Standard Classification of Occupations (ISCO) 1958, ISCO (1968), ISCO (1988), ISCO (2008)	https://www.bls.gov/emp/documentation/crosswalks.htm https://usa.ipums.org/usa/volii/occtooccsoc18.shtml https://www.ddorn.net/data.htm https://nces.ed.gov/ipeds/cipcode/post3.aspx?y=56 https://www.onetcenter.org/taxonomy.html
Merchandise Trade	Harmonized System (HS), North American Industry Classification System (NAICS), Standard International Trade Classification (SITC)	Pierce and Schott (2012)
Diseases	International Classification of Diseases (ICD), Revs 1-10 ICD-Clinical Modification (ICD-CM), Revs 1-10	https://seer.cancer.gov/tools/conversion/ https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings

TABLE II: Twitter Usage & Ideology of Congressional Representatives: Naïve Approach

	Land			Housing Units			Population (2010)			Population (2000)		
Dependent Variable	Actual (1)	RF (2)	IV (3)	Actual (4)	RF (5)	IV (6)	Actual (7)	RF (8)	IV (9)	Actual (10)	RF (11)	IV (12)
A. Ideology Score - Dimension 1												
`Ideology Score`	0.0173	0.0347	0.0850	0.0515	0.0377	0.1399	0.0474	0.0374	0.1392	0.0437	0.0370	0.1396
	(0.0085)	(0.0100)	(0.0245)	(0.0151)	(0.0105)	(0.0391)	(0.0149)	(0.0105)	(0.0391)	(0.0150)	(0.0105)	(0.0400)
`abs(Ideol Score)`	0.0023	0.0426	0.1042	0.0254	0.0488	0.1814	0.0243	0.0489	0.1820	0.0229	0.0492	0.1859
	(0.0057)	(0.0062)	(0.0169)	(0.0096)	(0.0065)	(0.0260)	(0.0095)	(0.0065)	(0.0262)	(0.0094)	(0.0065)	(0.0269)
$(\texttt{Ideol Score})^2$	0.0013	0.0398	0.0973	0.0248	0.0455	0.1692	0.0235	0.0457	0.1699	0.0222	0.0459	0.1734
	(0.0052)	(0.0057)	(0.0156)	(0.0084)	(0.0060)	(0.0241)	(0.0083)	(0.0060)	(0.0242)	(0.0082)	(0.0060)	(0.0249)
B. Ideology Score - Dimension 2												
`Ideology Score`	-0.0259	0.0142	0.0346	-0.0061	0.0170	0.0632	-0.0139	0.0161	0.0598	-0.0113	0.0150	0.0565
	(0.0098)	(0.0123)	(0.0303)	(0.0170)	(0.0128)	(0.0476)	(0.0169)	(0.0128)	(0.0477)	(0.0172)	(0.0128)	(0.0485)
`abs(Ideol Score)`	-0.0114	0.0142	0.1049	0.0119	0.0170	0.1784	0.0066	0.0161	0.1774	0.0083	0.0150	0.1804
	(0.0064)	(0.0123)	(0.0235)	(0.0121)	(0.0128)	(0.0364)	(0.0120)	(0.0128)	(0.0367)	(0.0122)	(0.0128)	(0.0374)
$(\texttt{Ideol Score})^2$	-0.0085	0.0142	0.0488	0.0006	0.0170	0.0850	-0.0026	0.0161	0.0842	-0.0023	0.0150	0.0862
	(0.0040)	(0.0123)	(0.0170)	(0.0080)	(0.0128)	(0.0280)	(0.0079)	(0.0128)	(0.0282)	(0.0081)	(0.0128)	(0.0288)

Notes: Column headers indicate crosswalk weighting scheme. RF = reduced form. IV = instrumental variable, where $ln(Twitter\ usage)$ is instrumented using the number of users who started following SXSW in March 2007. Ideology scores range from roughly -1 to 1, with higher values associated with more conservative positions. Robust standard errors in parentheses.

TABLE III: Twitter Usage & Ideology of Congressional Representatives: Model Averaging & Decomposition Approaches

	MA Optimal		MA Equal			Decomposition		
	Actual (1)	RF (2)	Actual (3)	RF (4)	IV (5)	Actual (6)	RF (7)	IV (8)
A. Ideology Score - Dimension 1								
`Ideology Score`	0.0458	0.0097	0.0399	0.0097	0.0450	-0.0003	-0.0002	-0.0184
	(0.0160)	(0.0075)	(0.0142)	(0.0072)	(0.0328)	(0.0002)	(0.0027)	(0.1514)
`abs(Ideol Score)`	0.0168	0.0319	0.0105	0.0281	0.1297	0.0000	0.0046	0.2357
	(0.0100)	(0.0045)	(0.0090)	(0.0043)	(0.0208)	(0.0001)	(0.0020)	(0.1438)
$(\texttt{Ideol Score})^2$	0.0175	0.0274	0.0097	0.0236	0.1091	-0.0000	0.0030	0.1507
	(0.0086)	(0.0039)	(0.0078)	(0.0037)	(0.0178)	(0.0001)	(0.0018)	(0.1157)
B. Ideology Score - Dimension 2								
`Ideology Score`	-0.0050	0.0215	-0.0075	0.0234	0.1080	0.0008	-0.0019	-0.0974
	(0.0187)	(0.0086)	(0.0164)	(0.0084)	(0.0395)	(0.0002)	(0.0058)	(0.3059)
`abs(Ideol Score)`	0.0199	0.0521	0.0191	0.0502	0.2322	0.0005	0.0140	0.7651
	(0.0132)	(0.0063)	(0.0114)	(0.0059)	(0.0298)	(0.0002)	(0.0047)	(0.4292)
$(\texttt{Ideol Score})^2$	0.0042	0.0268	0.0055	0.0260	0.1200	0.0003	0.0072	0.3960
	(0.0088)	(0.0047)	(0.0075)	(0.0043)	(0.0206)	(0.0001)	(0.0035)	(0.2788)

Notes: MA = moving average. MA Optimal reports the sum of the coefficients on the relevant covariates derived from each crosswalk. MA Equal reports the coefficient on a single covariate that is the average of $ln(Twitter\ usage)$ across the crosswalks. RF = reduced form. IV = instrumental variable, where $ln(Twitter\ usage)$ is instrumented using the number of users who started following SXSW in March 2007 aggregated over counties that lie entirely within a single congressional district. OLS and RF report the coefficients on $ln(Twitter\ usage)$ and the number of users who started following SXSW in March 2007 where each is aggregated over counties that lie entirely within a single congressional district. Actual and RF under the Decomposition approach report the coefficient on $ln(Twitter\ usage)$ aggregated over counties that lie entirely within a single congressional district. Ideology scores range from roughly -1 to 1, with higher values associated with more conservative positions. Robust standard errors in parentheses.

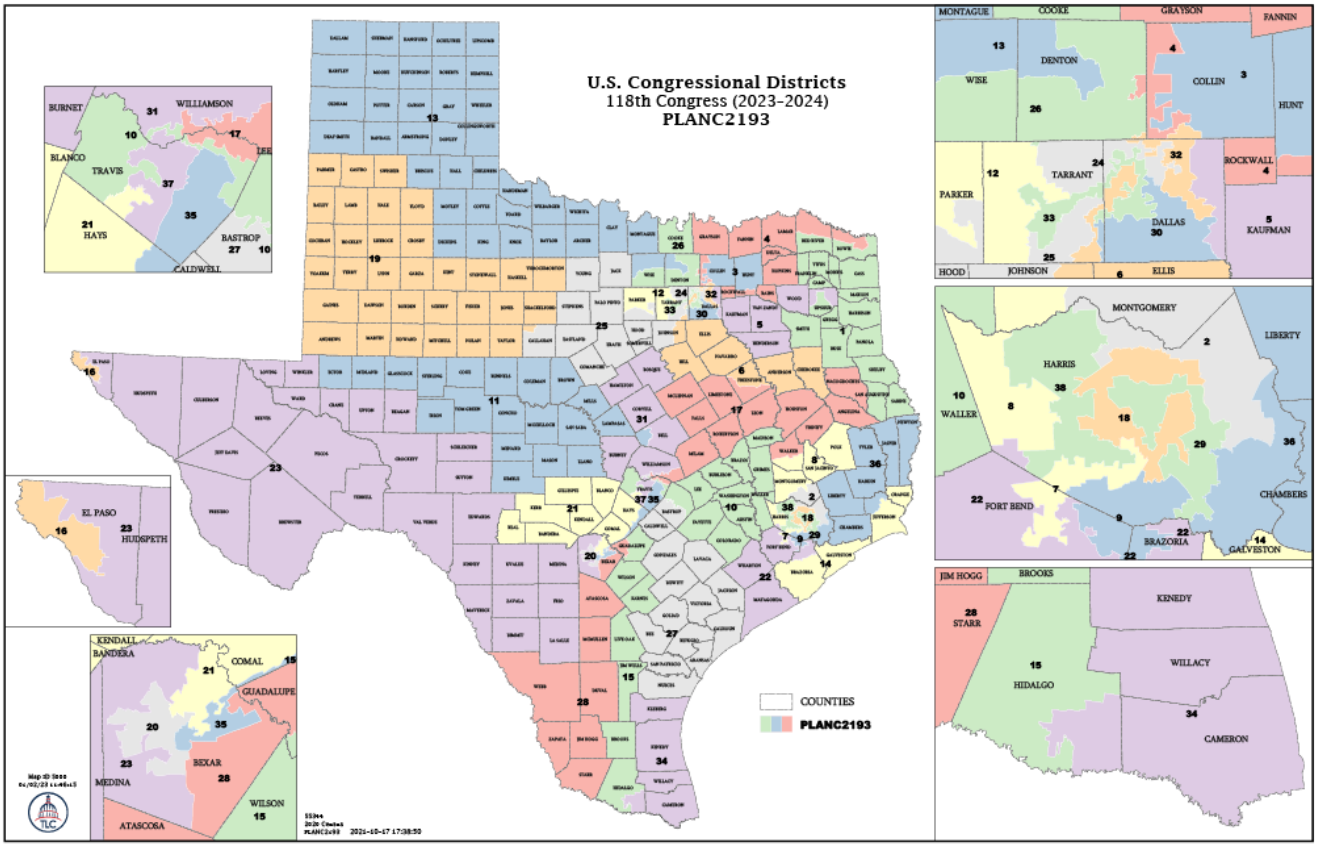


FIGURE I: Texas Congressional Districts in 2023-2024

Source: https://redistricting.capitol.texas.gov/Current-districts.

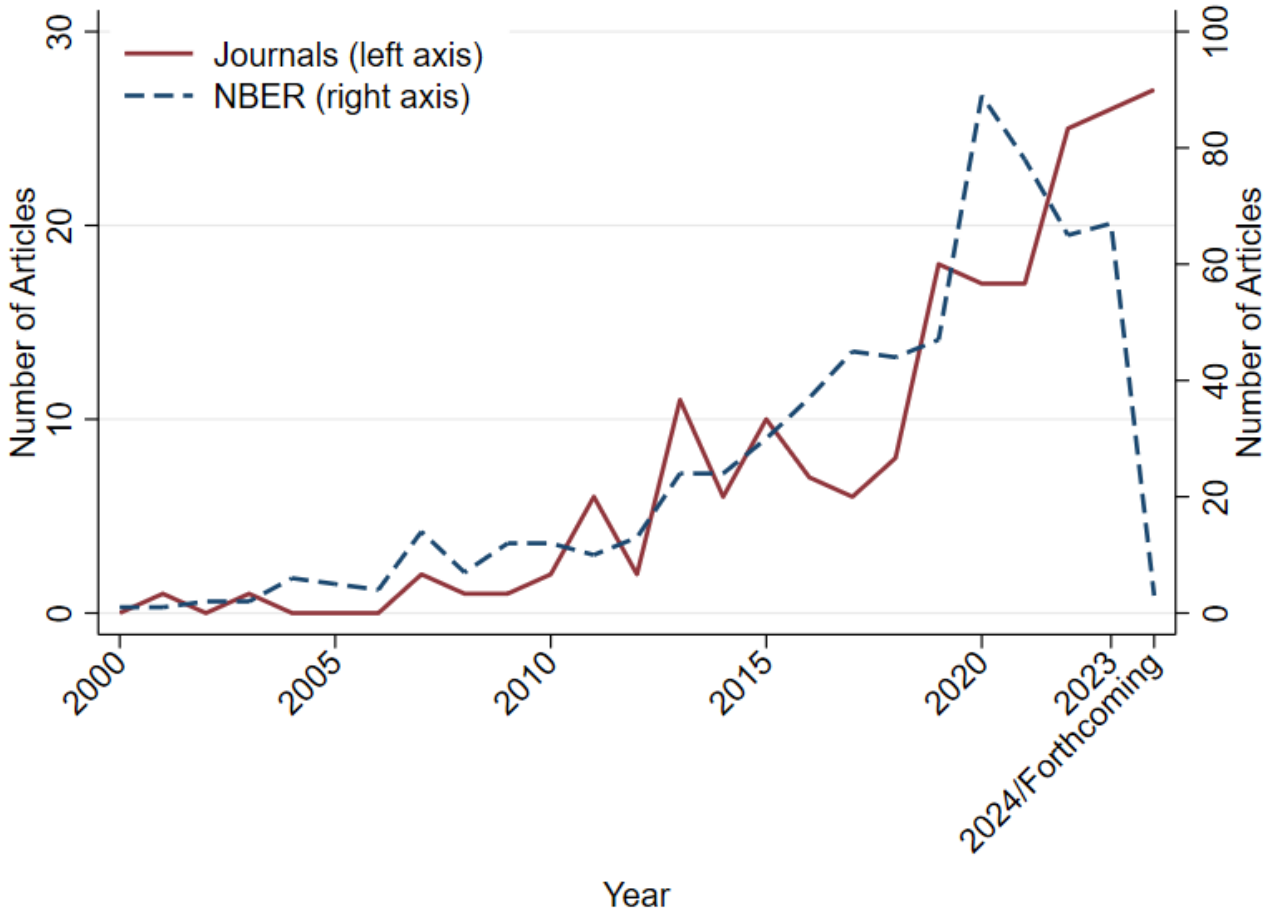


FIGURE II: Publications in Select Economics Journals Using Crosswalks

Notes: Journals included (from 2000-2023 unless otherwise noted): *AEJ: Applied* (2009-2023), *AEJ: Policy* (2009-2023), *American Economic Review*, *Econometrica*, *Journal of Applied Econometrics*, *Journal of Development Economics*, *Journal of Econometrics*, *Journal of International Economics*, *Journal of Labor Economics*, *Journal of Political Economy*, *Quantitative Economics*, *Quarterly Journal of Economics*, *Review of Economic Studies*, and *Review of Economics & Statistics*.

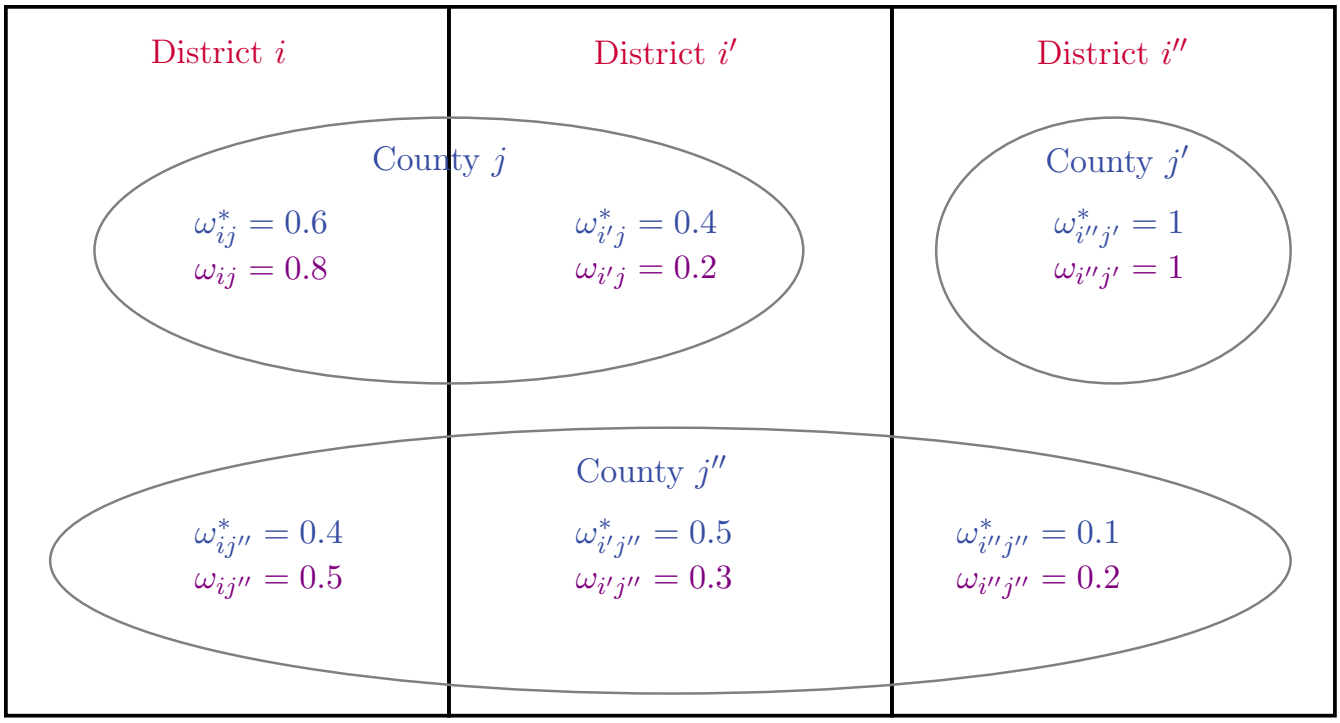


FIGURE III: Example of Mapping Counties to Congressional Districts.

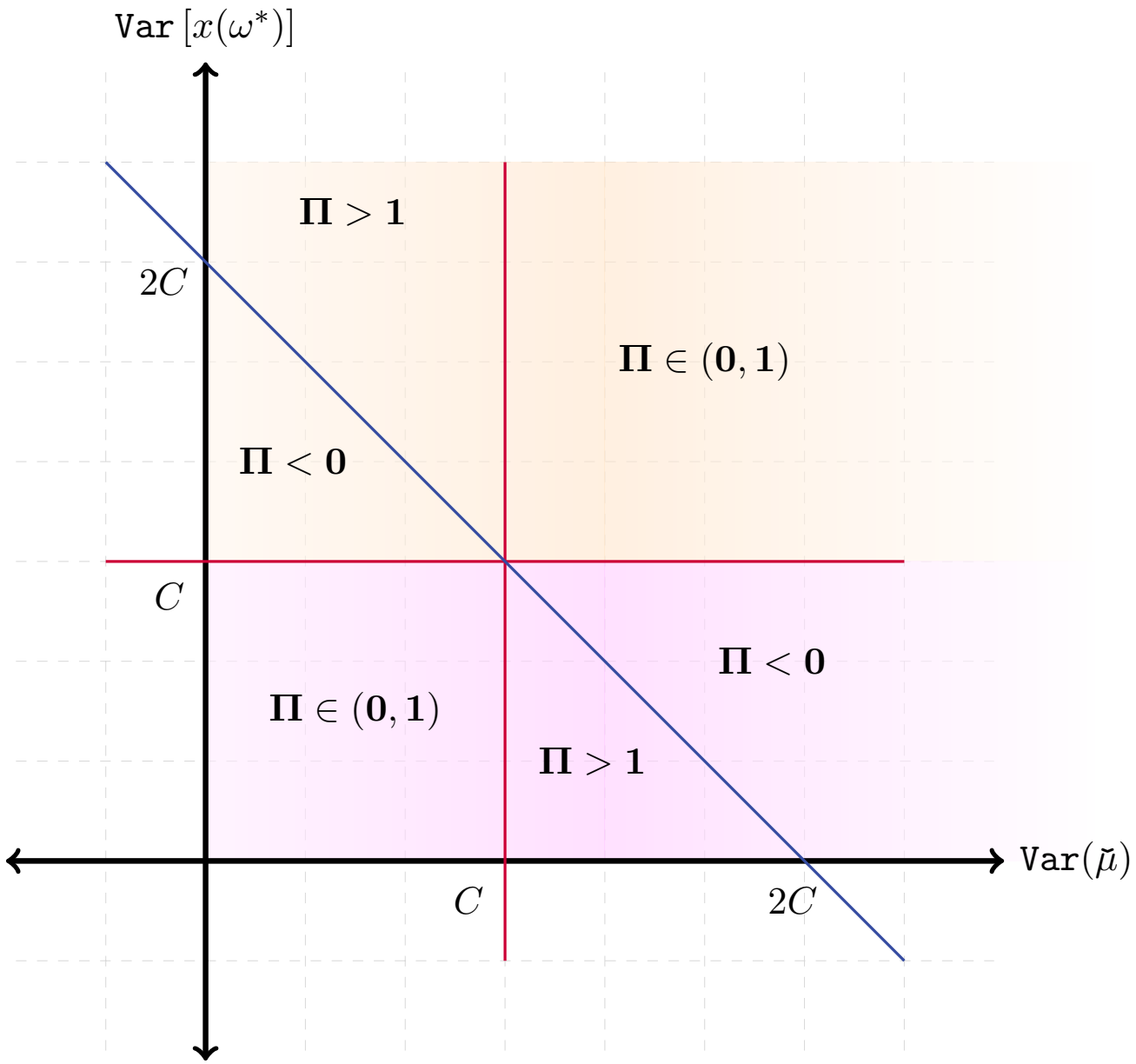


FIGURE IV: Magnitude and Direction of OLS Bias

Notes: Π is the proportional bias of OLS under weighting errors and measurement errors as defined in Proposition 1. $C = -\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^*\right)$. When z^* is observed, $\texttt{Var}\left[x(\omega^*)\right] + \texttt{Var}(\tilde{\mu}) < 2C$ is not feasible given that for two random variables, say P_1 and P_2, $2|\texttt{Cov}(P_1, P_2)| \leq \texttt{Var}(P_1) + \texttt{Var}(P_2)$. When z^* is unobserved, $\texttt{Var}\left[x(\omega^*)\right] + \texttt{Var}(\tilde{\mu}) < 2C$ is feasible under certain nonclassical errors. From Proposition E.1, the orange (magenta) region is where the reverse regression estimate suffers from expansion (attenuation) bias.

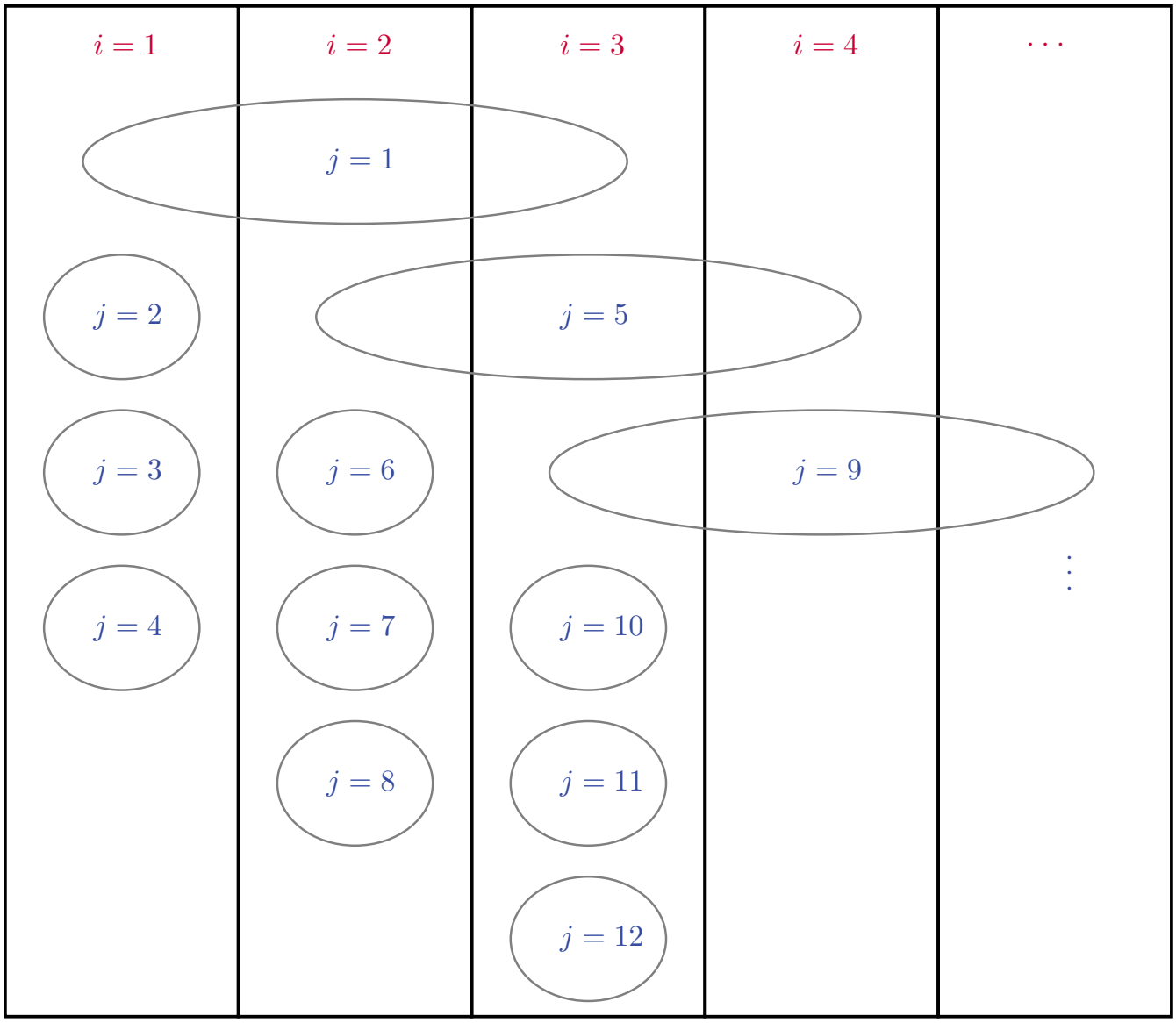
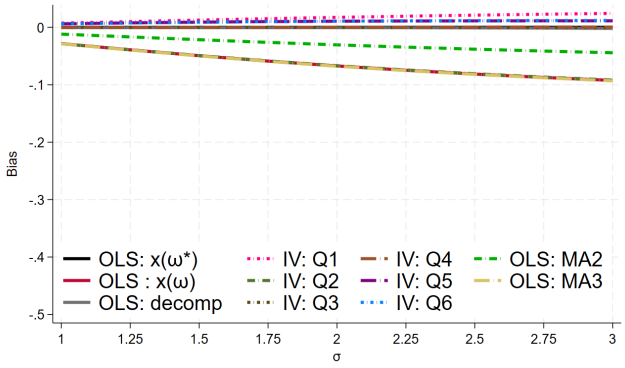
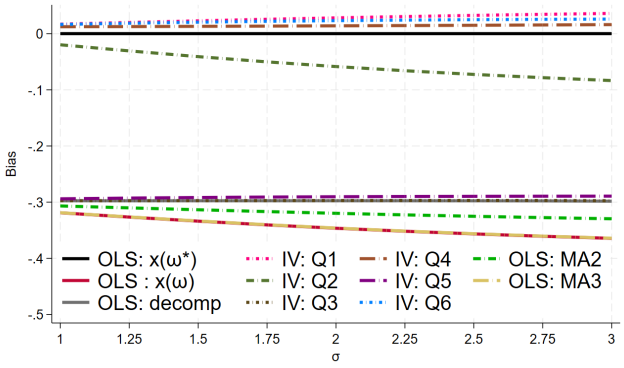


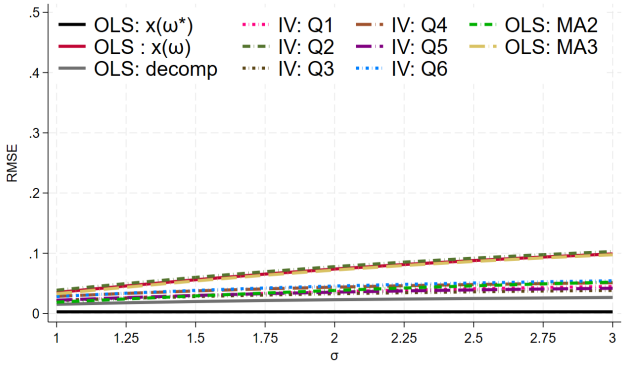
FIGURE V: Experimental Design



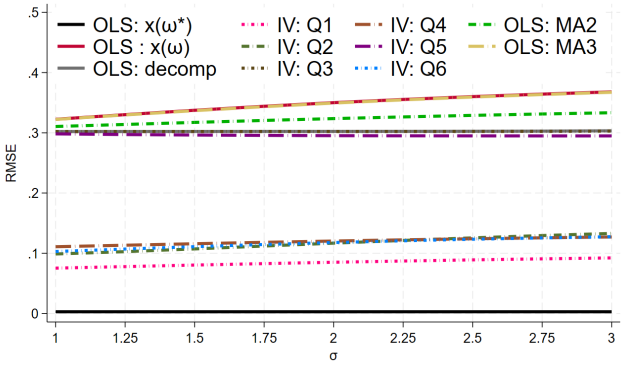
(A) Bias ($\texttt{RR} = 1$)



(B) Bias ($\texttt{RR} = 0.7$)



(C) RMSE ($\texttt{RR} = 1$)



(D) RMSE ($\texttt{RR} = 0.7$)

FIGURE VI: Simulation Results

Notes: $\texttt{RR}$ = reliability ratio and refers to the (classical) measurement error in z^*. See text for more details.

For Online Publication

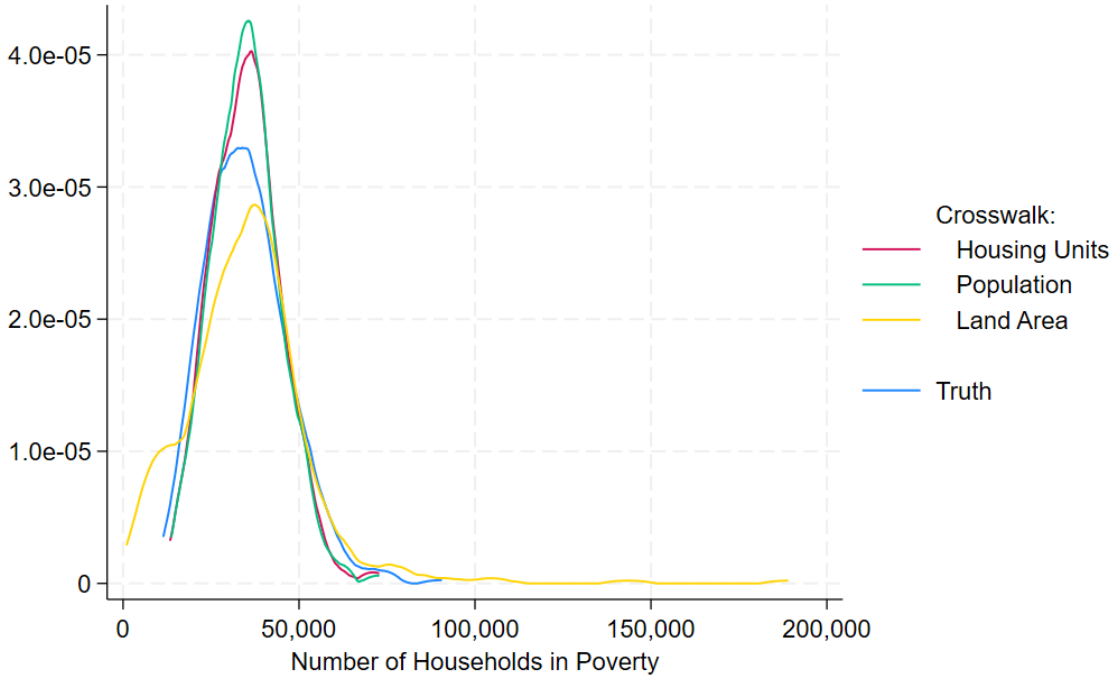
The Econometrics of Crosswalks

Supplemental Appendix

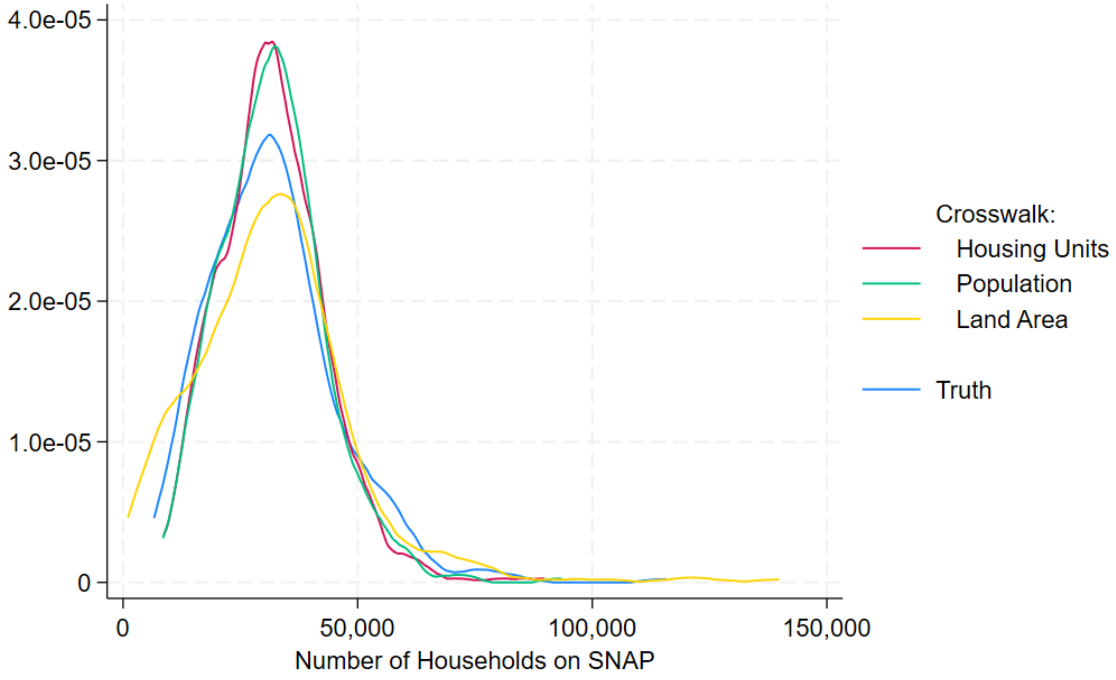
Daniel L. Millimet

July 16, 2024

A ACS Comparisons



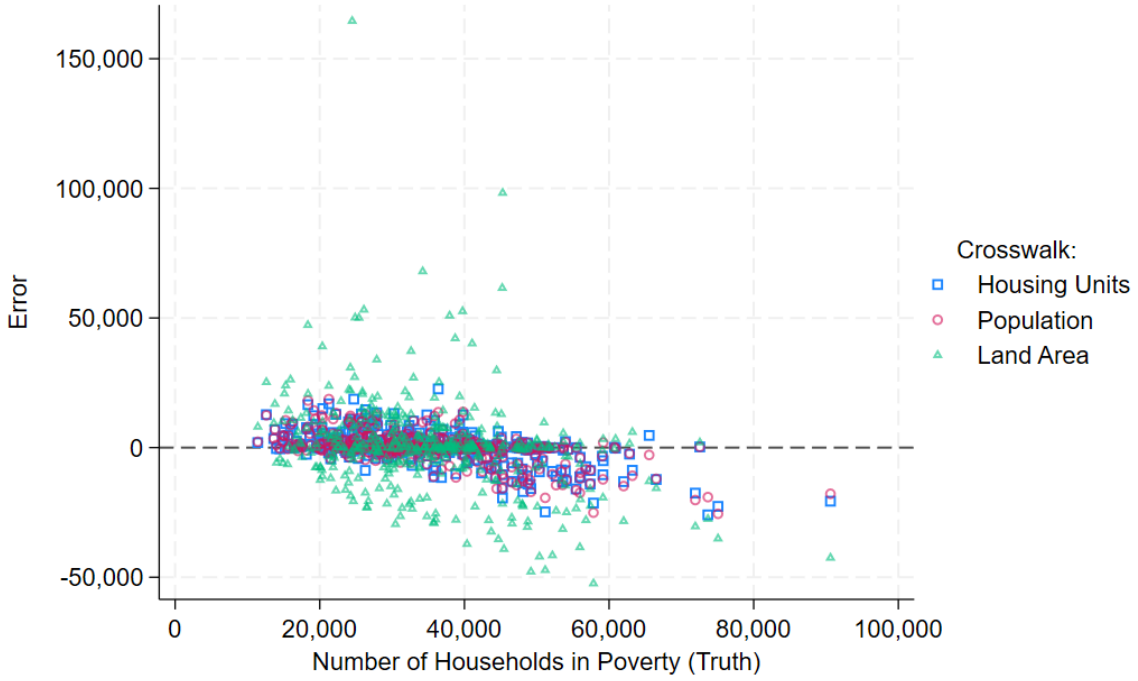
(A) Poverty



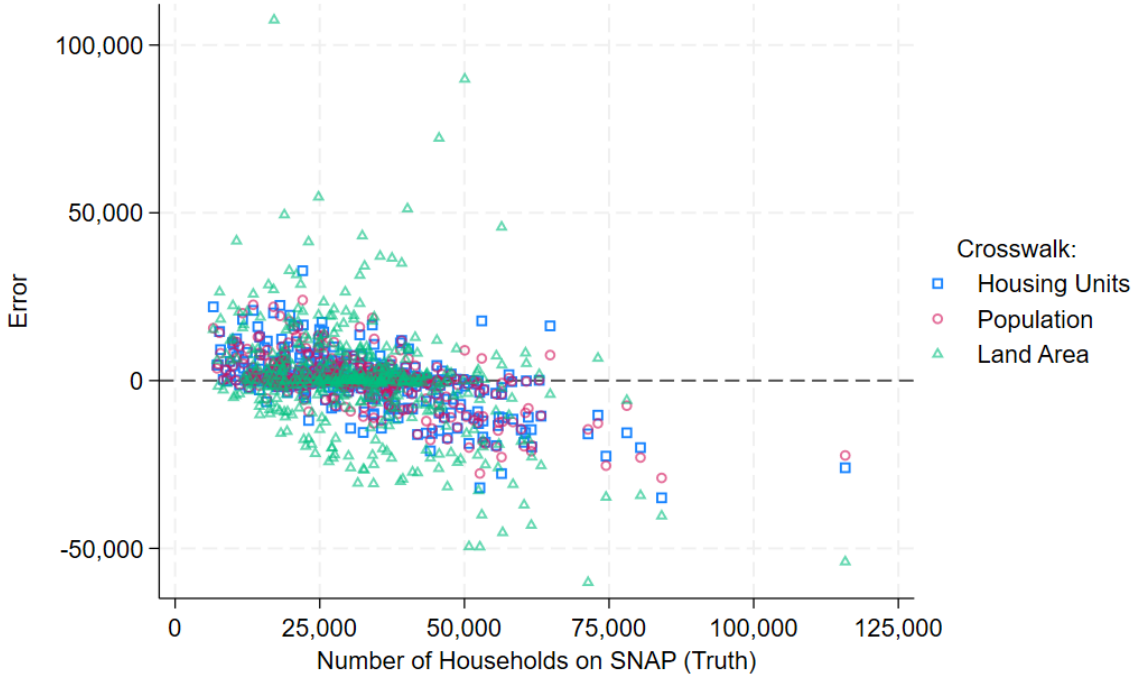
(B) SNAP

FIGURE A.I: Distribution of the Number of Households in Poverty and in SNAP Across Congressional Districts

Notes: Kernel density plots of the number of households in poverty and in SNAP across congressional districts. The crosswalked densities are obtained from county-level counts mapped into congressional districts using different weighting schemes. The true densities are directly obtained from the Census Bureau and based on aggregating census tracts which do not cross district borders. See text for more details.



(A) Poverty



(B) SNAP

FIGURE A.II: Crosswalk-Induced Errors Compared to the Truth Across Congressional Districts

Notes: Errors in the number of households in poverty and in SNAP across congressional districts due to the use of different crosswalk weighting schemes. The crosswalked counts (and, hence, errors) are obtained from county-level counts mapped into congressional districts using different weighting schemes. The true counts are directly obtained from the Census Bureau and based on aggregating census tracts which do not cross district borders. See text for more details.

The following simulation is conducted in Stata:

```
clear
set seed 248620
replace hh_pov=hh_pov/1000
replace xwhh_pov_h=xwhh_pov_h/1000
replace xwhh_pov_p=xwhh_pov_p/1000
replace xwhh_pov_l=xwhh_pov_l/1000
forval i=1/10 {
g x‘i’ = rnormal(hh_pov,5)
}
g y0 = hh_pov + rnormal(0,1)
g y1 = hh_pov + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + rnormal(0,1)
reg y0 hh_pov
reg y1 hh_pov x1-x10
reg y0 xwhh_pov_h
reg y1 xwhh_pov_h x1-x10
reg y0 xwhh_pov_p
reg y1 xwhh_pov_p x1-x10
reg y0 xwhh_pov_l
reg y1 xwhh_pov_l x1-x10
```

TABLE A.I: OLS Estimation of the Simulated Effect of the Number of Households in Poverty

	Simple Regression				Multiple Regression			
	Truth	Housing	Pop	Land	Truth	Housing	Pop	Land
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of Households	0.999***	1.035***	1.061***	0.216***	1.010***	0.064***	0.071***	0.009*
in Poverty	(0.004)	(0.030)	(0.030)	(0.029)	(0.032)	(0.017)	(0.017)	(0.005)
x_1					1.001***	1.107***	1.107***	1.109***
					(0.009)	(0.016)	(0.016)	(0.016)
x_2					0.989***	1.093***	1.093***	1.099***
					(0.009)	(0.016)	(0.016)	(0.016)
x_3					0.996***	1.089***	1.089***	1.099***
					(0.009)	(0.016)	(0.016)	(0.016)
x_4					0.981***	1.071***	1.069***	1.072***
					(0.009)	(0.016)	(0.016)	(0.016)
x_5					1.008***	1.077***	1.079***	1.088***
					(0.010)	(0.017)	(0.017)	(0.017)
x_6					0.999***	1.094***	1.094***	1.089***
					(0.010)	(0.017)	(0.017)	(0.017)
x_7					0.996***	1.097***	1.099***	1.101***
					(0.010)	(0.016)	(0.016)	(0.017)
x_8					1.008***	1.101***	1.100***	1.107***
					(0.010)	(0.017)	(0.016)	(0.017)
x_9					1.003***	1.104***	1.103***	1.105***
					(0.009)	(0.016)	(0.016)	(0.016)
x_{10}					1.003***	1.100***	1.096***	1.103***
					(0.010)	(0.017)	(0.017)	(0.017)
Constant	-0.037	-1.299	-2.224**	27.451***	0.141	0.015	-0.073	0.537*
	(0.149)	(1.089)	(1.094)	(1.166)	(0.151)	(0.324)	(0.329)	(0.288)
Observations	435	435	435	435	435	435	435	435

Column headers indicate true count or crosswalk weighting scheme used to derive the count. The true coefficient on all covariates is one; the true constant is zero. Standard errors in parentheses. * p <.10, ** p< .05, *** p<.01.

B Empirical Monte Carlo

As further motivation for the importance of thinking carefully about the role of crosswalks in statistical analyses, I revisit Che *et al.* (2022, hereafter CLPST). CLPST examine the impact of normalizing trade relations with China on US elections. To do so, they undertake the following steps:

Step 1. Compute the decrease in tariffs due to normalized relations in industry h, where industries are classified according to the Harmonized System

Step 2. Convert the decrease in tariffs in industry h to industry j, where j indexes four-digit SIC industries, using an industry crosswalk

Step 3. Convert the decrease in tariffs in industry j to county c using industry j employment shares in each county

Step 4. Convert all variables from county c to congressional district d using a geographic crosswalk

CLPST regress election outcomes using county-year level panel data after Step 3 and using district-year level panel data after Step 4. The (time invariant) geographic crosswalk used in Step 4 is based on 1990 population shares; the panel spans 1992 to 2008 or 2016. Thus, population shares in 1990 are used to allocate vote shares to districts roughly two decades later.

To be perfectly clear, the authors prefer the county-level analysis performed after Step 3 precisely to avoid the need to map counties to districts and deal with changes in districts over time. CLPST (p. 10) state:

> "One could construct district-level voting data that span a redistricting period using population-weighted averages of data for counties or county-district pairs. These weighted averages, however, may not accurately reflect votes in the redrawn districts if vote shares differ across portions of counties or county-district pairs that are split between multiple subsequent districts."

Later, the authors (p. 11) reiterate this concern, writing that "the accuracy of these district-level vote shares will depend on the extent to which county-level averages represent the portions of counties that map to different districts over time."

While the authors are careful to caveat the results, comparing the county- and district-level analyses is instructive. The authors note that the results are consistent in sign and statistical significance. Interestingly, however, the magnitude of the effect is much *larger* in the district-level analysis. In the county-level analysis,

experiencing a decline in local tariff exposure at the 75^{th} percentile instead of the 25^{th} under normalized relations led to a 2.2 percentage point (s.e. = 0.8) increase in the Democratic vote share in US House of Representative elections. In the district-level analysis, the corresponding effect is 7.0 percentage points (s.e. = 2.6). The threefold increase in the point estimate makes it clear that a superficial appeal to classical measurement error and attenuation bias is not useful.

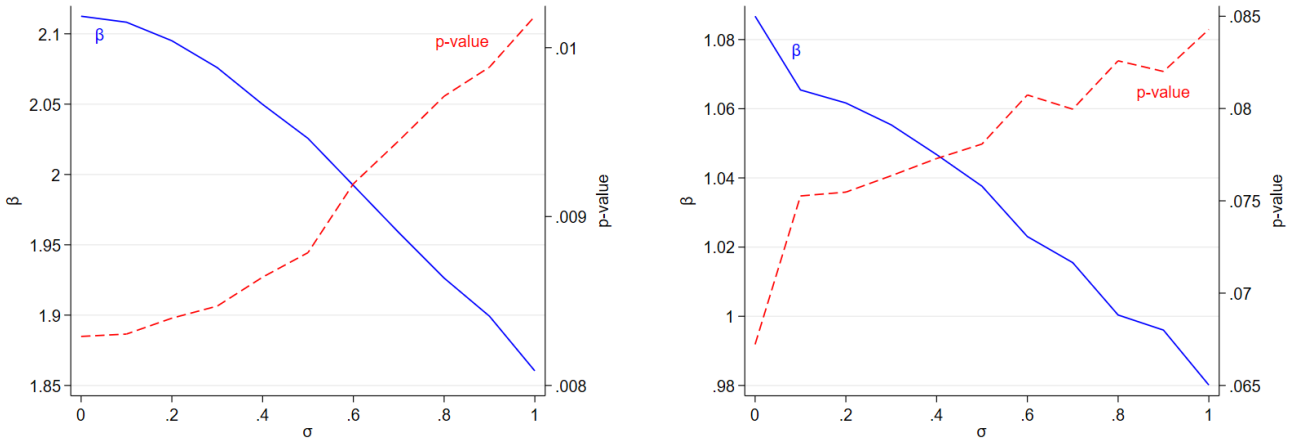
To further investigate the sensitivity of the results to the use of crosswalks, I use CLPST to perform a type of empirical Monte Carlo. Using their replication files[1], I do the following. Focusing solely on Step 4, I convert data from counties to districts using pseudo weights obtained by scaling the 1990 population shares used in CLPST by a random draw from a Gamma distribution with scale and shape parameters $\{1/\sigma^2, \sigma^2\}$ for different values of σ ranging from zero to one. These draws are non-negative, have a mean of one, and a variance of σ^2. I then normalize the pseudo weights such that they are between zero and one and sum to one.[2] I then re-estimate the district-level model from CLPST, retaining the estimates and p-values of the coefficient of interest.

The thought exercise here is the following. Suppose the geographic crosswalk used in CLPST is correct. If a researcher uses a crosswalk that deviates from this correct crosswalk, where σ controls the size of the deviations, how much will the estimates change? Note, however, that this is an extremely conservative exercise. First, I only create these pseudo weights for counties that span multiple districts; for counties entirely contained in a single district I continue to use the CLPST weights. Second, I only use the pseudo weights for the political variables; the demographic controls are fixed at the values in CLPST. Finally, I do not deviate from Steps 1-3 in CLPST, thus treating their industrial crosswalks as the truth. The focus here is solely on sensitivity to changes in the geographic crosswalk used in Step 4 for the covariate of interest.

For each draw of pseudo weights, I re-do the crosswalk from counties to districts and re-estimate the district-level model. I do this 1,000 times for each value of σ. Panel (A) in Figure B.III plots the median estimate of $\widehat{\beta}$ on the tariff exposure variable and the median p-value. Note, when σ is zero, the result is that reported in CLPST. This exercise shows that the magnitude of the estimate declines by roughly 12% as σ goes to one, although the estimates remain highly statistically significant.

[1] See https://www.dropbox.com/home/pntr_demovote/jie/replication/replication_clpsz_jie.

[2] When σ is 0.1 (1), the median correlation between the original and pseudo weights is 0.999 (0.918) and the standard deviation of the differences in the weights is 0.013 (0.129). This is within the range of differences across crosswalks discussed in Section 2.



(A) Population Weights

(B) Land Area Weights

FIGURE B.III: Sensitivity of Point Estimate to Alternative Weighting Schemes

Notes: Data from Che *et al.* (2022). σ is the standard deviation of random draws from a Gamma distribution with mean one used to create pseudo weights. Results are medians computed over 1,000 simulations. See text for more details.

As a second exercise, I create weights based on land area instead of population using the crosswalk from the MCDC.[3] I then perturb the land-based weights with similar draws from a Gamma distribution. The results are shown in Panel (B) in Figure B.III. The estimate of β using the true land weights (i.e., σ equal to zero) is roughly half the size as when using CLPST's population weights and the p-value exceeds 0.05. As σ increases, the median estimate declines by about 10% and the median p-value increases further.

To re-iterate, the point of this exercise is not to dispute the results in CLPST. The goal is simply to illustrate the sensitivity of the results to an alternative weighting scheme based on land shares rather than population shares, as well as plausible perturbations to the weights.

[3] See https://mcdc.missouri.edu/applications/geocorr.html.

C Application of the FWL Theorem

The model setup given in Equation (1) is a simple regression with x^* as the only covariate. It is alleged that the Frisch-Waugh-Lovell (FWL) theorem, which states that the estimates from a simple linear regression model after partialling out additional covariates are equal to the estimates obtained from a multivariate regression including all covariates (e.g., Davidson and MacKinnon, 2004). The fact that x is observed instead of x^* does not change this result under one additional, minor assumption.

Assume the correct data-generating process is given by the following multivariate regression

$$y_i = \alpha + \beta x_i^* + W_i\theta + \varepsilon_i, \quad i = 1, ..., N \tag{C.1}$$

where W_i is a $1 \times K$ vector of covariates and θ is a conformable vector of coefficients. The FWL theorem states that if x^* is observed, then the OLS estimates of $\{\alpha, \beta, \theta\}$ in Equation (C.1) can be identically obtained from the OLS estimates of

$$\widetilde{y}_i = \alpha + \beta \widetilde{x}_i^* + \varepsilon_i, \tag{C.2}$$

where $\widetilde{y}_i$ and $\widetilde{x}_i^*$ are the residuals from the linear projection of y_i and x_i^* on W_i, respectively.

When x is observed instead of x^*, where x is defined as in Equation (5), the multivariate model regression becomes

$$y_i = \alpha + \beta x_i + W_i\theta + (\varepsilon_i - \beta\mu_i)\,. \tag{C.3}$$

The residuals from the linear projection of y on W, $\widetilde{y}$, are unchanged. The linear projections of x^* and x on W produce the identical coefficients, $(W'W)^{-1}W'x^*$, assuming $W'\mu = 0$. The assumption that W is orthogonal to the measurement error, μ, is *minor* in the sense that it does not require W to be orthogonal to x^*, z^*, or the weighting errors, δ. It only requires W to be orthogonal to the product δz^*, where δ and z^* are assumed to be orthogonal in Assumption 4. Thus, the residuals from the linear projection of x_i on W_i are

$$\begin{aligned}
\widetilde{\widetilde{x}} &= x - W\left(W'W\right)^{-1}W'x \\
&= (x^* + \mu) - W\left(W'W\right)^{-1}W'(x^* + \mu) \\
&= x^* - W\left(W'W\right)^{-1}W'x^* + \mu \\
&= \widetilde{x}^* + \mu
\end{aligned} \tag{C.4}$$

if $W'\mu = 0$. As such, Equation (C.3) is identical to the model in Equation (1).

This can be verified in simulation in Stata:

```
clear
set obs 10000
g xs=rnormal(0,2)
g mu=rnormal()
g x=xs+mu
g w=rnormal(mu,1)
g y=1+xs+w+rnormal()
reg y xs w
reg y x w
qui reg y w
predict yt, res
qui reg xs w
predict xt, res
qui reg x w
predict xtt, res
reg yt xt, nocons
reg yt xtt, nocons
```

The estimate from `reg yt xtt, nocons` is identical to that on x in `reg y x w`. The residuals from the two regressions are also identical. While note shown, it also be verified by partialling out $x_1, ..., x_{10}$ in the simulation in Appendix A.

D IID Assumption

Even if ω_{ij}^* is known for all ij, analysis of the Ordinary Least Squares (OLS) estimates of Equation (1) is nonstandard as x^* and z^* cannot both be independent and identically distributed (iid). From Equation (2) it follows that if z^* is iid, then x^* will be heteroskedastic and cross-sectionally dependent if some units j span multiple units i. Alternatively, due to cross-sectional dependence (and other reasons), it is not realistic to assume x^* is iid. To simplify matters, I assume that x^* and z^* are themselves a function of an underlying iid random variable, W, which is unobserved. The unit of observation for W, say $\ell = 1, ..., \mathcal{L}$, is a unique combination ij, $\ell = (i, j)$. The population DGP is then defined over units ℓ. For example, let ℓ index areas of overlap between county i and congressional district j. Units ℓ are non-overlapping. This is analogous to the empirical strategy used in Autor *et al.* (2020).

With this, z_j^* and x_i^* are equivalent to

$$z_j^* = \sum\nolimits_{\ell \in \mathcal{L}_j} W_\ell \tag{D.5}$$

$$x_i^* = \sum\nolimits_j \omega_{ij}^* z_j^* = \sum\nolimits_{\ell \in \mathcal{L}_i} \omega_\ell^* W_\ell, \tag{D.6}$$

where $\mathcal{L}_j \coloneqq \left\{\ell : \omega_\ell^* = \omega_{ij}^* > 0\right\}$ and $\mathcal{L}_i$ is defined analogously.[4] $\mathcal{L}_j$ ($\mathcal{L}_i$) is the set of $\ell = (i, j)$ such that $\omega_{ij}^* > 0$ for a given unit j (i). By construction, the intersection of $\mathcal{L}_j$ and $\mathcal{L}_{j'}$ ($\mathcal{L}_i$ and $\mathcal{L}_{i'}$) is empty for all $j \neq j'$ ($i \neq i'$).[5]

Assumption 1 can be replaced with the following.

Assumption 1′.

(i) *The population model is* $y_i = \alpha + \beta x_i^* + \varepsilon_i$ *for all* i.

(ii) $x_i^* \coloneqq x_i(\omega^*) = \sum_j \omega_\ell^* z_j^* = \sum_{\ell \in \mathcal{L}_j} \omega_{ij}^* w_\ell$, *where* $\omega^* = (\omega_{11}^*, ..., \omega_{1M}^*, \omega_{21}^*, ..., \omega_{2M}^*, ..., \omega_{1M}^*, ..., \omega_{NM}^*)$ *is the true weighting scheme.*

(iii) $\{W_\ell\}$ *is independently and identically distributed (i.i.d.) across* $\ell = 1, ..., \mathcal{L}$ *with finite first and second moments.*

(iv) $X^* \coloneqq [\iota \; x_i^*]$ *is an* $N \times 2$ *matrix of full rank where* ι *is an* $N \times 1$ *column vector of ones and* $x(\omega^*)$ *is an* $N \times 1$ *column vector with representative element* $x_i(\omega^*)$.

[4] Because each w_ℓ maps to a unique ij combination, the values and units of W can be defined such that Equation (D.5) holds. In other words, I do not need to express z^* as a weighted sum of W.

[5] The intersection between $\mathcal{L}_j$ and $\mathcal{L}_i$ will not be empty for all i, j.

(v) $\texttt{plim}\ \frac{1}{\mathcal{L}}\left[X^{*\prime}X^{*}\right] = Q$, *where* Q *is a positive definite matrix.*

(vi) $\texttt{plim}\ \frac{1}{\mathcal{L}}\left[X^{*\prime}\varepsilon\right] = 0$.

(vii) $\texttt{plim}\ \frac{1}{\mathcal{L}}\#\mathcal{L}_j = q_1$, *where* $q_1 < \infty$.

(viii) $\texttt{plim}\ \frac{1}{\mathcal{L}}\sum_{\ell\in\mathcal{L}_i}\omega_\ell^* = q_2$, *where* $q_2 < \infty$.

Assumption 1′(ii)-(iii) imply that $\texttt{E}\left[z_j^*\right] = \#\mathcal{L}_j \times \overline{W}$ and $\texttt{Var}\left(z_j^*\right) = \#\mathcal{L}_j \times \sigma_W^2$ where $\overline{W}$ and σ_W^2 are the mean and variance of W, respectively. In addition, z_j^* is iid if $j \subseteq i$ implying that a single W_ℓ maps to z_j^*. It is also iid conditional on $M_j \coloneqq \#\mathcal{L}_j$. For x^*, $\texttt{E}\left[x_i^*\right] = \overline{W}\sum_{\ell\in\mathcal{L}_i}\omega_\ell^*$ and $\texttt{Var}\left(x_i^*\right) = \sigma_W^2\sum_{\ell\in\mathcal{L}_i}\left(\omega_\ell^*\right)^2$. Assumption 1′(vii)-(viii) exploit the fact that the exact nature of the non-iid-ness is known and such that it is reasonable to assume that x_i^* and z_j^* are iid as $\mathcal{L}\to\infty$. Moreover, the asymptotic bias of an estimator of Equation (1) will be unaffected by the fact that x_i^* and $x_{i'}^*$ contain the same z_j^* for some $i \neq i'$ since the number of independent observations still increases with $\mathcal{L}$.[6]

[6] A proof in a similar context is given in Poirier and Ziebarth (2019).

E Proofs

Proof of Proposition 1. Note, this result is identical to Proposition 1 in Black *et al.* (2000) except for a particular structure on the covariance between the true covariate and the measurement error.

$$
\begin{aligned}
\texttt{plim}\,\widehat{\beta}_{ols} &= \frac{\texttt{Cov}\left[x(\omega), y\right]}{\texttt{Var}\left[x(\omega)\right]} && \text{(E.7)}\\
&= \frac{\texttt{Cov}\left[x(\omega), \alpha + \beta x(\omega^*) + \varepsilon\right]}{\texttt{Var}\left[x(\omega)\right]} && \text{(E.8)}\\
&= \beta\frac{\texttt{Cov}\left[x(\omega), x(\omega^*)\right]}{\texttt{Var}\left[x(\omega)\right]} && \text{(E.9)}\\
&= \beta\frac{\texttt{Cov}\left[x(\omega^*) + \check{\mu}, x(\omega^*)\right]}{\texttt{Var}\left[x(\omega^*) + \check{\mu}\right]} && \text{(E.10)}\\
&= \beta\frac{\texttt{Var}\left[x(\omega^*)\right] + \texttt{Cov}\left[x(\omega^*), \check{\mu}\right]}{\texttt{Var}\left[x(\omega^*) + \check{\mu}\right]} && \text{(E.11)}\\
&= \beta\frac{\texttt{Var}\left[x(\omega^*)\right] + \texttt{Cov}\left\{\sum_j \omega_j^* z_j^*, \sum_j \left[\omega_j^*\psi_j + \delta_j(z_j^* + \psi_j)\right]\right\}}{\texttt{Var}\left[x(\omega^*)\right] + \texttt{Var}\left(\check{\mu}\right) + 2\texttt{Cov}\left[x(\omega^*), \check{\mu}\right]} && \text{(E.12)}\\
&= \beta\frac{\texttt{Var}\left[x(\omega^*)\right] + \overbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \omega_j^*\psi_j\right)}^{=0} + \overbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^*\right)}^{<0} + \overbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j\psi_j\right)}^{=0}}{\texttt{Var}\left[x(\omega^*)\right] + \texttt{Var}\left(\check{\mu}\right) + 2\underbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^*\right)}_{<0}} && \text{(E.13)}\\
&= \beta\left\{\frac{\texttt{Var}\left[x(\omega^*)\right] + \overbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^*\right)}^{<0}}{\texttt{Var}\left[x(\omega^*)\right] + \texttt{Var}\left(\check{\mu}\right) + 2\underbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^*\right)}_{<0}}\right\} && \text{(E.14)}
\end{aligned}
$$

which leads to the usual attenuation bias if $\texttt{Cov}\left[x(\omega^*), \check{\mu}\right] < \texttt{min}\left\{\texttt{Var}\left[x(\omega^*)\right], \texttt{Var}\left(\check{\mu}\right)\right\}$.

Frisch Bounds. Consider the reverse regression counterpart to Equation (8) given by

$$x_i(\omega) = -\frac{\alpha}{\beta} + \frac{1}{\beta} y_i + \left(\breve{\mu}_i - \frac{\varepsilon_i}{\beta} \right) \tag{E.15}$$

This leads to the following result.

Proposition E.1. *Let Assumptions 1-5 hold and let $\widehat{\gamma}_{ols}$ denote the OLS estimate of the coefficient on y_i in Equation (E.15). Then*

$$\texttt{plim}\, \widehat{\gamma}_{ols}^{-1} = \frac{\texttt{Var}\,(y)}{\texttt{Cov}\,[x(\omega), y]} = \frac{\beta \texttt{Var}\,[x(\omega^*)]}{\texttt{Var}\,[x(\omega^*)] + \underbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^* \right)}_{<0}} + \frac{\texttt{Var}\,(\varepsilon)}{\beta \left\{ \texttt{Var}\,[x(\omega^*)] + \underbrace{\texttt{Cov}\left(\sum_j \omega_j^* z_j^*, \sum_j \delta_j z_j^* \right)}_{<0} \right\}}.$$

The plim *is unchanged if z^* is observed.*

Proof. Note, this result is identical to Proposition 2 in Black *et al.* (2000) except for a particular structure

on the covariance between the true covariate and the measurement error.

$$
\begin{aligned}
\texttt{plim}\,\widehat{\gamma}_{ols}^{-1} &= \frac{\texttt{Var}\,(y)}{\texttt{Cov}\,[x(\omega),y]} && \text{(E.16)}\\
&= \frac{\texttt{Var}\,(y)}{\texttt{Cov}\,[x(\omega),\alpha+\beta x(\omega^*)+\varepsilon]} && \text{(E.17)}\\
&= \frac{\texttt{Var}\,(y)}{\beta\texttt{Cov}\,[x(\omega),x(\omega^*)]} && \text{(E.18)}\\
&= \frac{\beta^2\texttt{Var}\,[x(\omega^*)]+\texttt{Var}\,(\varepsilon)}{\beta\texttt{Cov}\,[x(\omega^*)+\check{\mu},x(\omega^*)]} && \text{(E.19)}\\
&= \frac{\beta^2\texttt{Var}\,[x(\omega^*)]+\texttt{Var}\,(\varepsilon)}{\beta\left\{\texttt{Var}\,[x(\omega^*)]+\underbrace{\texttt{Cov}\left(\sum_j\omega_j^*z_j^*,\sum_j\delta_jz_j^*\right)}_{<0}\right\}} && \text{(E.20)}\\
&= \frac{\beta\texttt{Var}\,[x(\omega^*)]}{\texttt{Var}\,[x(\omega^*)]+\underbrace{\texttt{Cov}\left(\sum_j\omega_j^*z_j^*,\sum_j\delta_jz_j^*\right)}_{<0}}+ && \text{(E.21)}\\
&\qquad\qquad \frac{\texttt{Var}\,(\varepsilon)}{\beta\left\{\texttt{Var}\,[x(\omega^*)]+\underbrace{\texttt{Cov}\left(\sum_j\omega_j^*z_j^*,\sum_j\delta_jz_j^*\right)}_{<0}\right\}}
\end{aligned}
$$

which is biased away from zero if the forward regression is biased toward zero.

Proof of Proposition 2.

$$\texttt{plim}\,\widehat{\beta}_{ols} = \frac{\texttt{Var}(\widetilde{x})\texttt{Cov}(\breve{x},y) - \texttt{Cov}(\breve{x},\widetilde{x})\texttt{Cov}(\widetilde{x},y)}{\texttt{Var}(\breve{x})\texttt{Var}(\widetilde{x}) - [\texttt{Cov}(\breve{x},\widetilde{x})]^2} \tag{E.22}$$

$$= \frac{1}{D}\left\{\texttt{Var}(\widetilde{x})\texttt{Cov}\left[\breve{x},\alpha+\beta x(\omega^*)+\varepsilon\right] - \texttt{Cov}(\breve{x},\widetilde{x})\texttt{Cov}\left[\widetilde{x},\alpha+\beta x(\omega^*)+\varepsilon\right]\right\} \tag{E.23}$$

$$= \frac{1}{D}\left\{\texttt{Var}(\widetilde{x})\texttt{Cov}\left\{\breve{x},\alpha+\beta\left[\widetilde{x}(\omega^*)+\breve{x}^*\right]+\varepsilon\right\} - \texttt{Cov}(\breve{x},\widetilde{x})\texttt{Cov}\left\{\widetilde{x},\alpha+\beta\left[\widetilde{x}(\omega^*)+\breve{x}^*\right]+\varepsilon\right\}\right\} \tag{E.24}$$

$$= \frac{1}{D}\left\{\texttt{Var}(\widetilde{x})\texttt{Cov}\left\{\breve{x},\beta\left[\widetilde{x}(\omega^*)+\breve{x}^*\right]\right\} - \texttt{Cov}(\breve{x},\widetilde{x})\texttt{Cov}\left\{\widetilde{x},\beta\left[\widetilde{x}(\omega^*)+\breve{x}^*\right]\right\}\right\} \tag{E.25}$$

$$= \frac{\beta}{D}\left\{\texttt{Var}(\widetilde{x})\left\{\texttt{Cov}\left[\breve{x},\widetilde{x}(\omega^*)\right]+\texttt{Cov}(\breve{x},\breve{x}^*)\right\} - \texttt{Cov}(\breve{x},\widetilde{x})\left\{\texttt{Cov}\left[\widetilde{x},\widetilde{x}(\omega^*)\right]+\texttt{Cov}(\widetilde{x},\breve{x}^*)\right\}\right\} \tag{E.26}$$

$$= \frac{\beta}{D}\left\{\texttt{Var}(\widetilde{x})\left\{\texttt{Cov}\left[\breve{x},\widetilde{x}(\omega^*)\right]+\texttt{Var}(\breve{x})-\texttt{Cov}(\breve{x},\breve{\mu})\right\} - \left[\texttt{Cov}(\breve{x},\widetilde{x})\right]^2 - \texttt{Cov}(\breve{x},\widetilde{x})\left\{\texttt{Cov}\left[\widetilde{x},\widetilde{x}(\omega^*)\right]-\texttt{Cov}(\widetilde{x},\breve{\mu})\right\}\right\} \tag{E.27}$$

$$= \frac{\beta}{D}\left\{D+\texttt{Var}(\widetilde{x})\left\{\texttt{Cov}\left[\breve{x},\widetilde{x}(\omega^*)\right]-\texttt{Cov}(\breve{x},\breve{\mu})\right\} - \texttt{Cov}(\breve{x},\widetilde{x})\left\{\texttt{Cov}\left[\widetilde{x},\widetilde{x}(\omega^*)\right]-\texttt{Cov}(\widetilde{x},\breve{\mu})\right\}\right\} \tag{E.28}$$

$$= \frac{\beta}{D}\left\{\begin{array}{l} D+\texttt{Var}\left[\widetilde{x}(\omega^*)+\widetilde{\mu}\right]\left\{\texttt{Cov}\left[\breve{x}^*+\breve{\mu},\widetilde{x}(\omega^*)\right]-\texttt{Cov}(\breve{x}^*+\breve{\mu},\breve{\mu})\right\} \\ \qquad -\texttt{Cov}\left[\breve{x}^*+\breve{\mu},\widetilde{x}(\omega^*)+\widetilde{\mu}\right]\left\{\texttt{Cov}\left[\widetilde{x}(\omega^*)+\widetilde{\mu},\widetilde{x}(\omega^*)\right]-\texttt{Cov}\left[\widetilde{x}(\omega^*)+\widetilde{\mu},\breve{\mu}\right]\right\} \end{array}\right\} \tag{E.29}$$

$$= \frac{\beta}{D}\left\{\begin{array}{l} D+\left\{\texttt{Var}\left[\widetilde{x}(\omega^*)\right]+\texttt{Var}\left(\widetilde{\mu}\right)+2\texttt{Cov}\left[\widetilde{x}(\omega^*),\widetilde{\mu}\right]\right\}\times \\ \qquad \left\{\texttt{Cov}\left[\breve{x}^*,\widetilde{x}(\omega^*)\right]+\underbrace{\texttt{Cov}\left[\widetilde{x}(\omega^*),\breve{\mu}\right]}_{=0}-\underbrace{\texttt{Cov}(\breve{x}^*,\breve{\mu})}_{=0}-\texttt{Var}(\breve{\mu})\right\} \\ -\left\{\texttt{Cov}\left[\breve{x}^*,\widetilde{x}(\omega^*)\right]+\underbrace{\texttt{Cov}(\breve{x}^*,\widetilde{\mu})}_{=0}+\underbrace{\texttt{Cov}\left[\widetilde{x}(\omega^*),\breve{\mu}\right]}_{=0}+\underbrace{\texttt{Cov}(\breve{\mu},\widetilde{\mu})}_{=0}\right\}\times \\ \qquad \left\{\texttt{Var}\left[\widetilde{x}(\omega^*)\right]+\texttt{Cov}\left[\widetilde{x}(\omega^*),\widetilde{\mu}\right]-\underbrace{\texttt{Cov}\left[\widetilde{x}(\omega^*),\breve{\mu}\right]}_{=0}-\underbrace{\texttt{Cov}(\breve{\mu},\widetilde{\mu})}_{=0}\right\} \end{array}\right\} \tag{E.30}$$

$$= \frac{\beta}{D}\left\{\begin{array}{l} D+\left\{\texttt{Cov}\left[\breve{x}^*,\widetilde{x}(\omega^*)\right]-\texttt{Var}(\breve{\mu})\right\}\left\{\texttt{Var}\left[\widetilde{x}(\omega^*)\right]+\texttt{Var}\left(\widetilde{\mu}\right)+2\texttt{Cov}\left[\widetilde{x}(\omega^*),\widetilde{\mu}\right]\right\} \\ \qquad -\texttt{Cov}\left[\breve{x}^*,\widetilde{x}(\omega^*)\right]\left\{\texttt{Var}\left[\widetilde{x}(\omega^*)\right]+\texttt{Cov}\left[\widetilde{x}(\omega^*),\widetilde{\mu}\right]\right\} \end{array}\right\} \tag{E.31}$$

$$= \frac{\beta}{D}\left\{D-\texttt{Var}(\breve{\mu})\texttt{Var}(\widetilde{x})+\texttt{Cov}\left[\breve{x}^*,\widetilde{x}(\omega^*)\right]\left\{\texttt{Var}\left(\widetilde{\mu}\right)+\texttt{Cov}\left[\widetilde{x}(\omega^*),\widetilde{\mu}\right]\right\}\right\} \tag{E.32}$$

$$= \frac{\beta}{D}\left\{\begin{array}{l} D-\texttt{Var}(\breve{\mu})\texttt{Var}(\widetilde{x})+\underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j^*,\sum_{j\in\mathcal{J}\setminus\mathcal{J}^1}\omega_j^* z_j^*\right)}_{?}\times \\ \qquad \left\{\texttt{Var}\left(\widetilde{\mu}\right)+\underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}\setminus\mathcal{J}^1}\omega_j^* z_j^*,\sum_{j\in\mathcal{J}\setminus\mathcal{J}^1}\delta_j z_j^*\right)}_{<0}\right\} \end{array}\right\} \tag{E.33}$$

where

$$D \;:=\; \texttt{Var}(\breve{x})\texttt{Var}(\widetilde{x}) - [\texttt{Cov}(\breve{x},\widetilde{x})]^2 \tag{E.34}$$

$$\widetilde{\mu} \;:=\; \widetilde{x}(\omega) - \widetilde{x}(\omega^*) = \sum\nolimits_{j\in\mathcal{J}\setminus\mathcal{J}^1}\left[\omega_j^*\psi_j + \delta_j(z_j^*+\psi_j)\right] \tag{E.35}$$

$$\breve{\mu} \;:=\; \breve{x} - \breve{x}^* = \sum\nolimits_{j\in\mathcal{J}^1}\psi_j. \tag{E.36}$$

If z^* is spatially independent such that

$$\texttt{Cov}\left(\sum\nolimits_{j\in\mathcal{J}^1} z_j^*, \sum\nolimits_{j\in\mathcal{J}\setminus\mathcal{J}^1} \omega_j^* z_j^*\right) = 0, \tag{E.37}$$

then $\widehat{\beta}_{ols}$ is consistent when z^* is observed and suffers from attenuation bias if z suffers from classical measurement error. If z^* exhibits positive (negative) spatial dependence, then a sufficient condition for attenuation bias is that $|\texttt{Cov}\left[\widetilde{x}(\omega^*), \widetilde{\mu}\right]| > (<)\texttt{Var}\left(\widetilde{\mu}\right)$.

Proof of Proposition 3.

$$\begin{aligned}
\texttt{plim}\,\widehat{\beta}_{iv} &= \frac{\texttt{Cov}\left[q, y\right]}{\texttt{Cov}\left[q, x(\omega)\right]} && \text{(E.38)}\\
&= \frac{\texttt{Cov}\left[q, \alpha + \beta x(\omega^*) + \varepsilon\right]}{\texttt{Cov}\left[q, x(\omega)\right]} && \text{(E.39)}\\
&= \frac{\beta\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left(q, \varepsilon\right)}{\texttt{Cov}\left[q, x(\omega)\right]} && \text{(E.40)}\\
&= \frac{\beta\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left(q, \varepsilon\right)}{\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left(q, \breve{\mu}\right)} && \text{(E.41)}\\
&= \frac{\beta\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left(q, \varepsilon\right)}{\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left[q, \sum_j \left[\omega_j^*\psi_j + \delta_j(z_j^* + \psi_j)\right]\right]} && \text{(E.42)}\\
&= \frac{\beta\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left(q, \varepsilon\right)}{\texttt{Cov}\left[q, x(\omega^*)\right] + \texttt{Cov}\left(q, \sum_j \omega_j^*\psi_j\right) + \texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right) + \texttt{Cov}\left(q, \sum_j \delta_j\psi_j\right)} && \text{(E.43)}
\end{aligned}$$

Proof of Corollary 1. Computation of $\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right)$ for **Q2**.

$$\begin{aligned}
\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right) &= \texttt{Cov}\left(\sum_j \omega_j b_j, \sum_j \delta_j z_j^*\right) && \text{(E.44)}\\
&= \texttt{Cov}\left[\sum_j \left(\omega_j^* + \delta_j\right) b_j, \sum_j \delta_j z_j^*\right] && \text{(E.45)}\\
&= \underbrace{\texttt{Cov}\left(\sum_j \omega_j^* b_j, \sum_j \delta_j z_j^*\right)}_{<0} + && \text{(E.46)}\\
&\qquad \underbrace{\texttt{Cov}\left(\sum_j \delta_j b_j, \sum_j \delta_j z_j^*\right)}_{>0} > 0 && \text{(E.47)}
\end{aligned}$$

as the second term will dominate. $\Rightarrow \texttt{plim}\ \widehat{\beta}_{iv} \neq \beta$.

Proof of Corollary 1. Analysis of **Q3**.

$$\begin{aligned}
\texttt{Cov}\left(q, \sum_j \omega_j^* \psi_j\right) &= \texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j, \sum_j \omega_j^* \psi_j\right) && \text{(E.48)}\\
&= \texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} (z_j^* + \psi_j), \sum_j \omega_j^* \psi_j\right) && \text{(E.49)}\\
&= \underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j^*, \sum_j \omega_j^* \psi_j\right)}_{=0} + && \text{(E.50)}\\
&\qquad \underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} \psi_j, \sum_j \omega_j^* \psi_j\right)}_{\neq 0} = 0.\\
\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right) &= \texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j, \sum_j \delta_j z_j^*\right) && \text{(E.51)}\\
&= \texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} (z_j^* + \psi_j), \sum_j \delta_j z_j^*\right) && \text{(E.52)}\\
&= \underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j^*, \sum_j \delta_j z_j^*\right)}_{=0} + && \text{(E.53)}\\
&\qquad \underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} \psi_j, \sum_j \delta_j z_j^*\right)}_{=0} = 0.\\
\texttt{Cov}\left(q, \sum_j \delta_j \psi_j\right) &= \texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j, \sum_j \delta_j \psi_j\right) && \text{(E.54)}\\
&= \texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} (z_j^* + \psi_j), \sum_j \delta_j \psi_j\right) && \text{(E.55)}\\
&= \underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} z_j^*, \sum_j \delta_j \psi_j\right)}_{=0} + && \text{(E.56)}\\
&\qquad \underbrace{\texttt{Cov}\left(\sum_{j\in\mathcal{J}^1} \psi_j, \sum_j \delta_j \psi_j\right)}_{=0} = 0.
\end{aligned}$$

$\Rightarrow \texttt{plim}\ \widehat{\beta}_{iv} = \beta$ only if z^* is observed.

Proof of Corollary 1. Computation of $\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right)$ for **Q4**.

$$\texttt{Cov}\left(q, \sum\nolimits_j \omega_j^* \psi_j\right) = \underbrace{\texttt{Cov}\left(\sum\nolimits_{j\in\mathcal{J}^1} b_j, \sum\nolimits_j \omega_j^* \psi_j\right)}_{=0} = 0. \tag{E.57}$$

$$\texttt{Cov}\left(q, \sum\nolimits_j \delta_j z_j^*\right) = \underbrace{\texttt{Cov}\left(\sum\nolimits_{j\in\mathcal{J}^1} b_j, \sum\nolimits_j \delta_j z_j^*\right)}_{=0} = 0. \tag{E.58}$$

$$\texttt{Cov}\left(q, \sum\nolimits_j \delta_j \psi_j\right) = \underbrace{\texttt{Cov}\left(\sum\nolimits_{j\in\mathcal{J}^1} b_j, \sum\nolimits_j \delta_j \psi_j\right)}_{=0} = 0. \tag{E.59}$$

$\Rightarrow \texttt{plim}\ \widehat{\beta}_{iv} = \beta$.

Proof of Corollary 1. Computation of $\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right)$ for **Q5**.

$$\texttt{Cov}\left(q, \sum\nolimits_j \delta_j z_j^*\right) = \texttt{Cov}\left(\sum\nolimits_j \omega_{g',j} z_j, \sum\nolimits_j \delta_{g,j} z_j^*\right) \tag{E.60}$$

$$= \texttt{Cov}\left(\sum\nolimits_j (\omega_j^* + \delta_{g',j})(z_j^* + \psi_j), \sum\nolimits_j \delta_{g,j} z_j^*\right) \tag{E.61}$$

$$= \texttt{Cov}\left(\begin{array}{c} \sum_j (\omega_j^* z_j^* + \delta_{g',j} z_j^* + \omega_j^* \psi_j + \delta_{g',j}\psi_j), \\ \sum_j \delta_{g,j} z_j^* \end{array}\right) \tag{E.62}$$

$$= \underbrace{\texttt{Cov}\left(\sum\nolimits_j \omega_j^* z_j^*, \sum\nolimits_j \delta_{g,j} z_j^*\right)}_{<0} + \tag{E.63}$$

$$\underbrace{\texttt{Cov}\left(\sum\nolimits_j \delta_{g',j} z_j^*, \sum\nolimits_j \delta_{g,j} z_j^*\right)}_{\geq 0} +$$

$$\underbrace{\texttt{Cov}\left(\sum\nolimits_j \omega_j^* \psi_j, \sum\nolimits_j \delta_{g,j} z_j^*\right)}_{=0} +$$

$$\underbrace{\texttt{Cov}\left(\sum\nolimits_j \delta_{g',j} \psi_j, \sum\nolimits_j \delta_{g,j} z_j^*\right)}_{=0} \neq 0.$$

$\Rightarrow \texttt{plim}\ \widehat{\beta}_{iv} \neq \beta$.

Proof of Corollary 1. Computation of $\texttt{Cov}\left(q, \sum_j \delta_j z_j^*\right)$ for **Q6**.

$$\begin{aligned}
\texttt{Cov}\left(q, \sum\nolimits_j \delta_j z_j^*\right) &= \texttt{Cov}\left(\sum\nolimits_j \omega_{g',j} b_j, \sum\nolimits_j \delta_{g,j} z_j^*\right) && \text{(E.64)}\\
&= \texttt{Cov}\left(\sum\nolimits_j (\omega_j^* + \delta_{g',j}) b_j, \sum\nolimits_j \delta_{g,j} z_j^*\right) && \text{(E.65)}\\
&= \texttt{Cov}\left(\sum\nolimits_j (\omega_j^* b_j + \delta_{g',j} b_j, \sum\nolimits_j \delta_{g,j} z_j^*\right) && \text{(E.66)}\\
&= \underbrace{\texttt{Cov}\left(\sum\nolimits_j \omega_j^* b_j, \sum\nolimits_j \delta_{g,j} z_j^*\right)}_{<0} + && \text{(E.67)}\\
&\qquad \underbrace{\texttt{Cov}\left(\sum\nolimits_j \delta_{g',j} b_j, \sum\nolimits_j \delta_{g,j} z_j^*\right)}_{\geq 0} \neq 0.
\end{aligned}$$

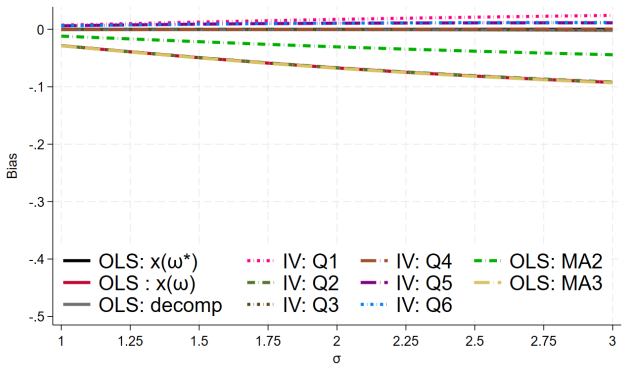
$\Rightarrow \texttt{plim}\, \widehat{\beta}_{iv} \neq \beta$.

F Simulation Results

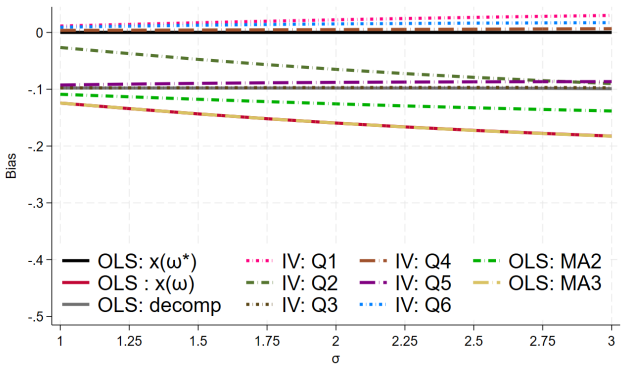
TABLE F.II: Attributes of the Experimental Design

	$\sigma = 1$	$\sigma = 1.25$	$\sigma = 1.50$	$\sigma = 1.75$	$\sigma = 2$	$\sigma = 2.25$	$\sigma = 2.50$	$\sigma = 2.75$	$\sigma = 3$
$\texttt{Corr}\,(x(\omega^*), x(\omega))$	0.983	0.976	0.970	0.964	0.959	0.954	0.950	0.946	0.943
$\texttt{RR}\,(x(\omega^*), x(\omega))$	0.978	0.971	0.964	0.957	0.951	0.946	0.941	0.937	0.933
$\texttt{Mean}\,(\omega^* - \omega)$	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
$\texttt{SD}\,(\omega^* - \omega)$	0.018	0.021	0.024	0.026	0.028	0.029	0.031	0.032	0.033
Frisch Coverage	0.577	0.642	0.672	0.698	0.716	0.732	0.738	0.741	0.747
Frisch Width	0.027	0.040	0.053	0.064	0.075	0.084	0.092	0.099	0.105
$\texttt{J} - \texttt{test}$	0.951	0.946	0.949	0.949	0.948	0.950	0.946	0.945	0.948
Alt Spec test	0.542	0.713	0.811	0.883	0.933	0.961	0.975	0.983	0.989
$\texttt{Med(F} - \texttt{stat)},\ \texttt{IV Q1}$	2548	1881	1492	1247	1081	962	878	814	764
$\texttt{Med(F} - \texttt{stat)},\ \texttt{IV Q2}$	129	129	129	129	129	129	129	129	129
$\texttt{Med(F} - \texttt{stat)},\ \texttt{IV Q3}$	555	529	503	479	461	445	434	426	421
$\texttt{Med(F} - \texttt{stat)},\ \texttt{IV Q4}$	92	91	89	87	86	85	84	84	83
$\texttt{Med(F} - \texttt{stat)},\ \texttt{IV Q5}$	1393	1002	793	663	578	516	472	441	416
$\texttt{Med(F} - \texttt{stat)},\ \texttt{IV Q6}$	111	105	100	96	93	90	87	85	84

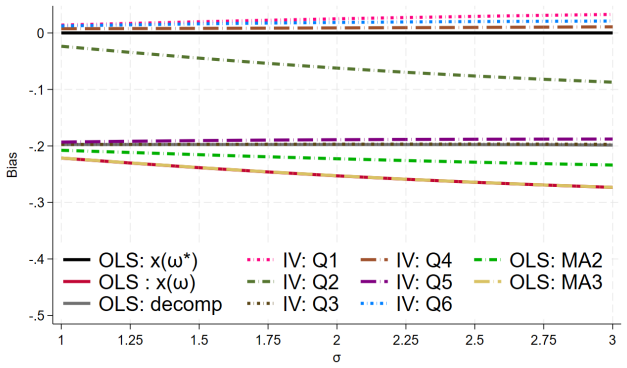
Notes: Results from experiments where z^* is observed. σ affects the severity of the measurement error in the weights. RR is the reliability ratio, defined as $\texttt{Var}(x(\omega^*))/\texttt{Var}(x(\omega))$. Frisch coverage is the proportion of Frisch bounds that include the true value of β. Frisch width is the width of the Frisch bounds. J-test is the proportion of samples where the specification test fails to reject that ω^* is the correct weighting at the $p < 0.05$ level, but rejects that either ω or the weighting scheme used for instrument **Q2** is the correct weighting scheme at the $p < 0.05$ level. Alt Spec Test test is the proportion of samples where the test of equality in Equation (14) is rejected. κ is the estimated weight placed on the weighting scheme ω relative to the weighting scheme used for instrument **Q2**. $\texttt{F} - \texttt{stat}$ refers to the first-stage strength of the instrument. IV Q1 = external instrument (**Q1**). IV Q2 = instrument based on alternative underlying covariate. IV Q3 = instrument based on underlying covariate when $\omega^* = 1$. IV Q4 = instrument based on alternative underlying covariate when $\omega^* = 1$. IV Q5 = instrument based on alternative weights. IV Q6 = instrument based on alternative weights and underlying covariate.



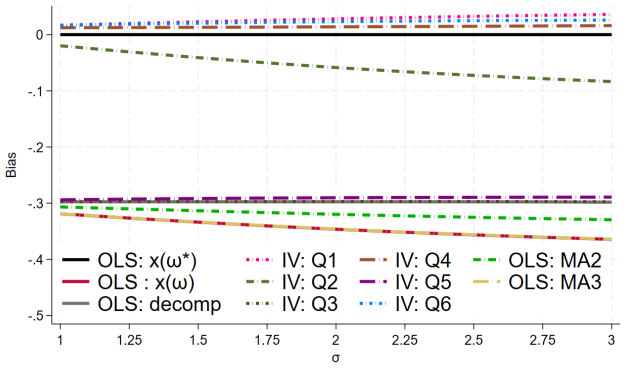
(A) RR = 1



(B) RR = 0.9



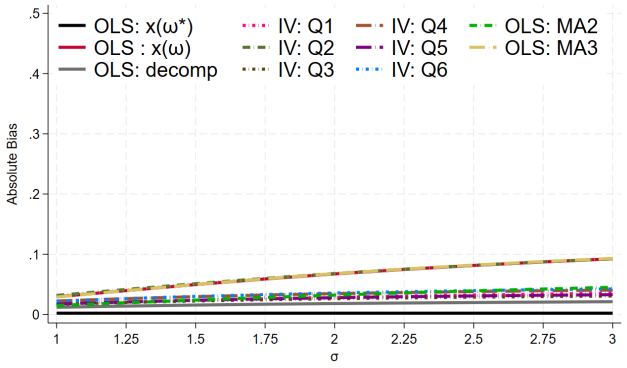
(C) RR = 0.8



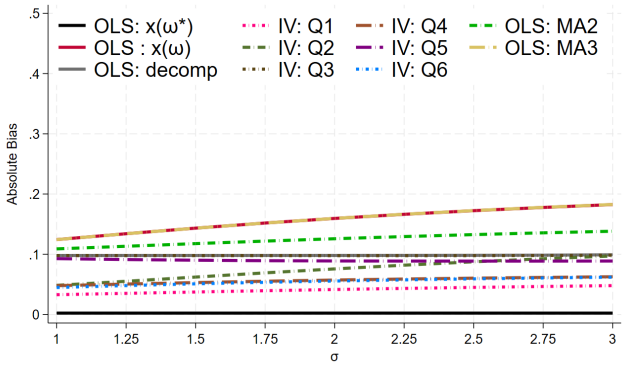
(D) RR = 0.7

FIGURE F.IV: Simulation Results: Bias

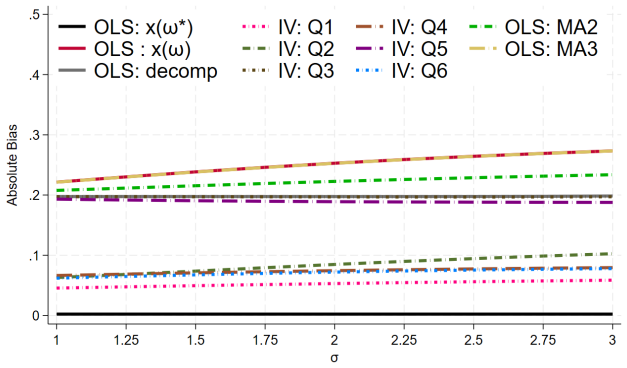
Notes: RR = reliability ratio and refers to the (classical) measurement error in z^*. See text for more details.



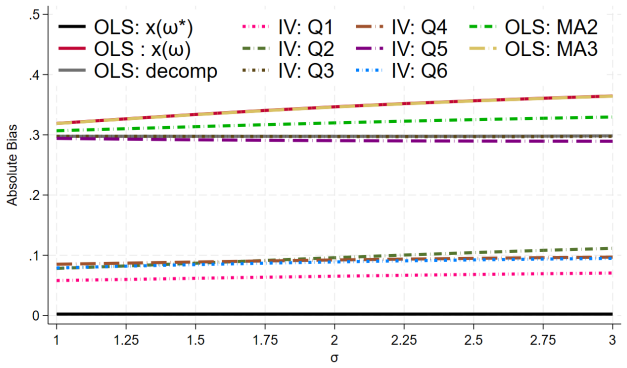
(A) RR = 1



(B) RR = 0.9



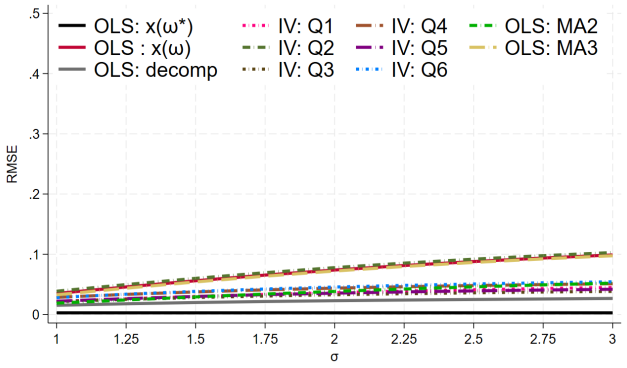
(C) RR = 0.8



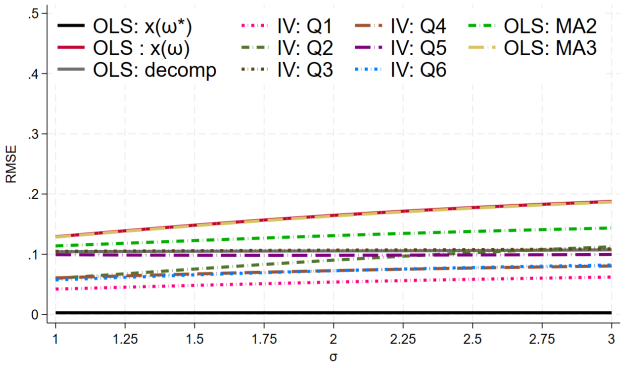
(D) RR = 0.7

FIGURE F.V: Simulation Results: Absolute Bias

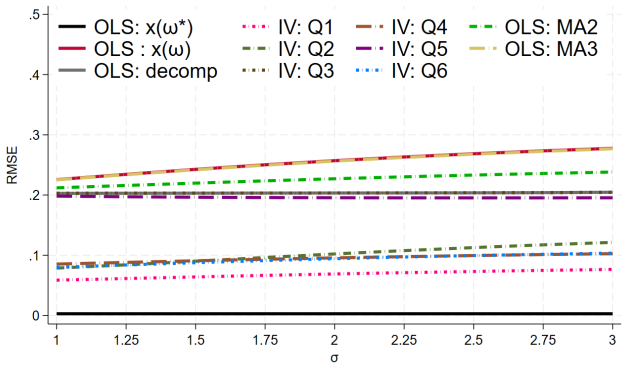
Notes: RR = reliability ratio and refers to the (classical) measurement error in z^*. See text for more details.



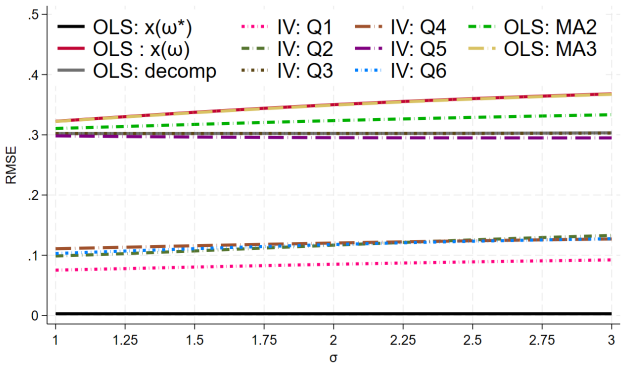
(A) RR = 1



(B) RR = 0.9



(C) RR = 0.8



(D) RR = 0.7

FIGURE F.VI: Simulation Results: Root Mean Squared Error

Notes: RR = reliability ratio and refers to the (classical) measurement error in z^*. See text for more details.