Using Digitized Newspapers to Address Measurement Error in Historical Data

ANDREAS FERRARA, JOUNG YEOB HA, AND RANDALL WALSH

This paper shows how to remove attenuation bias in regression analyses due to measurement error in historical data for a given variable of interest by using a secondary measure that can be easily generated from digitized newspapers. We provide three methods for using this secondary variable to deal with non-classical measurement error in a binary treatment: set identification, bias reduction via sample restriction, and a parametric bias correction. We demonstrate the usefulness of our methods by replicating four recent economic history papers. Relative to the initial analyses, our results yield markedly larger coefficient estimates.

The use of digitized newspaper data by economic historians has become more prominent in recent years. We propose a novel use of such data to overcome measurement error, a problem that is pervasive in the statistical analysis of historical data. Given that regression coefficients of mismeasured variables are attenuated (Aigner 1973), measurement error can lead promising research to be abandoned. A solution to such attenuation bias for continuous variables with classical measurement error is to use an instrumental variables approach leveraging a second mismeasured data source as the instrument. In the absence of other endogeneity concerns,[1] as long as the measurement error in the two

The Journal of Economic History, Vol. 84, No. 1 (March 2024). © The Author(s), 2024. Published by Cambridge University Press on behalf of the Economic History Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited. doi: 10.1017/S0022050723000530

Andreas Ferrara is Assistant Professor, University of Pittsburgh, Department of Economics, 4906 Wesley W. Posvar Hall, Pittsburgh, 230 South Bouquet Street, PA 15260 and NBER. E-mail address: a.ferrara@pitt.edu (corresponding author). Joung Yeob Ha is Ph.D. Candidate, University of Pittsburgh, Department of Economics, 4512 Wesley W. Posvar Hall, Pittsburgh, 230 South Bouquet Street, PA 15260. E-mail address: joh106@pitt.edu. Randall Walsh is Professor, University of Pittsburgh, Department of Economics, 4166 Wesley W. Posvar Hall, 230 South Bouquet Street, Pittsburgh, PA 15260 and NBER. E-mail address: walshr@pitt.edu.

We thank Karen Clay, James Feigenbaum, Patrick Testa, Alistair Wilson, as well as the editor, Eric Hilt, and three anonymous referees for helpful comments and suggestions. We are also grateful to seminar participants at Tulane University, the University of Pittsburgh, as well as at the Political Institutions & Political Economy workshop for helpful feedback. This version replaces an earlier one entitled “Using Digitized Newspapers to Refine Historical Measures: The Case of the Boll Weevil.”

[1] Many current papers in economic history seek to establish causal relationships for which instrumental variables from natural experiments are commonly used and because other endogeneity concerns, such as omitted variables, are a potential problem (see Dippel and Leonard 2021). These instruments also resolve classical measurement error, however, when treatment variables are not continuous, measurement error is non-classical by construction and the approach fails, an issue that we will discuss more fully later.

variables is uncorrelated, instrumenting for one mismeasured variable, X_1, with data from a second mismeasured source, X_2, recovers the true parameter (see Chalfin and McCrary 2018).[2] The main limitation of this approach is that it is difficult to find a second variable that is (i) measured with error, which is arguably uncorrelated with the error in X_1, and (ii) reasonably inexpensive to collect. Since economic historians often spend a significant amount of time and effort on original data collection, it is usually costly enough to just have X_1.

In this paper, we show how the second measure, X_2, can often be generated at a low cost from textual data available via digitized newspapers and how it can be used to resolve measurement error in the case where X_1 is continuous or binary.[3] The distinction between continuous and binary variables is important because using X_2 as an instrument for X_1 to recover the true parameter only applies to cases of classical measurement error, which requires X_1 to be continuous (Bingley and Martinello 2017).[4] If X_1 is binary *and* mismeasured, then IV estimates will be inflated by the inverse of the misclassification rate in X_1. This is true even when the instrument is generated by an otherwise perfectly valid natural experiment.

We provide three potential solutions when X_1 is binary. First, the treatment effect can be *set identified*. The OLS estimate using X_1 as treatment provides a lower bound, while the IV estimate using X_2 as an instrument for X_1 provides the upper bound such that $\hat{\beta}_{OLS} < \beta < \hat{\beta}_{IV}$. Second, we show that restricting the analysis to an *agreement sample* where $X_1 = X_2$ can substantially reduce the OLS bias. The probability that both variables are jointly misclassified is the product of the two variables' misclassification rates, and therefore the measurement error in the agreement sample tends to be much lower.[5] Third, we provide a *parametric bias correction* procedure that can recover the true parameter of interest as a nonlinear combination of the OLS and IV coefficients. All three procedures are fast and efficient, and given that newspaper data can be scraped in a reasonable amount of time, we hope to provide researchers who work with historical data with low-cost tools for dealing with measurement error. We begin the demonstration of our three procedures by replicating two recent papers that study the economic impact of the spread of the boll weevil across the U.S. South in the late nineteenth and early twentieth

[2] Both X_1 and X_2 seek to measure the true but unobserved quantity X^*.

[3] Examples of such databases include Chronicling America, Newspapers.com, and ProQuest.

[4] Classical measurement error requires that there is no correlation between the true value and the error. Suppose a binary treatment is misclassified, then the error has a perfect negative correlation with the true value by construction because if $X^* = 1$, then $u = -1$ and, vice versa, $u = 1$ if $X^* = 0$.

[5] For instance, suppose X_1 and X_2 have misclassification rates of 30 and 20 percent, respectively, where one minus the misclassification rate determines the OLS bias. The attenuation bias in the agreement sample will be $0.3 \times 0.2 = 0.06$.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

centuries, one by Clay, Schmick, and Troesken (2019) and one by Ager, Brueckner, and Herz (2017).[6]

To date, the sole source of data used by analysts to measure the timing of the boll weevil's arrival at the county-level comes from a U.S. Department of Agriculture (USDA) map by Hunter and Coad (1923), which documents the arrival date of the pest across Southern counties between 1892 and 1922. While the map itself is mostly accurate, it does contain errors.[7] Further, it does not necessarily measure what economists are typically interested in, namely the timing of the economic damage caused by the arrival of the boll weevil. As an example, if the weevil arrived late in the summer, it would typically hibernate soon after arrival, and thus the actual economic damage would not occur until the following year. The arrival date from the USDA map is therefore a mismeasured proxy for the date of the actual economic impact. And, as we document, this mismeasurement can markedly attenuate estimated effect sizes.

To produce a second measure for the arrival of the boll weevil, we collect data from Newspapers.com by jointly searching the database for pages containing "boll weevil" and each county's name in all newspapers in the county's state for each year between 1882 and 1932. Our arrival measure is then the peak salience of the weevil in the news as measured by the maximum five-year moving average of boll weevil-related pages.[8] We argue that errors in this newspaper-based measure are likely to be uncorrelated with errors on the USDA map, which was generated by trained USDA entomologists who reported back to the federal agency, whereas local newspaper reporters mainly wrote about salient issues in their home counties. Using an event study design, we also show that the newspaper-based salience peaks a year after the official USDA arrival date on average.

Our replications of Clay, Schmick, and Troesken (2019) and Ager, Brueckner, and Herz (2017) show that using our newspaper-based arrival measure can reduce measurement error and strengthen the results in both papers. In particular, our theory suggests a ranked pattern between the three proposed solutions, where $\hat{\beta}_{OLS} < \hat{\beta}_{X_1 = X_2} < \beta = \hat{\beta}_{\text{bias-corrected}} < \hat{\beta}_{IV}$.

[6] To show generalizability to other settings, we also replicate results from Hilt and Rahn (2020) on the effect of the liberty loan program on election outcomes, as well as the study by Howard and Ornaghi (2021) on the effect of local prohibition policies on agricultural outcomes.

[7] In some instances, the map reports inconsistent arrival dates. The map shows the arrival date with date borders that occasionally overlap in contradictory ways. See Figure 1 for examples of such overlaps.

[8] We use a moving average to additionally smooth out noise in the newspaper data and provide sensitivity checks to show that other transformations, such as using a three- or seven-year moving average or the raw data, give similar results.

While we do not observe the true coefficient, the estimated coefficients largely follow the prescribed pattern in both replication exercises. We find evidence that measurement error led to lower coefficient estimates in both studies, a finding that is robust across alternative specifications of our newspaper-based arrival date. However, the difference in the coefficients produced by our procedures was only statistically significant for Ager, Brueckner, and Herz (2017). We discuss the frequency of the time dimension as a potential reason for this finding, as Clay, Schmick, and Troesken (2019) use annual data while Ager, Brueckner, and Herz (2017) use data over five-year intervals.

We provide a broader discussion of when data generation from newspaper articles is a promising avenue, what settings are suitable for our approach, and the value of historical newspapers to generate novel data for research in economic history. Even though our newspaper-based measures were generated in a fast and arguably unrefined way, using this noisier measure still produces smaller but significant effects that are comparable to those in Clay, Schmick, and Troesken (2019) and Ager, Brueckner, and Herz (2017). Lastly, to show that our approach extends to other settings, we further replicate a study by Hilt and Rahn (2020) of the liberty loan program's effect on political outcomes, as well as a paper by Howard and Ornaghi (2021), which studies the impact of the adoption of local prohibition policies on population, agricultural outcomes, and investment. Their treatment measures are different in nature from the boll weevil, an arrival time measure, to provide additional examples for when our strategies can be gainfully applied to deal with measurement error in historical data.

Our paper highlights the usefulness of digitized newspapers to generate additional data to address measurement error. We extend the secondary measure IV framework in Chalfin and McCrary (2018) to the case where treatment is binary and when instrumenting ordinarily does not resolve measurement error (Bingley and Martinello 2017). We also contribute to a recent literature that uses digitized newspapers to generate novel data for research in economic history. This includes measures of media competition and partisan influence (Gentzkow, Shapiro, and Sinkinson 2014; Gentzkow et al. 2015), racial and anti-group sentiment (Ferrara and Fishback 2023; Ottinger and Winkler 2022; Bazzi et al. 2023), the spread of news relating to racial violence (Albright et al. 2021; Calderon, Fouka, and Tabellini 2023), technology diffusion (Feigenbaum and Gross 2022), the 1918 influenza (Beach, Clay, and Saavedra 2022), fertility restrictions (Beach and Hanlon 2023), advertisements for the movie "Birth of a Nation" (Esposito et al. 2023; Ang 2023), the price and types of available cotton seeds (Rhode 2021), among others.

BACKGROUND AND MEASUREMENT OF THE BOLL WEEVIL INFESTATION

We motivate the econometric theory by replicating two recent studies on the boll weevil infestation in the U.S. South to provide an example of how measurement error can be addressed with historical newspaper data. We first give a brief background on the boll weevil and measurement issues in the USDA data, which tracked the spread of the pest, followed by a discussion of how we use digitized newspaper data to generate a second boll weevil arrival measure before turning to the econometric theory.

The Spread of the Boll Weevil and Uses of the USDA Map

The boll weevil spread across the U.S. South starting in 1892 near Brownsville, Texas. The beetle, which gained its name because of its diet consisting mainly of cotton bolls and flowers, had infested all Southern cotton-growing regions by 1922. Given that cotton at the time was still the main cash crop in Southern agriculture (Wright 2013), the arrival of the pest had a substantial impact on the areas it infested. Consequently, the USDA traced the arrival of the weevil on a map in an annual report by Hunter and Coad (1923). A portion of this map is shown in Figure 1. During peak infestation in 1921, cotton acreage had declined by 31 percent (Ager, Brueckner, and Herz 2017), and the USDA estimated the average economic loss per year to be 200 to 300 million USD between 1916 and 1920 (Hunter and Coad 1923).[9] Given this substantial economic shock, a well-developed literature has studied the various impacts of the boll weevil on different aspects of the Southern economy.

Lange, Olmstead, and Rhode (2009) show the large negative impact of the pest on cotton production, yields, and land value. The drop in productivity also altered the structure of Southern agriculture with a reduced number of tenant farmers, farm wages, and female labor force participation (Ager, Brueckner, and Herz 2017). Ager, Herz, and Brueckner (2020) provide evidence that the lower returns to agriculture reduced fertility due to the opportunity cost of children and the decreased value of child labor. Also, Black Southerners tended to marry later after the pest arrived for the same reasons (Bloome, Feigenbaum, and Muller 2017). This fertility transition and the decline in the value of child labor in agriculture have also been linked to increased educational attainment (Baker 2015; Baker,

[9] The damage corresponds to $3.2–$4.8 billion in 2021 dollars.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

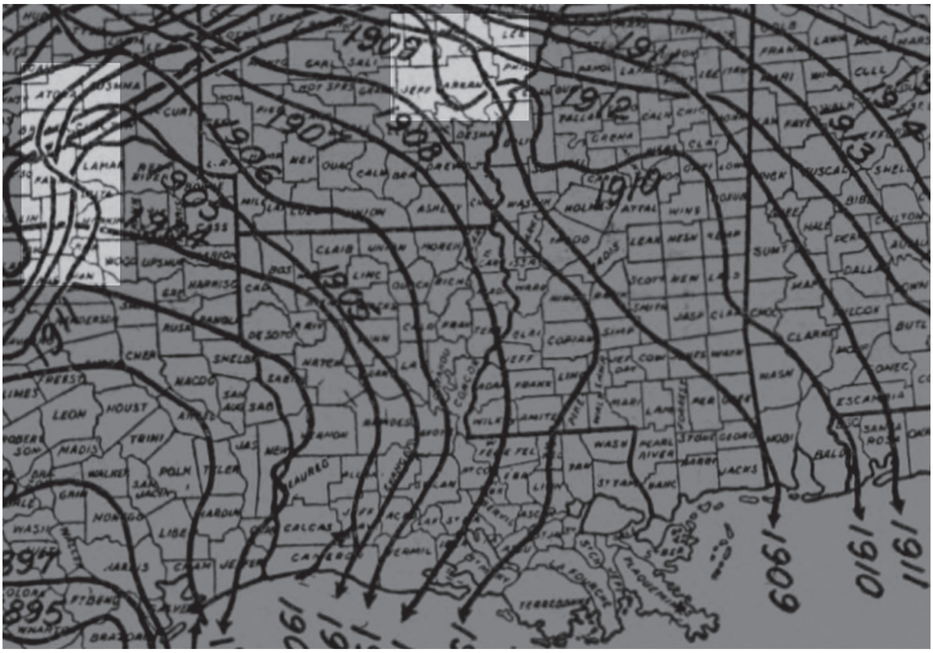

FIGURE 1
ERRORS ON THE USDA MAP FOR THE ARRIVAL OF THE BOLL WEEVIL

Notes: Snipped of the USDA map for the arrival of the boll weevil provided by Hunter and Coad (1923). Each solid line marks the arrival year of the pest. Researchers typically overlay the lines onto a map of Southern counties and determine the arrival date by the line that covers most of the county area. The highlighted areas are where date lines cross in contradictory ways.
Sources: Hunter and Coad (1923).

Blanchette, and Eriksson 2020). Another unintended consequence of the reduction in cotton production was increased food production. Clay, Schmick, and Troesken (2019) show that this significantly contributed to the reduction in pellagra deaths. In a later paper, the authors also found that the boll weevil spread reduced the racial income gap in the South (Clay, Schmick, and Troesken 2020). Similar to the population movements discussed in Lange, Olmstead, and Rhode (2009), Feigenbaum, Mazumder, and Smith (2020) show that the decline in cotton reliance also resulted in less violence against Black Southerners who saw an increased ability to move away from overtly discriminatory behavior.

Most of the papers noted previously either assign the arrival date for a county whenever the USDA map's first arrival year line crosses that county's area, or the arrival date is selected for the year line that contains most of the county's area (see Figure 1). What should be noted is that the solid lines in the map technically show the farthest extent of the boll weevil in any territory. This measure does not necessarily correlate with the exact timing of damage caused by the insect. Mature boll weevils hibernate during the winter and infest the cotton fields after the crop season in the subsequent year. Lange, Olmstead, and Rhode (2009) explicitly mention this caveat in their paper: "*First contact usually occurred during the August seasonal migration, too late to build up significant populations or do much damage in that year. Maximum damage occurred after the local weevil population became established and multiplied. Thus, the classic USDA maps detailing the spread of the weevil present a somewhat misleading picture of the area ravaged by the insect*" (p. 689).

Measuring the Boll Weevil's Arrival from Newspaper Data

Newspapers were the primary source of information in the late nineteenth and early twentieth centuries and mainly operated locally in the county where the paper was based (Gentzkow, Shapiro, and Sinkinson 2014). Newspapers published articles about the boll weevil's arrival as well as damages caused by the insects. An example of such reporting is shown in Online Appendix Figure A.1. Digitized newspaper data are a potential source to generate information on the arrival and damage extent caused by the pest, independent of the USDA map. We use Newspapers.com as our primary data source for digitized historical newspapers. To the best of our knowledge, this is the largest newspaper archive available online.[10]

For each county, in order to construct our newspaper-based boll weevil arrival and salience measure, we take all of the available newspapers from said county's state and identify the number of newspaper pages that include both the words "boll weevil" and the county's name for each year.[11] We use all newspapers from an individual county's state because no newspaper archive has information on the universe of newspaper pages. Thus, as described, our search not only considers pages in the county of interest but in all counties that are in the same state (e.g., Online Appendix Figure A.2). So, even if Autauga County in Alabama has no available newspaper pages for the search period but "Autauga County"

[10] Chronicling America is another digital archive for historical newspapers that is commonly used by researchers (e.g., Wang 2019; Ferrara and Fishback 2023). However, it has fewer volumes than Newspapers.com and does not contain many digitized newspapers that cover our sample period. As of 23 April 2023, there were 850,873,846 newspaper pages available on Newspapers.com compared to 20,389,221 pages in Chronicling America.

[11] In principle, one could search each article for a specific arrival date mentioned on the page for each county. However, this would be time-consuming and, therefore, costly. We instead use this simple search procedure to minimize the cost for researchers and later show that this quickly obtained raw measure of boll weevil activity is still a good proxy for the insect's arrival and salience in a county. The data are available in Ferrara, Ha, and Walsh (2023).

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

and “boll weevil” are mentioned in a newspaper based in Barbour County, Alabama, we are able to obtain data for Autauga County. Some counties may feature more prominently in the news than others, which is why we need to adjust these counts for the overall number of pages that mention the county. Thus, we apply the same search logic to generate the numerator in our boll weevil measure, which we compute as

$$\%BW_{ct} = \frac{No.\ of\ in-state\ newspaper\ pages\ mentioning\ "boll\ weevil"\ and\ a\ county's\ name_{ct}}{No.\ of\ in-state\ newspaper\ pages\ mentioning\ a\ county's\ name_{ct}} \quad (1)$$

where $\%BW_{ct}$ captures the salience of the boll weevil for county c in year t in the news. Our sample includes 911 infested counties from 13 Southern states between 1882 and 1932,[12] which is ten years before and after the time periods covered by the USDA map.

How does our salience measure relate to the official arrival date on the USDA map? To answer this question formally, we use an event study design and estimate the following equation:

$$\%BW_{ct} = \pi_c + \gamma_{st} + \sum_{\ell=-10}^{-2} \beta_\ell \cdot D(t - BW_c^{USDA} = \ell) + \sum_{\ell=0}^{10} \beta_\ell \cdot D(t - BW_c^{USDA} = \ell) + \varepsilon_{ct}, \quad (2)$$

where $\%BW_{ct}$ is our newspaper-based salience measure for county c in year t, and $D(t - BW_c^{USDA} = \ell)$ is an event indicator relative to the arrival of the boll weevil from the USDA map for the ten years before and after the official arrival date. The year before the arrival on the USDA map, $\ell = -1$, is omitted and serves as the baseline period. The county fixed effects π_c capture time-invariant unobservable county characteristics and aggregate time trends that affect counties jointly in each state are captured by state-by-year fixed effects γ_{st}. Standard errors are clustered at the county-level. Given the recent literature on issues related to event study designs, we use the estimator developed by Sun and Abraham (2021).

Our main interest is in the lag coefficients β_ℓ for $\ell \geq 0$. If salience in the news correlates highly with the USDA arrival date, then we should observe an immediate jump at the treatment date $\ell = 0$, followed by an either constant or slowly decaying coefficient pattern. Conversely, if the weevil tends to arrive later in the summer and hibernates, the more salient economic damage would occur in the following year, which implies that the main effect on salience in the news should occur after $\ell = 0$. The pattern of the coefficients should not only be informative about the decay

[12] The sample includes Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, Missouri, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, and Virginia.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

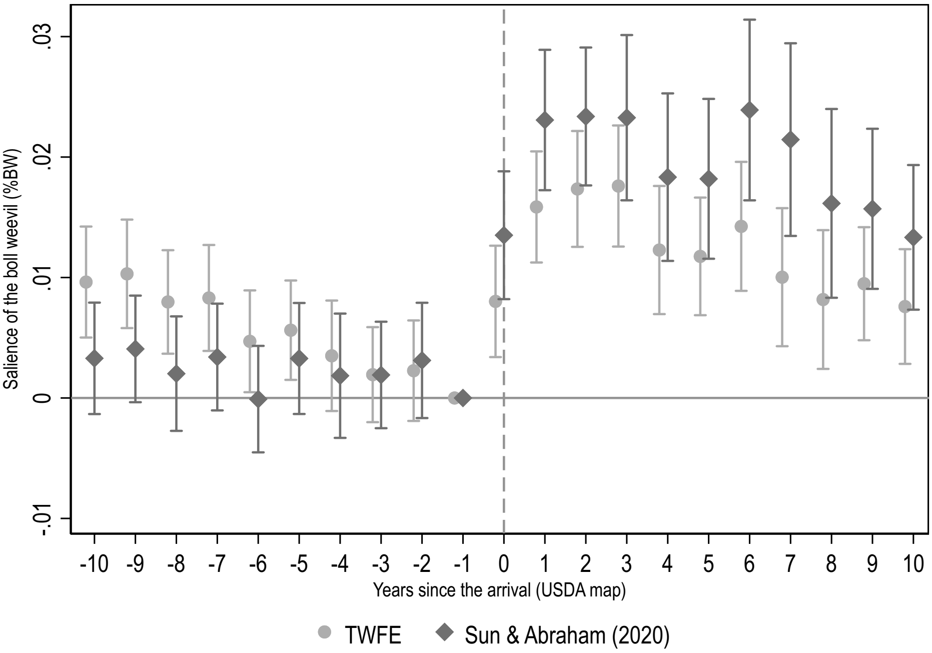

FIGURE 2

EVENT STUDY PLOT—TWFE AND SUN AND ABRAHAM (2021)

Notes: Coefficient plot from an event study regression of %*BW* on an event indicator relative to the arrival of the boll weevil from the USDA map as well as county and state-by-year fixed effects. Each circle and diamond present the estimates β_ℓ in Equation (2) using OLS and the estimator proposed by Sun and Abraham (2021), respectively. The sample consists of 911 infested counties in 13 Southern states. The omitted baseline period is $\ell = -1$, which is one year before the arrival of the USDA map. The relative time period for the latest-infested counties is omitted as well for the estimates using Sun and Abraham (2021) due to the lack of never-infested counties in our sample. Standard errors are clustered at the county-level, and 95 percent confidence intervals are reported around the point estimates.
Sources: Authors' calculations from data in Hunter and Coad (1923) and Newspapers.com.

in salience after arrival but also reveal potential anticipatory behavior if the lead coefficients are significant for $\ell < -1$.

Figure 2 plots the dynamic treatment effects for the 20-year event window around each county's boll weevil arrival date on the USDA map on our newspaper-based salience measure. The figure shows the coefficients from estimating Equation (2) via two-way fixed effects (TWFE) and with the estimator developed by Sun and Abraham (2021). We find that the salience measure significantly increases in counties after the boll weevil's arrival, based on the USDA map. More importantly, the effect is largest one year after the arrival date on the USDA map. This confirms the narrative that salience in the news and arrival are somewhat but not perfectly correlated due to the pests' hibernation if they arrive later in the

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

summer (see Harned 1910). While the post-arrival coefficients slowly decay, they are still statistically significant even ten years after the arrival of the weevil. We find no evidence for anticipatory reporting in the four years prior to the USDA map's arrival date. For earlier periods, there are significant coefficients in the TWFE results. We find no pre-trends using the estimator by Sun and Abraham (2021).

Prediction of the Boll Weevil Infestation Using Historical Newspapers

To generate a stable prediction of the boll weevil's arrival based on newspaper data that is less prone to outliers or noise, we first apply a five-year moving average

$$MA(5)_{ct} = \frac{1}{5}\sum_{k=-2}^{2} \%BW_{c,t+k}$$

and then assign the maximum of this smoothed variable as predicted year of infestation

$$Predicted\ year\ of\ infestation_c = \max_{t \in [1882,1932]} (MA(5)_{ct}). \tag{3}$$

For robustness, we later test alternative specifications such as the three- and seven-year moving averages, as well as the maximum salience measure $\%BW_{c,t}$ within a ten-year window around the USDA map arrival date. While our preferred specification is $MA(5)$, the results in the replication exercises are robust across alternative specifications. More details are discussed next.

To illustrate how our approach based on newspapers can predict a county's effective infestation, consider the following example for Marion County in Mississippi. The USDA map recorded that the boll weevil arrived in Marion in 1909. However, the damage caused by the insect was not severe. Harned (1910), the head of the department and entomologist for the Mississippi Agricultural Experiment Station, investigated the infestation in Mississippi during 1907 and 1909. For Marion County, he found that "*boll weevils probably spread entirely over this county during September, 1909, although not in large enough number to do serious damage*" (Harned 1910, p. 22). For each year between 1882 and 1932, we first calculate the salience of the boll weevil of Marion County using pages mentioning "boll weevil" and "Marion County." We calculate $MA(5)_{Marion,t}$ for each year and define the effective infestation of Marion County by choosing the year with the maximum $MA(5)_{Marion,t}$. Our newspaper-based approach predicts that the effective infestation was in

1910 in Marion County, which is one year after the boll weevil's arrival in 1909, according to the USDA map. In panel (a) of Figure 3, we plot our newspaper-based boll weevil salience measure (dashed line) and the smoothed version using its five-year moving average (solid line) over time for an example county. While our salience measure based on newspapers is noisy, the five-year moving average smooths out this noise. Peak salience in the news appears to be a reasonable approximation for the arrival of the pest. The raw correlation of the two measures is 0.7, and Online Appendix Figure A.3 provides a visualization of this correlation with a binned scatter plot.

Lastly, we provide a comparison between our predicted arrival date from Equation (3) and that provided by the USDA map. Panel (b) of Figure 3 plots the difference in the two arrival dates for the 911 counties in our sample. A positive difference means that the predicted year based on newspapers is later than the arrival of the boll weevil as presented on the USDA map. While the difference is typically small, less than four years for more than half of the sample counties (54.88 percent), we find that the difference is extreme for a small number of counties. This result is likely due to the noise in the newspaper data, such as cases where the search words appear in separate articles even though they appear on the same newspaper page.[13] It should be kept in mind that our measure is, in some ways, purposefully noisy simply to reduce the cost of collecting the data. More refined versions are possible by applying a visual inspection of the newspaper data, which would increase the cost and time of the data collection process.

Another reason for some of the extreme values in the difference is due to some newly constructed counties. An example is shown in Online Appendix Figure A.4. Dixie County, in Florida, was created in 1921 from the southern portion of Lafayette County. While the boll weevil arrived in Dixie County in 1916, according to the USDA map, our newspaper-based measure predicts its effective infestation in 1932. This is because our prediction is based on newspapers mentioning "Dixie County." Since Dixie County did not exist before 1921, the prediction is mostly based on newspapers after 1921, which is shown in panel (a) of Online Appendix Figure A.4. One possible solution is to aggregate those counties as "multi-counties," as in Lange, Olmstead, and Rhode (2009) and Ager, Brueckner, and Herz (2017), or assign the predicted year from its original county.[14]

[13] Online Appendix Figure A.5 shows that the search word "boll weevil" appears in one article and "Marion County" appears in another article.

[14] For available crosswalks to standardize county boundaries over time, see Ferrara, Testa, and Zhou (2021).

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

(a) Example of Salience Measures for Marion County, Mississippi

(b) Distribution of Differences between Boll Weevil Measures

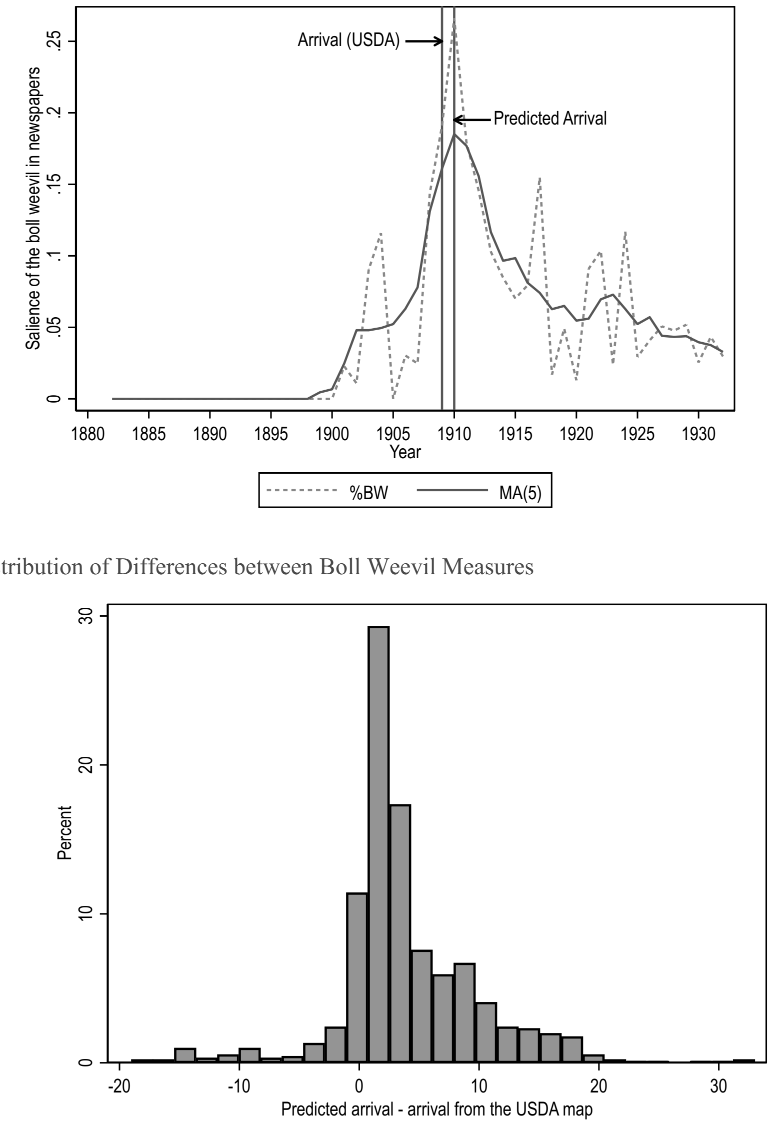

FIGURE 3

CHARACTERISTICS OF THE NEWSPAPER-BASED BOLL WEEVIL MEASURES

Notes: Panel (a) plots the newspaper-based boll weevil measure (dashed line) and its smoothed five-year moving average (solid line) for Marion County over time between 1882 and 1932. The vertical lines indicate the boll weevil's arrival from the USDA map and the predicted arrival, where *MA*(5) is the highest, respectively. Panel (b) shows a histogram of the distribution of the differences between the USDA map arrival year and the year of arrival predicted by the maximum of the *MA*(5) measure constructed from the newspaper data. This is a cross-sectional comparison between the two measures for the 911 counties in the South that were ever infested by the boll weevil to provide a summary measure of the average difference between the USDA and newspaper-based arrival date.

Sources: Authors' calculations from data in Hunter and Coad (1923) and Newspapers.com.

RESOLVING BIAS FROM MEASUREMENT ERROR USING SECONDARY MEASURES

Classical Measurement Error

How can the second measure for the boll weevil arrival from newspaper data be used to correct for measurement error on the USDA map arrival date? First, consider the case where the data is used as a continuous exposure measure, such as years since the arrival of the pest, for instance. Suppose a researcher wants to estimate the following linear equation by OLS, which is assumed to be unconfounded with a clear direction of causality but where the years since the arrival of the boll weevil, X_1, are continuous and measured with error,

$$y = \alpha + \beta X_1 + \epsilon \text{ and } X_1 = X^* + u,$$

where $Cov(X^*,u) = 0$, β is the true parameter, and X^* is the true measure (i.e., measured without error). The estimated coefficient will then suffer from the typical attenuation bias. Now suppose there is a second variable that seeks to capture X^* as well but that is also mismeasured, $X_2 = X^* + e$, and for which the same conditions apply as for X_1. We can then use X_2 as an instrument for X_1 to solve the measurement error problem (see Chalfin and McCrary 2018). The IV estimate will be

$$\hat{\beta}_{IV,X_1} = \beta \frac{Var(X^*)}{Var(X^*) + Cov(u,e)}, \tag{4}$$

where we denote the estimator and treatment variable of interest in the subscripts of $\hat{\beta}_{IV,X_1}$. In the absence of any other endogeneity problems and if the two measurement errors are uncorrelated such that $Cov(u,e) = 0$, the IV estimate will recover the true parameter. As with the exclusion restriction, one would then have to make an argument as to why the two errors should be uncorrelated or that this correlation is close to zero. In the case of the boll weevil, a possible argument would be that the USDA map was compiled by trained entomologists who primarily reported back to the agency, whereas the newspapers were written by journalists who reacted to local developments in their county. If journalists were basing their stories, and in particular the timing of their articles, on the USDA map, then this assumption fails, in which case $Cov(u,e) > 0$ and the estimated IV coefficient in Equation (4) would be biased downward.[15]

[15] For a method to deal with non-classical measurement errors in continuous variables see Hu and Schennach (2008).

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

Since applied economists tend to think hard about the exclusion restriction, we would like to highlight that this condition is satisfied in our case by assuming no endogeneity concerns other than measurement error. If X_2 affects y through channels other than X_1, such other channels must necessarily be in ϵ. Since X_2 and X_1 seek to measure the same quantity, this essentially also implies a correlation between X_1 and the error term as well. This is something that our approach cannot solve. At best, X_2 can remove biases relating to measurement error but not those stemming from omitted variables or reverse causality, for instance.

Non-Classical Measurement Error

Oftentimes, the arrival or presence of the boll weevil, however, is coded as a binary variable (e.g., Clay, Schmick, and Troesken 2019, 2020; Ager, Brueckner, and Herz 2017). In this case, the IV coefficient will no longer be unbiased because when the treatment variable is discrete or binary, measurement error is no longer classical by construction (Bingley and Martinello 2017).[16] Suppose that X_1 is now binary. When regressing y on X_1, the estimated OLS coefficient is still attenuated with $\hat{\beta}_{OLS,X_1} = \beta(1 - \theta)$, where θ is the misclassification rate in X_1 (Aigner 1973). If $\theta = 0$, then there is no measurement error, whereas $\theta = 1$ means that X_1 is entirely randomly misclassified, such that it is uncorrelated with X^* and therefore contains no usable information. Now suppose that X_2 is also binary and misclassified, but with an error γ that is uncorrelated with θ, and $\gamma < \theta$. If we then regress y on X_2, the estimated coefficient will also be biased, $\hat{\beta}_{OLS,X_2} = \beta(1 - \gamma)$, however, this attenuation bias will be smaller than for X_1 since $\beta(1 - \gamma) > \beta(1 - \theta)$ in absolute terms.

If we instrument X_1 with X_2, or vice versa X_2 with X_1, the estimated coefficient for those two cases will be

$$\hat{\beta}_{IV,X_1} = \beta\frac{1}{(1-\theta)} \text{and} \, \hat{\beta}_{IV,X_2} = \beta\frac{1}{(1-\gamma)}$$

depending on which variable was used as the treatment and the instrument. The IV bias is the inverse of the respective OLS bias.[17] Unlike

[16] A key assumption of classical measurement error is $Cov(X^*,u) = 0$, that is, the error is uncorrelated with the true value. Now suppose X^* is binary. If for a given observation $X^* = 1$, then the error can only be $u = –1$. Conversely, if $X^* = 0$ then $u = 1$, meaning that there is a perfect negative correlation between the true variable and the error.

[17] See Bingley and Martinello (2017) as well as Dupraz and Ferrara (2023) for measurement error in linked Census data. For a derivation, see the Online Appendix.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

OLS, which suffers from attenuation bias, the IV estimate will be inflated instead with $\beta\frac{1}{(1-\gamma)} < \beta\frac{1}{(1-\theta)}$.[18] Neither OLS nor IV yield an unbiased estimate; however, we now offer three potential approaches for identifying the treatment effect or for at least minimizing the attenuation coming from the misclassification.

Solution 1 - set identification: Even though the true parameter of interest cannot be directly point identified, the OLS and IV coefficients can be used as lower and upper bounds, respectively, to set identify β given that $\hat{\beta}_{OLS,X_1} < \hat{\beta}_{OLS,X_2} < \beta < \hat{\beta}_{IV,X_2} < \hat{\beta}_{IV,X_1}$. While it is not known a priori whether X_1 or X_2 has the higher measurement error, the inequality previously noted suggests that the set order can be inferred from the relative magnitudes of the OLS and IV coefficients. In the previous example, set identification implies that $\beta \in (\hat{\beta}_{OLS,X_2}, \hat{\beta}_{IV,X_2})$.Without additional assumptions, these bounds are tight and are informative as long as zero is not included in the set. To assess the latter condition, the OLS estimate provides the corresponding test that rejects non-informativeness when $\hat{\beta}_{OLS,X_2}$ is significantly different from zero.

Solution 2 - agreement sample: If instrumenting as described earlier is too complicated, for example, if researchers wish to estimate nonlinear treatment effects or their specification includes interactions of the treatment with other variables, the OLS bias can be reduced by considering only the part of the sample for which X_1 and X_2 both provide the same value. We call this an agreement sample.[19] The probability that both measures are jointly incorrect is $\theta \times \gamma = \delta$. For example, suppose the error rates are $\theta = 0.3$ and $\gamma = 0.2$, then $\delta = 0.06$, which substantially reduces the OLS bias for $\hat{\beta}_{OLS,X_1 = X_2} = \beta(1 - \delta)$, which will be closer to the true parameter.

Solution 3 - parametric bias correction: While neither OLS nor IV on their own identify the true parameter, their estimates can be used jointly to recover β. The bias-corrected (BC) estimate is

$$\hat{\beta}_{BC} = \sqrt{\hat{\beta}_{OLS,X_1} \times \hat{\beta}_{IV,X_1}} = \sqrt{\beta(1-\theta) \times \frac{1}{(1-\theta)}\beta} = \sqrt{\beta^2} = \beta. \tag{5}$$

Estimation of Equation (5) is straightforward, as the product of two coefficients from different equations can be readily estimated in standard

[18] Notice that this requires $\theta \neq 1$and $\gamma \neq 1$ as the IV estimator is not even defined otherwise.

[19] We provide a graphical illustration in Online Appendix Figure A.6. Suppose X_1 and X_2 assign treatment at time $t = 0$ and $t = 1$, respectively, where one is correct and the other is not, but it is unknown to the researcher which one value is true. The agreement sample excludes the shaded region of potential error that would bias the estimate. The coefficient estimate for the agreement sample is only based on the unbiased pre- and post-treatment periods, where both X_1 and X_2 report the same value of y.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

statistical software, with standard errors being estimated via the delta method or bootstrapping. One drawback of this bias correction is that it only works if both $\hat{\beta}_{OLS,X_1}$ and $\hat{\beta}_{IV,X_1}$ are of the same sign, which should be true in theory but may be violated in practice. This is another reason why we prefer the agreement sample as our main method of bias reduction. Taken together, our three possible solutions yield the following relationship,

$$\hat{\beta}_{OLS,X_1} < \hat{\beta}_{OLS,X_2} < \hat{\beta}_{OLS,X_1=X_2} < \beta = \hat{\beta}_{BC} < \hat{\beta}_{IV,X_2} < \hat{\beta}_{IV,X_1}, \qquad (6)$$

which is the pattern that we look for in the subsequent replication exercises.

Testing the Required Assumptions in Practice

A key assumption in our framework is that no other endogeneity issues aside from measurement error are present. A possible concern in this regard is that differential newspaper coverage could generate selectivity issues if such coverage correlates with problematic unobservables that also correlate with the outcome of interest. If this only affects the variable generated from the newspaper data, this has implications for the measurement error correction methods introduced in the previous section. Set identification $\beta \in (\hat{\beta}_{OLS,X_2}, \hat{\beta}_{IV,X_2})$ remains true as long as the bias in the IV is such that $\hat{\beta}_{IV,X_2} > \beta$. Bias in the opposite direction would imply a widening of the upper bound, making it less informative. Likewise, the parametric bias correction in Equation (5) will not recover the true parameter but an attenuated estimate if $\hat{\beta}_{IV,X_1} < \beta$, and an inflated estimate if the converse is true.

The method least affected by such biases is the agreement sample, which potentially generates a selected subsample that is not necessarily representative of the underlying population. One available correction is to apply inverse propensity score reweighting.[20] First, regress the indicator for being included in the agreement sample on a wide set of pretreatment county characteristics using a Probit regression. Second, obtain the predicted probability from the previous Probit regression. Lastly, run the regression of interest, weighting observations with the inverse of the estimated propensity score. The weights ensure that the estimation sample is more representative of observations in the entire sample.

Whichever method for bias correction is chosen by practitioners, they should always study whether differences between their original and the newspaper-based treatments are systematic by testing if pre-treatment characteristics can predict such differences. Table 1 provides an example of a

[20] A related application of this method is frequently employed when working with linked census data (see Bailey, Cole, and Massey 2019).

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

TABLE 1
TESTING FOR OBSERVABLE DETERMINANTS OF THE DIFFERENCES BETWEEN THE USDA AND NEWSPAPER-BASED BOLL WEEVIL ARRIVAL DATES

	No. of Years Difference	Difference > 1 Year	Difference > 2 Years
	(1)	(2)	(3)
Total population, 1890	–0.526	0.026	0.020
	(0.469)	(0.050)	(0.056)
Percent Black population, 1890	0.300	0.044	0.051
	(0.208)	(0.029)	(0.035)
Percent urban, 1890	0.049	–0.011	0.019
	(0.278)	(0.030)	(0.031)
Percent farmland in cotton, 1890	–0.212	–0.014	–0.038
	(0.186)	(0.026)	(0.033)
Number of farms per capita, 1890	–0.453**	0.006	–0.009
	(0.202)	(0.028)	(0.029)
Farm area, 1890	–0.041	–0.028	–0.011
	(0.239)	(0.027)	(0.030)
1(Late boll weevil cycle)	–2.020**	–0.220***	–0.050
	(0.796)	(0.079)	(0.087)
Pct. manufacturing empl., 1890	–0.298	–0.040	–0.047
	(0.267)	(0.035)	(0.033)
Ln(manufacturing wage per capita), 1890	0.045	0.019	0.000
	(0.233)	(0.030)	(0.031)
Total newspaper per capita, 1882–1932	0.341	–0.003	–0.034
	(0.315)	(0.035)	(0.041)
Obs.	627	627	627
Outcome mean	4.035	0.707	0.507
R-squared	0.500	0.254	0.286

Notes: Cross-sectional county-level regressions of the absolute difference in predicted boll weevil arrival year from the USDA map and the newspaper-based measure (Column (1)), and indicators for whether this difference is more than a year or two years (Columns (2) and (3)) on standardized county observables in 1890. Observable characteristics include total population (in 1,000), percent Black population, percent urban population, percent farmland in cotton, number of farms per capita, total acres in farms, an indicator for whether the boll weevil arrived late (i.e., if the arrival date is later than the average arrival across all counties), percent manufacturing employment, the log manufacturing wage per capita, and total newspaper pages in the Newspapers.com database by state (in 1,000,000) from 1882 to 1932 times 1890 county population. All observable characteristics are standardized to have mean zero and variance one, except indicator variables, such that coefficients can be interpreted in terms of a one standard deviation increase in the associated variable. Additional geographic controls include latitude, longitude, and state fixed effects. Robust standard errors in parentheses. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Sources: Authors' calculations from data in Hunter and Coad (1923), Haines (2010), and Newspapers.com.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

covariate balancing test where we regress the absolute value of the annual difference on the USDA map versus newspaper-based boll weevil arrival dates on various 1890 county-level characteristics. Variables that consistently generate significant coefficients in this exercise should be controlled for in the main regression of interest.[21] Depending on the context of their study, other tests and placebos may be applicable, and practitioners should think about possible implementations relevant to their setting.

REPLICATION OF CLAY, SCHMICK, AND TROESKEN (2019) AND AGER, BRUECKNER, AND HERZ (2017)

In this section, we replicate two recent papers that study the boll weevil's impacts on pellagra deaths (Clay, Schmick, and Troesken 2019) and cotton productivity (Ager, Brueckner, and Herz 2017). Implementing our suggested approaches to measurement error based on historical newspaper data, we demonstrate the potential for such data to markedly reduce attenuation bias. Our results suggest that the impact of the boll weevil was larger than previously documented. Further, our analysis largely confirms the ranked pattern for the different measurement error approaches as suggested by Equation (6) in the previous section. Results are robust across the alternative specifications discussed in the previous section.

Replication of Clay, Schmick, and Troesken (2019)

Using annual data between 1915 and 1925 for counties in North and South Carolina, Clay, Schmick, and Troesken (2019) show that pellagra deaths decreased following the boll weevil infestation. They argue that this outcome can be explained by the resulting diversification in food production. After the boll weevil infestation, the prevailing cotton monoculture was switched to more niacin-rich crops such as corn and sweet potatoes. This led to the fall of pellagra, which is a disease related to insufficient niacin consumption. Clay, Schmick, and Troesken (2019) estimate the following regression equation,

$$\ln\,[\mathit{pellagra}]_{ct} = \alpha + \theta_1 \mathit{boll\ weevil}_{ct} + \theta_2(\mathit{boll\ weevil}_{ct} \times \mathit{intensity}_{c,1909}) + \theta_c + \theta_t + \varepsilon_{ct}, \tag{7}$$

where $\ln[\mathit{pellagra}]_{ct}$ is the log number of pellagra deaths, or the log pellagra death rate in other specifications, and $\mathit{boll\ weevil}_{ct}$ is an indicator

[21] An additional layer of confidence could be created by performing additional tests for the sensitivity of the results to unobservables, such as the test developed by Oster (2019), comparing the outcome of the test in a regression that uses X_1 and a second regression that uses X_2 as the treatment variable.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

for whether or not the boll weevil has arrived in county c as of time t. They provide results with and without the additional interaction of the boll weevil variable and an intensity measure. The latter is an indicator for whether a county was in the top quartile of either (i) the pre-treatment pellagra death rates measured as average for 1915/16 or (ii) cotton acres per capita in 1909. County and year fixed effects are captured by θ_c and θ_t, and standard errors are clustered at the county-level.

Our Table 2 replicates the corresponding Table 3 in Clay, Schmick, and Troesken (2019) using the arrival date from the USDA map (X_2) and our predicted arrival from the newspaper data (X_1). We label the treatment variable used by Clay, Schmick, and Troesken (2019) as X_2, as the results presented in Table 2 suggest that, for their application, the map-based measure contains less measurement error than that based on our newspaper data.[22] Each column corresponds to different specifications in Table 3 of Clay, Schmick, and Troesken (2019). Columns (1)–(4) report the impact of the boll weevil on pellagra deaths, and Columns (5)–(8) repeat the same exercise using the log pellagra death rate as outcome. The table reports estimates of θ_1 in Equation (7), and we return to θ_2. The first row reports the OLS ($\hat{\beta}_{OLS,X_1}$) results for our newspaper-based arrival date treatment. These coefficient estimates are statistically significant and of the same sign as those provided by Clay, Schmick, and Troesken (2019), except for one statistically insignificant coefficient in Column (4) (same sign, p-value = .11). The second row for $\hat{\beta}_{OLS,X_2}$ is the replication of Table 3 in Clay, Schmick, and Troesken (2019). The following rows report the coefficient estimates for each specification using the agreement sample, the parametric bias correction, and the IV regressions, respectively. Due to the inclusion of the interaction term in Columns (2) to (4) and Columns (6) to (8), the bias-correction estimate using Equation (5) was only produced for the specifications in Columns (1) and (5). However, the agreement sample approach is still valid under the interaction term models. For the IV models, we follow the standard approach of using the interacted instrument to instrument for the interaction itself. While the IV interaction models do not technically fit the analysis in the theoretical section, the basic intuition still holds, and we believe that a comparison of the IV coefficients remains informative.

Focusing on the main effect, θ_1, we draw four main conclusions from our results. First, as might be expected, our newspaper-based arrival measure appears to be more noisy than that provided by the map. Nonetheless, we achieve similar, though smaller, results compared to those of Clay,

[22] This distinction is based on the relative differences between the OLS and IV estimates, as discussed in our theoretical section.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

TABLE 2
REPLICATION OF CLAY, SCHMICK, AND TROESKEN (2019)—MAIN EFFECTS

	Log Pellagra Deaths				Log Pellagra Death Rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\hat{\beta}_{OLS,X_1}$	–0.183***	–0.142*	–0.150**	–0.122	–0.151***	–0.113**	–0.144**	–0.125**
	(0.068)	(0.072)	(0.075)	(0.077)	(0.052)	(0.055)	(0.058)	(0.058)
$\hat{\beta}_{OLS,X_2}$	–0.283***	–0.197***	–0.237***	–0.202***	–0.235***	–0.161***	–0.212***	–0.185***
	(0.059)	(0.065)	(0.065)	(0.063)	(0.046)	(0.050)	(0.050)	(0.047)
$\hat{\beta}_{OLS,X_1=X_2}$	–0.396***	–0.310***	–0.333***	–0.278***	–0.326***	–0.251***	–0.295***	–0.256***
	(0.093)	(0.097)	(0.099)	(0.101)	(0.074)	(0.076)	(0.078)	(0.078)
$\hat{\beta}_{BC}$	–0.410***				–0.340***			
	(0.101)				(0.080)			
$\hat{\beta}_{IV,X_2}$	–0.595**	–0.460**	–0.427**	–0.346*	–0.493***	–0.371**	–0.401**	–0.346**
	(0.231)	(0.216)	(0.207)	(0.206)	(0.173)	(0.164)	(0.159)	(0.158)
$\hat{\beta}_{IV,X_1}$	–1.073***	–1.058***	–1.092***	–1.094***	–0.892***	–0.879***	–0.893***	–0.893***
	(0.260)	(0.275)	(0.259)	(0.269)	(0.208)	(0.221)	(0.202)	(0.207)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High pellagra		Yes				Yes		
BW × High cotton			Yes	Yes			Yes	Yes
Controls				Yes				Yes
Obs.	1,312	1,312	1,312	1,312	1,312	1,312	1,312	1,312
Counties	141	141	141	141	141	141	141	141
Obs. ($X_1 = X_2$)	1,051	1,051	1,051	1,051	1,051	1,051	1,051	1,051

Notes: Replication of Equation (1) in Clay, Schmick, and Troesken (2019) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). Columns (1) and (5) report OLS and IV regressions of deaths by pellagra on an indicator for whether the boll weevil has arrived in county *c*. The coefficients β_{BC} are estimated using Equation (5) and the delta method. The rest of the columns report OLS and IV regressions of deaths by pellagra on a boll weevil indicator and its interaction term with an indicator for whether county *c* was in the top 25 percent cotton production in 1909 (Columns (3), (4), (7), and (8)) or a dummy variable equal to one if county *c* was in the top 25 percent pellagra death rates in 1915/16 (Columns (2) and (6)). The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e., an agreement sample). In IV regressions, X_1 is instrumented with X_2 and vice versa. The sample is 141 counties in North Carolina and South Carolina between 1915 and 1925. All regressions include county and year fixed effects. Controls include county *c*'s malaria death rate in 1915 and the share of urban population in 1910, both of which interacted with a full set of year dummies. Standard errors are clustered at the county-level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sources: Authors' calculations from data in Hunter and Coad (1923), Clay, Schmick, and Troesken (2019), and Newspapers.com.

Schmick, and Troesken (2019). Thus, in the absence of the USDA map, Clay, Schmick, and Troesken (2019) could have successfully conducted their study using information from newspaper data alone—highlighting the usefulness of digitized historical newspapers as a potential data source for economic historians. Second, the relationship between the various coefficient estimates is consistent with the prediction provided in Equation (6) of our theoretical section. The pattern is more easily seen visually; hence, we provide a version of Column (1) of Table 2 as a bar chart in Online Appendix Figure A.7. Third, for all eight columns, coefficient estimates from the agreement sample and parametric bias correction models are on the order of 40–60 percent larger than the original estimates of Clay, Schmick, and Troesken (2019), suggesting marked gains from our measurement error corrections. Finally, we note that in the two cases where we can implement our parametric bias correction model, these coefficient estimates are quite similar in magnitude to the agreement sample estimates.

The earlier discussion focused on the estimated main effect, θ_1. To account for the interaction term, θ_2, in Table 3, we report the estimated marginal boll weevil impact for counties in the top 25th percentile of cotton production (Columns (3), (4), (7), and (8)) and pellagra deaths (Columns (2) and (6)).[23] These results mimic those from Table 2. In all models, we obtain slightly attenuated but significant results based solely on the newspaper data. The agreement sample estimates are highly significant and larger in magnitude than those reported by Clay, Schmick, and Troesken (2019). The pattern of the IV estimates exactly matches the predictions from our theoretical section. Additional results that implement propensity score reweighting for the agreement sample are provided in Online Appendix Table A.1.

Replication of Ager, Brueckner, and Herz (2017)

To further validate our approach, we replicate a second paper—that of Ager, Brueckner, and Herz (2017), which refines Lange, Olmstead, and Rhode (2009) by considering cotton intensity in each county.[24] They study the boll weevil's effect on Southern agriculture in terms of output, labor arrangements, and labor market outcomes using data from 13 Southern states between 1889 and 1929 in five- and ten-year intervals.[25] The authors show that the boll weevil reduced cotton output and productivity, the number of tenant farms, farm wages, and female labor force participation. They estimate the following linear regression model,

$$y_{ct} = \alpha_c + \beta_t + \gamma BollWeevil_{ct} + \delta BollWeevil_{ct} \times Cotton_{c,1889} + \epsilon_{ct}, \quad (8)$$

where y_{ct} is a given outcome variable for county c in a given five-year period t. As in the previous study, $BollWeevil_{ct}$ is an indicator of whether

[23] Here we are reporting on the linear combination $\theta_1 + \theta_2$. Thus, in Columns (1) and (5), we just replicate the exact results from Table 2.

[24] See Section 2.2 of Ager, Brueckner, and Herz (2017) for their reasoning. By replicating their paper, we are also essentially replicating the study by Lange, Olmstead, and Rhode (2009).

[25] These are 1889, 1899, 1909, 1919, 1924, and 1929.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

TABLE 3
REPLICATION OF CLAY, SCHMICK, AND TROESKEN (2019)—MARGINAL EFFECTS AT THE 75TH PERCENTILE

	Log Pellagra Deaths				Log Pellagra Death Rate			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\hat{\beta}_{OLS,X_1}$	–0.183***	–0.278***	–0.259***	–0.253***	–0.151***	–0.241***	–0.169***	–0.169***
	(0.068)	(0.096)	(0.083)	(0.084)	(0.052)	(0.070)	(0.059)	(0.058)
$\hat{\beta}_{OLS,X_2}$	–0.283***	–0.531***	–0.442***	–0.469***	–0.235***	–0.452***	–0.314***	–0.335***
	(0.059)	(0.086)	(0.083)	(0.080)	(0.046)	(0.067)	(0.063)	(0.060)
$\hat{\beta}_{OLS,X_1=X_2}$	–0.396***	–0.652***	–0.603***	–0.595***	–0.326***	–0.551***	–0.429***	–0.428***
	(0.093)	(0.120)	(0.110)	(0.110)	(0.074)	(0.094)	(0.084)	(0.082)
$\hat{\beta}_{BC}$	–0.410***				–0.340***			
	(0.101)				(0.080)			
$\hat{\beta}_{IV,X_2}$	–0.595**	–0.806***	–0.817***	–0.750***	–0.493***	–0.682***	–0.613***	–0.579***
	(0.231)	(0.280)	(0.269)	(0.268)	(0.173)	(0.205)	(0.199)	(0.196)
$\hat{\beta}_{IV,X_1}$	–1.073***	–1.476***	–1.221***	–1.271***	–0.892***	–1.247***	–0.900***	–0.938***
	(0.260)	(0.280)	(0.246)	(0.247)	(0.208)	(0.227)	(0.188)	(0.187)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High pellagra		Yes				Yes		
BW × High cotton			Yes	Yes			Yes	Yes
Controls				Yes				Yes
Obs.	1,312	1,312	1,312	1,312	1,312	1,312	1,312	1,312
Counties	141	141	141	141	141	141	141	141
Obs. ($X_1 = X_2$)	1,051	1,051	1,051	1,051	1,051	1,051	1,051	1,051

Notes: Replication of Equation (1) in Clay, Schmick, and Troesken (2019) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). Columns (1) and (5) report OLS and IV regressions of deaths by pellagra on an indicator for whether the boll weevil has arrived in county *c*. The coefficients β_{BC} are estimated using Equation (5) and the delta method. The rest of the columns report OLS and IV regressions of deaths by pellagra on a boll weevil indicator and its interaction term with an indicator for whether county *c* was in the top 25 percent cotton production in 1909 (Columns (3), (4), (7), and (8)) or a dummy variable equal to one if county *c* was in the top 25 percent pellagra death rates in 1915/16 (Columns (2) and (6)). The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e., an agreement sample). In IV regressions, X_1 is instrumented with X_2 and vice versa. The sample is 141 counties in North Carolina and South Carolina between 1915 and 1925. All regressions include county and year fixed effects. Controls include county *c*'s malaria death rate in 1915 and the share of urban population in 1910, both of which interacted with a full set of year dummies. Standard errors are clustered at the county-level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sources: Authors' calculations from data in Hunter and Coad (1923), Clay, Schmick, and Troesken (2019), and Newspapers.com.

a county is infested in the current five-year period. $Cotton_{c,1889}$ is the demeaned acreage share of cotton planted in 1889 as a measure of cotton intensity. County and time fixed effects are captured by α_c and β_t, and standard errors are again clustered at the county-level. Because Ager, Brueckner, and Herz (2017) estimate models incorporating interaction terms in all specifications, we are not able to implement the bias correction model, β_{BC}, and we thus focus attention on the agreement sample results as our preferred model.

Table 4 reports the resulting γ coefficients from estimating Equation (8).[26] Ager, Brueckner, and Herz (2017) find significant main effects in seven of the 12 models that they estimate. Using only our newspaper data, we also find significant results in each of these seven models—with our newspaper-based coefficient estimates being larger in magnitude for all but two of these models. The newspaper data leads to significant estimates of the main effect in three of the five models, where Ager, Brueckner, and Herz (2017) find no effect. For this reason, we keep the same notation in terms of X_1 and X_2 as in Tables 2 and 3 (with X_1 reflecting the newspaper-based data). In six of the seven models where Ager, Brueckner, and Herz (2017) find statistically significant main effects the agreement sample point estimates, $\beta_{X_1=X_2}$, are larger in magnitude than those based on either the map data or the newspaper data—the exception being the estimated effect on corn yield in Column (7). Notice that in all seven models, the overall pattern of the OLS and IV estimates matches the predictions of Equation (6). The only exception is Column (7), where the agreement sample estimate is slightly below that of the map-based OLS estimate.

To account for the continuous interaction terms in Ager, Brueckner, and Herz (2017), in Table 5 we present estimated marginal effects at the 75th percentile of cotton production.[27] The newspaper-based treatment yields significant OLS results in eight of the nine cases where the map-based data gives significant results. In five of these cases, the newspaper-based data leads to larger OLS estimates. The newspaper data also leads to significant OLS results in the three models, which were insignificant when using the map-based data. Estimates using the agreement sample were again larger in magnitude than either newspaper-based or map-based OLS estimates in ten of the 12 models. Within the eight models where both data sets have predictive power, agreement sample estimates are on average 37 percent larger than the original estimates of Ager, Brueckner, and Herz (2017). Additional results that implement the propensity score reweighting for the agreement sample are provided in Online Appendix Table A.2.

DISCUSSION OF PRACTICAL ISSUES AND FURTHER APPLICATIONS

Potential Gains, Future Applications, and Drawbacks

The replications have shown that newspaper data can be gainfully used for bias reduction in statistical analyses using historical data. We also

[26] Because Ager, Brueckner, and Herz (2017) demean the cotton production data before constructing their interaction measures, γ represents that marginal effect at the mean level of cotton production.

[27] The table summarizes the linear combination $\gamma + 0.165 \times \delta$.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

TABLE 4
REPLICATION OF AGER, BRUECKNER, AND HERZ (2017)—MAIN EFFECT

	Log Cotton Production				Log Corn Production				Log Other Outcomes			
	Bales (1)	Acres (2)	Yield (3)	Share (4)	Bushels (5)	Acres (6)	Yield (7)	Share (8)	Farm (9)	Farm Value (10)	Pop. (11)	Black Pop. (12)
$\hat{\beta}_{OLS,X_1}$	–0.486***	–0.245***	–0.248***	–0.066***	0.037	0.077***	–0.040**	0.053***	–0.025**	–0.038**	0.024*	–0.028
	(0.057)	(0.052)	(0.022)	(0.006)	(0.029)	(0.019)	(0.017)	(0.006)	(0.011)	(0.015)	(0.013)	(0.034)
$\hat{\beta}_{OLS,X_2}$	–0.386***	–0.173***	–0.208***	–0.061***	–0.032	0.045**	–0.077***	0.064***	–0.000	0.005	–0.002	0.028
	(0.055)	(0.049)	(0.024)	(0.007)	(0.037)	(0.023)	(0.021)	(0.007)	(0.012)	(0.018)	(0.014)	(0.030)
$\hat{\beta}_{OLS,X_1=X_2}$	–0.651***	–0.290***	–0.360***	–0.091***	0.055	0.118***	–0.062**	0.087***	–0.028**	–0.023	0.001	–0.028
	(0.060)	(0.054)	(0.025)	(0.008)	(0.041)	(0.023)	(0.026)	(0.007)	(0.013)	(0.017)	(0.016)	(0.042)
$\hat{\beta}_{IV,X_2}$	–1.069***	–0.532***	–0.551***	–0.141***	0.090	0.177***	–0.087**	0.114***	–0.052**	–0.075**	0.080**	–0.048
	(0.123)	(0.112)	(0.047)	(0.014)	(0.066)	(0.043)	(0.039)	(0.014)	(0.026)	(0.037)	(0.034)	(0.079)
$\hat{\beta}_{IV,X_1}$	–0.939***	–0.485***	–0.447***	–0.135***	–0.074	0.091**	–0.165***	0.139***	–0.005	0.001	–0.024	0.051
	(0.106)	(0.094)	(0.048)	(0.014)	(0.076)	(0.046)	(0.044)	(0.014)	(0.025)	(0.036)	(0.033)	(0.071)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High cotton	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weather controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	4,323	4,329	4,323	4,440	4,440	4,440	4,440	4,440	4,440	4,440	3,700	3,679
Counties	735	735	735	740	740	740	740	740	740	740	740	739
Obs. ($X_1 = X_2$)	3,927	3,933	3,927	4,032	4,032	4,032	4,032	4,032	4,032	4,032	3,328	3,311

Notes: Replication of Equation (1) in Ager, Brueckner, and Herz (2017) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). OLS and IV regressions of agricultural and demographic outcome variables on an indicator for whether the boll weevil has arrived in county *c* and its interaction term with county *c*'s demeaned acreage share of cotton in 1889. The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e., an agreement sample). In the IV regressions, X_1 is instrumented with X_2 and vice versa. The sample includes counties in the U.S. South between 1889 and 1929. All regressions include county and year fixed effects as well as weather controls. Weather controls are January's mean temperature and average summer precipitation from May to July. Standard errors are clustered at the county-level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sources: Authors' calculations from data in Hunter and Coad (1923), Ager, Brueckner, and Herz (2017), and Newspapers.com.

TABLE 5
REPLICATION OF AGER, BRUECKNER, AND HERZ (2017)—MARGINAL EFFECTS AT THE 75TH PERCENTILE

	Log Cotton Production				Log Corn Production				Log Other Outcomes			
	Bales (1)	Acres (2)	Yield (3)	Share (4)	Bushels (1)	Acres (2)	Yield (3)	Share (4)	Farm (1)	Farm Value (3)	Pop. (3)	Black Pop. (4)
$\hat{\beta}_{OLS,X_1}$	–1.008*** (0.058)	–0.713*** (0.054)	–0.311*** (0.023)	–0.116*** (0.007)	–0.006 (0.031)	0.056*** (0.021)	–0.062*** (0.018)	0.086*** (0.007)	–0.051*** (0.014)	–0.106*** (0.019)	–0.073*** (0.017)	–0.099** (0.038)
$\hat{\beta}_{OLS,X_2}$	–0.918*** (0.063)	–0.638*** (0.058)	–0.287*** (0.025)	–0.108*** (0.008)	–0.072* (0.039)	0.038 (0.026)	–0.110*** (0.021)	0.098*** (0.008)	–0.026 (0.017)	–0.039* (0.021)	–0.076*** (0.018)	–0.028 (0.033)
$\hat{\beta}_{OLS,X_1=X_2}$	–1.223*** (0.066)	–0.796*** (0.061)	–0.437*** (0.026)	–0.143*** (0.009)	0.011 (0.044)	0.103*** (0.027)	–0.092*** (0.026)	0.122*** (0.008)	–0.057*** (0.018)	–0.086*** (0.022)	–0.094*** (0.020)	–0.104** (0.047)
$\hat{\beta}_{IV,X_2}$	–1.712*** (0.119)	–1.098*** (0.109)	–0.639*** (0.048)	–0.203*** (0.015)	0.040 (0.067)	0.156*** (0.043)	–0.115*** (0.039)	0.156*** (0.015)	–0.084*** (0.026)	–0.158*** (0.039)	–0.034 (0.033)	–0.133 (0.082)
$\hat{\beta}_{IV,X_1}$	–1.566*** (0.113)	–1.037*** (0.100)	–0.537*** (0.048)	–0.190*** (0.016)	–0.121 (0.077)	0.082* (0.048)	–0.202*** (0.044)	0.178*** (0.015)	–0.036 (0.030)	–0.052 (0.040)	–0.119*** (0.036)	–0.022 (0.071)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
BW × High cotton	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weather controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	4,323	4,329	4,323	4,440	4,440	4,440	4,440	4,440	4,440	4,440	3,700	3,679
Counties	735	735	735	740	740	740	740	740	740	740	740	739
Obs. ($X_1 = X_2$)	3,927	3,933	3,927	4,032	4,032	4,032	4,032	4,032	4,032	4,032	3,328	3,311

Notes: Replication of Equation (1) in Ager, Brueckner, and Herz (2017) using the boll weevil's arrival from the USDA map (X_2) and the predicted arrival based on newspapers (X_1). OLS and IV regressions of agricultural and demographic outcome variables on an indicator for whether the boll weevil has arrived in county *c* and its interaction term with county *c*'s demeaned acreage share of cotton in 1889. The coefficients $\beta_{OLS,X_1=X_2}$ are estimated using a subset of the sample for which X_1 and X_2 both provide the same value (i.e., an agreement sample). In the IV regressions, X_1 is instrumented with X_2 and vice versa. The sample includes counties in the U.S. South between 1889 and 1929. All regressions include county and year fixed effects as well as weather controls. Weather controls are January's mean temperature and average summer precipitation from May to July. Standard errors are clustered at the county-level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sources: Authors' calculations from data in Hunter and Coad (1923), Ager, Brueckner, and Herz (2017), and Newspapers.com.

found that the predictions based on the inequality in Equation (6) tend to hold up in applied examples. The gains in bias reduction appear to have been larger in the replication of Ager, Brueckner, and Herz (2017) as compared to the replication of Clay, Schmick, and Troesken (2019). While we cannot offer a definitive explanation for this finding, a possible reason seems to be the difference in the frequency of the time dimension. The study by Clay, Schmick, and Troesken (2019) uses annual data, a much higher frequency than the five-year intervals in Ager, Brueckner, and Herz (2017), which potentially mitigated some of the measurement error bias. Nonetheless, results in both papers held up in our replications and could be strengthened using our methods.

Our newspaper-based boll weevil arrival measure was generated in a fast and low-cost way. Compared to the USDA measure used by Clay, Schmick, and Troesken (2019), it appears to be more noisy, which is to be expected. It would certainly be possible to refine the measure, but doing so would increase the time and cost of collecting the information. What we want to highlight instead is that our very coarse measure still managed to produce very similar results in the two replications, meaning that both studies could have been conducted had the USDA map never existed. For the purpose of the methods introduced in this paper, it does not matter whether the data from the newspapers or the original variable (here the USDA map arrival date) is noisier as long as the measurement errors in the two variables are uncorrelated. This assumption cannot be directly tested, just as the exclusion restriction in instrumental variable regressions, for instance, but institutional knowledge and the robustness checks suggested in previous sections should help to increase our confidence in this assumption. In the boll weevil case, we also argued that this assumption holds because newspapers reported any boll weevil-related events that were observed by newspaper reporters, whereas the USDA map was created by federal entomologists. The report by Hunter and Coad (1923), for which the USDA map was created, does not contain the words "newspaper," "news," "article," or "journalist." Conversely, searching Newspapers.com jointly for "USDA" and "boll weevil" only returned 59 hits. However, the majority of those hits were due to transcription errors by the character recognition software.[28]

Our approach is particularly suited for measures that can be easily generated or extracted using textual data. Simple n-gram or bag-of-words approaches, as in Beach, Clay, and Saavedra (2022), Ferrara and

[28] For example, the word "Wednesday" was flagged as a match with "USDA" in one article.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

Fishback (2023), Albright et al. (2021), Beach and Hanlon (2023), Bazzi et al. (2023), or Ottinger and Winkler (2022), are particularly promising. Anything that can be measured or extracted with a single search word or a combination of a few words lends itself to this approach and the generation of newspaper-based data. For variables such as prices, this approach is less promising because these can rarely be extracted in a low-cost way as they oftentimes require more careful extraction by hand. Generation of data from newspaper articles is likely impractical for variables that would not ordinarily be reported in the news or for which the non-random nature of the availability of digitized newspapers might be a concern. For example, measures relating to corruption or trade might be more difficult to find in newspapers. Large-scale or salient events tend to be covered in newspapers, and our boll weevil infestation example fits into this category as Lange, Olmstead, and Rhode (2009, p. 685) noted: "*the boll weevil is America's most celebrated agricultural pest.*" Other examples of such salient events studied in previous literature are the 1918 influenza pandemic (Beach, Clay, and Saavedra 2022), natural disasters across the United States (Boustan et al. 2020), labor strikes (Schmick 2018), the Tulsa race massacre in 1921 (Albright et al. 2021), or the Bradlaugh-Besant trial of 1877 (Beach and Hanlon 2023).[29] Among these examples, studies of the 1918 influenza, for instance, that likely could have gainfully applied our methods are Almond (2006), Hatchett, Mecher, and Lipsitch (2007), Hilt and Rahn (2020), or Beach, Clay, and Saavedra (2022), all of whom use an intensity measure of the flu at the local level.[30]

Newspaper information can also be used to generate data at the sub-county-level. Most online archives report the city, town, or place of publication. The data can then be combined with newly available crosswalks to sub-county locations for every individual in the census and consistently defined place names that are provided by the Census Place Project (Berkes, Karger, and Nencka 2023). Also, practitioners can simply search for newspaper pages containing the names of any sub-county areas for which they need to collect data. We provide an example using 960 sub-county areas (hereinafter towns) in North Carolina from Berkes, Karger, and Nencka (2023). Using newly scraped data for all pages from North

[29] The cited studies of the Tulsa race massacre and the Bradlaugh-Besant trial could not directly apply our methods because their only measure quantifying exposure to the events they study is already drawn from newspaper data. However, they could apply our methods with a different data source on exposure to the events, including measures based on a different newspaper archive.

[30] For a review of work related to the 1918 influenza pandemic, see Beach et al. (2022).

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

Carolina newspapers that mention "boll weevil" and each town's name, we compute the following town-level measure,

$$\%BW_{ot} = \frac{\textit{No. of in} - \textit{state newspaper pages mentioning "boll weevil" and a town's name}_{ot}}{\textit{No. of in} - \textit{state newspaper pages mentioning a town's name}_{ot}}, \quad (9)$$

where $\%BW_{ot}$ now captures the newspaper-based salience measure for the boll weevil in town o in year t. We then estimate the following equation,

$$\%BW_{ot} = \lambda_o + \theta_t + \sum_{\ell=-10}^{-2} \beta_\ell \cdot D(t - BW_{o(c)}^{USDA} = \ell) + \sum_{\ell=0}^{10} \beta_\ell \cdot D(t - BW_{o(c)}^{USDA} = \ell) + \epsilon_{ot}, \quad (10)$$

where $D(t - BW_{o(c)}^{USDA} = \ell)$ is an event indicator relative to the arrival of the boll weevil in town o in county c from the USDA map. Since the USDA map only provides the arrival date at the county-level, we assume that the map-based arrival date for town o is the same as its county c. The year before the arrival from the USDA map, $\ell = -1$, serves as the baseline period. We include town fixed effects λ_o and year fixed effects θ_t, as opposed to county and state-by-year fixed effects in Equation (2). Standard errors are clustered at the town-level.

The result is shown in Online Appendix Figure A.8. Similar to our county-level analysis, newspaper analysis at the town-level finds that salience increases significantly after the arrival of the boll weevil in a given town's county (as shown on the USDA map).[31] We also replicate the analysis in Figure 3 using three distinct towns in Alamance County, North Carolina. We find that the town-level salience measures strikingly resemble their county-level salience measure. Online Appendix Figure A.9 shows that the salience measures of Melvile (a township), Burlington (a city), and Patterson (an unincorporated community) follow a similar pattern to that of their county. Each town-level salience measure shows a small increase around 1904 and its peak around 1923. Furthermore, the maximum of $MA(5)$ predicts that all three towns were infested by the boll weevil between 1922 and 1923. These predicted years using towns are comparable to the arrival of the boll weevil in Alamance County based on the USDA map (1922) and our newspaper-based approach (1923). Notice that we do not take a stance regarding the interpretation of a boll weevil measure at the town-level since the weevil was mainly an issue

[31] Online Appendix Figure A.8 also reports county-level analysis based solely on the North Carolina data. Here the county-level salience measure is based on 77 infested counties in North Carolina using Equation (2) with county and year fixed effects instead of county and state-by-year fixed effects.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

in the country side. Newspapers in towns most likely reported about the surrounding areas and not the towns themselves. The exercise here is mainly to highlight the potential usefulness of generating town-level data from digitized newspaper archives.

Practitioners must also be aware of other flaws and shortcomings affecting digitized newspaper archives. These archives do not contain the universe of all newspapers in the United States, and they also do not contain the universe of all articles. Papers from more populated places, such as larger cities, tend to be overrepresented. Particular states, such as Massachusetts, are poorly represented on Newspapers.com. Beach and Hanlon (2022) discuss these issues in more detail and provide potential solutions for attrition and sample selectivity in the context of digitized newspaper archives using newspaper directories and other external sources. Even though newspaper data can be generated at the sub-county-level and at high time frequencies, the trade-off is that increased granularity comes at the expense of a higher chance of missingness in the data and noise.

Generalizing the Method to Other Settings

Both of our replications have focused on the boll weevil. This was to demonstrate that success in reducing bias in one study by employing our newspaper-based approach was not merely a fluke. However, one remaining question is how well the methods developed in this paper extend to other settings. We therefore replicated two additional studies where the treatment variables of interest are conceptually different in nature than the arrival of the boll weevil. The first of these two additional examples is a replication of the reduced form regression in Hilt and Rahn (2020), whose right-hand side variable is a county-level measure for the average distance to the nearest military camps that seeks to proxy for the severity of the 1918 influenza epidemic.[32] Even though a distance- rather than an arrival-based measure is conceptually different from our first set of replications, one may wonder whether our setting is solely applicable to natural events, such as agricultural pests or diseases, that spread in potentially similar fashions. We therefore also consider a human-made policy, namely the county-level adoption of prohibition policies in the early twentieth century, which were studied by Howard and Ornaghi (2021) with regards to the impact of such policies on population, farming,

[32] They study the effect of the liberty bond program on the county-level Democratic vote share and use distance to military camps during WWI as a proxy for the severity of the 1918 influenza epidemic to instrument for a county's liberty bond participation rate.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

and investment outcomes. The Hilt and Rahn (2020) measure was a proxy to start with; hence, an argument for how a secondary measure can be helpful is easy to imagine. The prohibition adoption data used by Howard and Ornaghi (2021) originally came from Sechrist (2012). In Online Appendix Figure A.10, we document cases of counties that were reported as being dry in newspaper articles but that were recorded as non-dry in the Sechrist data.

To generate a measure for Spanish flu severity from newspaper data, we searched Newspapers.com for articles containing the search words “flu” and the county name within each state, as before. After standardizing this measure by the total number of newspaper pages mentioning each county name, we then considered areas to be hotspots of the 1918 influenza if they were in the top decile of this measure. Lastly, we computed the average distance to influenza hotspots for each county to mimic the average distance to military camps proxied by Hilt and Rahn (2020). We use both the continuous distance measure as well as a binarized version that is equal to one for distances above the median distance. For the prohibition measure, we use the search terms “prohibition” and “dry” together with the county name in each year between 1890 and 1919 and divide this count variable by the total search hits for the county name in each year. We then predict the adoption of prohibition in each county by using the maximum of the five-year moving average of the share. We provide more detailed descriptions of how we generated our newspaper-based measures for the Spanish flu intensity distance measure and prohibition adoption in Online Appendix A.2.

Table 6 reports the results of these two replications. Column (1) shows our results using the Chalfin and McCrary (2018) approach for the continuous distance measure, and Column (2) shows our approach using the binary median split variable. Row 2 ($\hat{\beta}_{OLS,X_2}$) of Column (1) replicates the corresponding results in Online Appendix Table A.6 Column (2) in Hilt and Rahn (2020) with a coefficient of –0.473 (s.e. = 0.218). Row 1 ($\hat{\beta}_{OLS,X_1}$) of the same table estimates a coefficient of –0.565 (s.e. = 0.235), which uses our newspaper-based measure. This confirms that the original influenza proxy used by Hilt and Rahn (2020) was very close to other measures of influenza severity. When instrumenting their distance to military camp variable with our newspaper-based influenza measure, we estimate a coefficient of –0.627 (s.e. = 0.264), which is larger in absolute terms as the theory in Chalfin and McCrary (2018) would predict. The same is true when using the binarized version of the distance measure, where we can now also apply our preferred approach to reduce measurement error by using an agreement sample. Here, the Democratic vote

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

TABLE 6
REPLICATION OF HILT AND RAHN (2020) AND HOWARD AND ORNAGHI (2021)

	Hilt and Rahn (2020)		Howard and Ornaghi (2021)					
	Dem. Share (1)	Dem. Share (2)	Pop. (3)	Farm Value (4)	Productivity (5)	Implements (6)	Farm Share (7)	Banks (8)
$\hat{\beta}_{OLS,X_1}$	–0.565**	–1.437**	0.016	0.063**	0.048	–0.001	0.007	0.018
	(0.235)	(0.558)	(0.029)	(0.032)	(0.043)	(0.027)	(0.007)	(0.013)
$\hat{\beta}_{OLS,X_2}$	–0.473**	–3.555***	0.092**	0.127***	0.066	0.125***	0.013	0.022
	(0.218)	(0.854)	(0.037)	(0.036)	(0.053)	(0.033)	(0.008)	(0.016)
$\hat{\beta}_{OLS,X_1=X_2}$		–4.905***	0.066	0.269***	0.243***	0.209***	0.043**	0.069**
		(1.208)	(0.071)	(0.064)	(0.094)	(0.052)	(0.017)	(0.028)
$\hat{\beta}_{BC}$		–4.931	0.165	0.393	0.253		0.042	0.084
		(27.771)	(1.888)	(2.240)	(2.183)		(0.514)	(0.337)
$\hat{\beta}_{IV,X_2}$	–0.627**	–10.269**	0.295	1.210*	0.967	–0.026	0.140	0.328
	(0.264)	(4.887)	(0.544)	(0.722)	(0.936)	(0.519)	(0.147)	(0.260)
$\hat{\beta}_{IV,X_1}$	–0.492**	–16.956***	1.275*	1.778*	0.981	1.745*	0.179	0.289
	(0.226)	(5.395)	(0.764)	(0.928)	(0.892)	(0.904)	(0.139)	(0.241)
Treatment	Dist. mil.	Dist. mil.	Dry law	Dry law	Dry law	Dry law	Dry law	Dry law
Treatment binary		Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
State × Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	9,854	9,854	4,356	4,329	4,290	4,329	4,329	4,344
Counties	1,426	1,426	1,452	1,443	1,430	1,443	1,443	1,448
Obs. ($X_1 = X_2$)		8,587	3,746	3,724	3,690	3,724	3,724	3,738

Notes: Columns (1) and (2) replicate the reduced-form results in Online Appendix Table A.6 of Hilt and Rahn (2020) using the average distance to military camps (X_2) and the average distance to influenza hotspots based on newspapers (X_1). OLS and IV regressions of the Democratic Party vote share on a continuous (Column (1)) and binary (Column (2)) measure of the 1918 influenza epidemic. Both regressions are weighted by population in 1920. Columns (3)–(8) replicate Equation (1) in Howard and Ornaghi (2021) using the introduction of Prohibition from Sechrist (2012) (X_2) and the predicted year of adoption based on newspapers (X_1). The sample only includes counties that adopted Prohibition between 1900 and 1919, both in Sechrist (2012) and our newspaper data. Columns (3)–(8) report OLS and IV regressions of economic outcome variables on an indicator for whether county *c* adopted prohibition after 1900 but before 1910 interacted with an indicator for the post period. The coefficients β_{BC} are estimated using Equation (5) and the delta method. All regressions include county and state-by-year fixed effects as well as controls. Controls in Columns (1) and (2) are the share of population in urban areas and home ownership rate. Controls in Columns (3)–(8) include baseline religiosity and demographics. See Howard and Ornaghi (2021) for details. Standard errors are clustered at the county-level. Significance levels are denoted by * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.
Sources: Authors' calculations from data in Hunter and Coad (1923), Hilt and Rahn (2020), Howard and Ornaghi (2021), and Newspapers.com.

share is predicted to decline by 3.56 percentage points if a county had an above-median military camp distance. This coefficient is significant at the 1 percent level. When a county with an above-median military camp distance also had an above-median distance to the nearest influenza hotspots (i.e., if it was in the agreement sample), then the estimated reduction in the Democratic vote share was 4.91 percentage points.

Columns (3) to (8) in Table 6 report the results from the replication of Howard and Ornaghi (2021). The main takeaway from this exercise is that the agreement sample generates a significantly larger result for the estimated coefficients. We focus on results from the agreement sample, given that many of the instrumented coefficients are only noisily estimated. This highlights that certain measures are more precisely approximated with newspaper data, such as measures of distance or arrival dates and locations, especially when they are saliently featured in the news. Turning to the results, when considering log value of farm implements and log farm values as outcomes, the agreement sample estimates an effect of local prohibition policies of 0.209 (s.e. = 0.052) for the log value of farm implements and of 0.269 (s.e. = 0.064) for log farm values. This is 1.6 and 2.1 times larger than the estimates from using the Sechrist (2012) prohibition data. These are large effects that may seem implausible a priori, and Howard and Ornaghi (2021, p. 813) say little that puts their estimates in perspective other than the following: "*The increase in productivity is consistent with increased investment in labor-saving technology. The early twentieth century was a time of increased mechanization.*" However, when we dug deeper into the topic ourselves, we found a contemporaneous paper in the *Quarterly Journal of Economics* by Coulter (1912, p. 11), who found that: "*In 1900 the average value of all farm property per acre of land in farms was $24.37; in 1910 it was $46.64. This is an increase of 91.4 per cent during the decade.*" Considering the stark developments in American agriculture and land values at the time, a prohibition-induced farm value increase of 26.9 log points therefore appears much more reasonable. In summary, Howard and Ornaghi (2021) were potentially able to explain much more of the change in farm values at the time with their prohibition hypothesis than what their initial study had shown.

CONCLUSION

Measurement error in historical data is often a source of bias in statistical analyses that leads to attenuation bias in the relationships that researchers seek to identify. When measurement error is classical, it is known that this attenuation bias can be removed via an instrumental variable approach. A potential instrument is a second measure of the same variable with errors, as long as the errors in two variables are uncorrelated (Chalfin and McCrary 2018). Generating such a second measure tends to be expensive, and therefore measurement error tends to be ignored as long as some conventional level of statistical significance is achieved.

In this paper, we introduce the idea of inexpensively generating such a second measure from digitized newspapers, which can be scraped or

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

downloaded at low costs. We show how a newspaper-based secondary measure can be used to deal with measurement error when the variable of interest is either continuous or binary. The latter case is more challenging since measurement error in a binary variable is non-classical by construction, and therefore, an instrumental variable approach alone does not remove the associated bias (Bingley and Martinello 2017). Instead, we propose three alternative methods for dealing with measurement error in this setting based on (i) set identification, (ii) using an agreement sample where both the primary and secondary measure give the same answer, and (iii) a parametric bias correction that can be obtained as a nonlinear combination of the OLS and IV coefficients. Our theory predicts that OLS and IV provide the lower and upper bounds of the identified set that include the true parameter, and that the coefficients from the agreement sample and the parametric bias correction should lie in between these bounds. Also, the bias-corrected estimate should still be larger in magnitude than the OLS coefficient from the agreement sample.

To test this prediction as well as to showcase our methods, we replicate two recent papers by Clay, Schmick, and Troesken (2019) and Ager, Brueckner, and Herz (2017) on the impact of the boll weevil infestation in the U.S. South between 1892 and 1922. Like most studies on the boll weevil, the main treatment is measured from a map of the pest by Hunter and Coad (1923), which arguably is measured with error because of crossing lines and given that the arrival dates are an imperfect measure of the economic impact of the beetle. To produce a second measure for the boll weevil's arrival from digitized newspaper data, we scrape Newspapers.com and search for pages that mention "boll weevil" and each county's name from all newspapers in the county's state. This approach maximizes the chance to find articles related to the arrival of the weevil in that county. In both replications, we find larger coefficients than in the original studies that show the usefulness of our approach to dealing with measurement error and also reaffirm the main results of the two papers. In both cases, we also find the patterns prescribed in the theoretical section, where plain OLS yields the smallest coefficient, followed by the agreement sample and the parametric bias correction.

The main contribution of the paper is to provide an easy way to generate a secondary measure for a given mismeasured variable of interest and to show how this secondary measure can be used to remove attenuation bias resulting from measurement error. We extend the framework in Chalfin and McCrary (2018) for classical measurement error to the case where a variable is binary. The emphasis is on the newspaper data being easily available, which substantially reduces the cost of generating a secondary

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

measure for bias correction purposes, which is usually the main prohibitive factor for researchers to apply such methods. We also contribute to a recent literature that has highlighted the usefulness of historical newspapers to generate novel data for the purpose of research in economic history.

REFERENCES

Ager, Philipp, Markus Brueckner, and Benedikt Herz. "The Boll Weevil Plague and Its Effect on the Southern Agricultural Sector, 1889–1929." *Explorations in Economic History* 65 (2017): 94–105.

Ager, Philipp, Benedikt Herz, and Markus Brueckner. "Structural Change and the Fertility Transition." *Review of Economics and Statistics* 102, no. 4 (2020): 806–22.

Aigner, Dennis J. "Regression with a Binary Independent Variable subject to Errors of Observation." *Journal of Econometrics* 1, no. 1 (1973): 49–59.

Albright, Alex, Jeremy A. Cook, James J. Feigenbaum, Laura Kincaide, Jason Long, and Nathan Nunn. "After the Burning: The Economic Effects of the 1921 Tulsa Race Massacre." NBER Working Paper No. 28985, Cambridge, MA, July 2021.

Almond, Douglas. "Is the 1918 Influenza Pandemic Over? Long-Term Effects of In Utero Influenza Exposure in the Post-1940 U.S. Population." *Journal of Political Economy* 114, no. 4 (2006): 672–712.

Ang, Desmond. "The Birth of a Nation: Media and Racial Hate." *American Economic Review* 113, no. 6 (2023): 1424–60.

Bailey, Martha, Connor Cole, and Catherine Massey. "Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850–1930 IPUMS Linked Representative Historical Samples." *Historical Methods* 53, no. 2 (2019): 80–93.

Baker, Richard B. "From the Field to the Classroom: The Boll Weevil's Impact on Education in Rural Georgia." *Journal of Economic History* 75, no. 4 (2015): 1128–60.

Baker, Richard B., John Blanchette, and Katherine Eriksson. "Long-Run Impacts of Agricultural Shocks on Educational Attainment: Evidence from the Boll Weevil." *Journal of Economic History* 80, no. 1 (2020): 136–74.

Bazzi, Samuel, Andreas Ferrara, Martin Fiszbein, Thomas P. Pearson, and Patrick A. Testa. "The Other Great Migration: Southern Whites and the New Right." *Quarterly Journal of Economics* 138, no. 3 (2023): 1577–647.

Beach, Brian, Ryan Brown, Joseph Ferrie, Martin Saavedra, and Duncan Thomas. "Re-evaluating the Long-Term Impact of In Utero Exposure to the 1918 Influenza Pandemic." *Journal of Political Economy* 130, no. 7 (2022): 1963–90.

Beach, Brian, Karen Clay, and Martin H. Saavedra. "The 1918 Influenza Pandemic and Its Lessons for Covid-19." *Journal of Economic Literature* 60, no. 1 (2022): 41–84.

Beach, Brian, and Walker W. Hanlon. "Historical Newspaper Data: A Researcher's Guide and Toolkit." NBER Working Paper No. 30135, Cambridge, MA, June 2022.

———. "Culture and the Historical Fertility Transition." *Review of Economic Studies* 90, no. 4 (2023): 1669–700.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

Berkes, Enrico, Ezra Karger, and Peter Nencka. “The Census Place Project: A Method for Geolocating Unstructured Place Names.” *Explorations in Economic History* 87 (2023): 101477.

Bingley, Paul, and Alessandro Martinello. “Measurement Error in Income and Schooling and the Bias of Linear Estimators.” *Journal of Labor Economics* 35, no. 4 (2017): 1117–48.

Bloome, Deirdre, James Feigenbaum, and Christopher Muller. “Tenancy, Marriage, and the Boll Weevil Infestation, 1892–1930.” *Demography* 54, no. 3 (2017): 1029–49.

Boustan, Leah P., Matthew E. Kahn, Paul W. Rhode, and Maria Lucia Yanguas. “The Effect of Natural Disasters on Economic Activity in US Counties: A Century of Data.” *Journal of Urban Economics* 118 (2020): 103257.

Calderon, Alvaro, Vasiliki Fouka, and Marco Tabellini. “Racial Diversity and Racial Policy Preferences: The Great Migration and Civil Rights.” *Review of Economic Studies* 90, no. 1 (2023): 165–200.

Chalfin, Aaron, and Justin McCrary. “Are U.S. Cities Underpoliced? Theory and Evidence.” *Review of Economics and Statistics* 100, no. 1 (2018): 167–86.

Clay, Karen, Ethan Schmick, and Werner Troesken. “The Rise and Fall of Pellagra in the American South.” *Journal of Economic History* 79, no. 1 (2019): 32–62.

———. “The Boll Weevil’s Impact on Racial Income Gaps in the Early Twentieth Century.” NBER Working Paper No. 27101, Cambridge, MA, May 2020.

Coulter, John L. “Agricultural Development in the United States, 1900–1910.” *Quarterly Journal of Economics* 27, no. 1 (1912): 1–26.

Dippel, Christian, and Bryan Leonard. “Not-so-Natural Experiments in History.” *Journal of Historical Political Economy* 1, no. 1 (2021): 1–30.

Dupraz, Yannick, and Andreas Ferrara. “Fatherless: The Long-Term Effects of Losing a Father in the US Civil War.” *Journal of Human Resources* (2023): https://doi.org/10.3368/jhr.0122-12118R2.

Esposito, Elena, Tiziano Rotesi, Alessandro Saia, and Mathias Thoenig. “Reconciliation Narratives: The Birth of a Nation after the US Civil War.” *American Economic Review* 113, no. 6 (2023): 1461–504.

Feigenbaum, James, and Daniel P. Gross. “Answering the Call of Automation: How the Labor Market Adjusted to the Mechanization of Telephone Operation.” NBER Working Paper No. 28061, Cambridge, MA, April 2022.

Feigenbaum, James J., Soumyajit Mazumder, and Cory B. Smith. “When Coercive Economies Fail: The Political Economy of the US South After the Boll Weevil.” NBER Working Paper No. 27161, Cambridge, MA, May 2020.

Ferrara, Andreas, and Price V. Fishback. “Discrimination, Migration, and Economic Outcomes: Evidence from World War I.” *Review of Economics and Statistics*, forthcoming (2023).

Ferrara, Andreas, Joung Yeob Ha, and Randall Walsh. “Using Digitized Newspapers t Address Measurement Error in Historical Data.” Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2023-12-07. https://doi.org/10.3886/E195648V1

Ferrara, Andreas, Patrick A. Testa, and Liyang Zhou. “New Area- and Population-Based Geographic Crosswalks for U.S. Counties and Congressional Districts, 1790–2020.” CAGE Working Paper No. 588, University of Warwick, Coventry, UK, 2021.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press

Gentzkow, Matthew, Nathan Petek, Jesse M. Shapiro, and Michael Sinkinson. "Do Newspapers Serve the State? Incumbent Party Influence on the US Press, 1869–1928." *Journal of the European Economic Association* 13, no. 1 (2015): 29–61.

Gentzkow, Matthew, Jesse M. Shapiro, and Michael Sinkinson. "Competition and Ideological Diversity: Historical Evidence from US Newspapers." *American Economic Review* 104, no. 10 (2014): 3073–114.

Haines, Michael R. "Historical, Demographic, Economic, and Social Data: The United States, 1790–2002." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-05-21. https://doi.org/10.3886/ICPSR02896.v3

Harned, Robey W. "Boll Weevil in Mississippi, 1909." *Mississippi Agricultural Experiment Station Bulletin* 139 (1910).

Hatchett, Richard J., Carter E. Mecher, and Marc Lipsitch. "Public Health Interventions and Epidemic Intensity during the 1918 Influenza Pandemic." *PNAS* 104, no. 18 (2007): 7582–87.

Hilt, Eric, and Wendy Rahn. "Financial Asset Ownership and Political Partisanship: Liberty Bonds and Republican Electoral Success in the 1920s." *Journal of Economic History* 80, no. 3 (2020): 746–81.

Howard, Greg, and Arianna Ornaghi. "Closing Time: The Local Equilibrium Effects of Prohibition." *Journal of Economic History* 81, no. 3 (2021): 792–830.

Hu, Yingyao, and Susanne M. Schennach. "Instrumental Variable Treatment of Nonclassical Measurement Error Models." *Econometrica* 76, no. 1 (2008): 195–216.

Hunter, Walter D., and Bert R. Coad. *The Boll-Weevil Problem*. Washington, DC: U.S. Department of Agriculture, 1923.

Lange, Fabian, Alan L. Olmstead, and Paul W. Rhode. "The Impact of the Boll Weevil, 1892–1932." *Journal of Economic History* 69, no. 3 (2009): 685–718.

Oster, Emily. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal Business & Economic Statistics* 37, no. 2 (2019): 187–204.

Ottinger, Sebastian, and Max Winkler. "The Political Economy of Propaganda: Evidence from U.S. Newspapers." IZA DP No. 15078, Bonn, Germany, February 2022.

Rhode, Paul W. "Biological Innovation without Intellectual Property Rights: Cottonseed Markets in the Antebellum American South." *Journal of Economic History* 81, no. 1 (2021): 198–238.

Schmick, Ethan. "Collective Action and the Origins of the American Labor Movement." *Journal of Economic History* 78, no. 3 (2018): 744–84.

Sechrist, Robert P. "Prohibition Movement in the United States, 1801–1920." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2012-10-26. https://doi.org/10.3886/ICPSR08343.v2

Sun, Liyang, and Sarah Abraham. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Journal of Econometrics* 225, no. 2 (2021): 175–99.

Wang, Tianyi. "The Electric Telegraph, News Coverage and Political Participation." Working Paper, 2019. Available at https://drive.google.com/file/d/1AuWnM-gJR1yZ0yIqjpe4B9poM69sr3Hx/view.

Wright, Gavin. *Sharing the Prize: The Economics of the Civil Rights Revolution in the American South*. Cambridge, MA: Belknap Press, 2013.

https://doi.org/10.1017/S0022050723000530 Published online by Cambridge University Press