*Online Appendix*

A.1 Sensitivity Analyses for the Clay et al. (2019) and Ager et al. (2017) Replications

A.2 Data Collection for the Hilt and Rahn (2020) and Howard and Ornaghi (2021) Replications

List of Figures

A.1 Boll Weevil in Newspapers

A.2 Boll Weevil in Newspapers from Different Counties

A.3 Cross-Sectional Correlation Between the USDA and Newspaper-Based Boll Weevil Arrival Years

A.4 Salience of the Boll Weevil - Dixie and Lafayette Counties, Florida

A.5 Errors in Newspapers

A.6 Illustration of an Agreement Sample

A.7 Visualization of Column 1 in the Replication of Clay et al. (2019)

A.8 Event Study Plots - TWFE and Sun and Abraham (2021)

A.9 Salience of the Boll Weevil - Towns in Alamance County, North Carolina

A.10 Measurement Errors in Sechrist (2012)

A.11 Sensitivity Analysis

List of Tables

A.1 Replication of Clay et al. (2019) - Agreement Samples without and with Weights

A.2 Replication of Ager et al. (2017) - Agreement Samples without and with Weights

ADDITIONAL FIGURES AND TABLES

Figure A.1: Boll Weevil in Newspapers

Text, letter

Description automatically generated

*Notes*: An example of newspaper articles on the boll weevil infestation. Newspapers published articles about the boll weevil’s arrival as well as damages in cotton production caused by the insects.

*Sources*: *The Times-Democrat*, 8 June 1908.

Figure A.2: Boll Weevil in Newspapers from Different Counties

A page of a book with text

Description automatically generated with medium confidence

*Notes*: Newspapers reported the number of boll weevil cases not only in their own county, but also from other counties or even different states. Figure A.2 shows that the boll weevil infestation in Marion County was reported in Hinds County (left) and Attala County (right).

*Sources*: *Jackson Daily News*, 11 June 1910 (left); *The Star Ledger*, 8 July 1910 (right).

Figure A.3: Cross-Sectional Correlation Between the USDA and Newspaper-Based Boll Weevil Arrival Years



*Notes*: Binned scatter plot of the USDA boll weevil arrival date and the predicted arrival date based on the newspaper-based measure defined in equation (3). The data is a cross-section of the 911 counties in the South that have ever been infested by the boll weevil.

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

Figure A.4: Salience of the Boll Weevil - Dixie and Lafayette Counties, Florida

(a) Dixie County, Florida



(b) Lafayette County, Florida



*Notes*: The dashed line is the salience measure of Dixie County (in panel (a)) and Lafayette County (in panel (b)) over time, respectively. The salience measure is constructed based every available newspaper outlet in Florida between 1882 and 1932. In panel (a), missing values are shown in early periods because the search word “Dixie County” did not show in newspapers until 1921 (except for errors). Dixie County was created in 1921 from the southern portion of Lafayette County. The solid line is the five-year moving average of the salience measure () for each county. The vertical lines indicate the boll weevil’s arrival from the USDA map and the predicted arrival where is the highest, respectively.

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

Figure A.5: Errors in Newspapers

A black and white photo of a document

Description automatically generated with low confidence

*Notes*: An example of possible errors in our approach. The search word “boll weevil” shows in one article (top left) and “Marion County” shows in another article (bottom right). This page is still counted when we construct the salience measure of Marion County even though it did not report the boll weevil infestation of the county.

*Sources*: *The Lexington Advertiser*, 28 July 1904.

Figure A.6: Illustration of an Agreement Sample

(a) Visualization

Diagram

Description automatically generated

(b) Example Data Frame

Table

Description automatically generated

*Notes*: An illustration of an agreement sample where and measure that is treated at time and , respectively. The agreement sample excludes the shaded region of potential error. The coefficient estimate for the agreement sample is only based on the unbiased pre- and post-treatment periods (i.e. the part of the sample for both and provide the same value of ).

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

Figure A.7: Visualization of Column 1 in the Replication of Clay et al. (2019)



*Notes*: Regression of log pellagra deaths on an indicator for the boll weevil’s arrival from the USDA map () and the predicted arrival based on newspapers () for the replication of Clay et al. (2019). The figure visualizes the coefficients from our replication exercise in column 1 of Table 2 to show the ranked pattern according to our theory in the theoretical section. The variable subscript indicates which variable was used as treatment. The figure confirms the pattern described in the inequality in equation (6).

*Sources*: Authors’ calculations from data in Hunter and Coad (1923), Clay et al. (2019), and Newspapers.com.

Figure A.8: Event Study Plots - TWFE and Sun and Abraham (2021)

(a) Town-level Event Study



(b) County-level Event Study



*Notes*: Coefficient plots from an event study regressions of on an event indicator relative to the arrival of the boll weevil from the USDA map as well as year fixed effects and sub-county area (hereinafter town) fixed effects (in panel (a)) or county fixed effects (in panel (b)). Each circle and diamond presents the estimates using OLS and the estimator proposed by Sun and Abraham (2021), respectively. The sample consists of 954 towns in 77 infested counties in North Carolina. The omitted baseline period is , which is one year before the arrival from the USDA map. The relative time period for the latest-infested towns (in panel (a)) or counties (in panel (b)) is omitted as well for the estimates using Sun and Abraham (2021) due to the lack of never-infested areas in our sample. Standard errors are clustered at the town level (in panel (a)) or county level (in panel (b)) and 95% confidence intervals are reported around the point estimates.

*Sources*: Authors’ calculations from data in Hunter and Coad (1923), Berkes et al. (2023), and Newspapers.com.

Figure A.9: Salience of the Boll Weevil - Towns in Alamance County, North Carolina

(a) Melvile (Township)



(c) Patterson (Unincorporated Community)



(b) Burlington (City)



(d) Alamance County



*Notes*: The dashed line is the salience measure of Melvile (in panel (a)), Burlington (in panel (b)), Patterson (in panel (c)), and Alamance County (in panel (d)) over time, respectively. The salience measure is constructed based on every available newspaper outlet in North Carolina between 1882 and 1932. The solid line is the five-year moving average of the salience measure () for each sub-county area (hereinafter town) (in panels (a)-(c)) and county (in panel (d)). The vertical lines indicate the boll weevil’s arrival from the USDA map and the predicted arrival where is the highest based on town-level and county-level newspaper-based approach, respectively.

*Sources*: Authors’ calculations from data in Hunter and Coad (1923), Berkes et al. (2023), and Newspapers.com.

Figure A.10: Measurement Errors in Sechrist (2012)

A close-up of a document

Description automatically generated with medium confidence

*Notes*: The list of dry counties in Illinois published ina newspaper article. Boone, Jersey, and Johnson Counties are not recorded as dry in 1908 in Sechrist (2012).

*Sources*: *The Daily Journal*, 23 April 1908.

Table A.1: Replication of Clay et al. (2019) - Agreement Samples without and with Weights

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Log pellagra death | | | |  | Log pellagra death rate | | | |
|  | (1) | (2) | (3) | (4) |  | (5) | (6) | (7) | (8) |
|  | -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) |
|  | -0.371∗∗∗ | -0.286∗∗ | -0.301∗∗∗ | -0.230∗∗ |  | -0.292∗∗∗ | -0.214∗∗∗ | -0.257∗∗∗ | -0.218∗∗∗ |
|  | (0.107) | (0.112) | (0.113) | (0.108) |  | (0.075) | (0.078) | (0.079) | (0.079) |
| 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. | 1,051 | 1,051 | 1,051 | 1,051 |  | 1,051 | 1,051 | 1,051 | 1,051 |
| Counties | 141 | 141 | 141 | 141 |  | 141 | 141 | 141 | 141 |

*Notes*: Regressions of deaths by pellagra on a boll weevil indicator and its interaction term with an indicator for whether county was in top 25% cotton production in 1909 or a dummy variable equal to one if county was in top 25% pellagra death rates in 1915/16 using the agreement sample in Table 2 without and with the inverse propensity score weights. The weight is the inverse of the predicted value of a Probit regression of an indicator for whether county never has a disagreement observation on pre-treatment county characteristics in 1910 such as total population, urban population, total area in farms, acreage share of cotton, and the distance to the first infested county (Cameron County, Texas). When a value for county characteristics is missing, we replace it with its average value. We set the weight to one for never-infested counties. The sample includes counties in North Carolina and South Carolina between 1915 and 1925. All regressions include county and year fixed effects. Controls include county ’s malaria death rate in 1915 and the share of urban population in 1910 both 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), Haines (2010), Clay et al. (2019), and Newspapers.com.

Table A.2: Replication of Ager et al. (2017) - Agreement Samples without and with Weights

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Log cotton production | | | |  | Log corn production | | | |  | Log other outcomes | | | |
|  | Bales | Acres | Yield | Share |  | Bushels | Acres | Yield | Share |  | Farm | Farm value | Pop. | Black pop. |
|  | (1) | (2) | (3) | (4) |  | (5) | (6) | (7) | (8) |  | (9) | (10) | (11) | (12) |
|  | -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) |
|  | -0.652∗∗∗ | -0.296∗∗∗ | -0.355∗∗∗ | -0.091∗∗∗ |  | 0.029 | 0.110∗∗∗ | -0.081∗∗∗ | 0.085∗∗∗ |  | -0.028∗ | -0.030∗ | -0.004 | -0.042 |
|  | (0.060) | (0.056) | (0.025) | (0.008) |  | (0.044) | (0.025) | (0.027) | (0.008) |  | (0.015) | (0.018) | (0.017) | (0.043) |
| 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. | 3,927 | 3,933 | 3,927 | 4,032 |  | 4,032 | 4,032 | 4,032 | 4,032 |  | 4,032 | 4,032 | 3,328 | 3,311 |
| Counties | 733 | 733 | 733 | 740 |  | 740 | 740 | 740 | 740 |  | 740 | 740 | 740 | 739 |

*Notes*: Regressions of agricultural and demographic outcome variables on an indicator for whether the boll weevil has arrived in county and its interaction term with county ’s demeaned acreage share of cotton in 1889 using the agreement sample in Table 4 without and with the inverse propensity score weights. The weight is the inverse of the predicted value of a Probit regression of an indicator for whether county never has a disagreement observation on pre-treatment county characteristics in 1890 such as total population, urban population, total area in farms, acreage share of cotton, distance to the first infested county (Cameron County, Texas), a dummy for counties with unchanged boundaries, and average temperature and precipitation. When a value for county characteristics is missing, we replace it with its average value. 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), Haines (2010), Ager et al. (2017), and Newspapers.com.

A.1 Sensitivity Analyses for the Clay et al. (2019) and Ager et al. (2017) Replications

As a sensitivity analysis, we replicate our main analysis (Tables 2 and 4) under the alternative variable definitions discussed above in our theoretical section. In particular, we test different approaches to constructing the newspaper-based boll weevil arrival measure: including the three- and seven-year moving averages, as well as the maximum of the raw salience measure in equation (1) in a ten-year window around the USDA map arrival date. Because we feel the most important issue is the ability to substantially reduce measurement error, for parsimony, we focus our attention in the sensitivity analysis on the estimated coefficients based on the agreement sample.

The agreement sample coefficients using the different measures of the newspaper-based boll weevil arrival for the replication of Clay et al. (2019) are plotted in panel (a) of Figure A.11. The same exercise for the replication of Ager et al. (2017) is plotted in panel (b) of the same figure. Each bar presents the estimated coefficient using a given measure with error bars reporting their corresponding 95% confidence intervals. The black crosses indicate the coefficient value in the original study that was replicated.

Figure A.11 panel (a) shows coefficient estimates for each alternative measure of the newspaper-based boll weevil arrival for columns 1-8 of the corresponding Table 2. We observe that the estimates using the agreement sample, (the gray bars), are robust to different measures of the newspaper-based boll weevil arrival date. Moreover, the point estimates based on these measures are bigger than (the black crosses) regardless of specification. In addition to the agreement sample estimates, columns C1\* and C5\* of Figure A.11 present the bias-corrected estimates, , using alternative specifications corresponding to columns 1 and 5 of Table 2. While these coefficients are less precisely estimated in some specifications, the magnitude of each point estimate remains bigger than that of the OLS estimates from the USDA map.

Panel (b) of the same figure repeats the exercise using the agreement sample for Ager et al. (2017). We report as well as the OLS estimates based on the arrival from the USDA map, which is in Table 4. Again, the figure shows very similar estimates, regardless of different specifications and outcome variables. As in the main replication exercise, is bigger than (the black crosses). Here, not only are the agreement sample estimates uniformly larger in magnitude than the original estimates of Ager et al. (2017), but in all but one case the entire 95 percent confidence interval lies above the original point estimates.

Figure A.11: Sensitivity Analysis

(a) Sensitivity Checks for the Replication of Clay et al. (2019)



(b) Sensitivity Checks for the Replication of Ager et al. (2017)



*Notes*: In panel (a), each bar in C1 through C8 presents the estimates in in Table 2 (where C refers to the corresponding column in the table) for alternative specifications with the 95% confidence intervals. C1\* and C5\* show the estimates in Table 2 for different specifications. Alternative specifications of the newspaper-based boll weevil arrival date include the three- and seven-year moving averages as well as the maximum of the raw salience measure in a ten-year window around the USDA map. Black crosses indicate the OLS estimates based on the arrival from the USDA map. In panel (b), each bar presents the estimates in Table 4. Black crosses indicate the OLS estimates based on the arrival from the USDA map (i.e. in Table 4).

*Sources*: Authors’ calculations from data in Hunter and Coad (1923), Clay et al. (2019), Ager et al. (2017), and Newspapers.com.

A.2: Data Collection for the Hilt and Rahn (2020) and Howard and Ornaghi (2021) Replications

*Influenza hotspots*:Hilt and Rahn (2020) study the effect of the liberty bond program on the county-level Democratic vote share and use distance to military camps during World War I as proxy for the severity of the 1918 influenza epidemic to instrument for a county’s liberty bond participation rate. We replicate their reduced form regression (column 2 in their Table A.6). We constructed our newspaper-based proxy for influenza severity in the same way we did for the boll weevil exposure measure: first, we searched Newspapers.com for the joint occurrence of the word *flu* and each county’s name in 1918.[[1]](#footnote-1) We then standardized this count variable by the number of search hits for a given county name in the same year after which we identified influenza hotspots as the top decile of this share. Lastly, to mirror the Hilt and Rahn (2020) instrument, we computed the average “as the crow flies” distance to each of these hotspots.

*Prohibition policies*:To identify counties with a dry-law policy, we first search Newspapers.com for the terms *prohibition* and *dry*, and 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 for the same period. After constructing this share, we predict the adoption of prohibition in each county by using the maximum of the five-year moving average of the share, just as we did in the boll weevil example. Howard and Ornaghi (2021) consider all rural counties that adopted prohibition between 1900 and 1919. We impose a minimal sample restriction by which counties are kept that adopted prohibition in the Sechrist (2012) data and that we also found in the newspaper data. Otherwise we would not be able to apply our methods if one of the two measures for prohibition adoption is missing. This reduced their original sample of 2,315 counties to 1,452 counties (62.7%). While this is a sizable reduction of their original data set, we verify that their results hold in this subset of their data. For instance, they find coefficients of 0.090 (s.e. = 0.022) and 0.107 (s.e. = 0.024) for a regression of log value of farm implements and log farm values, respectively, on the prohibition adoption dummy in their original data set. Using their variables in the reduced data set, these coefficients are 0.125 (s.e. = 0.033) and 0.127 (s.e. = 0.036), respectively. This replication nonetheless provides an example of the trade-off between sample size and precision, where a smaller sample with the possibility for correcting measurement error can still outperform the larger sample without this option.

Measurement Error Derivations

For the continuous case where measurement error in and are classical and uncorrelated with each other, the IV estimator is

Which yields the true parameter if .

For the binary case, consider the first stage regression

Meyer and Mittag (2017) show that measurement error in a binary outcome yields a biased right- hand side coefficient. In the absence of measurement error in , the estimated first stage coefficient therefore would be . However, since is mismeasured, the coefficient is additionally attenuated as . Now consider the reduced form,

where , , and . Also here the measurement error in is reflected in the attenuation of the reduced form coefficient,. Lastly, the IV coefficient on , using as instrument, will be the estimated reduced form coefficient divided by the estimated first stage coefficient,

resulting in an inflated estimate of the true parameter unless, in which case the IV estimator is undefined.

1. The term *flu* was chosen over *Spanish flu* since newspapers would refer to the disease as *flu*, especially in the early months of the pandemic. [↑](#footnote-ref-1)